Measurement of innovation in South Africa: more than science and technology

A review of the control of schistosomiasis in South Africa

Housing sprawl near tailings storage facilities puts population at risk

Paper-based microfluidic devices for low-cost point-of-care diagnostics

Conflating expression and meaning in the mortuary behaviour of *Homo naledi*

NOVEMBER/DECEMBER 2015

ISSN: 0038-2353
eISSN: 1996-7489

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

SOUTH AFRICAN Journal of Science

volume 111
number 11/12

NOVEMBER/DECEMBER

ISSN: 0038-2353 (Print) | ISSN: 1996-7489 (Online)

SOUTH AFRICAN

Journal of Science

volume 111

number 11/12

Hassina Mouri
Department of Geology,
University of Johannesburg

Johann Mouton
Centre for Research on Science and Technology, Stellenbosch University

Maano Ramutsindela
Department of Environmental & Geographical Science, University of Cape Town

Published by
the Academy of Science of South Africa (www.assaf.org.za) with financial assistance from the Department of Science & Technology.

Design and layout
SUN MeDIA Bloemfontein
T: 051 444 2552
E: admin@sunbloem.co.za

Correspondence and enquiries
sajs@assaf.org.za

Submissions
Submissions should be made at http://mc.manuscriptcentral.com/sajs

Subscriptions
Subscriptions and print copies will be discontinued from 2016

Corrigendum

Errors appear in the Commentary on pages 6–9 of Volume 111 (9/10).

William J. Bond was erroneously reported as a B3 NRF-rated researcher. He is in fact an A1 NRF-rated researcher. The error appears on page 2, paragraph 4, line 1 and in Table 2. The NRF rating of Christopher S. Henshilwood was reported as N/A in Table 2. He is actually an A2 NRF-rated researcher with primary affiliation to the University of Bergen, Norway.

Research Letter

Is the decline and fall of South African universities looming?

Over the past few weeks, hundreds of thousands, if not millions, of words have been written or spoken about the student protests and marches that led to President Zuma's clearly panicked response, declaring that there would be no fee increases at South African universities in 2016.

There are three considerations that reveal the extent of the President's alarm. Firstly, it is not clear who in his cabinet he consulted before making his announcement. Secondly, it is very clear that, at the very least, he did not consult anyone in his financial cluster as to how he expected his country's universities to survive without the funds they need in order to operate effectively and to deliver decent tertiary-level education. And, thirdly, it is constitutionally unclear as to whether or not the President has the authority to arrogate, to himself, powers that are legally assigned to the councils of South African universities.

The most fundamental reason for the President's inopportune and alarmist response without (it would seem) any consideration of the short- and long-term consequences of his announcement, is that the government has, for 20 years, paid little or no attention to the critical role and significance of post-school and higher education, and scientific research, in the country. While other BRICS countries which are, apart from India, all in dire financial straits are, as a consequence, providing substantial funding for higher education as they understand the importance of education in creating and sustaining development, South Africa has shown short shrift for this sector. The history of neglect and its consequence – the absence of consultation and careful planning – resulted in newspaper headlines such as 'Scramble to fund 0% hike', 'Blade waits to hear from Zuma', 'Chaos rocks Nene's budget', 'Zuma slips away', 'Blade beats a retreat' and, perhaps the most telling of all, this statement from the Director General of the Department of Higher Education and Training (DHET), Mr Gwebs Qonde: 'We don't know where we'll get the money. That's a fact.'

Of course, one might point to what would seem to be policies and practices that are contrary to these three assertions. The Department of Science and Technology (DST) has, through a number of ventures, managed primarily by the National Research Foundation (NRF), invested funds in support of postgraduate students, institutional research projects and the Square Kilometre Array, amongst others. But the country's investment in research through the DST, the NRF and private sources remains below 1% (at 0.92%) while China's contribution is 1.98%, and Brazil's is 1.21%. The irony is that the National Development Plan, approved by the Cabinet, mentions (almost 150 times) research as being critical to economic growth and development, and specifically states that the National Treasury is to use fiscal instruments to subsidise research and development, and that research capacity in higher education must be nurtured and coordinated, and linked to postgraduate studies in universities.

In direct contradiction to this, the Minister of Finance has effectively frozen research and post-education budgets for 2016, and has reduced the funds available for technical and vocational education and training. At the same time, he has increased the budget for public service salaries by ZAR20 billion a year for 3 years, bringing the total to ZAR359 billion for 2016 or 35% of the total national budget.

One might also point to efforts made by the DHET to fund research publications, support improved teaching and help to develop research skills in universities. None of these is in any way insignificant – but the funds for these ventures come from an existing and diminishing budget rather than from additional funds. In other words, these admirable and necessary ventures serve to reduce the overall funds available for the general support for and operations of South Africa's universities, whose expenditures increase far above the official inflation rates.

The cost of electricity is increasing at close to 25% each year, and the increase in the cost of water is, on average, about 9%. Worse still, however, is that the cost of scholarly and scientific books is escalating rapidly; universities have to pay for the research that is core to their role, and then pay for that research to be published. As if that were not bad enough, the universities have then to pay, yet again, for students and staff to have access to the books and journals in which their researchers have published their findings, with each of these costs increasing beyond the so-called 'official' inflation rate. Operational costs – those which make it possible for universities to do their best by way of teaching well, supporting their students, and undertaking development-related research – keep going up and state contributions keep declining, in real terms.

And then there is a further challenge, lurking in the shadows of the future. Demands, backed by protests (some violent), that are so rashly and ignorantly acceded to, produce the understandable notion that the next round of demands will also result in acquiescence.

Here, then, are the remaining items on the 'demand list'. *Polity* recently reported[1] a meeting between members of the ANC National Executive Committee (NEC), and the Progressive Youth Association (PYA) chaired by ANC Secretary General Gwede Mantashe:

> *The* [NEC] *led by the Secretary General of the ANC… received a report from the PYA on the student protests against increments in university fees.*
>
> *The meeting acknowledged and commended students for running a predominantly peaceful protest campaign nationally that brought to the fore the critical challenges facing education, particularly in the context of fees that excludes many students from tertiary education.*
>
> *The meeting agreed to urgently raise…the demand for a 0% increase in university fees, the scrapping of registration fees, long-term solution to realise free education, a review of the autonomy of universities and the capping of earnings of academics.*

The agreement reached by the meeting formed part of a presentation and recommendations tabled before President Zuma ahead of his meeting with student representatives.

That the NEC commended the students for their lack of violence is, at the very least, disingenuous, as there were violent protests outside the Union Buildings in Pretoria, threats of violence on campuses, and the destruction of, or damage to, buildings on at least two university campuses. The agenda is, however, of greater concern.

The first item – a 0% increase in fees – has been achieved, with consequences yet to be understood and resolved, other than that universities have been asked to trim their budgets, while Minister Nzimande scrambles around to find ways of determining where the compensatory funds may be found – and the civil service salary budget is increased by ZAR20 billion a year for 3 years.

A start, of course, would be a compensatory reduction in the budgets of those state departments identified by the Auditor General as having spent large sums of monies fraudulently and unaccountably – funds which, then, they would not have needed had they been operating honestly and accountably in their use of public monies.

HOW TO CITE: Butler-Adam J. Is the decline and fall of South African universities looming? S Afr J Sci. 2015;111(11/12), Art. #a0132, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0132

That item one was met, makes the others cause for concern. Where will the Minister and/or the President draw the line? And why would they do so at that point?

The second item – scrapping registration fees – reduces university income in addition to the loss of fee increases – while institutional expenses continue to grow. The third item – free education – means one of two things: growing DHET support for universities through the state, or inevitable decline. Unless 'no fee' education is replaced by a continuing and growing state income that will retain and improve the quality of the country's world-class universities and, at the same time, develop and improve those institutions that were the victims of apartheid and extended periods of mismanagement, this item simply means that students (and the government) would be happy for their successors in universities to receive a low-quality education.

So to item four: a review of the autonomy of the universities. This item is difficult, because it is not clear what is meant by 'autonomy'. If it is just a matter of control over fee increases or no fees at all, then it is the same issue covered by item three. If, however, the 'review' means looking at autonomy more widely, as is envisaged in the *Higher Education Amendment Act* now before Parliament, then this could mean that universities might no longer select their own Chancellors, appoint their own Vice Chancellors, set staff salaries, and determine and teach their own curricula. This would turn South Africa's universities into the equivalent of those elsewhere in Africa 20 years ago, and end any notion of international standing and quality – and intellectual and scientific integrity. This is not to say that universities should not be expected to act in ways that are socially accountable – but the means for holding institutions to their legal and social responsibilities are already firmly in place.

Finally, item 5: capping earnings of academics. It is almost tempting to laugh at this notion; yet it is not a laughing matter. South Africa's academics are not particularly well paid, and in face of salary increases for the employees of the state sector amounting to ZAR20 billion a year for the next 3 years, this item is nothing less than insulting – and thoughtless. It is to be assumed that students who wish to pay no fees nonetheless wish to be educated in ways that make them effective players in the world economy. Capping the salaries of researchers and teachers will have two consequences: those staff who can, will leave; and those who are unable to do so will take on additional work – unpublishable contract research or multiple teaching responsibilities. Either or both consequences will result in the move from the decline to the fall of South Africa's universities.

Is this really what the PYA, the NEC and the government wish for? To move a reasonably sound system into the backwaters of international higher education, producing little or no serious research? And, worst of all, graduating students whose education will join the worst in the world?

Reference

1. ANC. Statement: ANC: Meeting with the Progressive Youth Alliance [homepage on the Internet]. c2015 [cited 2015 Oct 30]. Available from: http://www.polity.org.za/article/anc-meeting-with-the-progressive-youth-alliance-2015-10-22

Estimating the age and affinities of *Homo naledi*

AUTHOR:
J. Francis Thackeray[1]

AFFILIATION:
[1]Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
J. Francis Thackeray

EMAIL:
Francis.thackeray@wits.ac.za

POSTAL ADDRESS:
Evolutionary Studies Institute, Private Bag 3, Wits 2050, South Africa

KEYWORDS:
Rising Star; Plio-Pleistocene; least squares linear regression; morphometric analysis

HOW TO CITE:
Thackeray JF. Estimating the age and affinities of *Homo naledi*. S Afr J Sci. 2015;111(11/12), Art. #a0124, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0124

Recent discoveries of more than 1500 hominin fossils from the site of Rising Star in the South African Cradle of Humankind, attributed to a new species (*Homo naledi*),[1] have attracted global interest. As yet no secure date for this extraordinary material has been obtained, and the relationship of this species to other Plio-Pleistocene taxa has been greatly debated in the media. Here I report results of morphometric analyses that may facilitate an assessment of the age and affinities of crania attributed to *H. naledi*.

The method is based on a least squares linear regression analysis of mean values of measurements for crania of 12 hominin species (Table 1), as published by Berger et al.[1] The analyses were performed to obtain standard errors of *m*-coefficients (se_m) in regression equations of the form $y = mx + c$, based on pairwise comparisons of cranial data, as described elsewhere for pairs of hominin specimens,[2,3] taking into account criticisms raised by Gordon and Wood[4].

The degree of scatter around a regression line of pairwise comparisons is quantified by the se_m statistic. Log transformed se_m values for conspecific pairs of modern vertebrates (as well as invertebrates) display a normal distribution with a mean value of -1.61,[5] which has been considered to be an approximation of a biological species constant (T) with a standard deviation of circa 0.1,[3] and which has been used to facilitate a mathematical (probabilistic) definition of a species[5].

It is of great interest to use this approach by comparing cranial measurements of *H. naledi* to those of other species listed in Table 1, using data published by Berger et al.[1] Log se_m Hn (*x* axis) values refer to results of pairwise comparisons when the measurements for *H. naledi* are on the *x* axis, and measurements for other taxa are on the *y* axis in regression analyses. Log se_m Hn (*y* axis) values refer to results of pairwise comparisons when the measurements for *H. naledi* are on the *y* axis, and measurements for other taxa are on the *x* axis. 'Log se_m mean' refers to the mean of these two values, and 'delta log se_m' refers to the difference between the two values, which can also be used to assess degrees of similarity or dissimilarity in the context of log se_m values.[3]

In the context of results reported for conspecific pairs of modern taxa, it is relevant to report two results from the current study. Firstly, *H. naledi* appears to be significantly different (dissimilar) from other species listed in Table 1, because all of the mean log se_m values listed in Table 1 are outside the 95% confidence limits around the mean value of -1.61 +/- 0.1 for conspecifics. Secondly, *H. naledi* appears to be most similar to specimens attributed to early *Homo*, notably *H. habilis*, and (to a lesser extent) *H. rudolfensis* and *H. erectus* (see numbers listed in bold in Table 1).

Table 1: Results of pairwise comparisons between cranial measurements of *Homo naledi* and corresponding measurements of other species

	Log se_m Hn (x axis)	Log se_m Hn (y axis)	Log se_m mean	Delta log se_m
Paranthropus aethiopicus	-0.852	-0.980	-0.916	0.128
P. boisei	-1.098	-1.067	-1.082	-0.031
P. robustus	-1.156	-1.119	-1.137	-0.037
Australopithecus afarensis	-1.110	-1.139	-1.124	0.029
A. africanus	-1.189	-1.142	-1.165	-0.047
A. sediba	-1.192	-1.120	-1.156	-0.072
Homo habilis	-1.326	-1.331	**-1.328**	0.005
H. rudolfensis	-1.200	-1.272	**-1.236**	0.072
H. erectus	-1.131	-1.256	**-1.193**	0.125
H. sapiens (Middle Pleistocene)	-1.062	-1.285	-1.173	0.223
H. sapiens (modern)	-1.102	-1.202	-1.152	0.100

Log se_m Hn (x axis) values refer to results of pairwise comparisons when the measurements for H. naledi *are on the* x *axis, and measurements for other taxa are on the* y *axis. Log se_m Hn (y axis) values refer to results of pairwise comparisons when the measurements for* H. naledi *are on the* y *axis, and measurements for other taxa are on the* x *axis. Log se_m mean refers to the mean of these two values, and delta log se_m refers to the difference between the two values. Relatively low log se_m values in bold reflect the fact that* H. naledi *appears to be most similar to specimens attributed to* H. habilis, H. rudolfensis *and (to a lesser extent)* H. erectus, *associated with ages in the order of 2 (+/- 0.5) mya.*

A conclusion from this analysis is that the claim that *H. naledi* represents a distinct species appears to be warranted, at least from cranial data. Without assuming that log se_m values can provide accurate dates, the results presented in Table 1 may be used to provide an estimate for the age of *H. naledi*, here considered to be in the order of 2 million years (+/- 0.5 years), recognising that the maximum age for *H. rudolfensis* is circa 2.5 mya, the age for African *H. erectus*

in this comparative study is circa 1.5 mya, and the age for *H. habilis* from sites such as Olduvai Gorge in Tanzania is circa 1.8 mya.

Acknowledgements

This research was supported by the National Research Foundation (South Africa) and the A.W. Mellon Foundation. I am grateful to Sue Dykes for assistance and discussion.

References

1. Berger LR, Hawks J, De Ruiter DJ, Churchill SE, Schmid P, Delezene LK, et al. *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. eLife. 2015;4:e09560. http://dx.doi.org/10.7554/eLife.09560
2. Thackeray JF. *Homo habilis* and *Australopithecus africanus*, in the context of a chronospecies and climatic change. In: Runge J, editor. Changing climates, ecosystems and environments within arid southern Africa and adjoining regions: Palaeoecology of Africa 33. Forthcoming 2015.
3. Thackeray JF, Dykes S. Morphometric analyses of hominoid crania, probabilities of conspecificity and an approximation of a biological species constant. HOMO J Comp Hum Biol. Forthcoming 2015.
4. Gordon AD, Wood BA. Evaluating the use of pairwise dissimilarity metrics in paleoanthropology. J Hum Evol. 2013;65:465–477. http://dx.doi.org/10.1016/j.jhevol.2013.08.002
5. Thackeray JF. Approximation of a biological species constant? S Afr J Sci. 2007;103:489.

The mournful ape: Conflating expression and meaning in the mortuary behaviour of *Homo naledi*

AUTHOR:
Patrick S. Randolph-Quinney[1,2]

AFFILIATIONS:
[1]School of Anatomical Sciences, Faculty of Health Sciences, University of the Witwatersrand Medical School, Johannesburg, South Africa

[2]Evolutionary Studies Institute, Centre for Excellence in Palaeosciences, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Patrick Randolph-Quinney

EMAIL:
patrick.randolph-quinney@wits.ac.za

POSTAL ADDRESS:
School of Anatomical Sciences, Faculty of Health Sciences, University of the Witwatersrand Medical School, 7 York Road, Parktown 2193, South Africa

KEYWORDS:
Rising Star; Dinaledi Chamber; body disposal; taphonomy

HOW TO CITE:
Randolph-Quinney PS. The mournful ape: Conflating expression and meaning in the mortuary behaviour of *Homo naledi.* S Afr J Sci. 2015;111(11/12), Art. #a0131, 5 pages. http://dx.doi.org/10.17159/sajs.2015/a0131

As an article in the last issue of the *South African Journal of Science*[1] highlighted, in the last few months we have seen a bonanza of early hominin material from the Cradle of Humankind – *Homo naledi* – presented to the world, courtesy of Lee Berger, John Hawks, Paul Dirks and the Rising Star science team. Firstly, there were papers on the taxonomy[2], and the geological and taphonomic[3] context, followed shortly afterwards by the detailed functional anatomy of the hands and feet of *H. naledi* in papers led by Tracy Kivell[4] and Will Harcourt-Smith[5], respectively. The media attention surrounding the fossils, and inferred mortuary behaviours, has been intense, but it has definitely put South African palaeoanthropology back on the world stage, and more importantly, encouraged the public to engage directly with the science and – via social media and exhibitions – with the scientists themselves.

The metrics for the two primary papers[2,3] have been nothing short of astonishing: 243 485 views and 25 435 downloads for the taxonomy paper, and 82 399 views and 9207 downloads for the context paper at the time of writing. In addition, there has been over 5500 downloads of the 3D surface models of the Dinaledi fossils, which allow users (both in academia and the public) to generate their own models of the fossils, provided they have a suitable 3D printer (Figure 1). This open-access public science has been a triumph in the democratisation and dissemination of data.[6] However, along with the media attention has come more than a fair degree of scientific and professional scrutiny (some as measured responses, some far from it) which has exposed some ugly truths at the heart of what we might like to think is a dispassionate and logical scientific debate.

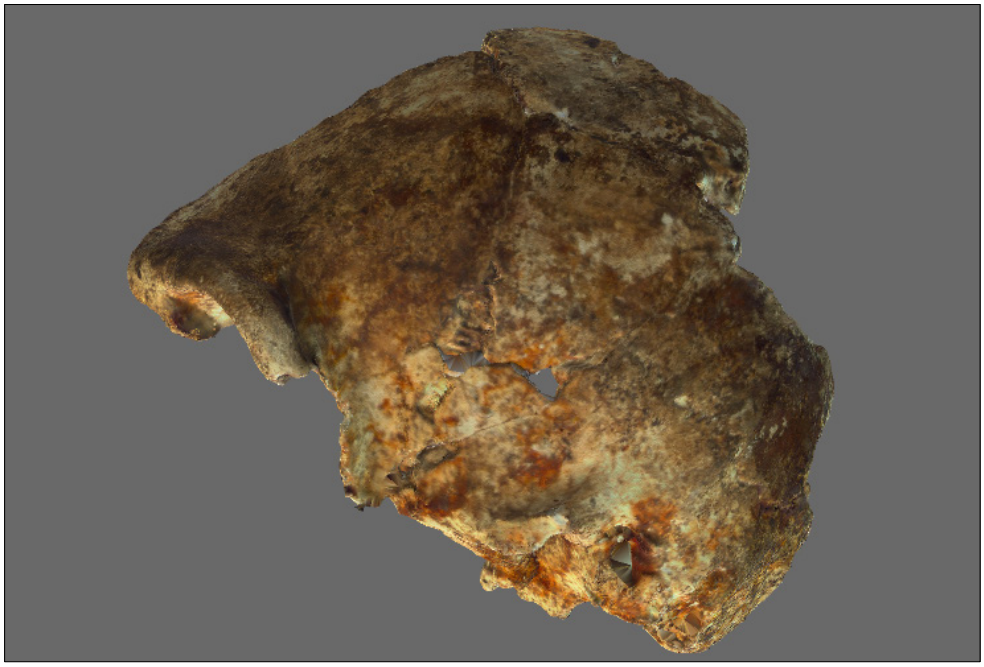


Photo: Patrick Randolph-Quinney.
Original specimen cc Evolutionary Studies Institute, University of the Witwatersrand

Figure 1: An example of 3D model sharing of the *Homo naledi* fossils. The figure shows a rendered surface scan of the U.W. 101-0396 (DH3) calvaria. The scan (3D Mesh, polygon file format) is free to download from *Morpho Source* (http://morphosource.org) and is displayed using *Microsoft 3D Builder*. The scan mesh is suitable for direct 3D printing.

First, the good scrutiny. The initial launch of the taxonomy and context papers was accompanied by a thoughtful and insightful commentary from Chris Stringer of the Natural History Museum in London. Stringer[7] highlights the issue of the (as yet) lack of radiometric dating from the site, and makes the important point that because *H. naledi* is currently only known from one site (as is also the case with *Australopithecus sediba*[8]), it is unclear whether or not the taxon was restricted to southern Africa. If *H. naledi* was more geographically widespread, its moderate body size may force palaeoanthropologists to re-examine other small-bodied fossils from across Africa, which have usually been attributed to a small form of *Homo erectus*.[7]

Other commentaries (all web based) followed from researchers such as Darren Curnoe (University of New South Wales) and Daniel Lieberman (Harvard University). Curnoe[9] states:

> *Reading the scientific article describing* Homo naledi *you realise that the work is detailed, rigorous and careful. It involved a large number of specialists covering a very wide set of physical features on the bones and teeth. The case for the new species is, in my opinion, detailed, compelling and praise worthy.*

Lieberman (interviewed by Allison Pohle[10] of Boston.com) was also broadly in support. He states:

> *The head of this thing is extremely like the* Homo erectus. *It has a brain that's a little larger than a chimpanzee, which is the smallest end of the range of brain sizes in the genus* Homo. *The shape of skull is* Homo erectus. *And its brow ridge, the shape of the face, and teeth, pretty much from the neck up, a lot looks like the* Homo erectus. *…From the neck down, there's a mixture of features….the foot was beautifully preserved and looks a lot like a human foot, except for the arch being a little flat. The upper body, arms and shoulders, look very primitive, like Lucy* [*Au. afarensis*]. *There was a beautifully preserved hand that was also very humanlike. The hands were humanlike in most regards except for the fingers and thumbs, and the shape of the wrist bones. The phalanges…are extremely curved, which you'd find in apes. It's an interesting mixture of stuff, some modern, some early* Homo, *and a few things you'd find in the* Australopithecus [sic]. *It's entirely reasonable for them to create a new species.*

Elsewhere science bloggers have described the *H. naledi* papers as a textbook example of how to do science.[11]

Now for the bad scrutiny. There were a number of negative commentaries, although it becomes abundantly clear that many of them are *ad hominem* attacks. The three primary nay-sayers to date have been Professors Christoph Zollikofer (University of Zurich), Jeffrey Schwartz (University of Pittsburgh) and Tim White (University of Berkeley). All three are senior scientists, and all three have profoundly negative views of the validity of *H. naledi* as a new and distinct species. Zollikofer, quoted in an interview with Johan von Mirbach[12], states:

> *The idea that this is a new genus is just another headline grabber. About 90% of this publication addresses the media and not the scientific community. I call this a 'media species', which is usually quite short-lived….My intuition says it is a primitive* Homo erectus. *But I'm just speculating, since nobody knows its exact biological age. Assuming that it is 2 million years old, you could say it is an early* Homo erectus*, but not a new genus.*

This statement gives pause for thought on two fronts: firstly, taxonomy should be divorced from chronology, and secondly, Berger and colleagues[2] do not name a new genus – *naledi* has been placed firmly within the existing genus *Homo*, thus joining a pantheon of taxa including manifestly primitive forms such as *H. habilis* and *H. rudolfensis*, evolutionary novelties such as *H. floresiensis*, and advanced (humanlike) morphs such as *H. ergaster, H. erectus, H. antecessor, H. heidelbergensis, H. neanderthalensis* and our own species, *H. sapiens*. Quite where the notion that *naledi* is representative of a 'new genus' comes from escapes me.

Writing for *Newsweek* on the day of the publication of the primary *eLife* papers[2,3], Jeffrey Schwartz[13] seems to suggest that the fossils should be placed in *Australopithecus*, and that several species are represented in the assemblage. He states:

> *Viewed from the side, two partial skulls are long and low, with a long gently sloping forehead that flows smoothly into the brow – nothing like us, or most specimens regarded as* Homo. *A third partial skull is very short and rounded, with a high-rising forehead that is distinguished from a distinct, well-defined brow by a shallow gutter – not like the other skulls, and not like us or most specimens regarded as* Homo. *The femur has a small head (the ball end that fits in the hip socket) that is connected to the shaft of the bone by a long neck, and, below the neck, is a 'bump' of bone that points backward. These features are seen in every australopith femur. In us, and all other living primates, the head of the femur is large and the neck short, and the 'bump' points inward. Further, the teeth are very similar to those from a nearby fossil site that has yielded various kinds of australopith. Even at this stage of their being publicized, the '*Homo naledi*' specimens reflect even greater diversity in the human fossil record than their discoverers will admit.*

In response, John Hawks highlights that *H. naledi* presents a *uniform* mix of primitive and derived traits, noting that every feature that is repeated in the sample is nearly identical in all individuals that preserve it. Hawks[14] states: '…It would be very strange to have a mix of different species where all seven proximal femora come from one species, while all of a dozen lower third premolars come from a different species.' One could also be a little less charitable, and suggest, given the rapidity with which the *Newsweek* article came out, that Professor Schwartz did not, perhaps, have sufficient time to fully absorb and assimilate the 35 pages of primary taxonomic description[2] and 26 pages of supplementary data and measurements before making his assessment.

Finally, Tim White has been the most prominent critic of both the taxonomy and the behavioural interpretation of *H. naledi.* White has challenged the primary nature of the deposit (he suggests it was mixed and disturbed), the care with which the fossils were recovered (he suggests that fresh breaks were caused by rushing the work, and by the excavators rather than the ingress of recreational cavers prior to the site being secured), and the specific taxonomy and composition of the assemblage (he attributes all to small-bodied *H. erectus*). This latter criticism may be considered somewhat ironic from a scientist who wrote (p. 291) that no one should publish on a fossil without seeing the original,[15] but who has not set foot in South Africa in a decade. Finally, White (along with Zollikofer) claim that the evidence for mortuary behaviours by the *naledi* hominins were specifically hyped for the press. In an interview with Glen Martin[16], White states, 'There is no evidence of burial rituals…the only evidence seems to be "We can't think of anything else". This is not evidence.'

I will leave it to others to address the criticism of the taxonomy and specific phylogenetic assessment of material from the Dinaledi Chamber, and instead concentrate on the issues raised about the inferred behaviour of *H. naledi* – deliberate body disposal. The case for this behaviour is based on geological, sedimentological, taphonomic and archaeological grounds; to contradict Professor White, what we present *is evidence,* and whilst we raise a number of alternative hypotheses to test against the physical data (hominin occupation of the cave, water transport of the remains, predator accumulation, mass fatality and death trap), the filter through which one assesses claims for each of these alternative scenarios simply does not fit the evidence at hand. In his commentary, Chris Stringer draws parallels between the depositional context of *H. naledi* and the 'sepulchral' pit from Sima de los Huesos (Atapuerca) in Spain, which provides evidence of at least 28 early Neanderthals who excavators suggest had been intentionally thrown into the pit, although it is worth noting that Sima does contain material from other large mammals, unlike the Dinaledi Chamber. There is general acceptance that Sima de los Huesos represents a charnel pit, for which large-brained archaic hominins (certainly more modern-looking than *H. naledi)* practised intentional disposal of the dead. Stringer highlights that such a mortuary behaviour in *H. naledi* is a surprisingly complex one for a hominin with a brain no bigger than that of *H. habilis* or a gorilla; others (myself included) would disagree.

The primatologist and professor of psychology at Emory University, Frans de Waal[17], writing in the *New York Times,* rightly takes Stringer to task for this assumption of linking brain size with complex social behaviour, and you may feel De Waal's frustration as he writes

> *…The problem is that we keep assuming that there is a point at which we became human. This is about as unlikely as there being a precise wavelength at which the color spectrum turns from orange into red. The typical proposition of how this happened*

> *is that of a mental breakthrough – a miraculous spark – that made us radically different. But if we have learned anything from more than 50 years of research on chimpanzees and other intelligent animals, it is that the wall between human and animal cognition is like a Swiss cheese. Apart from our language capacity, no uniqueness claim has survived unmodified for more than a decade since it was made. You name it – tool use, tool making, culture, food sharing, theory of mind, planning, empathy, inferential reasoning – it has all been observed in wild primates or, better yet, many of these capacities have been demonstrated in carefully controlled experiments.*
>
> *We know, for example, that apes plan ahead. They carry tools over long distances to places where they use them, sometimes up to five different sticks and twigs to raid a bee nest or probe for underground ants. In the lab, they fabricate tools in anticipation of future use. Animals think without words, as do we most of the time. Since they never stay in one place for long, they have no reason to cover or bury a corpse. Were they to live in a cave or settlement, however, they might notice that carrion attracts scavengers, some of which are formidable predators, like hyenas. It would absolutely not exceed the ape's mental capacity to solve this problem by either covering odorous corpses or moving them out of the way.*
>
> *The suggestion by some scholars that this requires belief in an afterlife is pure speculation. We simply don't know if* Homo naledi *buried corpses with care and concern or unceremoniously dumped them into a faraway cave to get rid of them. Apes appear to be deeply affected by the loss of others to the point of going totally silent, seeking comfort from bystanders and going into a funk during which they don't eat for days. They may not inter their dead, but they do seem to understand death's irreversibility. After having stared for a long time at a lifeless companion – sometimes grooming or trying to revive him or her – apes move on.*

Other researchers are more critical, but unfortunately seem fixated on the act of burial, on ritualistic symbol-laden interment, which we never suggest *H. naledi* as undertaking. This misunderstanding may come down to an inadvertent conflation of any form of mortuary behaviour with the notion of the burial act – they are categorically not the same. Curnoe[9] for instance, despite being happy with the taxonomic interpretation, comments: 'My 'nonsense-filter' also tells me that all the talk in the media about this new species *burying its dead and having human-like morality* [my emphasis], or that it dismantles one of the key pillars of human uniqueness, needs to be called out for what it truly is: absurd.' Views shared by Zollikofer who is quoted as saying[12]:

> *If you look more closely at the site where the skeletons were found, the cemetery theory becomes less probable. Think about it: according to the publication, there had never been direct access to the Dinaledi Chamber, where the bones were found. So our prehistoric human had to climb down there, squeeze through the narrow cave in complete darkness while dragging a corpse belonging to a member of its own species. From a purely practical standpoint, that makes no sense whatsoever.*

Well, archaeologists and primatologists would disagree with both Curnoe and Zollikofer, although specialists in palaeolithic burial have as yet not entered the fray…or perhaps they are winding up to address these issues through the correct medium – the peer-review process. Scientific discourse and peer-review is what drives the process of modern science, but being personally involved with the Dinaledi research has made me question the very nature of how the scientific community works, and in particular, how it fundamentally deals with evidence (ugly, inconvenient or otherwise) – the bedrock on which the modern scientific process is allegedly based.

Although I originally trained in archaeology and palaeoanthropology, and worked for many years in and around the sites of the Cradle of Humankind, the focus of my work in recent years has been forensic science, specifically anthropology and archaeology.[18] I am a specialist in the recovery and analysis of buried remains, burial environments (defined as the sedimentary and environmental context in which a body or bodies are contained post-mortem) and post-mortem processes.[19] I have recovered bodies from archaeological cemeteries, clandestine graves, homicides, fatal fires, and mass graves as the result of war crimes. My input into the Rising Star excavations was primarily as a forensic taphonomist and archaeologist.[3] As anyone with even a cursory interest in the forensic media circus which is 'real-life crime' or '*CSI*' will know, forensic science is deeply steeped in concepts of admissibility of evidence, and the application of basic sciences to the judicial or medico-legal process. In the USA, the admissibility of scientific evidence has been formalised through a number of legal case judgements, the most pertinent of which are those termed the Daubert Protocols or Daubert Standard.[20] In practice, Daubert is used by a trial judge to make an assessment of whether an expert's scientific evidence or testimony is based on scientifically valid reasoning or methodology and can properly be applied to the facts of the case.[21-23] Under Daubert, the factors considered in determining whether the science is valid are whether the hypothesis or technique in question (1) can be, and has been, tested; (2) has been subjected to peer review and publication; (3) has known or verifiable error rates; (4) has standards (professional or otherwise) controlling its operation; and (5) has attracted widespread acceptance within a relevant scientific community.[24]

So, does the science behind both the taxonomic and the context papers meet these definitions of admissibility? Of course it does. The science behind *H. naledi* is not controversial and is applied elsewhere in palaeoanthropology, evolutionary biology, geology, sedimentology, archaeology, etc., with widespread acceptance. Data are carefully collected (whether they be measurements of crania, teeth or femora, or elemental composition of cave sediments) and compared to existing standards and data sets. These data are subject to assessable error rates in terms of measurement error in data collection, or the effects of statistical sample size. Those undertaking the fieldwork and analyses are highly trained. And, most importantly, the work was peer reviewed before publication in a Thomson Reuters Web of Science accredited international journal. The evidence behind the science is sound – and presented in exceptional detail in the primary papers and supporting materials.[2,3] We present primary raw data, and *interpretations* of those data in such a way that scientists can use the published evidence to either accept the hypotheses presented or, if they so choose, re-analyse the data in a rigorous scientific fashion, and refute or (and this is the critical point) falsify our hypotheses. This process – observation, hypothesis, data collection, analysis and review, acceptance or rejection of hypothesis – is the cornerstone of how most modern science is conducted.

What critics perhaps fail to grasp is the difference between primary historical data (the physical evidence of the past – of which there is only one) and interpretation of intentionality or process in the past.[25] It is the creative tension between these two which basically defines the science of archaeology, particularly when it comes to understanding behavioural repertoires in hominins closely related to us.[26,27] Unfortunately this is complicated in species for which we have no clear modern analogue, falling as they are biologically somewhere between ape and human. To assist us in interpretation we triangulate data from a variety of disciplines – human archaeology, primate archaeology, ethology, evolutionary psychology, geology and taphonomy, to name a few.

Taphonomy is my area of primary interest, and an understanding of how we interpret the process of the introduction of bodies into the Dinaledi Chamber through taphonomic data is critically important. Because

of the confines of the Chamber, the Rising Star team applied forensic recovery and analytical techniques in order to extract the maximum amount of information about the formation of the assemblage, how the bodies had decomposed and skeletonised, and ultimately how they were introduced into the Dinaledi Chamber. As such, we adopted a multidisciplinary framework, bringing in a wide range of expertise in buried environments (more used to being applied to clandestine burials or forensic mass graves) to ensure that the most complete range of evidence was collected; the epistemological core of this is termed forensic taphonomy.[19] Whilst taphonomy can be considered the 'laws of burial',[28-33] forensic taphonomy[34] is perhaps unique in that the subject area marks a shift in the temporal nature of taphonomic studies, away from complex time-averaged assemblages accumulated over millennia, to shorter post-mortem timeframes spanning days to years, with the acknowledgment of humans as taphonomic agents and the emergence of the individual cadaver as the unit of analysis.[35] Much of the research is either derived from actual forensic casework, or applies neotaphonomy or actualistic taphonomy which concentrates on experimental work in the modern environment and applies its results to the past by analogy. This approach differs from 'classic' vertebrate taphonomy which is sometimes referred to as palaeotaphonomy[36]; this examines the context and content of depositional sites in great detail using temporo-spatial patterning, skeletal part representation and the pattern of skeletal damage as a means of interpreting formation processes. The neotaphonomic approach can be seen as primarily hypothetico-deductive in nature[19] and implicitly attempts to deal with issues of equifinality.[37] This is defined as reaching the same final state from different initial conditions and in different ways, without consideration of whether a system was open or closed. This is one of the great problems with taphonomic interpretation, in that whilst there is only one physical past, there may be multi-causative agents which produce that past, and as such can affect our reconstruction of an event or a taphonomic trace.[38-40]

To reiterate our interpretation based on physical evidence, the assemblage is unique by what it does not evidence – there is no evidence of peri-mortem breakage or trauma indicative of a fall or death trap as seen at sites such as Malapa[41], no carnivore modifications, no cut marks, no sub-aerial exposure or weathering indicative of death outside the cave, no evidence of water transportation of bodies or bones within the cave, and no evidence of burning or charring[3]. Despite an exhaustive search by a professional caving team and geologists, we failed to find any other plausible access points into the Dinaledi Chamber, and there is no evidence to suggest that an older, now sealed, entrance to the Chamber ever existed. Detailed surface mapping of the landscape overlying the cave system indicates that no large flowstone-filled fractures occur in the region above the Dinaledi Chamber. These findings, taken together with evidence that the Chamber in-fill was derived from in-situ weathering and from filtered exogenous clays and silts, indicates that there was not an easy or more accessible entrance from the surface into the Dinaledi Chamber at any time in the past (despite what some commentators choose to believe). No other animals found their way into the Dinaledi Chamber, indicating that it has always been a tortuous route. Hominins came into the Chamber as whole bodies, averaged over considerable time, and are not found distributed through the wider cave system as would be the case if they had wandered blindly into the labyrinth or been chased by predators. The accumulation of so many individuals in such an isolated specific locality, over a considerable period of time, suggests the route into the cave was intentional and deliberate. Following on from his comments on the taxonomy, Dan Lieberman suggests:

> *…It was a crazy deep cave, and getting in there wasn't easy. When they got in, there was nothing but the remains of this species. It's hard to imagine them getting there other than being intentionally deposited there. It smells to me like that's a form of burial, and it's a reasonable conjecture.*[10]

Again, Lieberman uses the word burial. This word is perhaps where part of the problem lies – with semantics – an entirely human problem, and one in which the meaning of 'burial' obviously causes confusion. Paul Pettitt[42] draws a very nice distinction with regard to mortuary behaviours in his excellent book *The Palaeolithic Origins of Human Burial*. He discusses the difference between non-human primate grief, suffering and loss as seen in the chimp communities of Gombe, Taï and Bossou, and Huntington and Metcalf's notion of the universality of human death. Huntington and Metcalf suggest that the diversity of cultural reactions to death and mortality is a measure of the universal impact of death – but that any reaction to it is not random; it is always meaningful and expressive. Pettitt draws the distinction that the chimpanzee reaction to death is certainly *expressive* (involving grooming, carrying, patting, vocalisations, etc.) but it is not *meaningful*. That then is the gulf between the mournful primate (expressive, even ritualistic at times) and the origins of complex mortuary behaviours (enriched with ritual *symbolism*, meaning and cognitive depth). Pettitt makes the point that mortuary practices form a wider set of transitions marked by ritual activity, but that how we contextualise them is important. Thus, there are a number of well-understood social concepts within our understanding of mortuary rituals, but these have differing physical expressions, functions and meanings. The most critical in the case of *H. naledi* is to differentiate among three most basic forms of mortuary practice: structured abandonment, funerary caching, and formal burial or inhumation. The former is the deliberate placement of a corpse at a certain point in the landscape, for simple functional reasons (protection from scavengers or predators). The second is structured deposition of a corpse, or parts of a corpse, in a chosen place, without modification of that environment, such as at the back of caves. Unlike structured abandonment, however, the place is given some 'meaning' beyond simple function. The third term is the creation of an artificial place for the purposes of containing a corpse, and involves at least three stages: excavation of an artificial pit or trench to serve as a grave, the interment of a body within the grave, and the covering of the body with the extracted sediment. Physically, and based on the contextual evidence, *H. naledi* may have practised funerary caching over multiple generations; we do not know where this behaviour fits on a scale from primate grief expression to symbolic meaning, but for palaeoanthropologists to dismiss the notion of a small-brained *naledi* showing a degree of social complexity in relation to mortality is arrogant and anthropocentric in the extreme.

None of the commentary and criticism in the media would lead me in any way to modify the working hypothesis of deliberate body disposal by *H. naledi.* Although research is ongoing on the assemblage and its context, no new data have come along to force us to reject our hypothesis. If at some point in the future such data do arise, then we will readdress the theory and, if appropriate, re-evaluate, rethink and raise new hypotheses which fit the data and present those to the scientific community through peer-reviewed publications. That process is how science works – it is provisional. But that provision is based on data and evidence…not belief. To reject a hypothesis because you simply do not like or do not believe the evidence presented to you is not science – it is pseudoscience at best and wilful ignorance at worst. It remains abundantly clear that many of the criticisms of the discovery and interpretation of *H. naledi* are not based on evidence; if they were, they would be published through the scientific peer-review process, and not through the popular media. They are either *ad hominem*, or perhaps caused by lack of appreciation or understanding of areas of science outside the comfort zone of the scientists concerned; I am not sure which I find the most depressing.

And with that, I leave you with closing remarks from Frans de Waal[17] – who views the discovery of *H. naledi* as an opportunity to re-contextualise our understanding of hominin behaviour back into the 'real' natural world, and for us not to view our own ancestral lineage as something unique and separate from our shared primate heritage. He closes his *New York Times* article with

> *…It is an odd coincidence that '*naledi*' is an anagram of 'denial'. We are trying way too hard to deny that we are modified apes. The discovery of these fossils is a major paleontological breakthrough. Why not seize this moment to overcome our anthropocentrism and recognize the fuzziness of the distinctions within our extended family? We are one rich collection of mosaics, not only genetically and anatomically, but also mentally.*

Well said.

References

1. Randolph-Quinney PS. A new star rising: Biology and mortuary behaviour of *Homo naledi*. S Afr J Sci. 2015;111(9/10), Art. #a0122, 4 pages. http://dx.doi.org/10.17159/sajs.2015/a0122
2. Berger LR, Hawks J, De Ruiter DJ, Churchill SE, Schmid P, Delezene LK, et al. *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. eLife. 2015;4, e09560, 35 pages. http://dx.doi.org/10.7554/eLife.09560
3. Dirks PHGM, Berger LR, Roberts EM, Kramers JD, Hawks J, Randolph-Quinney PS, et al. Geological and taphonomic context for the new hominin species *Homo naledi* from the Dinaledi Chamber, South Africa. eLife. 2015;4, e09561, 37 pages. http://dx.doi.org/10.7554/eLife.09561
4. Kivell TL, Deane AS, Tocheri MW, Orr CM, Schmid P, Hawks J, et al. The hand of *Homo naledi*. Nat Commun. 2015;6, Art. #8431, 9 pages. http://dx.doi.org/10.1038/ncomms9431
5. Harcourt-Smith WEH, Throckmorton Z, Congdon KA, Zipfel B, Deane AS, Drapeau MSM, et al. The foot of *Homo naledi*. Nat Commun. 2015;6, Art. #8432, 8 pages. http://dx.doi.org/10.1038/ncomms9432
6. Hawks J. *Homo naledi* fossil discovery a triumph for open access and education. The Conversation. 2015 September 28. Available from: https://theconversation.com/homo-naledi-fossil-discovery-a-triumph-for-open-access-and-education-47726.
7. Stringer C. The many mysteries of *Homo naledi*. eLife. 2015;4, e10627. http://dx.doi.org/10.7554/eLife.10627
8. Berger LR, De Ruiter DJ, Churchill SE, Schmid P, Carlson KJ, Dirks PHGM, et al. *Australopithecus sediba*: A new species of *Homo*-like australopith from South Africa. Science. 2010;328(5975):195–204. http://dx.doi.org/10.1126/science.1184944
9. Curnoe D. Did 'Rising Star' shine too bright? The Conversation. 2015 September 15. Available from: https://theconversation.com/did-rising-star-shine-too-bright-47501.
10. Pohle A. Will a new species change everything we thought we knew about human history? Boston.com. 2015 September 12. Available from: http://www.boston.com/news/science/2015/09/12/will-new-species-change-everything-thought-knew-about-human-history/3EdQluQFmChYNfDhm6ze9I/story.html.
11. Newton S. The cave of *Homo naledi*, or a textbook example of how to do science. National Center for Science Education Blog. 2015 October 20. Available from: http://ncse.com/blog/2015/10/cave-homo-naledi-textbook-example-how-to-do-science-0016693.
12. Von Mirbach J. *Homo naledi*: Scientific sensation or just a big show? DW.com. 2015 September 14. Available from: http://dw.com/p/1GVtx
13. Schwartz JH. Why the *Homo naledi* discovery may not be quite what it seems. Newsweek. 2015 September 10. Available from: http://europe.newsweek.com/why-homo-naledi-discovery-may-not-be-quite-what-it-seems-332804.
14. Hawks J. Is *Homo naledi* just a primitive version of *Homo erectus*? John Hawks Weblog. 2015 September 19. Available from: http://www.johnhawks.net/weblog/fossils/naledi/homo-naledi-homo-erectus-2015.html.
15. White TD. Physical anthropology at the millennium. Am J Phys Anthropol. 2000;113:287–292. http://dx.doi.org/10.1002/1096-8644(200011)113:3<287::AID-AJPA1>3.0.CO;2-8
16. Martin G. Bones of contention: Why Cal paleo expert is so skeptical that *Homo naledi* is new species. California Magazine. 2015 October 01. Available from: http://alumni.berkeley.edu/california-magazine/just-in/2015-10-05/bones-contention-why-cal-paleo-expert-so-skeptical-homo.
17. De Waal F. Who apes whom? The New York Times. 2015 September 15;Opinon pages. Available from: http://www.nytimes.com/2015/09/15/opinion/who-apes-whom.html?_r=1.
18. Randolph-Quinney P, Mallett X, Black SM. Forensic anthropology. In: Jamieson A, Moenssens A, editors. Wiley encylopedia of forensic science. London: John Wiley and Son; 2009. p. 152–178. http://dx.doi.org/10.1002/9780470061589.fsa058
19. Bristow J, Simms Z, Randolph-Quinney PS. Taphonomy. In: Ferguson E, editor. Forensic anthropology 2000–2010. Boca Raton, FL: CRC Press; 2011. p. 279–318. http://dx.doi.org/10.1201/b10727-10
20. Keierleber JA, Bohan TL. Ten years after Daubert: The status of the states. J Forensic Sci. 2005;50(5):1154–1163. http://dx.doi.org/10.1520/JFS2004241
21. Christensen AM. The impact of Daubert: Implications for testimony and research in forensic anthropology (and the use of frontal sinuses in personal identification). J Forensic Sci. 2004;49(3):427–430. http://dx.doi.org/10.1520/JFS2003185
22. Grivas CR, Komar DA. Kumho, Daubert, and the nature of scientific inquiry: Implications for forensic anthropology. J Forensic Sci. 2008;53(4):771–776. http://dx.doi.org/10.1111/j.1556-4029.2008.00771.x
23. Saks MJ, Faigman DL. Expert evidence after Daubert. Annu Rev Law Soc Sci. 2005;1:105–130. http://dx.doi.org/10.1146/annurev.lawsocsci.1.041604.115907
24. Risinger DM, Saks MJ, Thompson WC, Rosenthal R. The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. Calif Law Rev. 2002;90(1):1–56. http://dx.doi.org/10.2307/3481305
25. Trigger BG. A history of archaeological thought. Cambridge: Cambridge University Press; 1989.
26. Hodder I. Reading the past. Cambridge: Cambridge University Press; 1991.
27. Gamble C. The palaeolithic societies of Europe. Cambridge: Cambridge University Press; 1999.
28. Andrews P. Experiments in taphonomy. J Archaeol Sci. 1995;22:147–153. http://dx.doi.org/10.1006/jasc.1995.0016
29. Beherensmeyer AK, Kidwell SM, Gastaldo RA. Taphonomy and paleobiology. Palaeobiology. 2000;26(4):103–147. http://dx.doi.org/10.1666/0094-8373(2000)26[103:TAP]2.0.CO;2
30. Behrensmeyer AK. Taphonomy's contributions to palaeobiology. Palaeobiology. 1985;11:105–119.
31. Efremov JA. Taphonomy: A new branch of palaeontology. Pan Am Geol. 1940;74:81–93.
32. Fernández-Lopez SR. Taphonomic alteration and evolutionary taphonomy. J Taphonomy. 2006;4(3):111–142.
33. Lyman RL. Vertebrate taphonomy. Cambridge: Cambridge University Press; 1994. http://dx.doi.org/10.1017/cbo9781139878302
34. Haglund WD, Sorg MH. Introduction to forensic taphonomy. In: Haglund WD, Sorg MH, editors. Forensic taphonomy: The postmortem fate of human remains. London: CRC Press; 1997. p. 1–13.
35. Dirkmaat DC, Cabo LL, Ousley SD, Symes SA. New perspectives in forensic anthropology. Am J Phys Anthropol. 2008;47(suppl):33–52. http://dx.doi.org/10.1002/ajpa.20948
36. Quinney PS. Paradigms lost: Changing interpretations of hominid behavioural patterns since ODK. In: Rowley-Conwy P, editor. Animal bones, human societies. Oxford: Oxbow Press; 2000. p. 12–19.
37. Lyman RL. The concept of equifinality in taphonomy. J Taphonomy. 2004;2(1):15–26.
38. Bar-Oz G, Munro ND. Beyond cautionary tales: A multivariate taphonomic approach for resolving equifinality in zooarchaeological studies. J Taphonomy. 2004;2(4):201–220.
39. Marean CW, Domínguez-Rodrigo M, Pickering TR. Skeletal element equifinality in zooarchaeology begins with method: The evolution and status of the 'Shaft Critique'. J Taphonomy. 2004;2(2):69–98.
40. Munro ND, Bar-Oz G. Debating Issues of equifinality in ungulate skeletal part studies. J Taphonomy. 2004;2(1):1–13.
41. L'Abbé EN, Symes SA, Pokines JT, Cabo LL, Stull KE, Kuo S, et al. Evidence of fatal skeletal injuries on Malapa Hominins 1 and 2. Scientific Reports. 2015;5, Art. #15120, 11 pages. http://dx.doi.org/10.1038/srep15120
42. Pettitt P. The palaeolithic origins of human burial. London: Routledge; 2011. http://dx.doi.org/10.1093/oxfordhb/9780199232444.013.0022

Leadership: The invisibility of African women and the masculinity of power

AUTHOR:
Mamokgethi Phakeng[1]

AFFILIATION:
[1]Office of the Vice Principal, Research and Innovation, University of South Africa, Pretoria, South Africa

CORRESPONDENCE TO:
Mamokgethi Phakeng

EMAIL:
phakerm@unisa.ac.za

POSTAL ADDRESS:
Office of the Vice-Principal, Research and Innovation, PO Box 392, Unisa 0003, South Africa

KEYWORDS:
black African women; masculine power, transformation, academia

HOW TO CITE:
Phakeng M. Leadership: The invisibility of African women and the masculinity of power. S Afr J Sci. 2015;111(11/12), Art. #a0126, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0126

In a recent reply to the 'ideology of no transformation' in higher education, Cloete[1] shows that African women PhD graduates grew from 10 in 1996 to 106 in 2012 – a 960% increase. But read on, because this figure translates to a participation rate of only 2.4% per 100 000 of the relevant population. When white women are compared, their participation rate is 40 times higher. My argument, therefore, is that black African women are invisible in leadership and the 'masculinity of power' marginalises them.

The question to be asked then, is whether or not the picture changes for other areas of leadership beyond performance in advanced degrees? The answer is 'no'. At present, there is just one black African woman with an A rating from the National Research Foundation and only two who are B-rated in comparison to 12 white women who are A-rated and 67 who are B-rated. I am certainly not suggesting that black African women should be awarded A and B ratings even when they do not deserve them. Rather, I ask why is it that, after 21 years of democracy, only three black African women have risen to these levels. What is it that we are not doing to ensure that more black African women get to these levels? Although we thank Minister Pandor for the 42 SARChI chairs recently awarded to women, most of the recipients were white women. So irrespective of what the CREST data say, the point is that black African women remain invisible despite the increased number of black African women who hold doctorates.

Looking at the government's latest report on the status of women in South Africa, it is clear that despite educational gains, women are generally employed in low-skilled occupations. Gender inequality in South Africa varies across sectors, from 12th in the world in terms of political empowerment, to 83rd for economic participation, and 85th for educational attainment.

Although statistics are helpful, they are inadequate. On the matter of women and leadership, I believe that words are more important, perhaps because we have the figures and their power to speak often needs amplification. We know, for instance, that there are few women CEOs, that academic management is slow to appoint women, and that only the South African parliament can boast parity between men and women. In most cases, the numbers paint a dismal picture, despite the triumphalism of our democracy.

The representation of women in government has made us think that we are doing much better than we really are in the leadership stakes. We see rows of women in parliament, some taking prominent positions. In the media, women leaders in the political and judicial arena capture our imagination, and we congratulate ourselves on gender transformation. Of course these achievements are important, but in our space – research and higher education – the picture seems very different, especially when one begins to disaggregate both 'women' and 'leadership'. Which women? What kinds of leadership? My concern is that for all the demographic transformation in government, does it translate to anything approximating equity in leadership more broadly?

Even the relatively positive achievements of women in government mask the real relations of power within social and political movements, and Hassim[2] argues that 'questions concerning who has voice and agency …remain largely obscured'. This is not to say that women have not benefitted from the new procedural and institutional arrangements of the state after 1994: of course we have, but we have to see this transformation as very much still a 'work in progress'.

So today I choose words to help the numbers. In a recent book entitled *The Colour of Our Future*, Netshitenzhe[3] argues that black women have suffered 'triple oppression': first, as disadvantaged and marginalised in class terms; secondly as black in a racialised society that privileges the white minority; and thirdly as women in patriarchal African cultures.

It seems to me that black women are also affected by the assertions of Africanness which are a strong feature of contemporary politics, because their promotion of pristine forms of African culture generally valorises patriarchal practices. This important return to heritage as a way of restoring dignity to African identity sometimes depicts women in subject roles. Obbo[4], for example, cites the effect of modernity on women, as elite African men abandon their 'traditional' wives in villages to marry younger, educated women in urban centres. The irony, says Obbo, is that this is equally true of intellectuals who 'exhorted the celebration of African culture'.

Feminism among Africans has also not fared well, because of its epistemological whiteness which, understandably, has not had much purchase in South Africa. Feminism therefore often sets up a rivalry between black and white women, because it has generally been soft on racial and class inequality.[5]

Behind the numbers, therefore, there is an analysis of our historical and contemporary status as women. I can't speak for white women, but in academia they seem to have been the beneficiaries of democracy more than black African women, rising to senior professorial positions and even executive management, although there is a glass ceiling at vice-chancellor level for almost all women.

This brings me to my own experience, beginning in the township with its economic and educational deficits, making the best of apartheid higher education at the University of Bophuthatswana, and eventually earning a place at Wits. There, I had to contend with its forms of 'coloniality' in racial, gender, linguistic and curricular terms, until I graduated with a PhD. The academic totem pole has its own challenges, as I climbed the ranks to

professor, dean and deputy vice chancellor, against considerable odds. Hard work is important, but I have also learned the hard lesson that it is not sufficient in itself.

My experience has taught me that women have to be much better than men to land top jobs, they have to work doubly hard, and this gradually becomes an albatross as more women attain executive positions, because talented women are often constructed as ambitious rivals; and gender politics harshly depicts them in patriarchal stereotypes, caricaturing their personalities to curb their influence.

The 'Rhodes must fall' and the 'Open Stellenbosch' movements tackle the 'coloniality of power', as Grosfoguel[6] has coined it, but I don't see issues of gender equality being asserted in these current politics. In this sense, gender is trumped by race, which tends to make women compete with women, which often leaves black African women at the back of the colonial procession.

If I have lamented the invisibility of black African women in leadership, I am also challenging the 'masculinity of power' along with the coloniality of power in higher education and other sectors of the economy. Changing the numbers does not change influence, and leadership is not the outcome. So what do young up-and-coming women in science and research need to do to rise to leadership?

Firstly, they need to stay in the system. Often people do not think of research as a career. In fact, research and academia are not marketed as careers – perhaps because the salaries are not that great. Many of our young scientists aim to complete a PhD as a passport to another professional destination. This emigration of research capacity from universities weakens potential leadership in the academy and reduces our influence in knowledge creation and production.

It is true that a career in research is not immediately glamorous – it means exploring new, unknown paths with patience and persistence without losing focus or hope, even when setbacks and difficulties occur, as they frequently do. So my encouragement to all young researchers is to *stay the course*, no matter what the challenges might be – because challenges there will be. Young researchers, especially young, black African women, just stay the course!

Since black African women are almost entirely invisible, the possibilities of becoming, and then being, visible are an enormous challenge. Yet staying the course means not just your own progress and success, but the possible lowering of the glass ceiling for even younger potential black African women scholars. The prospects for leadership are immense: our critical mass efforts to attain improved higher education qualifications, greater research success and better positions open to us opportunities for intellectual growth, personal development and career advancement, and for feminising universities by changing ways of knowing, ways of doing and ways of leading to achieve a more equal society.

My advice to young black women scholars is to make the most of every opportunity that presents itself. Learn, connect with fellow researchers, make friends and take a firm hold on leadership possibilities, because that is how leadership starts – for all young people.

References

1. Cloete N. The PhD and the ideology of 'no transformation'. University World News. 2015 August 28; issue no: 379. Available from: http://www.universityworldnews.com/
2. Hassim S. Gender and policy-making: Terms of engagement. In: Gumede W, Dikeni L, editors. South African democracy and the retreat of intellectuals. Johannesburg: Jacana; 2009. p.183.
3. Netshitenzhe J. Interrogating the concepts and dynamics of race in public policy. In: Mangcu X, editor. The colour of our future: Does race matter in post-apartheid South Africa? Johannesburg: Wits University Press; 2015. p. 107–132.
4. Obbo O. But we know it all: African perspectives on anthropological knowledge. In: Ntarangwi M, Mills D, Babiker M, editors. African anthropologies: History, critique and practice. London: Zed Books; 2006. p. 154–169.
5. Marcus T. The women's question and national liberation in South Africa. In: Van Diepen M, editor. The national question in South Africa. London: Zed Books; 1988. p. 96–109.
6. Grosfoguel R, Georas CS. Coloniality of power and racial dynamics. Identities-Glob Stud. 2000;7(1):85–125. http://dx.doi.org/10.1080/1070289X.2000.9962660

Nurturing doctoral growth: Towards the NDP's 5000?

AUTHOR:
Nico Cloete[1,2,3]

AFFILIATIONS:
[1]Centre for Higher Education Trust, Cape Town, South Africa
[2]Department of Education, University of Oslo, Oslo, Norway
[3]DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Nico Cloete

EMAIL:
ncloete@chet.org.za

POSTAL ADDRESS:
CHET, PO Box 18094, Wynberg 7824, South Africa

KEYWORDS:
doctoral education; research intensive universities; differentiated doctoral production; knowledge economy

HOW TO CITE:
Cloete N. Nurturing doctoral growth: Towards the NDP's 5000?. S Afr J Sci. 2015;111(11/12), Art. #a0127, 3 pages. http://dx.doi.org/10.17159/sajs.2015/a0127

Although the matter still is contested, there is growing agreement amongst national policymakers and other socio-economic actors that the university is a driver for economic growth and development. This belief has to do with the role of the university in producing a highly-skilled and competent labour force, and in producing new knowledge. Both contributions are essential to the building of innovation and development for a national economy that aspires to be globally competitive. Both come together in the doctorate. There is also increasing recognition that research-intensive universities in low- and middle-income countries have an indispensable role to play in a differentiated academic system that can respond to the diverse requirements of contemporary development.[1]

According to Altbach[2], a differentiated academic system is needed for research-intensive universities to flourish:

> *The fact is that few if any developing countries have a differentiated academic system in place; and this central organisational requirement remains a key task…These institutions must be clearly identified and supported. There must be arrangements so that the number of research universities will be sufficiently limited so that funding is available for them and that other resources, such as well-qualified academics, are not spread too thinly.*

But as Altbach points out, research universities with strong doctoral programmes usually constitute a relatively small percentage of the higher education sector. In the USA, it is about 5% (220 research universities in a system of more than 4000 post-secondary institutions); in the UK, 15% (15 research universities amongst 100 universities); and in China, 3% (100 research universities out of more than 3000 institutions countrywide). In many developing countries there is often only one research university, and too many countries have none.[2]

A key component of a research-intensive university is a strong doctoral programme. Within the developed countries, with high doctoral production integral to their knowledge economy, there is a debate about whether an increase in doctoral graduates is required, and about what kind of contribution (research skills, innovation capacity, or something else) a doctorate makes to their knowledge economies. In sharp contrast, there are a range of developing countries, like Brazil and Mexico in Latin America, and China, Malaysia and Singapore in East Asia, where higher education, and specifically the doctorate, is seen as a development driver towards becoming a knowledge economy. These countries have invested massively in the expansion of doctoral programmes.[3]

In South Africa, the state, through its various bodies like the Department of Higher Education and Training (DHET), the Department of Science and Technology (DST) and the National Research Foundation (NRF), has responded to a knowledge economy discourse, even if the real economy is still mired in a low-skill, mineral extraction, export-dominated model. In a surprise move, in 2012, the National Planning Commission (NPC) made the target, of increasing the percentage of academic staff with a doctorate from the 2010 level of 34% to 75% by 2030, their number one priority in the section on higher education in their final report. This target replaces student enrolments and throughput rates as number one priority.

The rationale for this change between the draft and final report centres on the poor quality at the heart of poor performance in the higher education sector: 'The most important factor that determines quality is the qualifications of staff'[4]. The basic argument runs as follows: raise the qualifications of staff – in other words, increase the number of academics with doctorates – and the quality of the student outcomes will improve. This will also significantly improve throughput, the capacity to supervise higher degrees and, ultimately, the research productivity of the sector. In short, 'quality defined as having a doctorate is seen by the National Development Plan (NDP) as being the key that will unlock a virtuous cycle of effects'[5].

The NDP went further and set a target for the sector of producing more than 100 doctoral graduates per one million of the population by 2030. Broadly speaking, these numbers mean that the annual production of doctoral graduates will have to increase from 1420 per annum (in 2010) to 5000 per annum in 2030. This target was confidently repeated by Vice President, and Vice Chair of the NPC, Mr Cyril Ramaphosa, at the Transformation Summit in October 2015.

The DST, DHET and the NPC have all supported the notion of strengthening well-performing research-intensive universities in their policy documents. For example, the DHET White Paper[6] declared that 'in the university sector this continuum will range from largely undergraduate institutions to specialised, research-intensive universities which offer teaching programmes from undergraduate to doctoral level'.

In its rhetoric, the DHET is quite in line with the sentiments of the NDP[4] which urged government to:

> *… strengthen universities that have an embedded culture of research and development. They should be assisted to access private sector research grants (third stream funding) in addition to state subsidies and student fees, attract researchers, form partnerships with industry and be equipped with the latest technologies.*

In terms of financial policy, DHET has rewarded research output and doctoral enrolments and graduations through its subsidy formula. The DST and NRF have awarded scholarships, research chairs (through the South African Research Chairs Initiative) and Centres of Excellence based on merit and equity, and the research-intensive institutions have benefitted accordingly, as could be expected if merit is an important, if not the only, criterion. Whether these initiatives are enough to grow this sub-sector in line with the target set by the NDP is the question.

The question is not whether the higher education landscape should be differentiated; it is already differentiated in terms of a range of indicators, and has been so for some time.[7,8] The question is whether the government will support targeted differentiation in policy and fiscal terms. Left as it is, the system will grow modestly, or just drift. Policy inaction is also a choice.

A comprehensive study on doctoral education in South Africa offers an interesting lens on the differentiation debate, and on the categorisation of the university system.[7]

Differential doctoral production in South Africa

Table 1 shows that in 2012, seven universities produced 68% of all doctoral graduates and six universities produced only 1%. Table 1 also shows:

- Regarding graduation growth, during the 2008–2012 period, five universities grew their doctoral graduates by more than 20% annually while three posted 0% growth.
- Regarding efficiency, in 2006, four universities had a completion rate above 55% after 7 years, and seven universities had a completion rate lower than 35%. Another indicator of efficiency is the ratio of doctoral graduates as a percentage of academic staff with doctorates. In 2012, five institutions had a ratio higher than 0.3 and four universities had a ratio lower than 0.1.
- Regarding transformation, in 2012, five universities produced more than 90 black doctoral graduates each, while six between them produced only 15 in total. Five universities produced more than 75 female doctoral graduates each, and six universities just 13 in total.
- Regarding percentage of academic staff with a doctorate, in 2012, six universities had more than 50% of their staff with doctorates, five had fewer than 20%.

Of course, it is not the same institutions that perform well for all of the doctoral indicators. In 2012, Stellenbosch University and University of Pretoria both produced 200 or more doctoral graduates, while University of Fort Hare and Tshwane University of Technology grew at more than 35% annually during the 2008–2012 period. Stellenbosch University and the University of the Western Cape had a 60% completion rate for the 2006 cohort. Stellenbosch and Rhodes Universities had a ratio of above 0.39 doctoral graduates to academic staff with doctorates. Regarding transformation, the University of KwaZulu-Natal and Stellenbosch University produced more than 100 black doctoral graduates each and both of these two institutions each graduated more than 90 women. Finally, the Universities of Cape Town and the Witwatersrand had more than 55% of staff with a doctorate.

Looking at doctoral performance across the seven indicators, there is a grouping of at least seven traditional universities which consistently

Table 1: Indicators for performance in doctoral production

	GROWTH		EFFICIENCY		TRANSFORMATION		QUALITY
University	Doctoral graduates in 2012	Average annual growth rate: 2008–2012	Percentage of 2006 cohort graduating after 7 years	Ratio of doctoral graduates to academic staff with doctorates in 2012	Number of black doctoral graduates	Number of female doctoral graduates	Percentage of academic staff with doctorates
Cape Peninsula	24	16.6%	34.0%	0.19	19	5	16%
Cape Town	199	7.1%	55.8%	0.28	98	78	65%
Central	5	0.0%	30.8%	0.07	3	1	26%
Durban	6	18.9%	46.2%	0.07	4	5	15%
Fort Hare	43	40.6%	34.1%	0.36	41	9	38%
Free State	94	14.3%	50.7%	0.25	42	44	40%
Johannesburg	109	10.5%	55.0%	0.37	52	53	29%
KwaZulu-Natal	177	6.8%	50.3%	0.27	138	91	47%
Limpopo	17	5.0%	32.0%	0.13	16	7	16%
Mangosuthu	0	0.0%	0.0%	0.00	0	0	9%
Nelson Mandela	86	16.3%	51.4%	0.36	51	26	41%
North West	154	11.4%	52.1%	0.25	42	82	50%
Pretoria	200	2.7%	51.5%	0.32	83	99	49%
Rhodes	67	25.5%	50.6%	0.39	34	32	51%
South Africa	152	22.7%	24.5%	0.25	94	60	39%
Stellenbosch	240	18.9%	65.1%	0.46	107	96	53%
Tshwane	44	35.6%	51.1%	0.25	33	14	21%
Vaal	2	0.0%	0.0%	0.05	2	0	13%
Venda	4	18.9%	29.4%	0.04	4	1	31%
Walter Sisulu	3	10.7%	25.0%	0.04	2	1	12%
Western Cape	75	15.6%	59.8%	0.26	62	23	52%
Witwatersrand	150	9.1%	44.5%	0.25	92	56	55%
Zululand	28	21.1%	51.6%	0.35	26	11	27%

perform well: Stellenbosch, KwaZulu-Natal, Cape Town, Pretoria, Rhodes, Western Cape and Witwatersrand. There is a second grouping of nine institutions which are consistently in the top 10 in terms of at least five of the indicators: four traditional universities (Fort Hare, Free State, North West and Limpopo), four comprehensive universities (Johannesburg, Nelson Mandela Metropolitan, South Africa and Zululand), and one university of technology (Tshwane). The third grouping of seven universities performs indifferently across the indicators. This group consists of five universities of technology and two comprehensive universities.

It is worth noting that in terms of the official government classification of the system into traditional universities, comprehensive universities and universities of technology, four of the comprehensive universities (Johannesburg, Nelson Mandela Metropolitan, South Africa and Zululand) perform comparably with the second grouping of traditional universities as far as doctorate production is concerned. Regarding the universities of technology, Tshwane is the only one that could be classified as being in the doctorate-producing grouping.

High performance relative to indicators and goals has often been attributed to historical advantage (some universities are more than a 100 years old) and particularly to the apartheid practice of discriminatory allocation of resources and human capital.[6] Table 1 shows that in post-apartheid South Africa, some of these differences in performance – particularly in terms of the doctorate and research output – have persisted. More interestingly, the Table also shows that there is a differentiation occurring amongst the historically disadvantaged institutions. For instance, the Universities of the Western Cape and Fort Hare and the Mafikeng campus of North West University have become much more productive, and the Tshwane University of Technology is comparable to the traditional university grouping.

Also interesting to note, the 'transformationally challenged' universities Stellenbosch and Cape Town are, along with KwaZulu-Natal and Pretoria, the biggest producers of black and women doctoral graduates.

Tough questions

How will South Africa try to gear up the system to meet the target of 5000 new doctorates a year by 2030 set by Mr Ramaphosa and the NDP? These findings pose anew at least six policy questions that South Africa has struggled with since 1994 and continues to struggle with. Firstly, should the seven institutions that make up 30% of the system and produce 70% of the doctorates be regarded and recognised as having an 'embedded research culture', as research-intensive universities with strong doctoral programmes? If so, what are the policy levers for further strengthening such universities? Secondly, should nine institutions in the next cluster be encouraged and incentivised to develop and expand their research and doctoral education capacities? While this would broaden the base of the system, it would run counter to the international trend of singling out a smaller group of institutions worthy of high-level support. Thirdly, should the six institutions that produce 1% of the doctoral graduates be allowed to continue to offer doctoral programmes? In the USA and Norway for example, doctorate-awarding status is attained only after meeting fairly stringent conditions. Fourthly, if a decision is taken to increase full-time doctoral programmes (the main recommendation from Cloete et al.'s[7] study), with the attendant considerable costs involved, should these programmes be distributed across all institutions or be concentrated in the most efficient universities with demonstrated supervisory capacity? Fifthly, are we likely to meet the target without actively welcoming candidates and supervisors from the rest of Africa? Can this approach be encouraged in the face of the prevailing national mood? Last but certainly not least, can we afford not to incentivise (highly productive) universities to produce more black women doctoral graduates?

Acknowledgements

I thank Prof. Joe Muller for editorial assistance and Dr Charles Shephard for undertaking the statistical analysis for this Commentary.

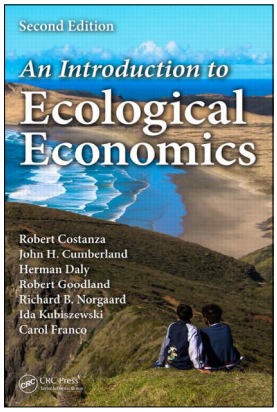
References

1. Olsson Å, Cooke N. The evolving path for strengthening research and innovation policy for development. Draft report. Paris: Organisation for Economic Cooperation and Development; 2013. Available from: http://www.oecd.org/sti/Session%204_The%20evolving%20path.pdf
2. Altbach PG. Advancing the national and global knowledge economy: The role of research universities in developing countries. Stud High Education. 2013;38(3):316–330. http://dx.doi.org/10.1080/03075079.2013.773222
3. Cyranoski D, Gilbert N, Ledford H, Nayar A, Yahia M. Education: The PhD factory. The world is producing more PhDs than ever before. Is it time to stop? Nature. 2011;472:276–279. http://dx.doi.org/10.1038/472276a
4. National Planning Commission. National development plan 2030: Our future, make it work [document on the Internet]. c2012 [cited 2015 Oct 23]. Available from: http://www.poa.gov.za/news/Documents/NPC%20National%20Development%20Plan%20Vision%202030%20-lo-res.pdf
5. Muller J. Discussion paper on the higher education and the NDP. Pretoria: Higher Education South Africa; 2013 [unpublished paper].
6. Department of Higher Education and Training (DHET). White paper for post-school education and training: Building an expanded, effective and integrated post-school system. Pretoria: DHET; 2013.
7. Cloete N, Mouton J, Sheppard C. Doctoral education in South Africa. Cape Town: African Minds; 2015.
8. Cloete N. Institutional clusters in higher education in South Africa. Presented at the Higher Education Summit of the Minister of Higher Education and Training; 2010 Apr 22–23; city, country. Available from: http://chet.org.za/files/DifferentiationChet_Web10May10.pdf

Ecological economics: From humanity's current dilemma to policies for a steady state economy

BOOK TITLE:
An introduction to ecological economics (2nd edition)

BOOK COVER:



AUTHORS:
Robert Costanza, John H. Cumberland, Herman Daly, Robert Goodland, Richard B. Norgaard, Ida Kubiszewski, Carol Franco

ISBN:
9781566706841 (hardcover)

PUBLISHER:
CRC Press, Boca Raton, FL; GBP54

PUBLISHED:
2014

REVIEWER:
Tony Leiman

EMAIL:
tony.leiman@uct.ac.za

AFFILIATION:
School of Economics, University of Cape Town, Cape Town, South Africa

POSTAL ADDRESS:
University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa

HOW TO CITE:
Leiman T. Ecological economics: From humanity's current dilemma to policies for a steady state economy. S Afr J Sci. 2015;111(11/12), Art. #a0128, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0128

Students entering a foundation course in economics are often told, 'economics is the study of choice; of decisions on the allocation of scarce resources between competing needs'. Ecological economics extends this definition to include choices made across time and between generations. In doing so it debates competing views of technological advances, questioning the security they can offer, and interrogates the natures of welfare, growth, development and the measurement of them. It is in the growth/development debate that ecological economics deviates furthest from the common conventions of environmental economics. Where modern economics still assumes a frontier economy, with boundless opportunities for growth, the ecological economist sees a world that is already nearly full.

It has to be stressed that this book is, as the title indicates, an *introductory* text. Anyone hoping for more will be disappointed. It is also a compendium with seven authors. This imposes a cost: the book feels like a series of linked pools rather than a unified stream. It is, however, a text whose authors have dominated the discipline for the last 30 years; it certainly summarises the 'state of the art' in the discipline.

The book has four sections. The first, 'Humanity's Current Dilemma', identifies the evidence that the current path of economic growth faces limits. It introduces the idea that economic policy should not simply target the expansion of outputs, but the reduction of raw inputs; making more with less.

The second section surveys the history of economics and ecology. Necessarily, much has been left out, and it would be easy to quibble about the balance, but for anyone new to the subject it manages to cover the basics in 60 pages. This section is not easy reading. It feels 'bitty', again probably as a result of multiple authorship; nonetheless, it does offer a reasonable summary of all but the contribution of conventional 'neoclassical' economics. The mention accorded to the neoclassical view is so brief that an uninformed reader might be forgiven for believing that modern economics has had nothing to say on the issue of sustainability. In reality, this topic has been a rapidly growing focus of study since the early 1970s, and a driving passion of major economists including the Nobel laureate Robert Solow.

The really interesting portion of the book is provided by the last two sections. From the perspective of a North American reader there is much in them that makes sense and is convincingly argued. However, viewed from a developing world standpoint, some of these arguments are far less persuasive.

The third section covers the principles and objectives of ecological economics. It compares growth and development, explaining that the former is merely increasing real GDP while the latter embodies improved national welfare. It also explains clearly the deficiencies in our current system of national accounts. Among the concerns raised are the ways in which both the market mechanism and the democratic process often undermine sustainability – both tending to support short-term advantage over long-term resilience. The authors discuss some of the alternatives to GDP and national income as measures of development from the original Tobin/Nordhaus 'measure of economic welfare' onwards, and include a very clear presentation of the theory underpinning Salah el Sarafy's proposed reforms to the national accounts.

The ecological threats posed by the modern market system and its use of advertising to promote consumption are extensively discussed. The authors argue that consumption and welfare are not closely linked, and indeed this seems to be the case in high-income states. However, in economies with high levels of poverty, this argument is more difficult to support. Also criticised is free trade. The authors seem to have adopted the narrow arguments made by the protestors at the 'Battle in Seattle' during the 1999 World Trade Organization meeting – trade does indeed cause job losses in the developed world, but the employment opportunities typically reappear in greater numbers in the developing world. They object that trade causes a pollution control 'race to the bottom' because dirty producers can undercut those bound to clean production techniques. Ironically, this argument ignores the World Trade Organization's willingness to allow tariff protection in such cases. They also complain that trade benefits those who pay low wages in the developing world, ignoring the way in which it has caused equalisation of worker incomes, driving up salaries in places like Japan, Singapore, Korea and now China, the very countries to which production has migrated. The authors have not addressed the basic notion that free trade rewards those who produce most efficiently, and that this should surely rationalise resource use and increase efficiency at a global level.

The final section, 'Institutions, Instruments and Policies', is the book's high point. Herman Daly has been writing on these issues for decades, and one sees his hand in much of the chapter. The 10-point synopsis of his policies for a 'steady state economy' could be usefully read to parliamentarians every year. He is the complete eclectic, happy to take good ideas from the workerist left (to limit the gap between worker and management incomes, provide more leave, etc.), from Keynes (to treat the International Monetary Fund as a bank clearing house with penalties imposed on those with substantial trade surpluses or deficits), from conservative Chicago monetarists (to do away with fraction reserve banking and to back all bank deposits 100%) and from modern China (to restrict population growth).

The three policy proposals for sustainability show similar eclecticism. The first, use of a natural capital (e.g. mineral deposits) depletion tax, adopts proposals made by Robert Solow. The size of the tax is not specified, but one assumes it would follow the Hartwick rule. The second policy proposal will be recognised by those familiar with current South African mining regulations, a 'precautionary polluter-pays principle' – specifically, a deposit system in which an operator is required to leave an interest bearing deposit sufficient to clear up expected damages in a worst-

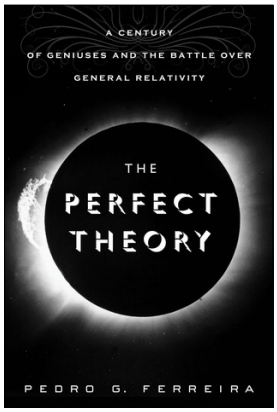
case situation. The third is the use of countervailing duties on goods produced using unduly 'cheap and dirty' technologies. Interestingly, it is recognised here that this proposal accords with longstanding GATT (General Agreement on Tariffs and Trade) guidelines, whereas the book's objections to trade mentioned earlier did not recognise these guidelines. Again, multiple authorship can lead to internal inconsistencies.

In summary, Costanza et al. have put together a useful handbook but not an enjoyable read. An author with a mission can educate and convert readers far more easily than can a consortium of writers. Aldo Leopold, Rachel Carson, Kenneth Boulding and Herman Daly all did this admirably in the past. This text unfortunately reads like the product of a committee, certainly serviceable, but not pretty.

A perfect theory

BOOK TITLE:
The perfect theory: A century of geniuses and the battle over general relativity

BOOK COVER:



AUTHOR:
Pedro G. Ferreira

ISBN:
9780547554891 (hardcover)

PUBLISHER:
Houghton Mifflin Harcourt, Boston, MA; USD28

PUBLISHED:
2014

REVIEWER:
Peter Dunsby

EMAIL:
dunsby@gmail.com

AFFILIATION:
Department of Mathematics and Applied Mathematics, University of Cape Town, Cape Town, South Africa

POSTAL ADDRESS:
Department of Mathematics and Applied Mathematics, University of Cape Town, Private Bag X1, Rondebosch 7701, South Africa

HOW TO CITE:
Dunsby P. A perfect theory. S Afr J Sci. 2015;111(11/12), Art. #a0129, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0129

Ask anyone on the street who Albert Einstein is or was and you will more often than not see their faces light up as they describe the eccentric man with crazy white hair and an infectious grin who changed our views about space and time, infinity and beyond – the very definition of genius. Karl Popper proclaimed that 'relativity' should be what all science should aspire to be. Like a work of art, Einstein's vision of the universe continues to fill us with wonder and delight. A reality filled with black holes and time warps, rapidly spinning neutron stars and gravitational waves, general relativity represents a magnificent unification of space and time, matter and energy, providing an explanation of some of nature's most profound mysteries and, as Pedro Ferreira describes it in his new book, very much 'The Perfect Theory'.

In *The Perfect Theory*, Professor Ferreira, a cosmologist at the University of Oxford tells the story of one of science's greatest accomplishments. He takes us from those early exploratory days of *gedanken* experiments, through the golden years of the late 1960s and 1970s when doing research in relativity was one of the trendiest things you could do as a theoretical physicist, all the way to the present day, when some are beginning to question its validity on the largest cosmological scales.

In the early 1900s, the scene was set for a scientific revolution. The clockwork deterministic world of Newton and Galileo, which had survived intact for more than 200 years, was about to be overturned by a 26-year-old patent clerk named Albert Einstein. The problem related to the way objects from elementary particles to rocket ships moved relative to each other. According to Newtonian mechanics, speeds should add up in a straightforward way. If two cars moved towards each other, for example, the driver of one car would observe the other car to be moving at their speed plus the speed of the approaching car. The consensus at the time was that this fundamental law of addition should apply to everything, even particles of light called photons. Unfortunately, Maxwell's theory of electricity and magnetism painted a very different picture. The speed of light in a vacuum seemed to be the exception to this basic rule of addition. All observers, irrespective of their state of motion, should observe light to be travelling at the same speed. From 1887 to 1902, Michelson and Morley and others confirmed this exception experimentally, which led to a fundamental and irreconcilable mismatch between two of the major pillars of classical physics – a crisis only Einstein with his natural intuition, imagination and a pinch of common sense could resolve. How did he resolve it? Simply, by keeping the speed of light constant and determining what the consequence would be for the Newtonian equations of motion. It seems almost obvious now, when one looks back, but the result of making this rather simple change to the framework of physics was revolutionary. Not only did it unify mechanics and electromagnetism, it banished the Galilean notions of absolute space and time to the dustbin of history.

It is worth mentioning at this point that Einstein's 1905 paper on special relativity was one of four remarkable pieces of work that shook the world of physics that year. One explained how to measure the size of molecules in a liquid, the second determined their movement – Brownian motion – and a third described how quanta of light could knock out electrons in metals (the photoelectric effect became the basis of quantum mechanics and electrochemistry). It was this idea, not special (or general) relativity, that eventually won him the Nobel Prize. At this point, most mortals would have put down their pen and taken a well-deserved break, but not Einstein. A few months later, he pointed out in a fifth paper that matter and energy are interchangeable at the atomic level through $E=mc^2$, arguably the most famous equation in history, providing the scientific foundations of the nuclear age. And with that, Einstein's 'miracle year' came to an end.

Einstein and others now turned their attention to the problem of how to make gravity fit into this new spacetime description of nature. As Ferreira explains, Einstein's real genius lay in his ability to take well-known observations about the way the universe worked and then use his imagination to construct outlandish thought experiments which revealed deep physical insight and concrete testable predictions. It was Galileo's fabled Pisa experiment which lay the foundations of Einstein's famous lift experiments. The fact that balls of different masses fell to the ground at the same rate implied that gravity was just another inertial effect and that experiments performed in small regions of spacetime should be indistinguishable from what would be obtained in special relativity (in the absence of gravity). This equivalence of gravity and acceleration was the critical ingredient for building a theory of gravity which included special relativity as a local limit. The lift experiments also pointed to some key predictions of this more general theory. For example, imagine accelerating a lift in empty space and observing the motion of a light ray within. The light ray would appear to curve down. Now magically transport the lift to the surface of the earth. The equivalence of acceleration and gravity would imply that the light ray would also bend in the earth's gravitational field. Without any equations, Einstein was able to predict gravitational lensing, a powerful method now used to constrain cosmological models. Finally, take the lift and throw it down a mine shaft. Two particles initially at rest appear to move towards each other, suggesting that gravity is described by spacetime curvature, so do not move on lines that remain parallel.

Thought experiments are one thing, but converting these visualisations into a set of precise equations is another matter. With help from Marcel Grossman, Einstein was able to build a mathematical framework based on Riemannian geometry able to describe how mass and energy distort the fabric of spacetime and how particles move on curved spacetime. Ferreira compares this undertaking to 'learning Sanskrit from scratch and then writing a novel in it'. With the gravitational field equations in the bag, general relativity could be used to determine the correct orbit of Mercury, which Newton's theory of gravity never quite got right because it did not take into account the curvature of spacetime. A paradigm shift in the understanding of gravity and how it affected the cosmos had occurred.

With general relativity largely accepted by the physics community and Einstein now one of the most famous men in the world, the exciting quest to discover its consequences started in earnest. From the trenches of the

First World War, Karl Schwarzschild discovered an exact solution of the field equations describing the spacetime around spherically symmetrical objects such as stars and the cores of galaxies. This discovery provided a way of determining the correct motion of planets around our sun, and how light moves on the curvature of spacetime – a key step in the development of the theory of gravitational lensing which is used today to map the distribution of dark matter. Even in weakly gravitating systems like the earth, this solution is required to calculate the precise orbits of GPS satellites. You see, general relativity is even important for navigating through cities using our smartphones!

The discovery of the Schwarzschild solution also represented the beginning of an exploration of the most exotic objects in the universe – black holes. The idea that the end state of stars could lead to a singular point, at which gravitational tidal forces become infinite and the laws of physics break down, was one of the most radical consequences of Einstein's theory. Even more mysterious was the fact that singularity was hidden from the rest of the universe by a surface (called the horizon) from which not even light could escape. It took decades of research before a complete understanding of this solution was obtained and even now there is still controversy over the exact nature of the surface that hides the singularity from prying eyes.

During the late 1960s and 1970s, Einstein's description of nature's most mysterious force reached maturity. By now numerous exact solutions had been found to describe objects from wormholes to cosmic strings and even the universe itself. However, it was on the smallest possible scales, where the determinacy breaks down and the mysterious quantum world rules supreme, that general relativity finally met its match. It seems that despite decades of searching for a self-consistent theory of quantum gravity, we are still a very long way from describing what happened at the moment of the Big Bang itself.

If this is not troubling enough, even the classical theory might have issues on the largest possible scales, when one deals with the universe as a whole. In 1998, our view of the expansion history of the universe was fundamentally changed – a result which rocked the cosmology community to its core. Two international groups of astrophysicists collected data on the light output from distant stellar explosions called type Ia Supernovae in order to measure how the universe has expanded over its 13.7-billion-year lifetime. Based on the simplest cosmological model at the time, which they originally had proposed, Einstein and de Sitter expected that the expansion would be slowing down as galaxies pulled toward each other as a result of their gravitational attraction. What they found was quite remarkable and such a surprise that initially both teams did not believe their results.

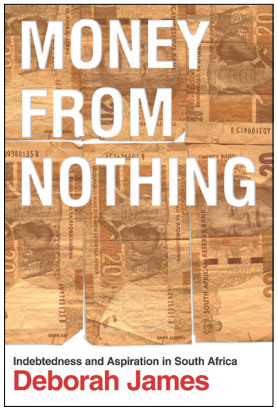
The interpretation of these observations was startling. Firstly, the Einstein–de Sitter model was completely ruled out, and secondly, if the geometry of the universe was still taken to have Friedmann–Roberton–Walker form, an almost preposterous conclusion resulted: the universe has an accelerated expansion today! Now, the problem is that ordinary matter cannot do this, so some new form of matter with negative pressure (known as dark energy) must be dominating the expansion history today. The nature of dark energy is without doubt the most puzzling mystery in cosmology and astrophysics today, and finding an explanation to what it is (or is not) is almost certainly the Holy Grail of theoretical cosmology. Part of the mystery is that we really do not have any clue what it is or whether we will even be able to find an answer to this deep fundamental question. New physics, or a change in how gravity behaves, might be required to resolve this impasse.

The last chapter of Ferreira's book, 'Something is going to happen', brings us almost full circle to the beginning of this story. Like then, the time is ripe for the next great revolution in physics. Only time will tell, but future developments in gravitational physics and cosmology might lead to an explanation of what the dark side of the universe is, whether the universe ends in a cataclysmic Big Rip or in frozen isolation, and whether we might again be forced to change the fundamental laws of physics. It is even possible that 'The Perfect Theory' might not be so perfect after all.

A race to the top and the fear of falling

BOOK TITLE:
Money from nothing: Indebtedness and aspiration in South Africa

BOOK COVER:



AUTHOR:
Deborah James

ISBN:
9780804791113 (cloth)
9780804792677 (paperback)
9780804793155 (digital)

PUBLISHER:
Stanford University Press, Stanford, CA
Rights for southern Africa: Wits University Press, Johannesburg, ZAR320

PUBLISHED:
2014

REVIEWER:
Iain Edwards

EMAIL:
edwardsi@gibs.co.za

AFFILIATION:
Gordon Institute of Business Science, University of Pretoria, Pretoria, South Africa

POSTAL ADDRESS:
GIBS, PO Box 787602, Sandton 2146, South Africa

HOW TO CITE:
Edwards I. A race to the top and the fear of falling. S Afr J Sci. 2015;111(11/12), Art. #a0125, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0125

The rapid rise of a growing African middle class is one of the most dramatic transformations of the post-1994 period. Its emergence has long been a cornerstone of African National Congress (ANC) policy – the 'patriotic bourgeoisie'– with its rise being a central feature of former President Mbeki's economic policies. This middle class has long been the focus of considerable political, popular and scholarly interest. Much has long been placed on the shoulders of this newly expanding vibrant and powerful class. A class that is viewed as: central to de-racialising the political economy of apartheid; the driving force in economic growth; key to the ANC's continued power in government (a very senior member of the ANC alliance and government minister presented his PhD on the very question of the patriotic bourgeoisie); the bulwark against revolutionary threats; and more.

Not surprisingly, much vital scholarly research, first led by anthropologists, has focused on this long aspirant class. Leo Kuper's scholarly and accessible *An African Bourgeoisie*, published in 1967, is rightly a classic. Given the centrality of this long-developing class's place in post-apartheid South African transformation, the contemporary period is also much studied in various criss-crossing disciplinary approaches ranging from economics and cultural anthropology to politics and sociology. James' work is an important monograph within this latter scholarly focus. As James, an anthropologist, sees matters, contemporary South African public culture offers three rather differing perspectives on this class. There is the notion of the 'black diamonds': ever more successful wannabe plutocrats, with access to hitherto undreamt of state power, political influence and, as a direct consequence, unimagined riches. A comment by Smuts Ngonyama, one of former President Mbeki's advisors, has entered the political lexicon: 'I didn't join the struggle to be poor' (James does not not mention him by name and uses 'revolution' not 'struggle'). Life-like characters from this new class also provide the dynamic drama in South Africa's television soapies. And, there is another publicly expressed view too: of avarice and brash self-seeking conspicuous celebrity. Yet increasing indebtedness will soon see their lifestyles crashing down in ruins. Despite entitlement and near limitless aspiration, profligacy will quickly result in wealth receding ever further from this self-absorbed parvenu class.

Beyond the plutocratic high end (it exists and shows itself too), soap opera dramas, and *Schadenfreude*, there are other increasingly more topical and pressing public perspectives, which also need important mention. Aimed specifically at this new consumer class, media, advertising and marketing campaigns, and magazines such as the high-end *Destiny*, present heroic, expensively dressed, urbanely chic icons of triumphal progress: here is the aspirational image of life fulfilment – destiny – at the top. Recently, persons from within this new class have provided vivid insights into the economic health and social places and responsibilities of this new class, including what is termed the 'black tax'. Interlinked to these insights is a further issue. As South Africans look towards local government elections in 2016, many, including respected political analysts, place great store in the black middle class's shifting political allegiances from the ANC, thereby allowing the increasingly multiracial Democratic Alliance and the newly formed Economic Freedom Fighters to gain increased influence, if not also control, over key urban metropolitan councils, including those in Gauteng – the country's economic heartland and the home of the densest concentration of this new class.

James' monograph is based on fieldwork, which included interviews, conducted during a pivotal time: 2008 and 2009, and in three areas. These areas are Langa, one of the first African townships in Cape Town; in the greater Bushbuckridge region in the Limpopo Province, where vibrant new sprawling towns have quickly grown; and in urban areas in Johannesburg and Midrand. Added to this fieldwork is good use of source material from South Africa's ubiquitous talk radio stations and other material from public media, including newsprint.

The key strength of this book lies in this fieldwork and James' ability to show and analyse the internal domestic financial, social and cultural dynamics of this class. With the global backdrop being the easy expansion of credit and consumer-led boom periods, here we get insights into South Africa's newest class, both before and after the 2008 financial crisis. Its roots are professional, nascent middle class, and aspirational plebeian too. But as a class it has risen quickly and established a set of often hugely diverse values. Religious senses range from Methodism to the *Prosperity Bible* – a very wide spectrum. Lifestyles have rapidly extended across the purchase of fixed property, house rentals, new household domestic goods, motor vehicles and other consumer goods, often of conspicuous kinds. But it is more complicated. This class is not based on a nuclear family. Its roots are as much single-parent and inter-generational extended family networks. Core concerns relate to needs and investments, not just wants, and duties and responsibilities. Social dynamics are not simply the pursuit of celebrity bling and brashness. Much stress is placed on the education of children and the social welfare needs of a wider extended family, including unemployed and other dependants. The realms of 'traditional culture', for example for bride wealth, weddings or funerals, become idealised and hyper valuated into often crippling cash payments. Not surprisingly, considering the apartheid past, credit is vital to this class. And this credit is rotating and revolving, with debtors often continuously seeking new sources of false liquidity in order to pay off often usurious existing credit debts. The democracy dividend is questioned.

And these characteristics reveal the critical roles of the state. The ANC government has ensured the rise of this new class, through black empowerment, tender policies and the direct employment of a new civil service and in related and dependent sectors. Yet, significantly, the state has left the formal banking and retail sectors, and partly regulated private credit market, to provide financing. With deep irony, James notes how the original sources of credit were largely recently retrenched white Afrikaans-speaking civil servants, amassing their retrenchment lump sums to form micro-lending enterprises. Their borrowers were the new civil service. The now well-known latter-day variant of those micro-lenders are the *mashonistas*, whose influence in plebeian debt-ridden credit cycles became tragically evident in the judicial commission of enquiry into the Marikana killings. Despite continual re-regulation,

this private credit market operates in conjunction with the larger formal credit, banking and retail institutions.

James is at her most powerful tracing the human dynamics of this class. Aware of the ever pervasive public examples of material success, there is a dynamic competitive edge to the pursuit of progress. Therein is a complicated mixture of seeking state grant or preferment, black employment equity policies and self-help. Attitudes towards debt span the range from negative to positive views. Throughout the interviews certain keywords are common; under the appearances, and indeed realities, of upwardly mobile success, words such as 'plight', 'problem', 'stress' and 'suffering' are oft used. The cynic would say, 'Welcome to the middle class', but given the historical legacies of this class's long struggles, and inabilities to secure fixed capital bases (for example, in real estate – even apartheid's 99-year-old leases were intended not to be freehold tenure) – the vicissitudes of aspiration, debt and progress make for precarious life choices.

In dealing with this complex world, James offers humane insights. But the monograph is often repetitive and overwritten. Differences between academic interpretations often appear forced. And the analysis of the new African middle class and the South African suburban housing market is weak, and ignores the greatest change in the South African contemporary real estate market: suburban African middle-class home ownership.

There are wider contexts to these important accounts and analyses, some of which South Africans are becoming all too aware of. Surprisingly so, James hardly touches on these. They include the failure of African Bank; the continual scandals and disasters relating to Ponzi schemes masquerading as stokvels or cooperatives; the *Prosperity Bible* and the popularity of the new evangelical churches; casino companies reporting record profits; increasing indebtedness of South African households; rising levels of labour militancy; how indebtedness is assuming an ever more central place in labour relations; the reluctance, despite militancy, of labour – including the civil service – to embark on legal but unaffordable strike action; and courageous battles to prevent usurious lending and impoverishment through garnishee orders. Granted, this latter long-fought battle – led by Wendy Appelbaum – one of the country's wealthiest persons, and not the ANC or its communist or trade union federation allies, has just been won. But case study monographs need contextual relevance, and the information is easily accessible.

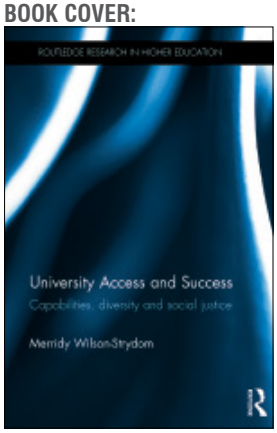
There also are historical contexts. Stokvels and cooperative societies have long been a central aspect of African rural and plebeian life. They were often linked to the *isolomuzi*, local level community associations with strong affiliation to the ANC, and church communities. Indeed, many of these links were part of the origins of the ANC itself. By the 1940s, as the pace of urbanisation increased so dramatically, such cooperative societies became recognised within African political circles as the sources of credit for ambitious self-help schemes: for artisan training, for the growth of African-owned enterprises and for African factory-owning industrialists. Here was the power of 'New Africa'. As is now recognised in political parties and government, this was precisely the way in which Afrikaner Nationalism acquired its powerful economic and financial meaning and muscle. And key to this acquisition was a bank: Volkskas (Peoples' Bank).

Despite great, often liberal, often patronising, or now presumptuous, political stress and expectations being placed on the growth and importance of the African middle class, this class has, and still struggles, against much adversity to develop. Why, after 20 years of ANC-led government, have more sustainable ways, including ones embedded in plebeian cycles of community aspiration and saving, joined to capital investment markets, and the burgeoning financial services sector, not been developed? How is it that an ethically based, properly financed and focused private sector banking initiative has not developed? Just how appropriate are the public and private sector's developmental and social policies? Can a properly funded and founded people's bank not be a sustainable alternative to social grants? And is the vital African middle class to be cast between the new 'race to the top' and the old *i-kasi* 'fear of falling'? Is this the rightful destiny of this class?

University access and success: An issue of social justice

BOOK TITLE:
University access and success: Capabilities, diversity and social justice

BOOK COVER:



AUTHOR:
Merridy Wilson-Strydom

ISBN:
9781138017771 (hardcover)

PUBLISHER:
Routledge, Oxford; GBP95

PUBLISHED:
2015

REVIEW TITLE:
University access and success: An issue of social justice

REVIEWER:
Alejandra Boni

EMAIL:
aboni@ingenio.upv.es

AFFILIATION:
INGENIO Institute, Universitat Politècnica de València, Valencia, Spain

POSTAL ADDRESS:
INGENIO, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

HOW TO CITE:
Boni A. University access and success: An issue of social justice. S Afr J Sci. 2015;111(11/12), Art. #a0130, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0130

University Access and Success represents an extremely valuable contribution to the higher education literature related to access to universities. But, contrary to the mainstream approaches to access which rely on school performance and admissions tests, Wilson-Strydom poses at the centre of the analysis *the issue of social justice*.

As Wilson-Strydom explains: 'When we consider issues of justice or injustice, we cannot merely ask whether different people have achieved the same outcome, but rather, whether different people have had the same opportunities to achieve this outcome' (p. 151). This shift from outcomes to opportunities is very different from other approaches to access and readiness in higher education (see Chapter 2).

The theoretical framework is provided by the capabilities approach, which is well known thanks to the writings of Nobel laureate Amartya Sen[1] and the contributions of philosopher Martha Nussbaum[2]. Capabilities could be defined as the substantive freedoms or real opportunities that people have to lead the kind of lives they have reason to value.[1] Those real opportunities are shaped and mediated not only by the personal characteristics of people (personal conversion factors in the capabilities literature), but also by the social context (social conversion factors), which is very different from other approaches to readiness which are more focused on the personal characteristics of the learners (see Chapters 3 and 4).

Wilson-Strydom explores 'the capability to participate in higher education' – the real opportunities first-year students have to participate in higher education. She conducted excellent research between 2009 and 2014 with high-school learners and first-year university students in the Free State Province in South Africa. The data collection methods included a quantitative survey, individual interviews, focus groups, visual methods and written reflections. From a methodological perspective, this book is an excellent contribution in the educational research domain and it brings a richness of data that allow the author to explore in a deeper way the personal and social conversion factors that influence the capability to participate (see Chapters 5 and 7).

Wilson-Strydom draws a vivid picture of students entering university, in which feelings such as loss, fear and confusion have a big impact on their confidence to learn. Students' personal conversion factors – such as the will to learn, the confidence to learn and home language – are affecting the readiness of the students and so need to be carefully considered by universities.

Moreover, Wilson-Strydom highlights the social conversion factors that affect students' opportunities. The school context is of paramount importance; and absent or poor-quality teachers, a general lack of resources, limited subject choices, and few opportunities to engage with a diversity of ideas and people are major influencing factors in the capabilities of learners.

Also, the socio-economic context has proved to be relevant in the way learners spend their time outside school. Learners living in township areas spent more time than other students walking to school and caring for family members than engaging in educational enrichment activities. This finding is especially remarkable for girls: they tend to have more responsibility than boys at home, but at the same time girls tended to be better prepared academically for university than boys; in that sense, gender appears to operate in 'somewhat conflicting ways' (p. 144).

It is not only the richness of data and its interpretations that make this book worthwhile, but its theoretical contributions are also extremely valuable. Rooted in previous work that has used the capabilities approach in higher education,[3-5] Wilson-Strydom proposes 'a list of capabilities for the transition to university' (p. 131) formulated using a comprehensive theoretical analysis and taking into account the voices of learners that emerged from the research.

The list, presented in Chapter 6, includes the following capabilities: (1) practical reason, (2) knowledge and imagination, (3) learning disposition, (4) social relations and social networks, (5) respect, dignity and recognition, (6) emotional health and (7) language, competence and confidence.

This list could provide a normative framework for understanding what is needed for access to university. This framework poses at its centre, the well-being of students, and raises issues of equity, participation and diversity. Also, as Wilson-Strydom points out (p. 132), the capability of language takes us beyond the traditional focus of access to research on measurable performance as a basis for making admission decisions and predicting the likelihood of success.

In that sense, this book offers interesting avenues for action in the university realm. As Wilson-Strydom notes in Chapter 8, universities should confront injustice, and even if the changes are complex and difficult to achieve, scholars and researchers committed to social justice should take action to focus on the 'remediable injustices we see around us'[6]. Coherently, Wilson-Strydom not only proposes some ideas, but she also translates them into practice and provides some preliminary results of a workshop with students in a Social Work department. Her honest and coherent commitment to social justice that you can perceive through the pages of this book, makes this contribution, in my view, even more valuable.

References

1. Sen A. Development as freedom. New York: Knopf; 1999.

2. Nussbaum M. Women and human development: The capabilities approach. Cambridge: Cambridge University Press; 2000. http://dx.doi.org/10.1017/CBO9780511841286

3. Walker M. Higher education pedagogies. Berkshire and New York: Society for Research into Higher Education and Open University Press; 2006.

4. Boni A, Walker M, editors. Universities and human development. A new imaginary for the university of the XXI century. London: Routledge; 2013.

5. Walker M, McLean M. Professional education, capabilities and the public good. London: Routledge; 2013.

6. Sen A. The idea of justice. London: Allen Lane Penguin Books; 2009.

A critical review of ionic liquids for the pretreatment of lignocellulosic biomass

AUTHOR:
Prashant Reddy[1,2]

AFFILIATIONS:
[1]Department of Chemistry, Durban University of Technology, Durban, South Africa
[2]Sugar Milling Research Institute NPC, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Prashant Reddy

EMAIL:
reddyp2@gmail.com

POSTAL ADDRESS:
Department of Chemistry, Durban University of Technology, PO Box 1334, Durban 4000, South Africa

DATES:
Received: 01 Mar. 2015
Revised: 21 Mar. 2015
Accepted: 25 Mar. 2015

KEYWORDS:
biomass; cellulose; biorefinery; lignin; fractionation

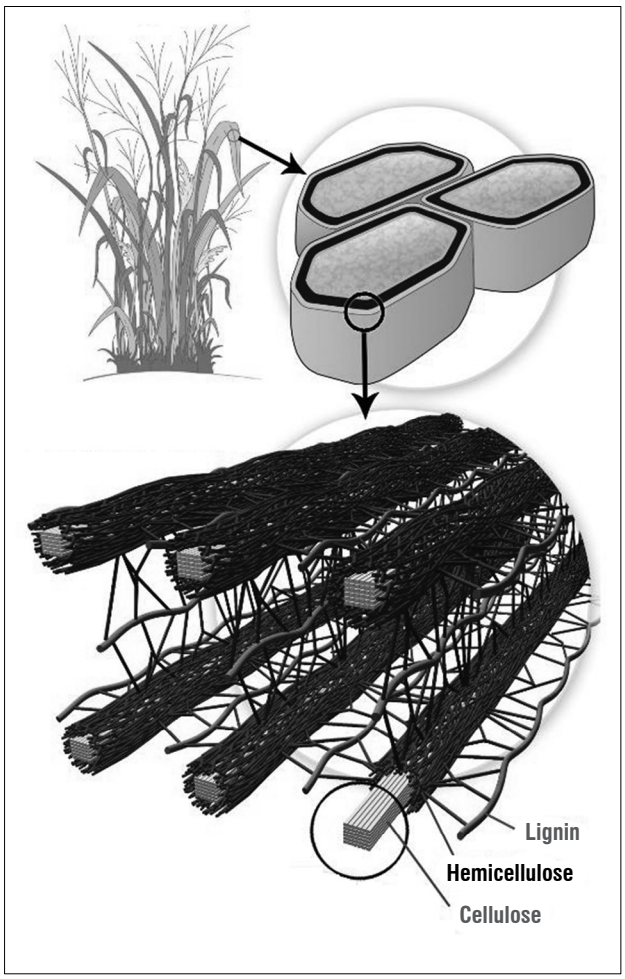
HOW TO CITE:
Reddy P. A critical review of ionic liquids for the pretreatment of lignocellulosic biomass. S Afr J Sci. 2015;111(11/12), Art. #2015-0083, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20150083

Ionic liquids have been the subject of active research over the course of the last decade and have in the past been touted as one of the most promising technologies for revolutionising the chemical and petrochemical industries. The sheer abundance of potential ionic liquid structures coupled with their tuneable physico-chemical properties has endeared ionic liquids to the scientific community across a broad range of disciplines with potential applications that include pharmaceuticals, electrolytes, thermal energy storage media and liquid mirror telescopes. Within the context of a biorefinery for the production of biofuels and other bio-based products from renewable resources, the unique abilities of some ionic liquids to selectively dissolve biomass components or whole native biomass have been demonstrated. This ability has sparked extensive investigations of ionic liquids for the pretreatment of different biomass types, particularly for the production of cellulosic biofuels. However, the esoteric nature of ionic liquids persists and constructing a fundamental framework for correlating ionic liquid structures with useful applications remains a significant challenge. In addition to the above, the more practical challenges of toxicity, high costs, high viscosities, low solids loading and complex recycling are key factors hindering the wide-scale uptake of ionic liquids as pretreatment solvents in a commercial biorefinery. This critical review provides insights from academic studies and the implications thereof for elevating ionic liquids from the status of 'promising' to 'commercialisable' in the pretreatment of biomass. It is vital that key hurdles for the commercialisation of ionic liquids in the form of high costs, high viscosities, poor water tolerance, toxicity, low solids loading and recovery/recycling be addressed.

Introduction

Lignocellulosic biomass is the most abundant plant material on the planet[1] and is a veritable treasure of renewable resources for the production of energy, chemicals and materials. The biomass-producing forestry and sugarcane industries in South Africa are mature industries that could benefit greatly from diversification of their product portfolios to achieve long-term sustainability. Based on the amount of sugar cane crushed in the last 5 years,[2] the South African sugarcane industry produces an average of ca. 2.6 million tonnes per annum of dry bagasse (Figure 1), the fibrous residue remaining after sucrose-rich juice has been extracted from sugarcane stalks. The major use of bagasse is boiler fuel with limited value addition for products that include furfural, animal feed and paper.[3] Deriving additional value from agro-industrial biomass such as bagasse involves addressing the inherent recalcitrance of biomass to access the lignocellulosic building blocks of cellulose, hemicellulose and lignin (Figure 2)[4] which possess useful chemical functional groups. Achieving an efficient, simple and cost-effective pretreatment step is a key enabler for the development of sustainable value chains for biomass-producing industries by unlocking value addition steps to produce saleable bio-based products. Amongst the various types of biomass pretreatments, chemical pretreatments (traditionally acid, alkali and organic solvents) have been quite popular[5], with one of the most studied classes of chemicals in recent times being ionic liquids.

Figure 1: Sugarcane bagasse.



Source: Oak Ridge National Laboratory[4]

Figure 2: Spatial representation of cellulose, hemicellulose and lignin in lignocellulosic biomass.

An ionic liquid (IL) is generally defined as a salt with a low melting point (at or close to room temperature)[1] that consists of an organic cation (e.g. pyridinium) and an organic (e.g. acetate) or inorganic (e.g. chloride) anion. A well-known property of most ILs is their 'negligible' volatility under ambient conditions. Mainly because of this property, ILs have frequently been described as environmentally friendly solvents. They have a host of potential applications from thermal energy storage to pharmaceuticals.[6] With the potential number of ILs that can be synthesised estimated as at least a billion,[7] coupled with the tuneable nature of their physico-chemical properties, the opportunities for designing task-specific ILs are abundant. However, the upscaling of ILs from being mere laboratory curiosities to wide-scale commercially viable technologies has remained to date a significant challenge. This challenge is primarily a result of the high costs of ILs and difficulties in addressing the conundrum of structure–property relationships.[8] However, a few ILs have achieved the 'holy grail' of commercial application, most notably for the BASIL™ (biphasic acid scavenging utilising ionic liquids) process of BASF[9] in which the formation of 1-methylimidazole chloride is an integral part of the cost-effective and efficient production of alkoxyphenylphosphines.

There has been considerable interest in ILs for the processing of lignocellulosic biomass derived from a variety of sources (woods, grasses, crop residues, etc.) over the course of the last decade,[1] but commercially significant progress has largely been hindered as a result of a combination of technical, economic and environmental factors to be expounded upon in this work. It is beyond the scope of this review to undertake any exhaustive surveys of ILs for biomass processing. To this end, there are recent reviews that can be consulted.[1,7,10-12] This review will focus specifically on the use of ILs as solvents for the pretreatment of lignocellulosic biomass and then provide insights into where a concentration of research efforts might be warranted to expedite the development of commercialisable IL pretreatment processes in a biorefinery.

Dissolution of lignocellulosic biomass in ionic liquids

There has been a plethora of studies on the selective and total dissolution of lignocellulosic biomass in ILs. Findings from these studies have stimulated interest in the use of ILs for biomass pretreatment in a biorefinery. Imidazolium ILs have been the most popular ILs for experimental investigations of lignocellulosic biomass treatments. The majority of the ILs of interest contain 1-ethyl-3-methylimidazolium or [EMIM], 1-allyl-3-methylimidazolium or [AMIM] and 1-butyl-3-methylimidazolium or [BMIM] cations. A selection of ILs that have been investigated extensively for biomass pretreatment is shown in Figure 3. Currently, the 1,3-dialkylimidazolium acetates, especially [EMIM][CH_3COO], are amongst the most widely investigated ILs.[1]

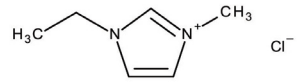
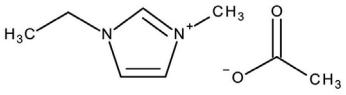
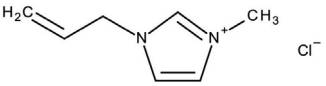
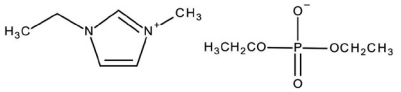
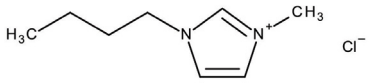
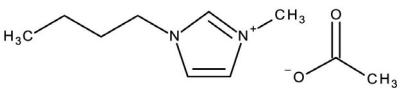
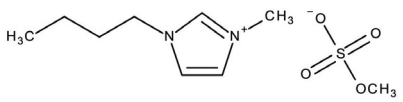
Name / abbreviation	Structure
1-ethyl-3-methylimidazolium chloride [EMIM][Cl]	
1-ethyl-3-methylimidazolium acetate [EMIM][CH_3COO]	
1-allyl-3-methylimidazolium chloride [AMIM][Cl]	
1-ethyl-3-methylimidazolium diethylphosphate [EMIM][DEP]	
1-butyl-3-methylimidazolium chloride [BMIM][Cl]	
1-butyl-3-methylimidazolium acetate [BMIM][CH_3COO]	
1-butyl-3-methylimidazolium methylsulfate [BMIM][$MeSO_4$]	

Figure 3: Ionic liquids commonly used for the pretreatment of lignocellulosic biomass.

Cellulose remains the most widely investigated lignocellulosic biomass component for IL dissolution.[13] In the pioneering study by Swatloski et al.[14], it was demonstrated that [BMIM][Cl] could dissolve up to 25% (w/w) of dissolving pulp cellulose with microwave heating. The regeneration of the dissolved cellulose in an amorphous form was also demonstrated through the use of anti-solvents (such as water) which were IL soluble. This work showcased the ability of ILs to not only dissolve cellulose but also to modify its crystalline structure. Also noted was the marked influence of water present in the ILs in amounts greater

than 1% (w/w) to diminish the ability of the ILs studied to dissolve cellulose. In addition to cellulose, the solubilisation of other lignocellulosic biomass components – hemicellulose[15] and lignin[16] – in ILs has been demonstrated.

Dissolution of whole biomass has been demonstrated across a wide variety of woody and non-woody biomass types. Sun et al.[17] demonstrated that [EMIM][CH_3COO] could achieve the complete dissolution of a softwood (southern yellow pine) and a hardwood (red oak). Rayne and Mazza[18] achieved the rapid (less than 2 min) dissolution of both woody (oak, ponderosa pine and sumac wood) and herbaceous (grape stem, flax shives and triticale straw) biomass materials in [BMIM][Cl] with microwave irradiation. In the study by Li et al.[19], it was shown that 0.5 g of sugarcane bagasse (<0.25 mm particle size) dissolved much more rapidly (15–16 h) than 0.5 g of pine (46 h) in 10 g of [EMIM][CH_3COO] at 100 °C, even compared with smaller particle sizes of the latter (<0.125 mm particle size). In addition, they showed that rapid dissolution of bagasse could be achieved with heating at 175 °C for 10 min or at 185 °C for 5 min. Edye and Doherty[20] also described a procedure in which bagasse with a larger particle size (2 mm) and lower biomass loading of 1.3% (w/w) in [EMIM][CH_3COO] was heated at a lower temperature (170 °C), resulting in complete dissolution of bagasse in 50 min.

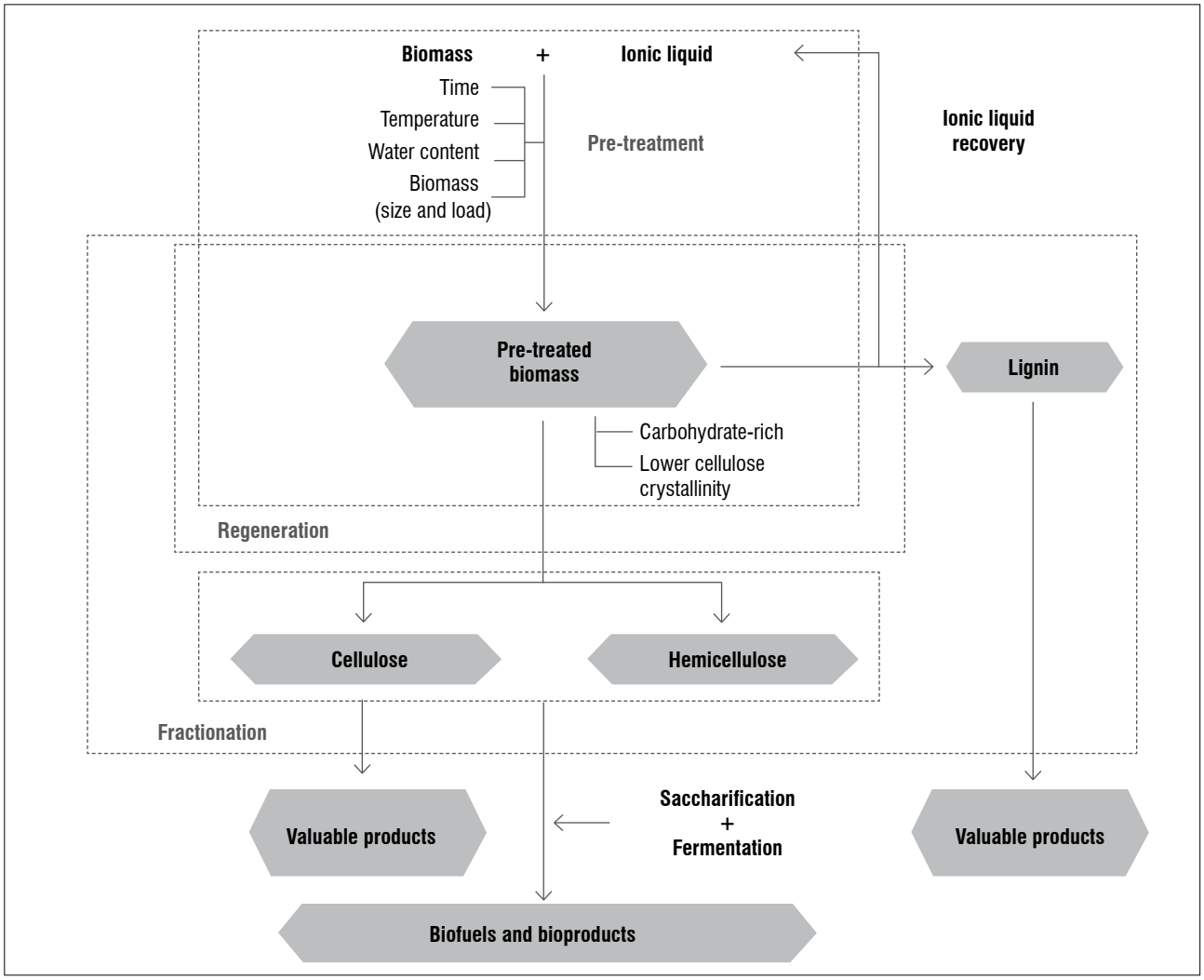
It has been shown in the above-mentioned studies that ILs are able to dissolve each of the major components of lignocellulosic biomass and are also able to completely dissolve the entire native lignocellulosic biomass itself. In general, herbaceous biomass is more easily dissolved than woody biomass (in particular softwoods) under similar conditions.[1]

Ionic liquids as solvents in biorefinery applications

The dissolution of lignocellulosic biomass and/or its constituents in ILs addresses the key issue of reducing biomass recalcitrance to facilitate the utilisation of biomass for bio-based products. A schematic representation of IL pretreatment and fractionation in a biorefinery process[10] is shown in Figure 4 and depicts the pretreatment of biomass with ILs, followed by fractionation, IL regeneration and downstream processing of the biomass components. The dissolution of the lignocellulosic biomass in the IL pretreatment allows for the cellulose and other lignocellulosic components to be accessible to external reagents and catalysts (chemical and biological) dissolved or dispersed in the same medium for conversion processes.[21] Current biorefinery research with ILs is heavily weighted towards the fermentation of cellulosic monosugars for the production of bioethanol.[22] For the hydrolysis of the cellulosic materials to yield fermentable sugars in ILs, there are three types of approaches that have been typically employed[23] as described below.

Enzymatic hydrolysis of materials regenerated from ionic liquids

The regeneration of dissolved cellulosic materials from an IL medium with anti-solvents such as water or ethanol was demonstrated in the pioneering work of Swatloski et al.[14] in which a reduction in cellulose



Source: Da Costa Lopes et al.[10]

Figure 4: Ionic liquid pretreatment and fractionation in a biorefinery.

crystallinity is made possible by the regeneration.[24] This alteration of the macromolecular structure of lignocellulosic components allows for more effective penetration of the enzymes for hydrolysis of the relevant chemical bonds in the lignocellulosic components.

Liu and Chen[25] investigated the enzymatic hydrolysis of regenerated cellulose from wheat straw and steam-exploded wheat straw after pretreatment with water or [BMIM][Cl]. The enzymatic hydrolysis rate of the IL-treated wheat straw was 70.37%, whereas complete hydrolysis was observed for the IL-treated steam-exploded wheat straw. The hydrolysis rates of the wheat straw and steam-exploded wheat treated only with water were 42.78% and 68.78%, respectively. The marked increase in hydrolysis rates for the biomass pretreated with [BMIM][Cl] was attributed to the decrease in the degrees of polymerisation in cellulose and hemicellulose, the reduction in the crystallinity of cellulose and the increased accessibility of reactive sites for the enzymes.

The necessity of anti-solvents to regenerate the dissolved biomass dictates that the anti-solvent must be removed to allow for the recovery of the ILs for the process. An alternative for the recovery of ILs from aqueous solutions is the use of solutions with kosmotropic (water-structuring) anions (such as phosphate, carbonate or sulfate) to form aqueous biphasic systems, as demonstrated by Gutowski et al.[26] and later by Shill et al.[27] for the enzymatic hydrolysis of [EMIM][CH_3COO]-pretreated cellulose (from Miscanthus, Avicel® or corn stover) through the addition of K_3PO_4 or K_2HPO_4. The triphasic process decreased the amount of water to be evaporated for the IL recycling, allowing for greater efficiency in water and energy usage. The major findings were that a more rapid and higher-yielding enzymatic conversion of cellulose to glucose is achieved in the three-phase system when compared to cellulose obtained from IL-pretreated biomass precipitated with water. Also, the kosmotropic salt allowed for partial delignification of the biomass, which made the substrate more accessible for enzymatic hydrolysis.

Enzymatic hydrolysis in ionic liquids

Direct use of ILs as reaction media for biocatalytic applications offsets the process complexity and high costs by eliminating the requirement of anti-solvents to regenerate the dissolved materials. However, this elimination places an additional strain on the selection of ILs for biomass pretreatment as the efficacy of the IL for the dissolution of the biomass must be considered synergistically with the biocompatibility of the IL with the enzymes used. Despite the fact that ILs are generally described as green solvents, toxicity of some ILs towards microorganisms has been demonstrated.[28] It is unfortunate that ILs that have been most studied based on their ability to dissolve cellulose – such as 1,3 dialkylimidazolium ILs with chloride, dicyanamide, formate and acetate anions – are not very compatible with enzymes.[23] These anions, which are strong hydrogen-bond acceptors, tend to denature the enzymes. Also the high viscosity of these ILs has an adverse effect on the in situ biocatalytic transformations. As a result, research efforts have been made to find ILs that are both enzyme compatible and able to dissolve cellulose.

Kamiya et al.[29] performed enzymatic saccharification studies with cellulase enzymes on Avicel® cellulose in an aqueous solution of [EMIM][DEP]. Direct correlation between the volumetric ratio of IL to water and the cellulase activity was observed. For ratios greater than 3:2 there was little cellulase activity observed and for lower ratios, cellulase activity was greatly enhanced where an IL to water ratio of 1:4 allowed for the conversion of over 70% of cellulose to glucose and cellobiose. A later study by Wang et al.[30] revealed that a mixture of cellulases retained high enzymatic activity in 15% (v/v) of [EMIM][CH_3COO] for the hydrolysis of Avicel® cellulose with a conversion efficiency of approximately 91%. This finding demonstrated the compatibility of [EMIM][CH_3COO] with the cellulase mixture under the conditions of investigation. Bose et al.[31] tested the reactivity and stability of a commercial mixture of cellulases in eight ILs (not including [EMIM][CH_3COO]). Of the ILs investigated, cellulose hydrolysis was observed only in 1-methylimidazolium chloride and tris-(2-hydroxyethyl)methylammonium methylsulfate (HEMA). The high viscosity of the ILs was seen as a retarding factor for the enzymatic hydrolysis, as it inhibited the diffusion of the enzyme, resulting in the lower enzymatic activity observed when compared to the cellulose hydrolysis in a buffer solution. The thermal stability of the cellulase enzymes in the two ILs was also investigated in terms of the reversibility of the folding and unfolding processes of the enzymes and it was found that HEMA imparted greater stability to the enzyme when compared to the imidazolium IL. The folding and unfolding processes were reversible even at temperatures up to 120 °C in HEMA, indicating that excellent cellulase thermal stability can be achieved in particular ILs.

Overall, the use of ILs as a medium for the pretreatment and in situ enzymatic hydrolysis of lignocellulosic materials holds great promise as a feasible technology for the application of ILs in biorefining. However, the realisation of a commercial process based on this technology has remained very challenging to date, as it requires the use of ILs that meet the very demanding requirements of good cellulose dissolution, enzyme compatibility and low viscosity.[23]

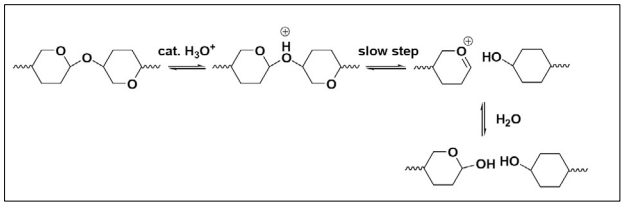
Acid-catalysed hydrolysis in ionic liquids

The list of shortcomings for the traditional acid hydrolysis of lignocellulosic materials includes harsh conditions, high capital costs and the inefficiency of the process with mineral acids.[32] The inherent advantage of acid hydrolysis in ILs is the homogeneity that ILs can potentially provide as a reaction medium for the lignocellulosic biomass and the reagents, in addition to the significant disruption of the macromolecular structure of cellulose through IL dissolution. Both solid-phase acid catalysts and liquid acids have been used for cellulose hydrolysis in ILs.[32]

Li and Zhao[33] studied cellulose hydrolysis by mineral acids (H_2SO_4, HCl, HNO_3 and H_3PO_4) in [BMIM][Cl]. It was shown that H_2SO_4 (with acid/cellulose mass ratios as low as 0.11) dissolved in [BMIM][Cl] was able to drive Avicel® cellulose hydrolysis reactions. For the Sigmacell® cellulose hydrolysis reactions, it was observed that HCl and HNO_3 behaved similarly to H_2SO_4, while H_3PO_4 was less effective. The results indicated that acid strength had an important role to play in the hydrolysis of cellulose in [BMIM][Cl]. A preliminary kinetic study of the experimental data indicated that cellulose hydrolysis proceeded much faster than total reducing sugar degradation. It was noted that excess acid loading in the IL system reduced sugar yields as a result of side reactions which consumed the hydrolysis products. In the study of Li et al.[34] for the acid catalysis of lignocellulosic biomass with various acids, total reducing sugar yields as high as 66%, 74%, 81% and 68% for the hydrolysis of corn stalk, rice straw, pine wood and bagasse, respectively, were achieved in the presence of only 7% (w/w) HCl at 100 °C within 60 min. Overall, the catalytic activity of the acids followed the hierarchy: hydrochloric acid > nitric acid > sulfuric acid > maleic acid > phosphoric acid.

Rinaldi et al.[35] analysed the mechanism for acid-catalysed cellulose hydrolysis, which is thought to consist of three sequential steps, as shown in Figure 5.[23] The first step is the cellulose activation through the protonation of the glycosidic oxygen, followed by formation of a cyclic carbocation via the slow unimolecular scission of the glycosidic linkage, and finally the nucleophilic attack on the cyclic carbocation species by a water molecule to create the hydroxyl group in the carbon at the C(1) position in the anhydroglucose unit. Even though the second step in the mechanism is rate limited, the protonation of the glycosidic oxygen has been reported to be difficult and slow with some electron-deficient acetals during the hydrolysis process.[35] Strong acids are required for the activation of cellulose for hydrolysis as a result of the weak basicity of the glycosidic oxygen in cellulose. A 'serious consequence'[35] of the strong acid requirement is the severe restriction imposed upon the pool of cellulose-dissolving ILs that can be used as it has been shown that 1,3-dialkylimidazolium ILs that make effective solvents for cellulose are often those with anions which are weak bases (acetate, organophosphates, etc.). Because these anions are weak bases, they are protonated quite easily by strong acids to produce weak acids. As a result, cellulose activation in these ILs is insufficient for reasonable rates of hydrolysis.[35] Another major challenge is that the often impressive solvation capabilities of ILs for a diversity of materials complicates the separation of the water-soluble monosugars and lower carbohydrate oligomers from the homogenous IL reaction medium introducing

additional steps and costs into the process. There are some advantages for the use of solid catalysts over molecular acids for the heterogeneous acid-catalysed hydrolysis of cellulose[36] such as the ease of recycling of the acid catalyst[32] and a more controllable process. However, there are practical issues associated with the use of solid acids such as the slower reaction times and the requirement of an additional step to further hydrolyse the cellooligomers to sugar monomers.



Source: ©2011 Haibo Xie, Zongbao K. Zhao. Adapted from Rinaldi et al.[35] Originally published in Xie and Zhao[23] under CC-BY-NC-SA 3.0 licence.

Figure 5: Proposed mechanism for acid-catalysed hydrolysis of cellulose.

Biomass fractionation with ionic liquids

From the preceding discussion, it can be noted that the overwhelming majority of research with ILs has been geared towards the deconstruction or dissolution of lignocellulosic biomass, principally for the production of fermentable sugars for biofuel production.[22] As a result, the dissolution of cellulosic materials in ILs has been investigated thoroughly. The achievement of a process design with ILs for the complete fractionation of lignocellulosic biomass into three distinct streams of cellulose, hemicellulose and lignin has received much less attention.[22,37]

Tan et al.[38] employed the hydrotropic pulping concept (aqueous solutions of sodium xylenesulfonate at elevated temperatures and under pressure) for the delignification of sugarcane bagasse. An IL mixture consisting of the [EMIM] cation and alkylbenzenesulfonates with xylenesulfonate as the main anion was employed to extract lignin from steamed bagasse at elevated temperatures (170–190 °C). The lignin fraction was recovered from the IL mixture by precipitation, allowing the IL to be recycled. An extraction yield higher than 93% could be achieved for which less than 50% of xylan was left in the pulp (for a 2-h treatment at 190 °C). A rough correlation was observed between the amount of hemicellulose removed during delignification and the amount of extraneous material in the recovered lignin. This correlation could be attributed to the reactive breakdown products of hemicellulose being incorporated within the recovered lignin. Recycling of the IL was demonstrated for which the regenerated IL showed good retention of structure and properties, based on nuclear magnetic resonance spectra.

Hamada et al.[39] investigated ILs for the thermally controlled selective dissolution of lignocellulosic biomass. *N*-methyl-*N*-(2-methoxyethyl) pyrolidin-1-ium 2,6-diaminohexanoate or $[P_{1ME}][Lys]$ was a good solvent for lignin below 60 °C; however, cellulose could only be dissolved at a temperature of 80 °C or higher. On this basis, the partial fractionation of 9% (w/w) mixture of Japanese cedar in $[P_{1ME}][Lys]$ was demonstrated through the direct extraction of lignin from the woody biomass under mild conditions (stirring at 60 °C for 12 h). Hamada et al.[39] reported that they were able to reuse the IL five times after recovery without any loss of its ability to dissolve lignin.

Diedericks et al.[37] investigated the fractionation of sugarcane bagasse into three streams of cellulose, hemicellulose and lignin with [EMIM] $[CH_3COO]$ and [BMIM][Cl]. The biomass dissolution was achieved through the use of the IL only or by using a dilute acid pre-extraction of the xylose prior to the IL dissolution of the remaining solids. The constituents were isolated from the reaction mixture with acetone, acetone-water mixtures or NaOH as the first anti-solvent and a sodium citrate buffer as the second anti-solvent. It was observed that delignification was enhanced through the use of NaOH; however, the addition of NaOH produced impure product streams. The xylose pre-extraction prior to IL treatment with dilute acid allowed for a 75% (w/w) recovery of the xylose content of bagasse and served to improve the lignin purity after anti-solvent separation. The fractionation efficiency of the combined process (with xylose pre-extraction) was maximised (84%) by IL treatment at 125 °C for 120 min for a removal of 80.2% (w/w) of the lignin content of bagasse and 76.5% (w/w) lignin recovery. Both ILs displayed similar degrees of delignification; however, only [EMIM] $[CH_3COO]$ treatment was able to produce enzyme-digestible cellulose-rich solids.

Commercialisation of ionic liquids for biomass pretreatment

There are major technical, economic and environmental hurdles that still remain for the industrial-scale design of pretreatment processes with ILs, as shown in Figure 6. The key considerations inherent in these hurdles and their implications for advancing IL research on the path towards achieving commercialisation are discussed below.

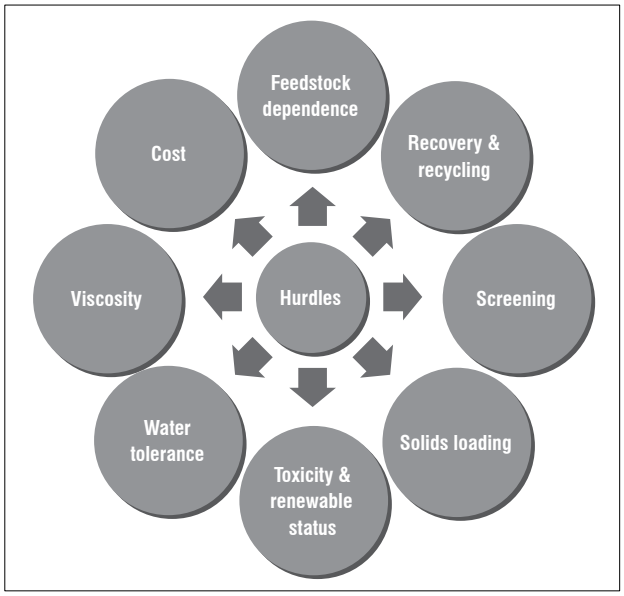


Figure 6: Key hurdles for the commercialisation of ionic liquid pretreatment of biomass.

Cost

ILs remain to date the most expensive research-grade solvents under investigation for the dissolution of biomass.[1] Brandt et al.[1] consider the ideal scenario as one in which ILs are produced for a cost in the region of USD2.50/kg or less. The current bulk cost for industrial organic solvents such as acetone is less than USD1.50/kg.[40] The best-case bulk estimate cost of [EMIM] $[CH_3COO]$ is USD20/kg.[41] Recently, work has been performed targeting the production of low-cost ILs, and in particular, those based on hydrogen sulfate anions for production from sulfuric acid. Based on the work of Chen et al.[40], the prices of triethylammonium hydrogen sulfate and 1-methylimidazolium hydrogen sulfate were estimated at USD1.24/kg and USD2.96–5.88/kg, respectively. The triethylammonium hydrogen sulfate IL was also tested against the [EMIM] $[CH_3COO]$ benchmark for the pretreatment of switchgrass and was able to achieve a 75% efficacy when compared to [EMIM] $[CH_3COO]$.[42] These findings bode well for the production of 'cheap' ILs; however, industrial-scale production of ethylammonium hydrogen sulfate ILs has not been achieved as yet to verify the estimated costs.

Viscosity

In addition to high costs, the high viscosity of ILs is another major factor limiting the uptake of ILs as they have much higher viscosities than industrial organic solvents. The viscosities of ILs are 1–3 orders of magnitude higher than conventional organic solvents, with values ranging as widely as 4.8–1110 mPa/s,[43] depending on the cation and anion. The high viscosity has direct implications not only in terms of increased process costs and complexity for operations such as mixing and pumping, but also for the efficacy of ILs as reaction media in terms of mass transfer and reaction rates, as is the case for diffusion-

controlled processes such as enzymatic hydrolysis. The IL currently of most interest in biomass treatment, [EMIM][CH_3COO], has a viscosity of 162 mPa/s at 20 °C[43], and can be classified as a high-viscosity fluid. The viscosity of ILs can be addressed either by designing ILs with lower viscosities or through the use of ILs in mixtures with aqueous or organic solvents.[44] The first option involves designing ILs to achieve viscosity reduction. However, this reduction must be achieved without diminishing the efficacy of the IL in the intended application. This task is a very challenging one indeed because of the abundance of potential IL structures from various cation–anion pairings and a lack of a fundamental understanding of the correlation of the IL structure with efficacy in the intended application. The second option is perhaps more attractive for biorefinery applications with ILs, as the use of less costly organic solvents or water with ILs will reduce the amount of IL required, thus resulting in cost savings. Work has been completed for the dissolution of cellulose in organic electrolyte solutions containing various amounts of ILs[45] in solvents such as dimethylformamide, in which only 10% (w/w) of [EMIM][CH_3COO] was required for the complete dissolution of cellulose. The usage of cheaper solvents, especially benign solvents such as water, would be a recommended strategy as long as there is no significant reduction in the efficacy of the IL mixture for the biomass pretreatment. This strategy has been demonstrated in the work by George et al.[42], in which IL-water mixtures with 20% (w/w) water were successfully employed for the pretreatment of switchgrass with [EMIM][CH_3COO] and ethylammonium hydrogen sulfate ILs.

Water tolerance

The investigation of ILs as pure solvents in many academic studies has necessitated scrupulous drying of ILs to provide results with 'dry' ILs. However, for the real-world application of ILs in biorefining, in addition to the water associated with the IL itself, the processing environment and the biomass will be major contributors to increasing the water content of the IL. Drying of the biomass and ensuring a processing environment free of water will place significant demands upon the process economics. It has been shown that the presence of water in some ILs can produce a very significant impairment of the ability of the IL to dissolve cellulose.[1] However, Abe et al.[46] discovered that hydroxide-containing ILs such as tetrabutylphosphonium hydroxides with 40% (w/w) water content could dissolve up to 20% (w/w) of cellulose within 5 min with mild stirring at 25 °C. Consequently, it is recommended that emphasis be placed upon the design and investigation of ILs for improved water tolerance in process applications. As discussed above, the current IL of choice for pretreatment, [EMIM][CH_3COO], does not require anhydrous conditions to be effective in biomass pretreatment.

Toxicity, biodegradability and renewability

While initial investigations focused on chloride-based ILs, the transition towards ILs with lower toxicities has been supported. BASF utilised [EMIM][CH_3COO] over [EMIM][Cl] in their commercial CELLIONIC™ formulation of IL solutions of cellulose.[22] Chloride-containing ILs possess toxicities and corrosive properties that would probably not suit most biorefining operations. It is highly likely that potential IL biorefineries of the future will involve the use of biological agents (such as enzymes) that are exposed to ILs in fermentation or biocatalytic conversions, in which ILs could even serve as the media for the microbial processes. As a result, biocompatibility of ILs will be an important factor to offset process complexities and additional costs to actively purify streams of IL-pretreated biomass for downstream microbiological processes. In addition, non-toxic ILs would widen the pool of potential applications for ILs such as the production of food-grade chemicals. It has also been shown that, at elevated temperatures, some ILs are not inert solvents for lignocellulosic materials. Even for the favoured [EMIM][CH_3COO], acetylation of cellulose was observed to occur at elevated temperatures.[21] For the dissolution of lignocellulose in some chloride-containing ILs, the thermal stability of the IL is also affected.[21] There are also concerns over the effects of residual amounts of impurities from the synthesis of the ILs, such as 1-methylimidazole[47] from the synthesis of imidazolium ILs. As a result, greater demands will be placed on the purification methods, which would most likely increase costs.

Biodegradable ILs are desirable to minimise the environmental impact of chemicals and for ILs to live up to their billing as green solvents, especially with the ever-tightening grip of environmental regulations on the use of industrial solvents. Cholinium ILs are a class of ILs that have low toxicity and are readily biodegradable.[48] Nockemann et al.[49] demonstrated that the cholinium ILs, choline saccharinate and choline acesulfamate, can be prepared from easily available starting materials such as chicken feed additives (choline chloride) and artificial sweeteners (saccharinate and acesulfamate salts). In addition to biodegradability, the production of ILs from renewable materials is a very important consideration in a biorefinery and in this regard bio-based ILs have been prepared and tested for biomass pretreatment. Cholinium amino acids, such as choline lysine, choline glycine and choline alanine were developed and tested by Hou et al.[50] for the pretreatment of sugarcane bagasse. Cholinium lysine or [Ch][Lys] was shown to be the most effective of all the IL solvents. Pretreatment of sugarcane bagasse (unmilled and cut into 60-mm lengths) at 90 °C for 6 h with a 5% (w/w) loading in a 50% [Ch][Lys]-water mixture provided yields of 80% for glucose and 84% for xylose from the enzymatic hydrolysis of the residue. The above findings indicate that trends towards renewable or bio-based ILs are emerging in the field; however, it should be borne in mind that this should not be achieved by increased cost of production of starting materials (in cases in which the renewable starting materials are more expensive).

Solids loading and size reduction

In the majority of studies involving IL pretreatment of biomass, solid loadings have typically been about 5% (w/w).[1] This value is low when compared to the solids loadings of other chemical pretreatment methods. The solids loading has important repercussions for the throughput and economics of the pretreatment process as there would be cost savings if less of the IL solvent can be used and more biomass processed per pretreatment. In addition to solids loading, comminution of biomass has frequently been applied in studies involving ILs with limited work performed on biomass that has not been reduced in size. The energetic requirements for the comminution of biomass can adversely affect the economics of biorefinery processes to a significant extent. Hou et al.[50] employed biomass that was not milled and a fairly long pretreatment time of 6 h was required for the pretreatment at 90 °C with a solids loading of 5% (w/w). If a higher solids loading were to be used, the pretreatment time would most likely have to be increased to achieve the desired results. Work has been performed on the impact of high solids loading on the performance of ILs in biomass pretreatment. Wu et al.[51] pretreated corn stover (1–2 mm particle size) in [EMIM][CH_3COO] at 125 °C for 1 h with biomass loadings of 4.8–50% (w/w). It was shown that over a wide range of biomass loadings from 4.8–33% (w/w), the yields of glucose and xylose from enzymatic hydrolysis remained similar at 80% and 50%, respectively. For loadings above 33% (w/w), there was less lignin extracted and less disruption of cellulose crystallinity, resulting in a decrease in enzymatic hydrolysis efficiency. In another study by Cruz et al.[52], switchgrass was pretreated with [EMIM][CH_3COO] at 160 °C for 3 h with biomass loadings in the IL varying between 3% and 50% (w/w). The relative glucan content of the recovered solid samples was observed to decrease with increasing levels of lignin and hemicellulose recovered with increased biomass loading. Maximum glucose recovery (nearly 100%) was obtained for the solid recovered after pretreatment with a 10% (w/w) biomass loading but this amount dropped significantly at higher biomass loadings. However, an unexpected advantage of increasing the loading levels to 40–50% (w/w) was the decrease in the cellulose crystallinity which produced faster enzymatic hydrolysis kinetics and could potentially lower enzyme loading levels to improve the economic viability of IL pretreatment.

Use of a priori methods for screening ionic liquids

To date, the 'trial and error' approach has for the most part evolved the landscape of IL pretreatment towards ILs such as [EMIM][CH_3COO]. However, with the potential number of ILs that can be synthesised in theory being close to a billion[7], the use of a priori methods provides a much-needed and very useful means to investigate ILs for the pretreatment or dissolution of biomass. The empirical Kamlet–Taft model parameters[53] and the quantum mechanical COSMO-RS model[54]

are two popular approaches for predicting or understanding the solubility of lignocellulosic components in ILs.[1] The Kamlet–Taft parameters are useful for understanding IL-lignocellulose interactions through arguments relating to hydrogen-bond acceptor and donor strengths (usually for IL anions). However, concepts such as hydrogen bonding may not be sufficient to understand and predict interactions between ILs and lignocellulose, for which there are many other factors that require consideration.[55] The COSMO-RS model has been shown to be very useful for screening ILs for lignocellulose treatment[1,56] as only structural information is typically required for the prediction of solubility and other thermodynamic properties. Being able to screen ILs through the use of predictive methods requiring a minimal amount of input properties would allow for much-needed savings in time, cost and effort in laboratory studies.

Ionic liquid recovery and recycling

In addition to increasing biomass loading in ILs, the recovery and recycling of ILs represents an important means for cost savings in the process. The recycling of the ILs is facilitated by their negligible volatilities under ambient conditions. However, there are some ILs that are 'distillable' such as those based on the combination of dimethylamine with CO_2 (2:1 ratio) to produce a dimethylammonium cation and a dimethylcarbamate anion.[57] These types of ILs can be distilled at temperatures as low as 45 °C, which opens up the avenue for new applications and options for recycling ILs.[57] It is quite commendable that an increasing number of researchers are demonstrating the recycling of the ILs used in their studies.[1] Hou et al.[50] recovered and reused their 50% [Ch][Lys]-water mixture directly (not separating out extracted lignin) for three batches in their bagasse pretreatment. After recovering and using the IL three times, the glucose and xylose yields decreased to 62.7% and 73.0%, respectively, probably as a result of the decrease in the delignification capacity of the IL–water mixture as a result of accumulation of lignin in the solvent. Wu et al.[51] also investigated the reuse of [EMIM][CH_3COO] for their pretreatment of corn stover. The extracted lignin decreased from 44% in the first batch to 20% in the tenth batch. However, the ability of the IL to disrupt the crystallinity of the corn stover cellulose remained largely unaffected. There was also no significant loss in the glucose and xylose yields of corn stover pretreated with [EMIM][CH_3COO] recycled after ten batches. It is recommended that, as in the above-mentioned studies, the recycling of the IL features as an integral part of the outcomes. This approach would serve to demonstrate that the integrity of the IL can be preserved in the process through a fairly undemanding regeneration sequence, as an important factor in the design of an IL process will be the IL recycling.

Feedstock dependence

To ensure a steady supply of biomass throughout the year and in sufficient quantities for sustainable biorefinery value chains for fuels, chemicals and materials, the use of mixed biomass will most likely be required to take advantage of the regional availability of cheap feedstocks.[58] Biomass feedstocks can differ significantly in lignocellulosic composition and bulk density depending on their origin, resulting in challenges for processing mixed biomass feedstocks. Although it has been shown above that ILs are capable of processing various types of herbaceous and woody biomass efficiently as single feedstocks, the conditions for pretreatment are dependent upon the biomass type, with softwoods recognised as the most difficult to process when compared to herbaceous biomass.[1] To address this issue, Shi et al.[58] investigated the densification and blending of switchgrass, lodgepole pine, corn stover and eucalyptus in flour and pellet form (in 1:1:1:1 ratios) with [EMIM][CH_3COO] at 160 °C for 3 h with 10% (w/w) biomass loading. There was no significant difference observed between the enzymatic digestibility of the pretreated solids from the mixed biomass pellets and flour starting materials. The IL-pretreated mixed pellets and flour had similar trends in terms of fast saccharification rates by reaching 90% digestibility in 24 h. The success of [EMIM][CH_3COO] in effectively pretreating mixed biomass feedstocks augers well for demonstrating that IL pretreatment can be feedstock-independent. However, particle size reduction and densification are high-energy and costly processes necessitating that the economics for the preparation of mixed feedstocks be closely scrutinised.

Conclusions

The unique ability of ILs to selectively dissolve biomass components and whole biomass has been demonstrated for both woody and herbaceous biomass types, and is of relevance to both the South African forestry and sugarcane industries. Biorefinery applications for IL pretreatment have focused on the production of cellulosic biofuels in which the usefulness of ILs has been demonstrated. However, there are significant technical, economic and environmental barriers for the wide-scale commercialisation of ILs in biorefineries. To overcome these barriers, it is recommend that investigations are undertaken for the development of ILs that

- can combine efficacy for biomass pretreatment with the ability to be used under hydrous conditions (which also reduces their viscosity);
- are thermally stable and inert at pretreatment conditions;
- can be recycled and reused;
- are non-toxic (to human life and microbiological organisms); and
- are low-cost (when compared to current industrial solvents).

In addition, for ILs to achieve an identity as sustainable technologies in a biorefinery context, it would be only fitting for ILs of the future to be manufactured from the very same renewable resources they pretreat for the production of renewable chemicals and materials.

References

1. Brandt A, Gräsvik J, Hallett JP, Welton T. Deconstruction of lignocellulosic biomass with ionic liquids. Green Chem. 2013;15(3):550–583. http://dx.doi.org/10.1039/c2gc36364j
2. South African Sugar Association (SASA). The SA sugar industry facts and figures [homepage on the Internet]. No date [cited 2015 Jan 31]. Available from: http://www.sasa.org.za/sugar_industry/FactsandFigures.aspx
3. Walford SN, Morel du Boil PG. A survey of value addition in the sugar industry. Proc S Afr Sug Technol Ass. 2006;80:39–61. Available from: http://www.sasta.co.za/wp-content/uploads/Proceedings/2000s/2006_Walford_a%20survey%20of%20value%20addition.pdf
4. Oak Ridge National Laboratory [homepage on the Internet]. No date [cited 2015 Jan 31]. Available from: https://public.ornl.gov/site/gallery/gallery.cfm?topic=53&restsection=public
5. Behera S, Arora R, Nandhagopal N, Kumar S. Importance of chemical pretreatment for bioconversion of lignocellulosic biomass. Renew Sust Energy Rev. 2014;36:91–106. http://dx.doi.org/10.1016/j.rser.2014.04.047
6. Olivier-Bourbigou H, Magna L, Morvan D. Ionic liquids and catalysis: Recent progress from knowledge to applications. Appl Catal A-Gen. 2010;373(1–2):1–56. http://dx.doi.org/10.1016/j.apcata.2009.10.008
7. Mora-Pale M, Meli L, Doherty TV, Linhardt RJ, Dordick JS. Room temperature ionic liquids as emerging solvents for the pretreatment of lignocellulosic biomass. Biotechnol Bioeng. 2011;108(6):1229–1245. http://dx.doi.org/10.1002/bit.23108
8. Niedermeyer H, Ashworth C, Brandt A, Welton T, Hunt PA. A step towards the a priori design of ionic liquids. Phys Chem Chem Phys. 2013;15(27):11566–11578. http://dx.doi.org/10.1039/C3CP50521A
9. Plechkova NV, Seddon KR. Applications of ionic liquids in the chemical industry. Chem Soc Rev. 2008;37(1):123–150. http://dx.doi.org/10.1039/b006677j
10. Da Costa Lopes AM, João KG, Morais ARC, Bogel-Łukasik E, Bogel-Łukasik R. Ionic liquids as a tool for lignocellulosic biomass fractionation. Sustain Chem Proc. 2013;1:3. http://dx.doi.org/10.1186/2043-7129-1-3
11. Gericke M, Fardim P, Heinze T. Ionic liquids: Promising but challenging solvents for homogeneous derivatization of cellulose. Molecules. 2012;17(6):7458–7502. http://dx.doi.org/10.3390/molecules17067458
12. Stark A. Ionic liquids in the biorefinery: A critical assessment of their potential. Energy Environ Sci. 2011;4(1):19–32. http://dx.doi.org/10.1039/C0EE00246A
13. Wang H, Gurau G, Rogers RD. Ionic liquid processing of cellulose. Chem Soc Rev. 2012;41(4):1519–1537. http://dx.doi.org/10.1039/c2cs15311d

14. Swatloski RP, Spear SK, Holbrey JD, Rogers RD. Dissolution of cellose with ionic liquids. J Am Chem Soc. 2002;124(18):4974–4975. http://dx.doi.org/10.1021/ja025790m

15. Wei L, Li K, Ma Y, Hou X. Dissolving lignocellulosic biomass in a 1-butyl-3-methylimidazolium chloride–water mixture. Ind Crop Prod. 2012;37(1):227–234. http://dx.doi.org/10.1016/j.indcrop.2011.12.012

16. Pu Y, Jiang N, Ragauskas AJ. Ionic liquid as a green solvent for lignin. J Wood Chem Technol. 2007;27(1):23–33. http://dx.doi.org/10.1080/02773810701282330

17. Sun N, Rahman M, Qin Y, Maxim ML, Rodríguez H, Rogers RD. Complete dissolution and partial delignification of wood in the ionic liquid 1-ethyl-3-methylimidazolium acetate. Green Chem. 2009;11(5):646–655. http://dx.doi.org/10.1039/b822702k

18. Rayne S, Mazza G. Rapid dissolution of lignocellulosic plant materials in an ionic liquid [document on the Internet]. c2007 [cited 2015 Jan 31]. Available from: http://precedings.nature.com/documents/637/version/1/files/npre2007637-1.pdf

19. Li W, Sun N, Stoner B, Jiang X, Lu X, Rogers RD. Rapid dissolution of lignocellulosic biomass in ionic liquids using temperatures above the glass transition of lignin. Green Chem. 2011;13(8):2038–2047. http://dx.doi.org/10.1039/c1gc15522a

20. Edye LA, Doherty WOS. Fractionation of a lignocellulosic material. World Patent W0/2008/095252 A1; 2008.

21. Zhang ZC. Catalytic transformation of carbohydrates and lignin in ionic liquids. WIREs Energy Environ. 2013;2(6):655–672. http://dx.doi.org/10.1002/wene.67

22. Sun N, Rodríguez H, Rahman M, Rogers RD. Where are ionic liquid strategies most suited in the pursuit of chemicals and energy from lignocellulosic biomass? Chem Commun. 2011;47(5):1405–1421. http://dx.doi.org/10.1039/C0CC03990J

23. Xie H, Zhao ZK. Selective breakdown of (ligno)cellulose in ionic liquids. In: Kokorin A, editor. Ionic liquids: Applications and perspectives. Rijeka: InTech; 2011. p. 61–80. http://dx.doi.org/10.5772/14681

24. Zhu S. Use of ionic liquids for the efficient utilization of lignocellulosic materials. J Chem Technol Biot. 2008;83(6):777–779. http://dx.doi.org/10.1002/jctb.1884

25. Liu L, Chen H. Enzymatic hydrolysis of cellulose materials treated with ionic liquid [BMIM]Cl. Chinese Sci Bull. 2006;51(20):2432–2436. http://dx.doi.org/10.1007/s11434-006-2134-9

26. Gutowski KE, Broker GA, Willauer HD, Huddleston JG, Swatloski RP, Holbrey JD, et al. Controlling the aqueous miscibility of ionic liquids: Aqueous biphasic systems of water-miscible ionic liquids and water-structuring salts for recycle, metathesis, and separations. J Am Chem Soc. 2003;125(22):6632–6633. http://dx.doi.org/10.1021/ja0351802

27. Shill K, Padmanabhan S, Xin Q, Prausnitz JM, Clark DS, Blanch HW. Ionic liquid pretreatment of cellulosic biomass: Enzymatic hydrolysis and ionic liquid recycle. Biotechnol Bioeng. 2011;108(3):511–520. http://dx.doi.org/10.1002/bit.23014

28. Romero A, Santos A, Tojo J, Rodríguez A. Toxicity and biodegradability of imidazolium ionic liquids. J Hazard Mater. 2008;151(1):268–273. http://dx.doi.org/10.1016/j.jhazmat.2007.10.079

29. Kamiya N, Matsushita Y, Hanaki M, Nakashima K, Narita M, Goto M, et al. Enzymatic in situ saccharification of cellulose in aqueous-ionic liquid media. Biotechnol Lett. 2008;30(6):1037–1040. http://dx.doi.org/10.1007/s10529-008-9638-0

30. Wang Y, Radosevich M, Hayes D, Labbé N. Compatible ionic liquid-cellulases system for hydrolysis of lignocellulosic biomass. Biotechnol Bioeng. 2011;108(5):1042–1048. http://dx.doi.org/10.1002/bit.23045

31. Bose S, Armstrong DW, Petrich JW. Enzyme-catalyzed hydrolysis of cellulose in ionic liquids: A green approach toward the production of biofuels. J Phys Chem B. 2010;114(24):8221–8227. http://dx.doi.org/10.1021/jp9120518

32. Xie H, Liu W, Beadham I, Gathergood N. Biorefinery with ionic liquids. In: Xie H, Gathergood N, editors. The role of green chemistry in biomass processing and conversion. Hoboken, NJ: John Wiley and Sons; 2012. p. 75–133. http://dx.doi.org/10.1002/9781118449400.ch3

33. Li C, Zhao ZK. Efficient acid-catalyzed hydrolysis of cellulose in ionic liquid. Adv Synth Catal. 2007;349(11–12):1847–1850. http://dx.doi.org/10.1002/adsc.200700259

34. Li C, Wang Q, Zhao ZK. Acid in ionic liquid: An efficient system for hydrolysis of lignocellulose. Green Chem. 2008;10(2):177–182. http://dx.doi.org/10.1039/B711512A

35. Rinaldi R, Meine N, Vom Stein J, Palkovits R, Schüth F. Which controls the depolymerization of cellulose in ionic liquids: The solid acid catalyst or cellulose? ChemSusChem. 2010;3(2):266–276. http://dx.doi.org/10.1002/cssc.200900281

36. Rinaldi R, Palkovits R, Schüth F. Depolymerization of cellulose using solid catalysts in ionic liquids. Angew Chem Int Edit. 2008;47(42):8047–8050. http://dx.doi.org/10.1002/anie.200802879

37. Diedericks D, Van Rensburg E, Görgens JF. Fractionation of sugarcane bagasse using a combined process of dilute acid and ionic liquid treatments. Appl Biochem Biotech. 2012;167(7):1921–1937. http://dx.doi.org/10.1007/s12010-012-9742-4

38. Tan SSY, MacFarlane DR, Upfal J, Edye LA, Doherty WOS, Patti AF, et al. Extraction of lignin from lignocellulose at atmospheric pressure using alkylbenzenesulfonate ionic liquid. Green Chem. 2009;11(3):339–345. http://dx.doi.org/10.1039/b815310h

39. Hamada Y, Yoshida K, Asai R, Hayase S, Nokami T, Izumi S, et al. A possible means of realizing a sacrifice-free three component separation of lignocellulose from wood biomass using an amino acid ionic liquid. Green Chem. 2013;15(7):1863–1868. http://dx.doi.org/10.1039/c3gc40445e

40. Chen L, Sharifzadeh M, MacDowell N, Welton T, Shah N, Hallett JP. Inexpensive ionic liquids: $[HSO_4]^-$-based solvent production at bulk scale. Green Chem. 2014;16(6):3098–3106. http://dx.doi.org/10.1039/c4gc00016a

41. Simmons BA. Ionic liquid pretreatment. Energy Efficiency and Renewable Energy (EERE) Webinar, 2013 June 24 [document on the Internet]. c2013 [cited 2015 Jan 31]. Available from: http://energy.gov/eere/bioenergy/downloads/ionic-liquid-pretreatment-technologies

42. George A, Brandt A, Tran K, Zahari SMSNS, Klein-Marcuschamer D, Sun N, et al. Design of low-cost ionic liquids for lignocellulosic biomass pretreatment. Green Chem. 2015;17:1728–1734. http://dx.doi.org/10.1039/c4gc01208a

43. Zhang S, Sun N, He X, Lu X, Zhang X. Physical properties of ionic liquids: Database and evaluation. J Phys Chem Ref Data. 2006;35(4):1475–1517. http://dx.doi.org/10.1063/1.2204959

44. Khupse ND, Kumar A. Ionic liquids: New materials with wide applications. Ind J Chem. 2010;49A(05–06):635–648.

45. Rinaldi R. Instantaneous dissolution of cellulose in organic electrolyte solutions. Chem Commun. 2011;47(1):511–513. http://dx.doi.org/10.1039/c0cc02421j

46. Abe M, Fukaya Y, Ohno H. Fast and facile dissolution of cellulose with tetrabutylphosphonium hydroxide containing 40 wt% water. Chem Commun. 2012;48(12):1808–1810. http://dx.doi.org/10.1039/c2cc16203b

47. Chidambaram M, Bell AT. A two-step approach for the catalytic conversion of glucose to 2,5-dimethylfuran in ionic liquids. Green Chem. 2010;12(7):1253–1262. http://dx.doi.org/10.1039/c004343e

48. Hou XD, Smith TJ, Li N, Zong MH. Novel renewable ionic liquids as highly effective solvents for pretreatment of rice straw biomass by selective removal of lignin. Biotechnol Bioeng. 2012;109(10):2484–2493. http://dx.doi.org/10.1002/bit.24522

49. Nockemann P, Thijs B, Driesen K, Janssen CR, Van Hecke K, Van Meervelt L, et al. Choline saccharinate and choline acesulfamate: Ionic liquids with low toxicities. J Phys Chem B. 2007;111(19):5254–5263. http://dx.doi.org/10.1021/jp068446a

50. Hou XD, Li N, Zong MH. Facile and simple pretreatment of sugar cane bagasse without size reduction using renewable ionic liquids-water mixtures. ACS Sustainable Chem Eng. 2013;1(5):519–526. http://dx.doi.org/10.1021/sc300172v

51. Wu H, Mora-Pale M, Miao J, Doherty TV, Linhardt RJ, Dordick JS. Facile pretreatment of lignocellulosic biomass at high loadings in room temperature ionic liquids. Biotechnol Bioeng. 2011;108(12):2865–2875. http://dx.doi.org/10.1002/bit.23266

52. Cruz AG, Scullin C, Mu C, Cheng G, Stavila V, Varanasi P, et al. Impact of high biomass loading on ionic liquid pretreatment. Biotechnol Biofuels. 2013;6:52. http://dx.doi.org/10.1186/1754-6834-6-52

53. Spange S, Fischer K, Prause S, Heinze T. Empirical polarity parameters of celluloses and related materials. Cellulose. 2003;10(3):201–212. http://dx.doi.org/10.1023/a:1025197520736

54. Klamt A. COSMO-RS: From quantum chemistry to fluid phase thermodynamics and drug design. Amsterdam: Elsevier; 2005. http://dx.doi.org/10.1016/B9-78-044451-9/94850-0080

55. Long JX, Li XH, Wang LF, Zhang N. Ionic liquids: Efficient solvent and medium for the transformation of renewable lignocellulose. Sci China Chem. 2012;55(8):1500–1508. http://dx.doi.org/10.1007/s11426-012-4633-7

56. Balaji C, Banerjee T, Goud VV. COSMO-RS based predictions for the extraction of lignin from lignocellulosic biomass using ionic liquids: Effect of cation and anion combination. J Solution Chem. 2012;41(9):1610–1630. http://dx.doi.org/10.1007/s10953-012-9887-3

57. De María PD. Recent trends in (ligno)cellulose dissolution using neoteric solvents: Switchable, distillable and bio-based ionic liquids. J Chem Technol Biotechnol. 2014;89(1):11–18. http://dx.doi.org/10.1002/jctb.4201

58. Shi J, Thompson VS, Yancey NA, Stavila V, Simmons BA, Singh S. Impact of mixed feedstocks and feedstock densification on ionic liquid pretreatment efficiency. Biofuels. 2013;4(1):63–72. http://dx.doi.org/10.4155/BFS.12.82

A review of the control of schistosomiasis in South Africa

AUTHORS:
Kundai Magaisa[1]
Myra Taylor[2]
Eyrun F. Kjetland[2,3]
Panjasaram J. Naidoo[1]

AFFILIATIONS:
[1]Pharmaceutical Sciences, University of KwaZulu-Natal, Durban, South Africa
[2]Public Health Medicine, Howard College Campus, University of KwaZulu-Natal, Durban, South Africa
[3]Norwegian Centre for Imported and Tropical Diseases, Department of Infectious Diseases, Oslo University Hospital, Oslo, Norway

CORRESPONDENCE TO:
Myra Taylor

EMAIL:
Taylor@ukzn.ac.za

POSTAL ADDRESS:
Public Health Medicine, 2nd floor, George Campbell Building, Science Drive, Howard College Campus, University of KwaZulu-Natal, Durban 4001, South Africa

DATES:
Received: 02 Feb. 2015
Revised: 19 May 2015
Accepted: 21 May 2015

KEYWORDS:
neglected tropical disease; praziquantel; *Schistosoma haematobium*; female genital schistosomiasis; schoolchildren

HOW TO CITE:
Magaisa K, Taylor M, Kjetland EF, Naidoo PJ. A review of the control of schistosomiasis in South Africa. S Afr J Sci. 2015;111(11/12), Art. #2014-0427, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20140427

Schistosomiasis is the second most important tropical disease in the world in terms of public health impact. In South Africa, more than 4 million people are estimated to be infected with schistosomiasis. School-age children usually have the highest prevalence and intensity of infection. *Schistosoma haematobium* may result in female genital schistosomiasis which presents as inflammation and ulceration of the genital mucosa and pathological blood vessels. These effects may increase the susceptibility of women with female genital schistosomiasis to HIV. Praziquantel is the drug used to treat schistosomiasis and it is best to treat people during the early stages of infection, before female genital schistosomiasis presents as lesions and sandy patches, as there currently is no treatment for these symptoms. Schistosomiasis is not regarded as a serious public health issue in South Africa despite evidence revealing the seriousness of the disease. In areas endemic for schistosomiasis, the World Health Organization recommends regular mass treatment of all school-age children. In 2001, South Africa became a signatory to the World Health Assembly resolution 54.19 which urged all member states to achieve the minimum goal of 75% treatment coverage in school-age children at risk by 2010. This goal was not achieved in South Africa, despite efforts made by the Department of Health, such as the first statutory school-based geohelminth control programme in the province of KwaZulu-Natal. However, this programme has not been continued. Therefore, there is still much work to be done in order to control and decrease the prevalence of schistosomiasis in endemic areas.

Introduction

Epidemiology

Schistosomiasis (also known as bilharzia) is the second most important tropical disease after malaria, in terms of public health impact.[1] Over 800 million people are at the risk of acquiring schistosomiasis in 76 endemic countries, with 46 of these countries in Africa.[2,3] Approximately 207 million people worldwide are infected with schistosomiasis[4] and 85% of the estimated number of infections are in the less developed areas of sub-Saharan Africa. Worldwide, *Schistosoma haematobium* causes more than half of the schistosomiasis infections, that is, approximately 112 million infections.[4] *S. haematobium* and *S. mansoni* are responsible for about 280 000 deaths annually.[4] In South Africa, more than 4 million people are estimated to be infected with schistosomiasis.[4] The endemic areas in South Africa are in the north and the east, covering about a quarter of the country.[5] The most common species in South Africa is *S. haematobium* . School-aged children usually present the highest prevalence and intensity of infection.[5]

Morbidity caused by schistosomiasis

Schistosomiasis is a systemic helminth infection caused by the *Schistosoma* blood fluke and is a disease of major public health concern.[1,6] Morbidity is determined by the type of schistosome species involved; for example, *S. haematobium* infection may cause haematuria, dysuria, nutritional deficiencies, lesions in the bladder, kidney failure, an elevated risk of bladder cancer, reproductive tract morbidity, and growth retardation in children.[7] Schistosomiasis is found in areas with poor sanitation where humans are in contact with water contaminated by human waste (urine, genital tract excretions and faeces) either as part of their daily lives or during recreational or professional activities.[4,8] Schistosomiasis is transmitted to humans via cercariae released from the *Bulinus* snail.[8] Upon contact with contaminated water, the parasite enters the human body by penetrating the skin. The parasite is then transported via the lymphatic system and the lungs to the liver. Final maturation into dimorphic worms takes place in the liver, with male and female adult worms growing to about 12–20 mm in size.[8] The adult worm pair migrates to the different organs in the body. *S. haematobium* affects the pelvic organs, whereas *S. mansoni* affects the intestines.[8]

Female genital schistosomiasis

According to community based studies in *S. haematobium* endemic areas, 33–75% of female individuals have genital schistosomiasis.[8] Female genital schistosomiasis is the presence of schistosome eggs and/or worms in the female genital organs, in the epithelium of the urinary bladder or genital mucosa where erosions, so-called sandy patches, contact bleeding and oedema may be found.[9-11]

Adult *S. haematobium* worms can be found in the venules surrounding any organ of the upper and lower genital tract.[9] These adult *S. haematobium* worms lay between 20 and 200 eggs per day.[4] A large number of the eggs penetrate the vessel wall of the urogenital tract and become sequestered in the tissues of the urinary bladder, cervix, vagina and fallopian tubes.[4,8] The *S. haematobium* eggs may induce blood vessel proliferation in the tissues surrounding the eggs.[8] The eggs, whether viable or calcified, may also induce an influx of immune cells, such as lymphocytes, macrophages, eosinophils, neutrophils, foreign body giant cells, Langerhans cells and fibroblasts.[11,12] This influx may result in chronic inflammation in the affected organs.[10,11]

The symptoms of female genital schistosomiasis include pelvic pain, postcoital bleeding, inter-menstrual bleeding, genital itch and abnormal vaginal discharge.[4,10] Female genital schistosomiasis may therefore easily be mistaken for a sexually transmitted infection like chlamydia and women may be reluctant to seek medical attention. Case reports and community based studies have further indicated decreased fertility and abortions in some women.[8] Female genital schistosomiasis is also associated with increased frequency of urination and stress incontinence.[8]

Female genital schistosomiasis and HIV infection

It has been hypothesised that other sexually transmitted infections increase the susceptibility of women to HIV infection.[8,13] Increased susceptibility may be because of an influx of immune cells but could also be because of breaches in the integrity of the mucosal barrier of the vagina and the cervix. Similarly, female genital schistosomiasis has been hypothesised to increase HIV susceptibility of women. HIV in semen may have easier access to the systemic circulation of women because of these entry points on the disrupted genital tract epithelium. Contact bleeding during sexual intercourse and other genital mucosal bleeding tendencies associated with female genital schistosomiasis might increase HIV susceptibility in infected women.[14]

Cross-sectional studies conducted in Zimbabwe and Tanzania indicated that women with female genital schistosomiasis had a higher HIV prevalence than women without female genital schistosomiasis.[4] Female genital schistosomiasis was associated with a 2.9–4-fold higher risk of HIV seropositivity. *S. haematobium* infection typically occurs during childhood, preceding HIV exposure, and may pose a risk for HIV in young girls when they become sexually active[4] because the vaginal epithelium would be disrupted before first sexual intercourse in *S. haematobium* infected girls. HIV prevalence seems to peak at a younger age in women in rural areas compared to women in urban areas. It is well known that there is often poor sanitation and a lack of piped water in rural areas and these conditions may increase a woman's risk of contracting schistosomiasis.

Immune cells in the genital lesions caused by *S. haematobium* infection may provide easy points of entry for HIV to attach to susceptible cells.[12] The eggs trapped in the genital mucosa, whether viable or calcified, may both induce an influx of immune cells. CD4 cells are the receptors used by HIV to enter the host cells and a high concentration of CD4 cells is contained in the egg granulomas and the adjacent areas.[13]

It has further been hypothesised that women co-infected with HIV and *S. haematobium* infection may transmit HIV to their partners more easily than HIV-infected women without *S. haematobium* infection.[4] Leukocytes may harbour HIV and conditions that result in recruitment of leukocytes to the genital tract, for example genital ulcer disease or *S. haematobium* infection, may result in an increase in viral shedding of HIV into the female genital tract.[13]

Young people and HIV in South Africa

According to the Joint United Nations Programme on HIV/AIDS (UNAIDS) 2012 fact sheet on adolescents, young people and HIV, there are about 1.6 billion people between the ages of 12 and 24 in the world today.[15] In 2010, 42% of new HIV infections were among youths between the ages of 15 and 24.[15] Among young people living with HIV, approximately 4 of the 5 million live in sub-Saharan Africa.[15] Statistics also show that, globally, women aged 15–24 have HIV infection rates twice as high as young men in the same age group. In South Africa, the National Antenatal Sentinel HIV and Syphilis Prevalence Survey, conducted in 2011 among pregnant women attending public sector clinics and hospitals, reported KwaZulu-Natal as the province with the highest HIV prevalence in the world.[16] KwaZulu-Natal had an HIV prevalence of 37.4% in 2011. Of the women aged between 15 and 24 years tested at antenatal care facilities, 25.5% were HIV positive.[16] The Ugu district in KwaZulu-Natal had the highest HIV prevalence – 41.7%.

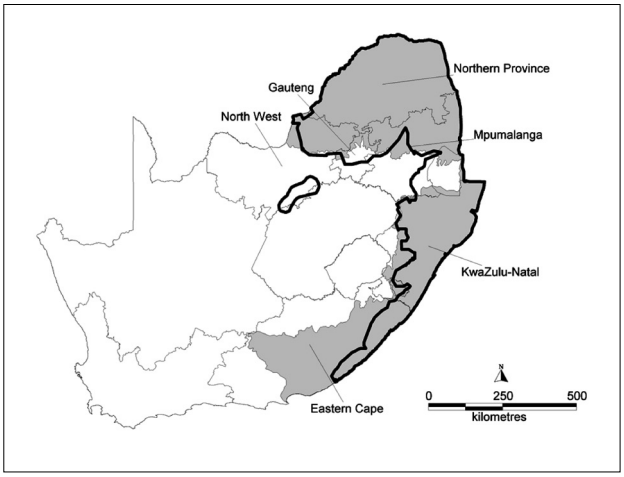
Method

Information in this article was obtained from peer-reviewed articles, published World Health Organization (WHO) reports and WHO statements, exclusively.[17,18] All articles were accessed online on PubMed. Of 106 studies reviewed, similar information was reported in several sources and hence only the references which were significantly relevant to this paper were selected and are quoted.

A brief history of schistosomiasis in South Africa

The first record of urogenital schistosomiasis in South Africa was in 1863. Dr J Harley, a general practitioner in Port Elizabeth, diagnosed boys from Uitenhage in the Eastern Cape with schistosomiasis.[19] This diagnosis was only 12 years after *Schistosoma haematobium* had been discovered by the German physician Bilharz in Egypt in 1851.[20] Case reports between 1864 and 1899 revealed how children, mostly boys between the ages of 3 and 16, were infected with schistosomiasis in the Eastern Cape.[19] Transmission among the boys was as a result of recreational swimming, contact with infested water in open pools, farm dams and furrows. The most common symptom was haematuria. In the 1920s and 1930s, Leipoldt and colleagues reported the wide spread of schistosomiasis in South Africa and negative consequences of the disease, particularly in children.[19]

By 1934, sufficient information had been gathered on urinary schistosomiasis to produce the first map of the geographical distribution of the disease in South Africa.[21] The map was drawn from data collected between the 1860s and 1932 and was published in 1938.[21] Further research and surveys on schistosomiasis were carried out in South Africa after the first map was published and using disease prevalence data collected between the early 1950s through to the late 1970s, the *Atlas of Bilharzia in South Africa* was published by Gear et al.[22] This map was quite similar to the first map published by Porter in 1938.[21]



Source: Appleton and Naidoo[19]

Figure 1: The areas of South Africa endemic for urinary schistosomiasis, as mapped by Porter[21] in 1938 (shaded area) and Gear et al.[22] in 1980 (area within the solid black line).

Efforts to get schistosomiasis recognised as a serious public health issue and to persuade health authorities to take measures to control the disease started as early as the 1960s.[23] After noticing in other surveys how serious the effects of the disease were, Bhagwandeen[24] carried out an investigation of the clinical pathology of urogenital schistosomiasis in Durban. High prevalence as well as high intensity infections were noted for both urinary and intestinal schistosomiasis among Africans.[25] The study also showed severe hepatic damage among the patients. However, despite this study by Bhagwandeen and several other studies, schistosomiasis was still not regarded as a serious health problem by the public health authorities. Therefore, no government-run control programme was carried out in the 1960s, 1970s or 1980s.[25] To justify this omission, the authorities often asked the question, 'How many people die of urogenital schistosomiasis?'[24] By 1989, Utroska estimated the average prevalence of schistosomiasis in the country to be 17.5% with 1 575 000 infected children under 15 years.[25]

Unique cases of schistosomiasis recorded

A study by Pitchford in 1966 reported evidence that *Schistosoma mattheei*, a form of intestinal schistosomiasis found mainly in sheep, had been found in people in Humansdorp, South Africa. However, in 1986, Pitchford reported that the human infection rates caused by *S.mattheei* were not as high as those caused by *S. haematobium* and *S. mansoni*.[26]

There were 14 cases of spinal cord schistosomiasis recorded between August 1985 and March 1987 in KwaZulu-Natal. Twelve patients were of Zulu origin and two patients were Caucasians.[27] Of these, two patients required surgery and three patients were treated with praziquantel only and recovered.[27] The other nine patients were treated with praziquantel and corticosteroids; six patients responded well to treatment (although one patient relapsed) and three patients showed a reduction in conus size.[27] The reported prevalence of spinal schistosomiasis is low and is believed to be an under-reported condition.[27]

In 2006, an 11-year-old girl in Port St Johns (Eastern Cape Province) presented with an itchy genital rash; she had no history of sexual abuse and was in good health.[28] She was later diagnosed with vulval schistosomiasis. This case occurred after a study carried out in the district from 1987 to 1989 revealed a prevalence of 42% of *S. haematobium* infection among schoolchildren in the area.[28] Vulval schistosomiasis is classified under female genital schistosomiasis and is not a common diagnosis in the endemic areas of South Africa. At the Department of Anatomical Pathology at Walter Sisulu University in the Eastern Cape, only 20 cases of this condition were recorded from 1986 to 2000.[28] At the Department of Anatomical Pathology at Nelson R Mandela School of Medicine (also in the Eastern Cape), 27 cases of vulval schistosomiasis were recorded from 1994 to 1998.[28]

Control programme efforts in South Africa

Plans made by South Africa to control schistosomiasis

The success of any schistosomiasis control programme requires long-term commitment. Three important aspects have contributed to the success of control programmes in other countries over the years.[29] These aspects are (1) recognising the public health importance of schistosomiasis, (2) the political will and commitment to utilising local resources for control and (3) readily available public health facilities for the delivery and maintenance of control interventions.[30-32] In line with these control programmes, the Department of Health in South Africa put together a document titled 'The Primary Health Care Package for South Africa – A set of norms and standards', in an attempt to incorporate schistosomiasis management into the primary healthcare system.[33] The norms and standards were intended to be used as planning guidelines by district and provincial health planners to help assess the unmet needs of their population and to draw up plans to bring services up to national standards. The document includes guidelines on the management of schistosomiasis, i.e. treatment, prevention and active surveillance of the disease in primary healthcare facilities in endemic areas in South Africa.[33]

These guidelines[33] on treatment and control strategies require every clinic in endemic areas to have the latest copy of *Bilharzia in South Africa* by Gear et al.[22] Each clinic is required to be well stocked with anthelminthic drugs and must have posters and information leaflets to educate the public on the disease.[33] Any dam, river or pond near a clinic in a schistosomiasis endemic area must have a notice board about the danger of children swimming there if the Environmental Health Officer has identified it as a habitat for host snails.[33] The staff are expected to know whether or not the clinic is in an area endemic for schistosomiasis. The staff should know basic information about schistosomiasis and be able to give correct information to patients on the life cycle of the schistosomiasis host snail and how to prevent future infections.[33] The staff are expected to help with mass prevention or treatment projects initiated by the district, e.g. treating schoolchildren in areas endemic for schistosomiasis.[33] The staff are also expected to educate schoolteachers, pupils and parents on the district advocated control.[33]

This strategy provides an integrated approach involving treatment, prevention and active surveillance, but unfortunately, it has not been implemented in primary healthcare facilities. Schistosomiasis control continues to be a low priority among health authorities in South Africa[34,35] despite the fact that schistosomiasis is endemic in South Africa and remains a public health concern among high risk groups, i.e. pre-schoolers, school-age children, women of childbearing age and adolescents.[22,34] South Africa harbours over 25.7 million people at risk of schistosomiasis, with an estimated 4.5 million infections annually.[35] In order for South Africa to show that schistosomiasis has been successfully controlled, no new infections should be observed over a given period of time, or the prevalence should significantly be reduced.[36] These outcomes can only be assured if a surveillance system such as the one described above is properly implemented in primary healthcare facilities in endemic areas.

WHO recommendations

Globally, schistosomiasis is regarded as an important health issue which requires attention because of the significant negative consequences it has both on the economy and on public health.[32,37] Interest in schistosomiasis control was revived during the 54th World Health Assembly held in 2001.[38] Although WHO recommended large-scale distribution of praziquantel to at-risk populations living in endemic areas in the 1970s and 1980s, it was only at the 54th World Health Assembly that chemotherapy was endorsed officially as the key public-health strategy to fight schistosomiasis. South Africa endorsed resolution WHA54.19, which calls on member states to take effective steps to control schistosomiasis and soil transmitted helminths.[38] The following is an excerpt from the resolution WHA54.19.[38] All member states were urged to[38]:

1. sustain successful control activities in low-transmission areas in order to eliminate schistosomiasis and soil-transmitted helminth infections as a public-health problem
2. give high priority to implementing or intensifying control of schistosomiasis and soil-transmitted helminth infections in areas of high transmission while monitoring drug quality and efficacy
3. ensure access to essential drugs against schistosomiasis and soil-transmitted helminth infections in all health services in endemic areas for the treatment of clinical cases and groups at high risk of morbidity such as women and children, with the goal of attaining a minimum target of regular administration of chemotherapy to at least 75% and up to 100% of all school-age children at risk of morbidity by 2010
4. promote access to safe water, sanitation and health education
5. ensure that any development activity likely to favour the emergence or spread of parasitic diseases is accompanied by preventive measures to limit its impact

In 2009, urinary schistosomiasis was renamed 'urogenital schistosomiasis' because in 75% of infected people, the disease affects both the urinary and genital tracts.[4]

What South Africa has achieved

In October 1997, South Africa set up its first helminth control programme in KwaZulu-Natal. This programme was funded and run entirely by the government.[23] The programme targeted both geohelminths (*Ascaris lumbricoides*, *Necator americanus*, *Trichuris trichiura*) and schistosomaiasis.[23] This was a primary school based programme and about 1.5 million primary school children were to be treated with praziquantel on a regular basis in high intensity areas.[31,36] The programme was in line with WHO guidelines. Children are likely to benefit from treatment because of the early stage of their chronic lesions and because schoolgoing children present with the highest prevalence and infection intensities of schistosomiasis.[2,36,37] The increasing number of children attending school in KwaZulu-Natal provided a good opportunity for a cost-effective programme.[39]

The overall aim of the first helminth control programme in KwaZulu-Natal was to significantly reduce prevalence and morbidity resulting from schistosomiasis and soil-transmitted helminths by the end of 2010.[23] The goals of the programme were to[23]:

1. permanently reduce morbidity caused by helminth infections in children living in endemic areas
2. improve sanitation (e.g. build school toilets, build latrines in the community, identify cheap building materials)
3. create links between parents, teachers and health personnel in the area in order to fight poor health amongst children
4. improve the children's learning potential by improving their health status

Children are especially vulnerable to schistosomiasis and may become physically and intellectually compromised because infection may result in anaemia, attention deficit disorder, learning disabilities, school absenteeism and high school dropout rates.[3,5,34,40] The effects on child development may result in a generation of adults disadvantaged by the irreversible consequences of this infection.[42] Endemic areas were identified using the *Atlas of Bilharzia in Southern Africa*[23] which helped to provide estimates of the population at risk of the disease and to adequately plan for the control programme.

A high total prevalence and a large proportion of high intensity infections of *S. haematobium* were observed.[24,40] This observation indicated that, according to the WHO guidelines, it is necessary to regularly treat schoolchildren in KwaZulu-Natal with praziquantel.[24,40] According to Appleton and Kvalsvig[23], a decline in infection was recorded during this intervention. A significant reduction (95.3%) in egg excretion 3 weeks after treatment and a 94.1% cure rate for heavy infections were recorded.[3] It was proposed that one treatment per year, at the end of summer, was sufficient to keep infection and intensities low.[23] Unfortunately, this programme did not gain the necessary support and there is currently no treatment strategy of any kind in any area of the country.[3]

Praziquantel

Praziquantel is the drug of choice for the treatment of all forms of schistosomiasis.[3] It is highly effective in killing adult worms.[3] With praziquantel, the cure rates for *S. haematobium* are 75–85% and 63–85% for *S. mansoni*.[3] In studies conducted in North, West and East Africa, South America and Asia, mass treatment with praziquantel in endemic areas has been shown to reduce the burden of urogenital schistosomiasis.[1,3,41] Studies have shown that a single treatment given in childhood seems to prevent half the cases of genital schistosomiasis in women.[8] Therefore, it is vital to do an early intervention through mass treatment programmes in endemic communities in South Africa, to prevent the serious impact of schistosomiasis.

Currently, praziquantel is given as a single oral dose of 40 mg/kg, with food and drink, in order to minimise gastrointestinal side effects.[3,42] In South Africa, only Bayer's Biltricide® (praziquantel) is available for the treatment of schistosomiasis.[3] The Department of Health purchases praziquantel for schistosomiasis at a cost of USD4.49 per tablet making mass treatment programmes unaffordable and almost impossible to run in South Africa.[3] Praziquantel is not even fully stocked in local clinics at the moment as it is too expensive.[3]

Contrary to the current practice in South Africa, other countries accept WHO-accredited generic praziquantel.[3] These simple, effective, cheap and commercially available generics are of excellent quality, have been on the market for more than 20 years[3] and have been used widely in schistosomiasis control programmes. Countries with successfully established schistosomiasis national control programmes using generic praziquantel supplied free of charge include Zambia, Nigeria, Mozambique and Uganda.[43] There has been no evidence of development of resistance to praziquantel. These generics are not available in South Africa because only one brand of praziquantel is registered for the treatment of schistosomiasis.[3] Therefore, for South Africa to be able to successfully control schistosomiasis and its debilitating consequences, donated or inexpensive and affordable generic medication must be made available for mass treatment programmes in endemic areas.

WHO strategic plan to eliminate schistosomiasis by 2025

Over the years, several countries have implemented schistosomiasis control programmes successfully. These countries include Brazil, Burkina Faso, China, Egypt and Uganda.[44] Significant reduction in urogenital schistosomiasis infection and morbidity has been achieved by mass treatment. In countries that have not been able to successfully control schistosomiasis, the major obstacle has been limited access to praziquantel as well as unavailability of water, sanitation, environmental health infrastructure or the necessary resources to implement schistosomiasis control.

The WHO came up with a strategic plan to eliminate schistosomiasis in endemic countries by 2025.[44] This plan can be used as a guide by the South African government. One of the goals is to successfully control schistosomiasis morbidity by 2020. This goal has been defined as 75% national coverage for schistosomiasis control and less than 5% prevalence of heavy infections recorded in endemic countries.[44] If this goal is achieved by 2020, it may be possible to eliminate schistosomiasis as a public-health problem in South Africa by 2025. By increasing schistosomiasis control activities in endemic areas and ensuring an adequate supply of praziquantel and the other resources required, this goal is attainable. The monitoring and evaluation of schistosomiasis control can be undertaken by local health facilities. Primary health facilities and school health teams across the country should be supplied with praziquantel for treating infected patients. The communities should also be provided with potable water in order to avoid contact with contaminated water as well as adequate sanitation to decrease contamination of water bodies.[44] Educational programmes are needed to encourage less recreational water contact, and bridges or stepping stones across rivers need to be provided.

Conclusion and recommendations

South Africa still has much to achieve in order to successfully control and decrease the prevalence of schistosomiasis in endemic areas. South Africa was unable to achieve the minimum goal of 75% treatment coverage in school-aged children at risk by 2010 as stated by resolution WHA54.19 of the World Health Assembly. Several factors contributed to this failure: the lack of support from public health authorities as the government still underestimates the seriousness of schistosomiasis, the price of praziquantel in South Africa and the lack of availability of praziquantel in public health facilities as well as the lack of blanket permission to use WHO-accredited generic drugs. Failure to incorporate schistosomiasis control into the general healthcare system, e.g. failure of primary healthcare systems to perform specific control tasks, is another contributing factor. Successful schistosomiasis control can only be achieved by taking into account other factors such as health education, improvements in sanitation (including preventing urination and defecation in or near the water body by building latrines and educating people on the importance of using them) and lastly, mass-treating the population in endemic areas. The latter method has been found to have a dual effect, by also reducing short- and even longer-term morbidity.[1]

To improve the control of schistosomiasis, a comprehensive primary healthcare district based schistosomiasis control programme needs to be established within the primary healthcare system. This programme could be led by school health teams at district level. It is important that there is increased awareness of the deleterious effects of schistosomiasis and the school health teams should provide health education to ensure health literacy about schistosomiasis amongst educators and learners and increase such awareness in the community. This programme could contribute to strengthening the health system as it offers the opportunity for linking mobile teams undertaking preventive work with schools, thus enabling the health service to reach many South African children before they need to access a primary healthcare clinic. The current initiative to re-engineer primary healthcare provides an opportunity for such efforts.

As a Schedule 4 drug, praziquantel currently must be dispensed by a professional nurse and the rescheduling of this drug, which does not

have serious side effects, should be considered. Urgent action is also required to register the less expensive generic forms of praziquantel tablets in South Africa as many of these generics have been, and are currently used in other countries throughout the world. The negative effects of female genital schistosomiasis on the genital system is now recognised, as is the association with HIV infection. A well-structured schistosomiasis control programme therefore has the potential to contribute to a reduction in the prevalence of HIV amongst young women.

Acknowledgements

We acknowledge those who are working to ensure that South African children are treated timeously for schistosomiasis.

Authors' contributions

K.M. drafted the article; P.J.N. was the project leader; M.T., P.J.N. and E.F.K. made conceptual contributions, added to the content and approved the final manuscript.

References

1. Fenwick A, Webster JP, Blair L, Fleming FM, Zhang Y, Garba A. The schistosomiasis control initiative (SCI): Rationale, development and implementation from 2002–2008. Parasitology. 2002;136(13):1719–1730. http://dx.doi.org/10.1017/S0031182009990400
2. Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: Systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis. 2009;6(7):411–425. http://dx.doi.org/10.1016/S1473-3099(06)70521-7
3. Berge S, Kabatereine N, Gundersen SG, Taylor M, Kvalsvig JD, Mkhize-Kwitshana Z, et al. Generic praziquantel in South Africa: The necessity for policy change to provide cheap, safe and efficacious schistosomiasis drugs for the poor, rural population: Review. South Afr J Epidemiol Infect. 2011;26(1):22–25.
4. Mbabazi P, Andan O, Fitzgerald DW, Chitsulo L, Engels D, Downs JA. Examining the relationship between urogenital schistosomiasis and HIV infection. PLoS Negl Trop Dis. 2011;5(12):e1396. http://dx.doi.org/10.1371/journal.pntd.0001396
5. Saathoff E, Olsen A, Magnussen P, Kvalsvig JD, Becker W, Appleton CC. Patterns of *Schistosoma haematobium* infection, impact of praziquantel treatment and re-infection after treatment in a cohort of schoolchildren from rural KwaZulu-Natal/South Africa. BMC Infect Dis. 2004;4(1):40. http://dx.doi.org/10.1186/1471-2334-4-40
6. World Health Organization. Schistosomiasis fact sheet no. 115 [document on the Internet]. c2012 [cited 2014 Apr 04]. Available from: http://www.who.int/mediacentre/factsheets/fs115/en/index.html
7. King CH, Dangerfield-Cha M. The unacknowledged impact of chronic schistosomiasis. Chronic Illn. 2008;4(1):65–79. http://dx.doi.org/10.1177/1742395307084407
8. Kjetland EF, Leutscher P, Ndhlovu P. A review of female genital schistosomiasis. Trends Parasitol. 2012;28(2):58–65. http://dx.doi.org/10.1016/j.pt.2011.10.008
9. Poggensee G, Kiwelu I, Weger V, Göppner D, Diedrich T, Krantz I, et al. Female genital schistosomiasis of the lower genital tract: Prevalence and disease-associated morbidity in northern Tanzania. J Infect Dis. 2000;181(3):1210–1213. http://dx.doi.org/10.1086/315345
10. Kjetland EF, Kurewa EN, Ndhlovu PD, Midzi N, Gwanzura L, Mason PR, et al. Female genital schistosomiasis – A differential diagnosis to sexually transmitted disease: Genital itch and vaginal discharge as indicators of genital *Schistosoma haematobium* morbidity in a cross-sectional study in endemic rural Zimbabwe. Trop Med Int Health. 2008;13(12):1509–1517. http://dx.doi.org/10.1111/j.1365-3156.2008.02161.x
11. World Health Organization (WHO). Report of an informal working group on urogenital schistosomiasis and HIV transmission. Geneva: WHO; 2009. Available from: http://whqlibdoc.who.int/hq/2010/WHO_HTM_NTD_PCT_2010.5_eng.pdf
12. Harms G, Feldmeier H. Review: HIV infection and tropical parasitic diseases – Deleterious interactions in both directions? Trop Med Int Health. 2002;7(6):479–488. http://dx.doi.org/10.1046/j.1365-3156.2002.00893.x
13. Johnson LF, Lewis DA. The effect of genital tract infections on HIV-1 shedding in the genital tract: A systematic review and meta-analysis. Sex Transm Dis. 2008;35(11):946–959. http://dx.doi.org/10.1097/OLQ.0b013e3181812d15
14. Jourdan PM, Roald B, Poggensee G, Gunderson SG, Kjetland EF. Increased vascularity in cervicovaginal mucosa with *Schistosoma haematobium* infection. PLoS Negl Trop Dis. 2011;5(6):e1170. http://dx.doi.org/10.1371/journal.pntd.0001170
15. UNAIDS. Fact sheet: Adolescents, young people and HIV [document on the Internet]. c2012 [cited 2014 Apr 04]. Available from: http://www.unaids.org/en/media/unaids/contentassets/documents/factsheet/2012/20120417_FS_adolescentsyoungpeoplehiv_en.pdf
16. Department of Health. 2011 National Antenatal Sentinel HIV and Syphilis Prevalence Survey, South Africa. Pretoria: Department of Health; 2012. Available from: http://www.health-e.org.za/2012/12/11/2011-national-antenatal-hiv-syphilis-prevalence-survey-released/
17. World Health Organization (WHO). Making health services adolescent friendly. Geneva: Department of Maternal, Child and Adolescent Health, WHO; 2012. Available from: http://apps.who.int/iris/bitstream/10665/75217/1/9789241503594_eng.pdf
18. World Health Organization (WHO). Adolescent friendly health services: An agenda for change. Geneva: WHO; 2002. Available from: http://whqlibdoc.who.int/hq/2003/WHO_FCH_CAH_02.14.pdf
19. Appleton CC, Naidoo I. Why did schistosomiasis disappear from the southern part of the Eastern Cape? S Afr J Sci. 2012;108(1/2), Art. #411, 11 pages. http://dx.doi.org/10.4102/sajs.v108i1/2.411
20. Rabello A. Diagnosing schistosomiasis. Mem Inst Oswaldo Cruz. 1997;92(5):669–676.
21. Porter A. The larval trematodes found in certain South African Mollusca with special reference to schistosomiasis (bilharziasis). Johannesburg: South African Institute of Medical Research; 1938.
22. Gear JHS, Pitchford RJ, Van Eede JA. Atlas of bilharzia in South Africa. Johannesburg: South African Institute for Medical Research; 1980.
23. Appleton CC, Kvalsvig JD. A school-based helminth control programme successfully implemented in KwaZulu-Natal. South Afr J Epidemiol Infect. 2006;21(2):55–67.
24. Bhagwandeen SB. The clinico-pathological manifestations of schistosomiasis in the African and the Indian in Durban. Pietermaritzburg: University of Natal Press; 1968.
25. Utroska JA, Chen MG, Dixon H, Yoon S, Helling-Borda M, Hogerzeil HV, et al. An estimate of global needs for praziquantel within schistosomiasis control programmes. Geneva: WHO; 1989.
26. Pitchford RJ. Some aspects of bilharzia in southern Africa. S Afr Med J. 1986;70(suppl):80–82.
27. Haribhai HC, Bhigjee AI, Bill PLA, Pammenter MD, Modi G, Hoffmann M, et al. Spinal cord schistosomiasis: A clinical, laboratory and radiological study, with a note on herapeutic aspects. Brain. 1991;114(2):709–726. http://dx.doi.org/10.1093/brain/114.2.709
28. O'Mahony D, Banach L, La Grange LL. A case of vulval schistosomiasis. S Afr Fam Pract. 2007;49(9):42–43. http://dx.doi.org/10.1080/20786204.2007.10873640
29. Utzinger J, Bergquist R, Shu-Hua X, Singer BH, Tanner M. Sustainable schistosomiasis control – The way forward. Lancet. 2003;362(9399):1932–1934. http://dx.doi.org/10.1016/S0140-6736(03)14968-9
30. Chitsulo L, Engels D, Montresor A, Savioli, L. The global status of schistosomiasis and its control. Actatropica. 2000;77(1):41–51. http://dx.doi.org/10.1016/s0001-706x(00)00122-4
31. Engels D, Chitsulo L, Montresor A, Savioli L. The global epidemiological situation of schistosomiasis and new approaches to control and research. Actatropica. 2002;82(2):139–146. http://dx.doi.org/10.1016/s0001-706x(02)00045-1
32. World Health Organization (WHO). Prevention and control of schistosomiasis and soil-transmitted helminthiasis. A report of the WHO Expert Committee. Technical report series 912. Geneva: WHO; 2002.
33. Department of Health. The primary health care package for South Africa – A set of norms and standards. Pretoria: Department of Health; 2000. Available from: http://www.hsph.harvard.edu/population/vaccination/southafrica2.doc
34. Johnson CL, Appleton CC. Urban schistosomiasis transmission in Pietermaritzburg, South Africa. South Afr J Epidemiol Infect. 2005;20(3):103–107.
35. Wolmarans CT, De Kock KN. The influence of health education on the prevalence, intensity and morbidity of *Schistosoma haematobium* infections in children over a two-year period in the Limpopo Province, South Africa. South Afr J Epidemiol Infect. 2009;24(1):13–17.

36. Moodley I, Kleinschmidt I, Sharp B, Craig M, Appleton CC. Temperature-suitability maps for schistosomiasis in South Africa. Ann Trop Med Parasitol. 2003;97(6):617–627. http://dx.doi.org/10.1179/000349803225001445

37. Taylor M. Global trends in schistosomiasis control. Bull World Health Organ. 2008;86(10):738. http://dx.doi.org/10.2471/BLT.08.058669

38. World Health Organization (WHO). WHA54.19 Schistosomiasis and soil-transmitted helminth infections. Resolutions and decisions 1 [document on the Internet]. No date [cited 2014 Apr 04]. Available from: http://www.who.int/neglected_diseases/mediacentre/WHA_54.19_Eng.pdf

39. Taylor M, Jinabhai CC, Couper I, Kleinschmidt I, Jogessar VB. The effect of different anthelmintic treatment regimens combined with iron supplementation on the nutritional status of schoolchildren in KwaZulu-Natal, South Africa: A randomized controlled trial. Trans R Soc Trop Med Hyg. 2001;95(2):211–216. http://dx.doi.org/10.1016/S0035-9203(01)90171-3

40. Guyatt HL, Brooker S, Kihamia CM, Hall A, Bundy DA. Evaluation of efficacy of school-based anthelmintic treatments against anaemia in children in the United Republic of Tanzania. Bull World Health Organ. 2001;79(8):695–703.

41. King CH. Parasites and poverty: The case of schistosomiasis. Actatropica. 2010;113(2):95–104. http://dx.doi.org/10.1016/j.actatropica.2009.11.012

42. King CH. Lifting the burden of schistosomiasis – Defining elements of infection-associated disease and the benefits of antiparasite treatment. J Infect Dis. 2007;196(5):653–655. http://dx.doi.org/10.1086/520522

43. World Health Organization (WHO). Report of the WHO informal consultation on the use of praziquantel during pregnancy/lactation and albendazole/mebendazole in children under 24 months. Geneva: WHO; 2002. p. 1–34.

44. World Health Organization (WHO). Schistosomiasis: Progress report 2001–2011 and strategic plan 2012–2020. Geneva: WHO; 2013. Available from: www.who.int/iris/bitstream/10665/78074/1/9789241503174_eng.pdf

Sweet potato (*Ipomoea batatas* L.) as a drought tolerant and food security crop

AUTHORS:
Nozipho M. Motsa[1]
Albert T. Modi[1]
Tafadzwanashe Mabhaudhi[1]

AFFILIATION:
[1]School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa

CORRESPONDENCE TO:
Nozipho Motsa

EMAIL:
mgcibelomotsa@yahoo.co.uk

POSTAL ADDRESS:
School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Private Bag X1, Scottsville 3209, South Africa

DATES:
Received: 28 July 2014
Revised: 09 Oct. 2014
Accepted: 16 June 2015

KEYWORDS:
physiology; water stress; phytochemicals; nutrition; utilisation

HOW TO CITE:
Motsa NM, Modi AT, Mabhaudhi T. Sweet potato (*Ipomoea batatas* L.) as a drought tolerant and food security crop. S Afr J Sci. 2015;111(11/12), Art. #2014-0252, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140252

Sweet potato is an important 'indigenised' root crop in South Africa. It features prominently in smallholder cropping systems because of its versatility, drought tolerance and positive role in food security. It outranks most staple crops in vitamins, minerals, dietary fibre and protein content. Much information is available on sweet potato as a drought tolerant and food security crop, but critical reviews that link its drought tolerance with food security are lacking. We review sweet potato as a food security crop, focusing on mechanisms associated with drought. We conclude that the crop has great potential in the light of imminent challenges associated with drought as a negative effect of climate change.

Introduction

Globally, sweet potato (*Ipomoea batatas* L.) is considered an important, versatile and underutilised food security crop.[1,2] It belongs to the second most important set of food crops in developing countries, namely root crops.[1] The crop is highly nutritive, and it outranks most carbohydrate foods in terms of vitamin, mineral, dietary fibre and protein content.[2,3] A recent report from the Food and Agriculture Organization (FAO)[4] indicated that 24.8% of the population in southern Africa remain undernourished and agricultural projects and crops that can deliver quick results were needed to meet the first 2015 millennium development goal. Sweet potato is one crop that has a wide ecological adaptation, drought tolerance and a short maturity period of 3 to 5 months.[3] It can also be harvested sequentially, thus ensuring continuous food availability and access, an important dimension of food security.

Prior to 1986, sweet potato was regarded as a low value, low status and highly perishable commodity in southern Africa.[5] It received little research attention until the Southern African Root Crops Research Networks (SARRNET) intervened by distributing germplasm and encouraging demand-led research. This initiative prompted the release of several new varieties within southern Africa.

Orange-fleshed and yellow-fleshed cultivars have been recognised as good sources of β-carotene, a precursor of vitamin A[6,7] and are promoted across the developing world. According to Laurie[8], vitamin A deficiency in South Africa is still a serious health problem. Studies by Labadarios[9] reported that 64% of children between the ages of 1 and 9 years, and 27% of women of reproductive age, were vitamin A deficient in South Africa. The highest prevalence was concentrated in Limpopo (43.5%) and KwaZulu-Natal (38.9%). Poor education of mothers was highlighted as the main cause of vitamin A deficiency in rural communities.[6] This deficiency suggests that nutrition education, not just literacy, needs to be part of participatory studies on food security in the rural areas affected by poverty and malnutrition.

In South Africa, sweet potato is grown throughout the tropical regions, including KwaZulu-Natal.[8] Its consumption is steadily increasing despite relatively low production levels.[10] Average yields are ten times lower for small-scale farmers compared with commercial growers. Laurie and Magoro[11] reported mean yields of 3.9–9.5 t/ha on communal gardens compared with 25.2 t/ha at experimental stations. These yields confirm the large yield gap that exists between poor and well-equipped environments. Because sweet potato is considered an important source of food, particularly in the dry season,[12] its importance for improving food security cannot be overemphasised. Despite the reported low yields among smallholder farmers, the crop has flexible planting and harvesting schedules, especially in frost-free areas.[13]

Sweet potato is a drought tolerant crop with the potential to enhance food and nutrition security, especially for subsistence and small-scale farmers in South Africa. For sweet potato, drought stress accounts for 25% total annual yield loss[14] compared to more than 50% yield loss or complete failure in staple crops such as maize.[15] Despite this advantage, its importance as a food security crop is still underestimated and fails to attract sufficient attention from agricultural researchers. The aim of this review was to illustrate, from an agronomic perspective, the importance of sweet potato as a traditional drought tolerant and potential food security crop.

Sweet potato as a traditional crop

Sweet potato is amongst the oldest crops in the world, especially in the wet tropics, and was among the first staple crops before the introduction of cereals. Today it is counted among root crops (cassava, sweet potato, yams and aroids) that represent the second most important set of staple foods in developing countries, closely following cereal crops.[1,2] According to Battisti[16], sweet potato is a misunderstood vegetable, often confused with yams. Portuguese explorers are believed to have introduced sweet potato to Africa in the 16th century.[1,17] From there it was taken to the eastern and western parts of the world in the form of storage roots by Spanish explorers. When it failed to yield in European countries because of low temperatures, it was brought back to Africa and was planted in the warm coastal regions where it spread rapidly.[1,18]

According to Laurie[8], sweet potato was introduced to South Africa in 1652 around the time when the Dutch, under the leadership of Jan van Riebeeck colonised the Cape. It is now referred to as one of the traditional crops[19] and a plant of the past[1] because of the time it arrived in Southern Africa. Sweet potato has been indigenised[13,19] and classified as a neglected and underutilised crop species. About 12 cultivars were released by the Agricultural Research Council between 1959 and 1989: 7 new (cream-fleshed) varieties and 1 locally bred orange-fleshed variety (Serolane) were released (in South Africa) for local production during the 2003/2004 season.[11,20] According

to Laurie[8], nine of the released varieties to date are orange-fleshed sweet potato (OFSP) varieties. Production of these varieties has also spread throughout the tropical and warm regions of the country, the main producing regions being the Western Cape, northeast Limpopo, Free State, Eastern Cape, Gauteng, Mpumalanga lowveld, and northern KwaZulu–Natal Provinces.

KwaZulu-Natal environment for sweet potato production

KwaZulu-Natal is a subtropical region situated on the east coast of South Africa with diverse climate, soils and topography, virtually making it a 'world-in-one'.[21] It is a summer rainfall area (600–2000 mm/annum) with weather that ranges from extremely hot along the coast in summer to heavy snow in the mountainous midlands in winter.[21,22] KwaZulu-Natal has three climate zones of which 1% is tropical, 44% subtropical and 55% temperate.[22] The environment in the tropical and subtropical parts of the province are regarded as suitable for sweet potato production throughout the year. Depending on whether farmers produce the white or OFSP varieties, the high prevalence of vitamin A deficiency[7,9] in the province could be controlled through the consumption of this vegetable (Figure 1).

Variations in the climatic zones of KZN are bound to bring about several abiotic and biotic stress conditions, which have direct and indirect effects on the crop.[23] Sweet potatoes are very sensitive to frost and require hot days and warm nights. The suitable conditions prevailing in the bioresource group areas along the coast, in parts of the midlands and the northern areas of Zululand,[21,22] are dominated by smallholder farmers. The crop has a cropping season of 4–5 months so smallholder farmers can plant sweet potatoes from August to March, thus spreading the cropping season. Underutilised crops such as sweet potato tend to possess characteristics adaptable to abiotic stress conditions. These conditions are predicted to become even more prevalent in the coming decades as a result of the effects of climate change.[24] The current and predicted challenges call for an agricultural system that is able to provide food crops that are capable of meeting current and future food demands and nutrition security. It is necessary to investigate whether the province that has a high population of rural poor people, namely KwaZulu-Natal, can be used as the major production areas for drought tolerant cultivars of sweet potatoes for food security and commercialisation in the near future.

Importance and utilisation

Sweet potato is regarded as the most important root crop of the tropics because of its flexibility in a number of production aspects.[13] It can be planted and harvested at any time of the year, especially in frost free areas. It has a short cropping season, uses non-edible parts as planting material, and has a non-trellising growth habit as well as low requirement for soil nutrients.[1,5,10] Bovell-Benjamin[2] reported that it produces more edible energy than any other major food crop. It is more productive within short periods on marginal lands and plays an important role in the economy of poor households.[12] The most important edible sweet potato parts are the roots and immature leaves, which are used for human consumption, animal feed and to some extent, for industrial purposes.[2,25]

It has been reported that in some African countries, starchy crops such as sweet potato are the staple food while other countries use it as an additional or as a food security crop.[14,26] The use of sweet potato for food security concurs with earlier reports that sweet potato storage roots were often kept in the ground and harvested when needed.[26] Keeping the roots in the field not harvested is an indication that sweet potato can provide a continuous food supply during the off season with no requirement for expensive storage infrastructure. It becomes important in the diet seasonally, typically in the month or two before the major grain harvest or when the grain stock from the previous year has been exhausted. Sweet potato also provides a food reserve when the major grain crop fails because of drought[15] and pest infestation.

Storage roots can be boiled, baked or roasted with some people preferring to eat them raw. The leaves can also be consumed as a green leafy vegetable.[10,25] Sweet potato leaves may possibly have some medicinal properties as the polyphenol-rich green extracts are reported to play a role in reducing prostate cancer.[25,27,28] Reports from China, the leading global sweet potato producer, indicate that they also use sweet potato for starch production and as a raw material for biofuel production.[1,29] The plant sap (juice) of red sweet potato has been reported to play a vital role in producing dye for cloth in South America. The attractive foliage adds aesthetic value[30] and is another non-nutritive use of sweet potato.

Flavonoid compounds of fruit and vegetables are considered therapeutic agents because they have health benefits such as their supposed protection against certain cancers, cardiovascular diseases and aging.[27,28] It is necessary to investigate if and how drought affects these properties because phenols and flavonoids may be responsive to stress.

Nutritional value

The sweet potato root provides almost a balanced diet for the human body. The crop contains significant amounts of carbohydrates when compared to other starchy crops such as rice, maize and sorghum porridge, although the protein content is slightly lower than in potatoes and other grain crops (Table 1).[1] It contains almost all the macro- and micro-nutrients, substantial quantities of vitamin C, moderate amounts of vitamin B complex (vitamins B1, B2, B5 and B6) and folic acid, as well as satisfactory amounts of vitamin E.[2,31]

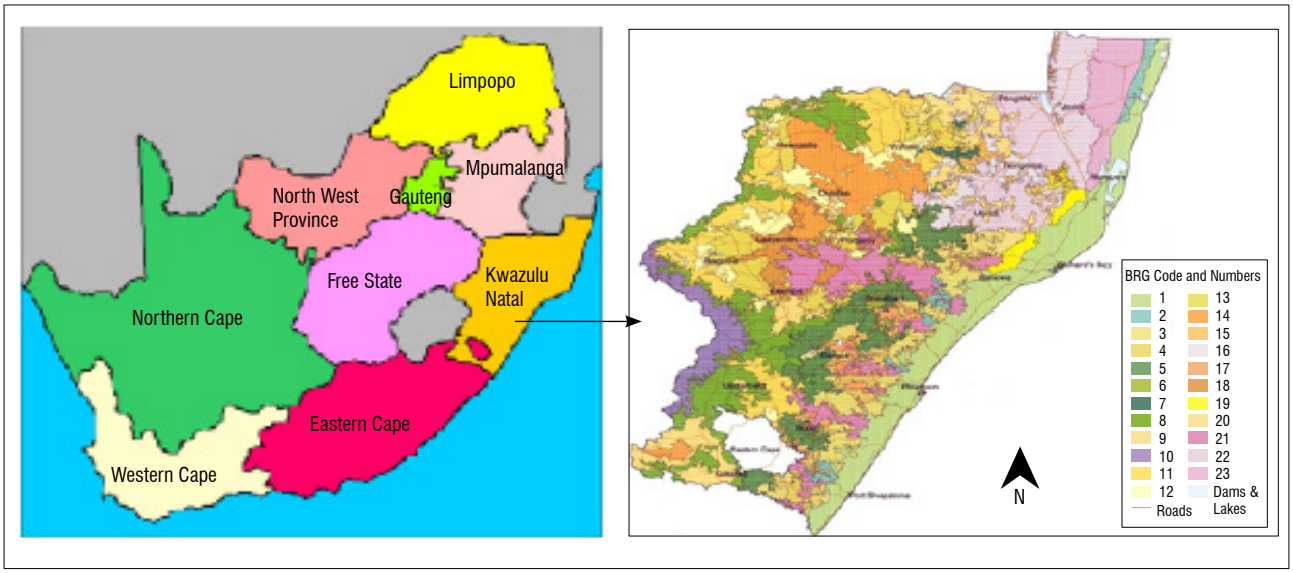


Figure 1: Map of KwaZulu-Natal and the different agro-ecological zones.

Table 1: Macro-nutrient content per 100 g boiled portion of sweet potato compared to other crops

Food	Moisture content (%)	Energy		Dietary fibre	Total carbohydrates	Fat	Protein
		kcal	KJ	(g)			
Sweet potato	71	114	477	2.4	26.3	0.4	1.7
Rice	68	135	565	0.8	28.0	0.3	2.3
Maize porridge	81	76	318	–	15.6	0.8	1.8
Sorghum porridge	80	85	356	–	17.0	0.5	2.7
Pasta	66	132	552	–	26.7	0.7	4.1
Potato	78	80	335	2.1	18.5	0.1	2.1

Source: Adopted from Woolfe[1]

Orange-fleshed sweet potato has been reported by several authors to contain high amounts of vitamin A and is currently being promoted by the FAO and other in-country programmes as a supplementary food to combat vitamin deficiencies in children.[2,3,7,20] According to Laurie[20], South Africa is also using orange-fleshed sweet potato varieties in food diversification programmes to alleviate vitamin A deficiencies. Such programmes include school and community garden projects within the integrated nutrition programme.[32,33]

Sweet potato leaves are also nutritious (Table 2). They are a good source of protein and vitamins A, C and B_2 (riboflavin).[1] According to Bovell- Benjamin[2], sweet potato leaf protein content is twice as much as that from the storage roots. Leaves are also an excellent source of lutein.[34,35] Lutein is responsible for central vision in the human eye and helps to keep eyes safe from oxidative stress and high-energy photons of blue light. Sweet potato is also popular in low fat diets and is recommended as a low glycaemic index (GI) food.[2,13] Given the ailing nutritional status of many South African communities,[11] plus the fact that many food sources which typically provide these nutrients are not as drought tolerant, the sweet potato, with its high nutritional content, is a very important food source.[10]

General food security aspects

The world is reported to produce enough food to feed the human population, yet approximately 30% of sub-Saharan Africa's population (218 million people) suffers from chronic hunger and malnutrition.[4] According to the World Health Organization[36], food security is built upon three pillars: availability, access and use (utilisation and stability). In trying to address the food availability pillar of food security, South Africa has identified small-scale agriculture as an important sector. This sector still needs to be investigated as little effort has been made in channelling investment towards it. Rising food prices, particularly of maize and wheat, which are the staple diet for the poor in South Africa, poses serious problems for urban and rural poor people.[37]

From the perspective of food security, sweet potato is an excellent crop as it often survives where staple crops fail.[1-3,13] It contributes to food availability by providing high output per unit of land (yields about 3.9–9.5 t/ha under communal agriculture compared to less than 1–3 t/ha reported for maize under similar management) and can produce on marginal soils.[1,11] The fact that it has higher energy (Table 1) than maize, also adds more strength to its potential role as a food security crop. Moreover, the fact that new OFSP varieties also contribute significantly towards vitamin A supplementation in rural households, where vitamin A deficiency is more prevalent, makes it a crop of choice. Good health is considered an outcome indicator for food utilisation in food security measurement,[4] and the high nutritional status of sweet potato qualifies the crop as a food security alternative. Improved early maturing varieties are often ready for harvest in 3–5 months and can be harvested as needed over several months.[26,38] The benefit of harvesting early and consuming over several months implies that there is a quick turnover and a lasting source of food, hence, improved food access, availability and stability. The food security benefits of sweet potato mentioned above could also lead to the crop being removed from the 'underutilised' category.

Accessibility to sweet potato is also simple because the vines and roots can be easily stored or multiplied. In drought prone areas, roots can be sprouted to produce vines so that communities have access to planting material. According to Lebot[39], sweet potato has often been a lifesaver, for example, it saved the Japanese nation when typhoons destroyed all their rice fields just before World War I. During the early 1960s, China was

Table 2: Nutrient composition of sweet potato leaves compared to other common vegetables

Crop	Total protein	β-carotene	Calcium	Iron	Riboflavin	Ascorbic acid
	(g/100 g)	(μg/100 g)	(mg/100 g)			
Sweet potato leaves	29	2700	75–183	1.8–3.9	0.35	32–136
Amaranth	28	6545	176	2.8	0.22	23
Cabbage	1.9	–	44	0.4	–	–
Carrot	0.7	–	48	0.6	–	–
Spinach	3.2	–	93	3.1	–	–
Tomato	0.9	–	13	0.4	–	–
Lettuce	1.0	1000	23	0.9	0.08	15

Source: Adopted from Woolfe[1]

plagued by famine and the availability of sweet potatoes saved millions of the population from starvation. The OFSP varieties played a major role in addressing hunger and drought mitigation after biotic and abiotic stresses in different parts of the world, including after the floods in Mozambique in 2000.[40] Its use as a lifesaver is evidence that sweet potato is indeed a food security crop.[2,12,13,26,39]

Sweet potato for combating vitamin A deficiencies

Raw and cooked sweet potato roots are highly nutritious.[1,2,20,39] Moreover, orange-fleshed sweet potato is reported to contain high amounts of β-carotene, a precursor for vitamin A. The β-carotene content increases with the intensity of orange colour of the sweet potato flesh.[2,20,41] Leighton[41] tested for β-carotene in different South African orange-fleshed sweet potato varieties (Table 3) and found that indeed, β-carotene concentration varied with depth of colour. Low et al.[42] reported that studies conducted in sub-Saharan Africa demonstrated that consumption of boiled orange-fleshed sweet potato improved vitamin A status of children. A separate study by Amagloh[43] confirmed that orange-fleshed sweet potato was high in vitamin A and could be used as a complementary food in infant feeding. To improve the consumption of this vitamin A rich type of sweet potato, Laurie and Van Heerden[44] prepared different food types (chips, doughnuts, juice from the roots and a green leafy dish from the leaves) from orange-fleshed sweet potato to determine consumer acceptability. They reported that 92% of the consumers liked the colour of the food products while 88% indicated that they would buy sweet potato products or prepare them at home.

Table 3: Concentration of β-carotene in different varieties of orange-fleshed sweet potato in South Africa

Variety	Flesh colour	μ RAE*/100 g (uncooked)	μ RAE/100 g (boiled root)
Resisto	Dark orange	1371	1165
Khano	Dark orange	1170	995
2001-5-1	Dark orange	983	836
W-119	Orange	872	741
Beauregard	Dark orange	804	683
1999-1-7	Dark orange to pale orange	797	677
Excel	Orange, yellow cortex	433	368
Serolane	Orange with yellow ring	426	362
Impilo	Orange to pale orange	424	360

Source: Adopted from Laurie et al.[13] and Laurie[8].

RAE, retinol activity equivalents

Drought and water stress on plants

Stress is defined as an alteration in physiological condition caused by factors that tend to disrupt the equilibrium.[45,46] Strain, on the other hand, is any physical and chemical change produced by stress.[46] Drought is a meteorological term commonly defined as a period without significant rainfall.[45] The crop physiologists' understanding of drought stress is that it occurs when available water in the soil is reduced and atmospheric conditions cause continuous loss of water by transpiration and evaporation.[23,45] Drought stress is considered to be a moderate loss of water, which leads to stomatal closure and limitation of gas exchange. Desiccation is a much more extensive loss of water which can potentially lead to gross disruption of metabolism and cell structure and eventually to the cessation of enzyme catalysed reactions.[46-48] It is characterised by a reduction in soil water content, diminished leaf water potential and turgor loss, closure of stomata and a decrease in cell enlargement and growth.[48] According to Sairam et al.[49], drought is not only the lack of rainfall; it can also be defined by its impacts such as crop shortages and indirect impact, for example, food price increases. Sairam et al.[49] further explained drought as 'an exogenous supply side-shock that causes marked declines in agricultural output, export earnings, employment and income levels'. It affects all facets of economy because of close sectorial linkages.

The four main categories of drought (meteorological, agricultural, hydrological and socio-economic) have a direct impact on food security in the sense that they challenge the whole notion of food availability, accessibility, utilisation and stability. The fact that South Africa is a drought prone region already provides reason to focus on drought. Additionally, climate change predictions that there will be increased frequency and severity of such droughts[25] gives an even greater sense of urgency to identify crops that are resilient and can produce under such adverse conditions[50]. These crops should not only possess drought tolerant characteristics, but should also be nutritious. It is in this context that sweet potato is being touted as a possible fit. However, before one can successfully promote the crop as drought tolerant and nutritious, one needs to generate empirical information detailing the crop's responses to such conditions.

Effect of drought on plant growth

Plant roots interconnect the physiological and biochemical reactions in stems and general plant growth to changes in plant nutrition, photosynthesis, water relations, respiration, nutrient metabolism and growth promoters.[23,46] Water stress inhibits cell enlargement more than cell division. A prolific root system has the advantage of supporting accelerated plant growth during early crop growth stages and extracts water from shallow soil layers that is otherwise easily lost to evaporation.[23]

The prolific root system of sweet potato makes it a drought tolerant crop, although supplementary irrigation is required at the time of planting for proper sprouting and establishment.[13,41] Water stress reduces stem extension and internode diameter of sweet potato cultivars. The severity of this reduction is reported to differ with different cultivars.[51] Inhibition of cell expansion and cell growth is mainly the result of low turgor pressure under water stress conditions. Osmotic regulation can enable the maintenance of cell turgor for survival or can assist plant growth under severe drought conditions.[52] Water stress reduces leaf growth which in turn reduces the leaf area of plants.[23,46,52]

Diminished biomass caused by water stress was observed in almost all cultivars of sweet potato; however some cultivars showed better stress tolerance than others.[53,54] The fact that water stress inhibits leaf expansion and vine extension[39,51] suggests that general vegetative growth in sweet potato would be compromised because the canopy represents the only source of biomass for subsequent partitioning to the storage root.

Studies by Jaleel et al.[23] indicated that water stress can reduce yield in many crop plants and that different crops respond differently to water stress. As the canopy represents the only source of biomass for subsequent partitioning to the storage root in sweet potato, if biomass is limited, it would consequently impose a source limitation of assimilates to storage roots thus lowering final yield. Little information or literature is available which details the extent to which sweet potato yield would be affected if vegetative growth was inhibited by water stress. Saraswati et al.[51] only reported on morpho-physiological characteristics of sweet potato when exposed to water stress; they did not report on the effects of water stress on yield. According to Lebot[39], attempts to demonstrate relationships between yields and particular morphological or physiological characteristics have so far been unsuccessful. However, varieties with small canopies, short stem length and small leaves were reported to have the capability of giving higher yields than those with long stems and numerous broad leaves.[55] The higher yields in small canopy varieties might be a result of plants maximising storage root development and expansion over shoot extension. Given that different genotypes will respond differently, there is a need to evaluate available genotypes across the range of climates where sweet potato can be grown, to allow for the formulation of better recommendations to assist farmers with variety selection. Plants exposed to environmental stresses such as drought and heat stress respond by showing ingenious adaptations at physiological level, which may be accompanied by

changes in various gene expressions.[56] The biological response and tolerance of plants to drought stress such as physiological, biochemical and transgenic tolerance also need to be understood.

Physiological responses to drought and water stress

Sweet potato exposed to drought stress recorded low photosynthesis by reducing chlorophyll 'a' fluorescence, stomatal conductance, intercellular carbon dioxide (CO_2) concentration and CO_2 assimilation rate primarily through stomatal closure.[57] Photosynthetic pigments are important for harvesting light and producing reducing powers in plants.[48] Chlorophyll 'a' and 'b' are reported to be susceptible to soil drying.[48,54] Foliar photosynthetic rate of higher plants is known to decrease as relative water content and leaf water potential decrease. Severe levels of drought are capable of causing irreversible damage to photosynthetic apparatus. Both stomatal and non-stomatal limitations were generally accepted to be the main determinants of reduced photosynthesis under drought stress conditions.[28,46,48] Bray[56] also reported a reduction of both chlorophylls a and b, and total chlorophyll at high water stress. This reduction was reported to be associated with an increase in electrolyte leakage caused by leaf senescence and reduced water use in plants.

The C3 plants of tropical origin, such as sweet potato, colonise easily in a drought environment because they have high phenotypic plasticity. Phenotypic plasticity refers to the ability of the plant to alter its characteristics in order to acclimate to variable environmental conditions. In sweet potato and other C3 plants, all photosynthetic cells are functionally equivalent, thus allowing each cell to acclimate to new environments in a more autonomous manner than those of C4 plants. This functional equivalence allows for photosynthetic plasticity at cellular rather than tissue level, resulting in a greater acclimation ability of C3 plants in general. This C3 metabolism allows sweet potato to be classified as a drought-tolerant plant.[54,58] Earlier studies on C3 plants also indicated that maximum rate of assimilation was more strongly temperature dependent than water dependent and that considerable photorespiration takes place in sunlit photosynthetic organs. Where light intensity is low, there would be 1% decrease in light use efficiency for every degree of temperature increase.[59] This report emphasises the importance of light and temperature in sweet potato production. Increases in photosynthetic plasticity and assimilation results in higher yields and this plasticity may also explain why sweet potato possesses wide adaptability when compared to common C4 staple crops such as maize.

Other anatomical adaptation characteristics of drought tolerance in plants include the development of spongy tissues that act as water reservoirs. Impaired growth and reduced folia area limit water loss through transpiration.[52,54] Some plants may have floral induction which is associated with long-distance movement of Flowering Locus T (FT) proteins, a trait observed in sweet potato plants grown under water stress.[3,13]

Plant adaptive responses to drought and water stress are based on complex changes mainly focused on maintaining water potential in key tissues. Earlier reports indicated that osmotic adjustment[50,55] is a common biochemical adaptation strategy in plants as it results in newly synthesised metabolites. These metabolites include hydrophilic (highly soluble) molecules that are able to produce a resuscitating surface that captures water molecules to be available later during water limitations.[52] Examples of these osmolytes are proteins (amino acids), sugars and sugar alcohols.[54,57] Some of the osmolytes have additional functions such as coping with oxidative stress by suppressing reactive oxygen species.[48,57]

Protein expression, accumulation and synthesis are modified in plants exposed to drought stress. According to Bray[56], these drought-induced proteins are important for physiological adaptation of plants to water stress. The drought-induced proteins are referred to as dehydrins and belong to the group II late embryogenesis abundant (LEA) proteins.[56,60] Studies performed on plants (wild watermelon and maize) under mild water stress reported more than 16 proteins involved in water stress responses.[60,61] Proline is one of the proteinogenic amino acids with exceptional conformational rigidity and is essential for primary metabolism in plants.[62] It is involved in the synthesis of proteins that are necessary for a stress response. This amino acid is of particular interest in stressed plants and is usually considered an osmoprotectant and accumulates in cells to obtain suitable conditions for water uptake under limited water conditions in the roots. It is also involved in reducing oxidative damage by scavenging and/or reducing free radicals.

Proline was found to be involved in tolerance mechanisms against oxidative stress and was the main strategy by which plants avoided detrimental effects of water stress. A study conducted on the influence of water stress on proline accumulation concluded that increases in proline levels helped protect cell membranes from oxidation instead of osmotic regulation as an initial response to water stress.[54,62] Proline accumulation can influence stress tolerance in multiple ways because it functions as a molecular chaperone, able to protect protein integrity and enhance the activities of different enzymes.[61,62] While there are two schools of thoughts in as far as proline accumulation is concerned, the bottom line is that it will accumulate in response to environmental stresses. Whether or not its accumulation is a sign of stress tolerance or rather a symptom of the severity of the stress is still debatable and is an avenue that needs to be explored, especially in food security crops like sweet potato. Modi and Mabhaudhi[50] suggested that, despite the controversy surrounding its exact role, proline accumulation could still be a useful index in drought tolerance. However, the authors went on to state that its value as an index could only be fully understood if other stress metabolites were also quantified alongside it.

Phytochemicals are secondary metabolites that are widespread in the plant kingdom. Phytochemicals protect plants against bacteria, viruses and fungi invasion.[63] The different compounds constituting phytochemicals are considered to have biological significance but are not established as essential nutrients in human or animal diets. They have, however, been reported to possess bioactive properties with potential health benefits.[25] Phytochemicals such as antioxidants, carotenoids, anthocyanins and flavonoids can be classified into different categories depending on their structure and have been associated with flavour and colour characteristics of fruits and vegetables.[36] Antioxidants such as ascorbic acid and glutathione, commonly found in chloroplasts and other cellular compartments in plants are crucial for plant defence against oxidative stress and have been reported to be indicators of plants' hypersensitivity to stress conditions.[25,36,63] Earlier findings by Sairam et al.[49] indicated that enzymatic and non-enzymatic antioxidants in plant cells played an important role in avoiding detrimental effects of free radicals. In addition to their high antioxidant and antiradical activities in humans, the health beneficial properties have also been attributed to many other mechanisms such as anti-inflammatory properties, inhibition of enzymes and induction of detoxification enzymes. Antioxidants were reported to be vital in protection against cancer and cardiovascular diseases in humans.[28,29,36]

Carotenoids are plant pigments that can partially help plants withstand the adversities of drought.[23] They form a key part of the plant's antioxidant defence system.[48] Carotenoids are divided into hydrocarbon carotenes (lycopene and β-carotene) and xanthophylls (lutein).[36,48] β-carotene is present in all chloroplasts of green plants and is exclusively bound to the complexes of photosystems one (PSI) and two (PSII). β-carotene functions as an accessory pigment, an effective antioxidant and plays a unique role in protecting and sustaining photochemical processes.[23] The protective role of β-carotene in photosynthetic tissue is achieved through direct quenching of triplet chlorophyll, thus preventing the generation of singlet oxygen and protecting against oxidative damage.[52]

As a phytochemical, carotenoids were reported to protect against the eye disease, macular degeneration.[36] β-carotene is of particular importance because it is a precursor for vitamin A. Vitamin A is important in human nutrition because of the role it plays in improving vision, bone growth, reproduction, maintenance of epithelia and overall growth.[64] The orange-fleshed sweet potato varieties contain high amounts of β-carotene and are being promoted in vitamin A deficiency alleviation programmes in South Africa and other developing countries. The fact that sweet potato can synthesise β-carotene to protect itself from oxidation stress and further store the β-carotene in both leaves and storage roots, is an indication that it is an important food security crop which can withstand drought stress conditions and still alleviate chronic malnutrition diseases associated with vitamin A deficiencies.

Genes expressed during drought stress

Many responses to water stress are controlled by a number of genes with different functions. Many regulatory processes are initiated the minute water is lost from plant cells, to help adjust cellular metabolism, and result in changes in gene expression.[56] Changes in expression patterns have been monitored on the plant when drought stress is perceived. These changes range from genes whose products are involved in early responses such as signal transduction, transcription and translation factors; to late response genes, such as water transport, osmotic balance, oxidative stress and damage repair.[54] Whether these water stress induced genes perform an adaptive role or not is still to be investigated.

Genetic engineers have tried to develop transgenic sweet potato plants by gene coding for spermidine synthase which is used to improve environmental stress. The transgenic sweet potato from this genetic coding showed high tolerance to drought, salt, chilling and heat stresses.[65] Recent reports indicate that transgenic sweet potato, containing the gene which is important for osmoregulation and abiotic stress, showed tolerance to multiple environmental stresses such as protection against cell damage, improved photosynthetic activity and increased activity of free radical scavenging enzymes.[60] These reports are proof of good progress in genetic engineering in as far as sweet potato drought tolerance is concerned, but there is still a shortage of successful screening methods and multidisciplinary approaches and genotype by environment interaction.[14] It is therefore believed that with the existence of these strength-enhanced drought tolerant varieties, food security will improve as long as the genotype by environment interaction selection is carried out in the environment in which the varieties will be released and grown.

Crop management for drought mitigation

Increased productivity in rain-fed agriculture reduces pressure on the limited arable land and water resources. Knowledge of weather and climate can reduce the impact of drought on communities, and this can be achieved by employing two distinct strategies. The first is long term planning in which strategies can be devised, and precautions taken to reduce impact. This planning includes climate and weather evaluation using scientific systems research tools such as simulation models, optimisation techniques, geographic information systems and databases.[66] The FAO AquaCrop model has already been calibrated and validated on OFSP varieties and proved to be useful in exploring irrigation options in sweet potato.[67] Adjusting planting dates (or timing of planting) and harvesting, substituting cultivars and, where necessary, modifying or changing the cropping system altogether, has also been identified as a long term strategy.[68] Adjusting planting dates for sweet potato production in KZN so as to allow yield maximisation could be a suitable management strategy.

The second phase where application of weather and climate knowledge can reduce impacts of drought is by taking action during the onset of the event (planting/growing season) to reduce adverse effects.[68,69] This action may include, but not be limited to, adopting proper crop management strategies such as water conservation strategies, manipulation of plant population, water recycling, rainfall harvesting into ponds or village tanks and even mid-season corrections.[23,69] Farooq et al.[52] further recommended adoption of strategies such as mass screening and breeding, marker-assisted selection and exogenous application of hormones and osmoprotectants to seed or growing plants, as well as engineering for drought tolerance. The two phases suggested by Das[66] of using weather and climate to mitigate and plan for drought; adaptation of agriculture to stability and resilience of production and conservation tillage[68,69] and other best practices, all form part of climate-smart agriculture. According to the FAO[69], the aim of climate-smart agriculture is to increase productivity in an environmentally and socially sustainable way, strengthen farmer's resilience to climate change, and reduce agriculture's contribution to climate change by reducing greenhouse gas emissions and increasing carbon storage of farmland.

Conclusions

Sweet potato remains an important root crop with an inherent ability to produce more edible energy than most major food crops. It is suitable for production on marginal lands and can play an important role as a food security crop. The importance of OFSP varieties in supplementing vitamin A deficiencies cannot be overemphasised. And production of the OFSP varieties in areas with a prevalence of vitamin A deficiency should be promoted. It is possible to produce this crop and have it available all year round in the tropical and subtropical areas of South Africa where warm conditions prevail, even if drought is a challenge for staple crops such as cereals. Rain-fed crops are always subjected to water stress during the course of growth because of the unpredictability of rainfall in subtropical environments. Sweet potato has the advantage of being drought tolerant after establishment. Hence, its yield potential is generally greater than that of popular staple crops of South Africa, such as maize. In addition to yield advantage, the higher nutritional value of sweet potato is of benefit to farmers producing under conditions that are susceptible to drought stress. While this review has identified sweet potato as a potential food security crop for dryland production, there are still challenges that may limit its potential to fully contribute towards food security. There is a need to introduce more participatory research and extension to educate smallholder farmers about the value of sweet potato as a major crop for subsistence and rural economic development.

Acknowledgements

We acknowledge the Organization for Women in Science from the Developing World (OWSD) and The World Academy of Science (TWAS) for funding.

Authors' contributions

N.M.M. performed and wrote the literature review. A.T.M. was the project leader and supervisor and provided conceptual direction for the review. T.M. provided guidance and mentorship.

References

1. Woolfe J. Sweet potato an untapped food resource. Cambridge, UK: Cambridge University Press; 1992.
2. Bovell-Benjamin AC. Sweet potato: A review of its past, present and future roles in human nutrition. Advanc Food Nutr Res. 2007;52:1–59. http://dx.doi.org/10.1016/S1043-4526(06)52001-7
3. Agili S, Nyende B, Ngamau K, Masinde P. Selection, yield, drought tolerance indices of orange-fleshed sweet potato (*Ipomoea batatas* Lam) hybrid clone. J Nutr Food Sci. 2012;2:1–8.
4. Food and Agriculture Organization (FAO). The state of food insecurity in the world: The multiple dimensions of food security [homepage on the Internet]. c2013 [cited 2014 Mar 06]. Available from: www.fao.org/publications
5. Laurie SM, Van Den Berg AA, Stork PO. Progress by the South African sweet potato breeding and multiplication program in support of resource-poor farmers. In: Food security and crop diversification in SADC countries: The role of cassava and sweet potato. Proceedings of the Scientific Workshop of the Southern Root Crop Research Network (SARRNET); 1998 Aug 17–19; Lusaka, Zambia. Lilongwe: SARRNET; 1999.
6. Van Jaarsveld PJ, De Wet M, Hermse E, Nestel P, Rodriguez-Amaya DB. Retention of β-carotene in boiled, mashed orange-fleshed sweet potato. J Food Comp Analy. 2006;19:321–329. http://dx.doi.org/10.1016/j.jfca.2004.10.007
7. Wenhold F, Faber M. Nutritional status of South Africans: Links to agriculture and water. Report no. TT362/P/08. Pretoria: Water Research Commission; 2008.
8. Laurie SM. Agronomic performance, consumer acceptability and nutrient content of new sweet potato varieties in South Africa [PhD thesis]. Bloemfontein: University of the Free State; 2010. Available from: http://etd.uovs.ac.za/ETD-db/theses/
9. Labadarios D, Steyn NP, Mgijima C, Dladla N. Review of the South African nutrition policy: 1994–2004 and targets for 2007: Achievements and challenges. Nutrition. 2005;21:100–108. http://dx.doi.org/10.1016/j.nut.2004.09.014

10. Department of Agriculture, Forest and Fisheries (DAFF). Sweet potato (*Ipomoea batatas* L.) production guide. Pretoria: DAFF; 2011.

11. Laurie SM, Magoro MD. Evaluation and release of new sweet potato varieties through farmer participatory selection. Afr J Agric Res. 2008;3:672–676.

12. International Potato Centre (CIP). Sweet potato and malnutrition in Africa [document on the Internet]. c2012 [cited 2014 Jan 12]. Available from: www.cipotato.org/esearch.cip.cgiar.org

13. Mukhopadhyay SK, Chattopadhyay A, Chakraborty I, Bhattacharya I. Crops that feed the world 5: Sweet potato. Sweet potatoes for income and food security. Food Secur. 2011;3:283–305. http://dx.doi.org/10.1007/s12571-011-0134-3

14. Placide R, Shimelis H, Laing M, Gahakwa D. Physiological mechanisms and conventional breeding of sweet potato (*Ipomoea batatas* (L) Lam) to drought-tolerance. Afr J Agric Res.2013;8:1837–1846.

15. Earl HJ, Davis RF. Effect of drought stress on leaf and whole canopy radiation use efficiency and yield of maize. Agron J. 2003;95:688–696. http://dx.doi.org/10.2134/agronj2003.0688

16. Battisti DS, Naylor RL. Historical warnings of future food insecurity with unprecedented seasonal heat. Science. 2009;323:240–244. http://dx.doi.org/10.1126/science.1164363

17. Allemann J, Laurie SM, Thiart S, Vorster HJ. Sustainable production of root and tuber crops (potato, sweet potato, indigenous potato, cassava) in southern Africa. S Afr J Bot. 2004;70:60–66. http://dx.doi.org/10.1016/S0254-6299(15)30307-0

18. O'Brian P. The sweet potato: Its origin and dispersal. Am Antropol. 1972;74(3):342–365.

19. Schippers RR. African indigenous vegetables: An overview of cultivated species [revised version on CD-ROM]. Aylesford: Natural Resources International Limited; 2002. p. 214.

20. Laurie SM, Van Den Berg AA, Magoro MD, Kgonyane MC. Breeding of sweet potato and evaluation of imported cultivars in South Africa. Afr Crop Sci J. 2004;12:189–196. http://dx.doi.org/10.4314/acsj.v12i3.27877

21. Joubert R. Huge projects promised for KwaZulu-Natal. In: Van der Walt A, editor. Farmers Weekly Magazine. 2012 March 26; Business section [cited 2013 Aug 07]. Available from: http://www.farmersweekly.co.za/ariticle.aspx?id=16512&h=Huge-projects-promised-for-KZN

22. Nqobo H, Dladla B. Provincial report on education and training for agriculture and rural development in KwaZulu-Natal, South Africa. Pietermaritzburg: KZN Department of Agriculture; 2002. Available from: www.nda.agric.za/doadev/sidemenu/educationandtraining/kzn.pdf

23. Jaleel CA, Manivannan P, Wahid A, Farooq M, Al-Juburi HJ, Samasundaram et al. Drought stress in plants: A review on morphological characteristics and pigment composition. Int J Agric Biol. 2009;11:100–105.

24. Schulze R. Approaches towards practical adaptive management options for selected water-related sectors in South Africa in a context of climate change. WaterSA. 2011;37:621–645. http://dx.doi.org/10.4314/wsa.v37i5.1

25. Islam S. Sweet potato (*Ipomoea batatas* L.) leaf: Its potential effect on human health and nutrition. J Food Sci. 2006;71:13–21. http://dx.doi.org/10.1111/j.1365-2621.2006.tb08912.x

26. Ebregt E, Struik PC, Odongo B, Abidin PE. Piecemeal versus one time harvesting of sweet potato in north-eastern Uganda with special reference to pest damage. J Life Sci. 2007;55(1):75–92. http://dx.doi.org/10.1016/S1573-5214(07)80005-4

27. Karna P, Gundala SR, Gupta MV, Shamsi SA, Pace RD, Yates C, et al. Polyphenol-rich sweet potato greens extract inhibits proliferation and induces apoptosis in prostate cancer cells in vitro and in vivo. Carcinogenesis. 2011;32:1872–1880. http://dx.doi.org/10.1093/carcin/bgr215

28. Ross JA, Kasum CM. Dietary flavonoids: Bioavailability, metabolic effects, and safety. Annu Rev Nutr. 2002;22:19–34. http://dx.doi.org/10.1146/annurev.nutr.22.111401.144957

29. Liu Q. Sweet potato omics and biotechnology in China. Plant Omics J. 2011;4(6):295–301.

30. Anonymous. Non-culinary uses of sweet potato [homepage on the Internet]. c2011 [cited 2012 July 10]. Available from: www.sweetsp.com/discovering-some-non-culinary-uses-of.html

31. Walter WM, Catignani GL, Yow LL, Porter DH. Protein nutritional value of sweet potato flour. J Agric Food Chem.1983;31:947–949. http://dx.doi.org/10.1021/jf00119a006

32. Maduna M. School gardens within the national school nutrition programme. In: Faber M, Laurie S, Van Jaarsveld P, editors. Proceedings of the Orange-fleshed Sweet Potato Symposium; 2007 Oct 03; Pretoria, South Africa. Pretoria: ARC; 2008. p.1–3.

33. Sikhakhane B. Garden projects within the integrated nutrition programme. In: Faber, M, Laurie S, Van Jaarsveld P, editors. Proceedings of the Orange-fleshed Sweet Potato Symposium; 2007 Oct 03; Pretoria, South Africa. Pretoria: ARC; 2008. p. 4–6.

34. Khachatryan A, Basonde RR, Labonte DR, Losso JN, Bansode RR, Labonte DR, et al. Identification of sweet potato leaves (*Ipomoea batatas*) as an excellent source of lutein. Session 88 of Nutraceutical and Functional Foods. Chicago, IL: McCormick Place; 2003.

35. Mozaffarieh M, Sacu S, Wedrich A. The role of the carotenoids, lutein and zeaxanthin, in protecting against age-related macular degeneration: A review based on controversial evidence. Nutr J. 2003;2:1–8. http://dx.doi.org/10.1186/1475-2891-2-20

36. World Health Organization (WHO). Food security [homepage on the Internet]. c2013 [cited 2013 Aug 14]. Available from: www.who.int/trade/glossary/story028/en/

37. Altman M, Hart T, Jacobs P. Food security in South Africa. Pretoria: Human Sciences Research Council; 2009.

38. Mbabu A, David S. Can OFSP enhance food security in sub-Saharan Africa? CIP Newsletter. 2012 March [cited 2013 Nov 12]. Available from: www.sweetpotatoknowledge.org/sweetpotato-introduction/ofsp-advocacy-toolkit/flyer-5-can-ofsp-ehnance-food-security-in-africa.pdf

39. Lebot V. Tropical root and tuber crops: Cassava, sweet potato, yams and aroids. Cambridge, MA: CABI North American Office; 2009.

40. Kapinga R, Andrade M, Lemaga B, Gani A, Crissman C, Mwanga R. Role of orange fleshed sweet potato in disaster mitigation: Experiences from east and southern Africa. Afr Crop Sci Conf Proc. 2005;7:1321–1329.

41. Leighton CS. Nutrient and sensory quality of orange-fleshed sweet potato [MSc dissertation]. Pretoria: University of Pretoria; 2007.

42. Low JW, Arimond M, Osman N, Cunguara B, Zano F, Tschirley DA. Food-based approach: Introducing orange-fleshed sweet potato increases vitamin A intake and serum retinal concentration in young children in rural Mozambique. J Nutr Comm Int Nutr. 2007;137:1320–1327.

43. Amagloh FK, Hardacre A, Mutukumira AN, Weber JL, Brough L, Coad J. A household-level sweet potato-based infant food to complement vitamin A supplementation initiatives. Mater Child Nutr. 2011;8:1–10. http://dx.doi.org/10.1111/j.1740-8709.2011.00343.x

44. Laurie SM, Van Heerden SM. Consumer acceptability of four products made from beta-carotene-rich sweet potato. Afr J Food Sci. 2012;6:96–103. http://dx.doi.org/10.5897/AJFS12.014

45. Whilhite DA. Drought as a natural hazard: Concepts and definitions. In: Whilhite DA, editor. Drought volume 1. A global assessment. London: Routledge; 2000. p. 245–255.

46. Blum A. Effective use of water (EUW) and not water-use efficiency (WUE) is the target of crop yield improvement under drought stress. Field Crop Res. 2009;112:119–123. http://dx.doi.org/10.1016/j.fcr.2009.03.009

47. Smirnoff N, Conklin PL, Loewus FA. Biosynthesis of ascorbic acid: A renaissance. Annu Rev Plant Physiol Plant Mol Biol. 2001;52:437–467. http://dx.doi.org/10.1146/annurev.arplant.52.1.437

48. Jaleel CA, Manivannan P, Lakshmanan GMA, Gomathinayagam M, Panneerselvam R. Alterations in morphological parameters and photosynthetic pigment responses of *Catharathus roseus* under soil water deficits. Colloids Surf B Biointerfaces. 2008;61:298–303. http://dx.doi.org/10.1016/j.colsurfb.2007.09.008

49. Sairam RK, Saxana DC. Oxidative stress and antioxidants in wheat genotypes: Possible mechanism of water stress tolerance. J Agron Crop Sci. 2000;184:55–61. http://dx.doi.org/10.1046/j.1439-037x.2000.00358.x

50. Modi AT, Mabhaudhi T. Drought tolerance and water use of selected South African traditional and indigenous crops. WRC report no. 1771/1/13. Pretoria: Water Research Commission; 2013.

51. Saraswati P, Johnson M, Coventry R, Holtum J. Identification of drought tolerant sweet potato (*Ipomoea batatas* (L) Lam) cultivars. In: Proceedings of the 4th International Crop Science Congress; 2004 Sep 26–Oct 01; Brisbane, Australia. Available from: www.cropscience.org.au/icsc2004/poster/1/1/1400_saraswatip.htm

52. Farooq M, Wahid A, Kobayashi N, Fujita D, Basra SMA. Plant drought stress: Effects, mechanisms and management. Agron Sustain Dev. 2009;29:185–212. http://dx.doi.org/10.1051/agro:2008021

53. Shao HB, Chu LY, Shao MA, Jaleel C, Hong-Mei M. Higher plant antioxidants and redox signalling under environmental stresses. Comp Rend Biol. 2008;331:433–441. http://dx.doi.org/10.1016/j.crvi.2008.03.011

54. Xoconostle-Cazares B, Ramirez-Ortega FA, Flores-Elenes L, Ruiz-Medrano R. Drought tolerance in crops. Am J Plant Physiol. 2010;5(5):1–16. http://dx.doi.org/10.3923/ajpp.2010.241.256

55. Wilson JE, Pole FS, Smith NEJM, Taufatofua P. Sweet potato breeding. Apia, Western Samoa: IRETA-USP Agro-Facts Publications; 1989.

56. Bray E. Classification of genes differentially expressed during water-deficit stress in *Arabidopsis thaliama*: An analysis using microarray and differential expression data. Ann Bot. 2002;89:803–811. http://dx.doi.org/10.1093/aob/mcf104

57. Van Heerden PDR, Laurie R. Effects of prolonged restriction in water supply on photosynthesis, shoot development and storage root yield in sweet potato. Phys Plantar. 2008;134:99–109. http://dx.doi.org/10.1111/j.1399-3054.2008.01111.x

58. Sage RF, McKown AD. Is C4 photosynthesis less phenotypically plastic than C3 photosynthesis? J Exp Bot. 2006;57:303–317. http://dx.doi.org/10.1093/jxb/erj040

59. Haxeltine A, Prentice IC. A general model for light-use efficiency of primary production. Funct Ecol. 1996;10:551–561. http://dx.doi.org/10.2307/2390165

60. Fan W, Zhang M, Zhang H, Zhang P. Improved tolerance to various abiotic stress in transgenic sweet potato (*Ipomoea batatas*) expressing spinach betaine aldehyde dehydrogenase. PLoS One. 2012;7:37344–37358. http://dx.doi.org/10.1371/journal.pone.0037344

61. Close TJ, Fenton RD, Moonan F. A review of plant dehydrins using antibodies specific to the carboxyl terminal peptide. Plant Mol Biol.1993;23:279–286. http://dx.doi.org/10.1007/BF00029004

62. Szabados L, Savoure A. Proline: A multifunctional amino acid. Trends Plant Sci. 2009;15:89–97. http://dx.doi.org/10.1016/j.tplants.2009.11.009

63. Karter N, Liu RH. Health benefits of wholegrain phytochemicals. Crit Rev Food Sci Nutr. 2010;50:193–208. http://dx.doi.org/10.1080/10408390802248734

64. Tee E. The medical importance of vitamin A and carotenoids (with particular reference to developing countries) and their determination. Mal J Nutr.1995;1:179–230.

65. Kasukabe Y, He L, Watakabe Y, Otani M, Shimada T, Tachibana S. Improvement of environmental stress tolerance of sweet potato by introduction of genes for supermidine synthase. Plant Biotechnol. 2006;23:75–83. http://dx.doi.org/10.5511/plantbiotechnology.23.75

66. Das HP. Agricultural drought mitigation and management of sustained agricultural development in India. In: Sivankumar MVK, Motha RP, Das HP, editors. Natural disaster and extreme events in agriculture. Berlin: Springer; 2005. p. 277–303. http://dx.doi.org/10.1007/3-540-28307-2_16

67. Beletse YG, Laurie R, Du Plooy CP, Laurie SM, Van den Berg A. Simulating the yield response of orange fleshed sweet potato 'Isondlo' to water stress using the FAO AquaCrop model. Acta Hort. 2012;1007:12. http://dx.doi.org/10.17660/actahortic.2013.1007.112

68. Rosenzweig C, Tubiello FN. Adaptation and mitigation strategies in agriculture: An analysis of potential synergies. Mitig Adapt Start Glob Change. 2007;12:855–873. http://dx.doi.org/10.1007/s11027-007-9103-8

69. Food and Agriculture Organization (FAO). Climate-smart agriculture sourcebook: Executive summary [document on the Internet]. c2014 [cited 2014 May 02]. Available from: www.fao.org/climatechange/climatesmart

In-vitro effects of protease inhibitors on *BAX, BCL- 2* and apoptosis in two human breast cell lines

AUTHORS:
Gbenga A. Adefolaju[1,2]
Katherine E. Theron[3]
Margot J. Hosie[1,3]

AFFILIATIONS:
[1]School of Anatomical Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]Department of Medical Sciences, Public Health and Health Promotion, University of Limpopo, Polokwane, South Africa
[3]Newcastle University Medicine, Johor, Malaysia

CORRESPONDENCE TO:
Gbenga Adefolaju

EMAIL:
Gbenga.Adefolaju@ul.ac.za

POSTAL ADDRESS:
School of Anatomical Sciences, University of the Witwatersrand Medical School, 7 York Road, Parktown 2193, South Africa

DATES:
Received: 26 Nov. 2014
Revised: 19 Jan. 2015
Accepted: 09 Feb. 2015

KEYWORDS:
lopinavir; ritonavir; apoptosis; breast cancer

HOW TO CITE:
Adefolaju GA, Theron KE, Hosie MJ. In-vitro effects of protease inhibitors on *BAX, BCL-2* and apoptosis in two human breast cell lines. S Afr J Sci. 2015;111(11/12), Art. #2014-0417, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140417

Currently, the treatment of choice of HIV/AIDS in South Africa is the multidrug combination regimen known as HAART (highly active antiretroviral treatment). HAART, which commonly consists of nucleoside or non-nucleoside reverse transcriptase inhibitors and protease inhibitors, has radically decreased mortality and morbidity rates among people living with HIV/AIDS. The emphasis of the original development of the antiretroviral drugs was on clinical effectiveness (reducing mortality). Presently, emphasis has shifted from the initial short- term considerations to the long-term undesirable or harmful effects induced by this treatment regimen. Whether antiretroviral compounds are oncogenic is widely speculated, which led to this investigation into the effects of protease inhibitors on the expression of key apoptotic regulatory genes, *BAX* and *BCL-2,* in two human breast cell lines, MCF-7 and MCF-10A by real-time qPCR gene expression and immunofluorescence. The anti-apoptotic effects of the protease inhibitors – LPV/r were also investigated by cell death detection ELISA and acridine orange staining. This study also evaluated the cytotoxicity of the antiretroviral drugs in normal and cancer cell lines of the breast (at clinically relevant concentrations of the drugs and at different time points, 24–96 h), employing the neutral red uptake assay. The drugs and combinations tested did not alter *BAX* and *BCL-2* gene expression and protein expression and localisation in both cell lines. In addition, the protease inhibitors–LPV/r did not inhibit camptothecin-induced apoptosis in both cell lines. We have shown that the protease inhibitors demonstrated varying degrees of cytotoxicity in the breast cells. The resulting DNA damage associated with cytotoxicity is strongly implicated in the processes of tumour initiation.

Introduction

The HIV/AIDS pandemic has had a catastrophic effect on the world population. In 2011, approximately 34 million people were recorded as living with HIV globally. Sub-Saharan Africa, which is home to just 10% of world's population contains a huge 69% of people living with HIV or AIDS.[1] South Africa has the highest number of people living with HIV in sub-Saharan Africa and in the world.[1] It is for this reason that in November 2003, South Africa started the world's largest public sector rollout of highly active antiretroviral treatment (HAART) to date.[2]

HAART has radically decreased mortality and morbidity among people living with HIV and AIDS[3,4,5] The emphasis of the original development of the antiretroviral (ARV) drugs was on clinical effectiveness (reducing mortality) with all other considerations secondary.[6] Presently, emphasis has shifted from the initial short-term considerations to the long-term undesirable or harmful effects induced by this treatment regimen.[6] Whether antiretroviral compounds are oncogenic remains to be fully elucidated.[7] Torres et al.[8] suggest that the clinical use of nucleoside/nucleotide reverse transcriptase inhibitor (NRTI) drug pairs may lead to additive or synergistic effects compounding long-term risk for cancer gene mutations and potential carcinogenesis.

Breast cancer is the most common cancer seen in women worldwide.[9] HIV-infected patients on antiretroviral treatment are reported to survive longer due to the treatment and therefore are at danger of developing breast cancer because its risk increases with age.[9,10,11] Other studies suggest that the use of ARV drugs might increase breast cancer risk by fat redistribution from peripheral and gluteal tissues to the breast as a result of increased production of oestrogen.[9] The total influence of HAART on the risk for non-AIDS defining malignancies (NADM) is far from concluded. In a meta-analysis,[12] the standardised incidence ratio of some non-AIDS defining malignancies, including breast cancer, was substantially increased prior to and in the HAART era. In a Swiss HIV cohort study, in which a mega prospective cohort of people were tracked for an average of about 5 years after HAART, no definite influence of HAART on standardised incidence ratios of NADMs was revealed.[13] However, another HIV/AIDS cohort conducted in England[7], showed that a substantial total rise in risk for NADMs objectively accompanied HAART treatment. Non-AIDS defining cancers now generally account for more deaths in HIV-infected individuals than AIDS defining cancers. Although traditional risk factors may account for some of this discrepancy, they certainly do not explain the explosion of these cancers in HIV-affected people on HAART.[14] Engels et al.[15] found that non-AIDS defining cancers comprise 58% of all cancer deaths post-HAART (1996–2002) in comparison to 31.4% in the pre- HAART era (1991–1995). These data strongly suggests that antiretroviral drugs may influence cancer development and progression.

In many cancers, there is a consistent pattern of apoptosis inhibition and deregulation of cell and tissue homeostasis. Numerous studies implicate apoptosis-related genes and their products in the development of cancer.[16-18] The induction of apoptosis occurs via two major pathways: the extrinsic (death receptor) pathway and the intrinsic (mitochondria) pathway.[16] The extrinsic pathway is activated by the attachment of the Fas plasma-membrane death receptor (and related receptors like tumour necrosis receptor 1 and its family) to its extracellular ligand, Fas-L. The intrinsic pathway also leads to apoptosis but under the control of mitochondrial pro-enzymes. The majority of cellular death in vertebrates follows the mitochondrial pathway of apoptosis.[17] This pathway to cell death is controlled by the B-cell CLL/lymphoma 2 *(BCL-2)* family of proteins. They operate to regulate the integrity of the mitochondrial outer membrane. When apoptosis arises as a result of cooperation among these proteins, the two pro-death BCL-2 effector proteins; BCL-2-associated X protein (BAX) and BCL-2 antagonistic killer (BAK), disrupt this membrane in a process known as mitochondrial outer membrane permeabilisation.[17] If mitochondrial outer membrane permeabilisation takes place, proteins located in the mitochondrial

inter-membrane compartment enter the cytosol and activate caspases and cysteine proteases that coordinate the disassembling of the cell.[17] The process of mitochondrial outer membrane permeabilisation is opposed by the pro-survival *BCL-2* proteins, such as BCL-2, BCL-W and BCL-xL, which impedes the capacity of BAX and Bak to permeabilise. The relevance of this impact in cancer is highlighted by the remark that oncogenes, like *Myc*, which advances proliferation, also encourage cell death that is inhibited by the anti-apoptotic *BCL-2* proteins.[17] The widely studied genes associated with apoptosis include *BCL-2*; the anti-apoptotic gene and *BAX*; a pro-apoptotic gene.[18] Hypothetically *BCL-2* over-expression renders a survival advantage for cancer cells; in contrast, *BAX* stimulation restores sensitivity to apoptosis induced by drugs or radiation.[18] In this study, we hypothesised that HIV protease inhibitors might inhibit the pro-apoptotic *BCL-2* family member *BAX*, or activate the anti-apoptotic *BCL-2* family member thereby promoting cell survival and lead to cancer development and/or progression.

In the majority of published studies, the in-vitro effects of antiretroviral drugs on immune cells were investigated. Few studies on non-immune cells were found. In the current study, we therefore investigated the effects of some of the drugs in the South African HAART treatment guidelines, individually and in combination, on *BAX* and *BCL-2* gene expression in two breast cell lines at clinically relevant concentrations. BAX and BCL-2 proteins were localised with immunofluorescence. We also determined whether the protease inhibitors – LPV/r, reported to have anti-apoptotic effects,[19-20] can inhibit drug-induced apoptosis in these non-immune cell lines. We also evaluated the cytotoxicity of the antiretroviral drugs in normal and cancer cell lines of the breast at clinically relevant concentrations of the drugs and at different time points (24 h, 48 h, 72 h and 96 h) by employing the neutral red uptake assay.

Materials and methods

Cells

Human breast carcinoma MCF-7 cells (American Type Culture Collection, Rockville, MD, USA) were cultured in DMEM medium (Gibco BRL, Gaithersburg, MD, USA) containing 10% foetal bovine serum (Gibco).

Immortalised human breast epithelial cells, MCF-10A (American Type Culture Collection) were cultured in mammary epithelium growth medium (MEGM; Lonza, Walkersville, MD, USA) containing human recombinant epidermal growth factor, hydrocortisone, insulin and bovine pituitary extract (Lonza).

Cells were cultivated as a stationary monolayer in plastic tissue-culture dishes (Nunclon, Denmark) and were incubated at 37 °C in a 5% CO_2 humidified environment. The cells were incubated with the drugs individually and in combination at the indicated physiologically relevant concentrations for 24–96 h. The normal breast MCF-10A cells were cultured in serum-free media. MCF-7 cells were synchronised by serum starvation for 24 h[21,22] prior to all experiments. After being cultured in complete medium for overnight attachment, MCF-7 cells were incubated with serum-free media for 24 h. The cells were then treated with the antiretroviral drugs (as described in Table 1) in medium supplemented with serum for the indicated periods.

This study was approved by the Human Research Ethics Committee of the University of the Witwatersrand (Clearance Certificate reference: M10905).

Table 1: Treatment groups and drug concentrations administered for each cell line

GROUP	TREATMENT
1	Growth medium
2	0.01% Methanol
3	9.8 μg/mL LPV
4	0.6 μg/mL RTV
5	LPV/r (Kaletra®)

LPV, lopinavir; RTV, ritonavir

Drugs and treatment

The drugs were administered at clinically relevant concentrations, which reflect their steady-state plasma peak concentration.[23,24] Lopinavir (LPV) and ritonavir (RTV) were purchased from Toronto Research Chemicals (Toronto, Ontario, Canada) and were dissolved in methanol. The cells were incubated with the drugs individually and in combination at the indicated concentrations (Table 1) for 24 h, 48 h, 72 h and 96 h. Normal and vehicle control groups were exposed to growth medium and vehicle, respectively (Groups 1 and 2).

The cells were exposed to 9.8 μg/mL LPV (Group 3) and LPV/r (Kaletra® Group 5). Even though the absolute bioavailability of ritonavir had not been determined in humans at the time of this study[25], across studies, daily administration of Kaletra® 400/100 mg yields mean steady-state lopinavir plasma concentrations 15–20 times higher than those of ritonavir in individuals living with HIV-1.[26] Therefore a 9.8 μg/mL LPV:0.6 μg/mL RTV ratio was used. In Group 4, cells received 0.6 μg/mL RTV alone.

Kits, antibodies and reagents

The neutral red TOX-4 kit, methanol, acridine orange (AO), ethylenediaminetetraacetic acid (EDTA) and camptothecin (CPT) were purchased from Sigma Chemical Co. (St. Louis, MO, USA). Rabbit polyclonal anti-BAX antibody and mouse monoclonal anti-BCL-2 were purchased from DAKO (DK-2600, Glostrup, Denmark). Secondary antibodies, goat anti-rabbit rhodamine-conjugated and goat anti-mouse FITC-conjugated were purchased from Abcam (Cambridge, UK). The GeneJET RNA Purification Kit, DNase I, RNase-free kit and O'GeneRuler Low Range DNA Ladder were purchased from Thermo Scientific (Pittsburgh, PA, USA). High Capacity cDNA Reverse Transcription Kit and Power SYBR® Green PCR Master Mix were purchased from Life Technologies (California, CA, USA). Oligos for quantitative polymerase chain reaction (qPCR) was purchased from Integrated DNA Technologies, Inc. (Coralville, IA, USA). Agarose D-1 Low EEO-GQT was purchased from Conda Laboratories (Madrid, Spain).

Neutral red uptake assay

The neutral red uptake cytotoxicity assay is a precise, easy and reproducible assay that allows for distinguishing between viable and non-viable cells.[27] Neutral red is a weak cationic supravital dye that is actively transported across the undamaged cell membrane of viable cells and accumulates in the lysosomes. Alterations to the lysosomal membrane integrity arising from the toxicity of substances reduces neutral red uptake thereby allowing the distinction between viable and damaged cells.[27] After the viable cells have incorporated the dye, the dye is subsequently liberated from the cells and the degree of cytotoxicity, quantified by a spectrophotometer, is a measure of how many cells excluded the dye.[27] This assay was used to evaluate the cytotoxic effects of the protease inhibitors used in this study. MCF-7 and MCF-10A cells were cultured in the presence of the drugs, or diluent at the indicated concentrations for periods of time ranging from 24 h to 96 h. For each cell line, a 96-well Falcon tissue culture plate was seeded with cells in replicates for each group including negative and vehicle controls (Table 1). After 24 h, 48 h, 72 h and 96 h of treatment with antiretroviral drugs, the medium was removed and the manufacturer's protocol[27] was followed for the assay. Cells were rinsed in phosphate buffered saline (PBS). Neutral red medium was added to each well and plates were incubated for 2 h at 37 °C for the neutral red dye to be taken up by viable cells. Thereafter, the neutral red solution was removed and cells were rinsed with PBS. The solubilising solution (1% acetic acid in 50% ethanol) was added to each well for 10 min in order to extract the dye. The absorbance of the extracted dye was measured in a microplate reader (Anthos 2010 Model 17-550, Salzburg, Austria) at a wavelength of 540 nm. Background absorbance was read at 690 nm and subtracted from the 540 nm measurement. All experiments were repeated three times on three different days.

Total RNA extraction, cDNA synthesis, and real-time qPCR analysis

Cells were exposed to the drugs for up to 96 h and total RNA extraction was performed using the GeneJET RNA Purification kit according to the manufacturer's instructions. RNA concentration and purity were

determined by using the Nanodrop-1000 spectrophotometer. RNA integrity was checked by gel electrophoresis. According to the manufacturer's instructions, genomic DNA was removed from total RNA using the DNase I, RNase-free kit. The DNAase I treated RNA was again cleaned with the GeneJET RNA Purification kit, re-quantified and stored at -80 °C until used. According to the manufacturer's instructions, cDNA was synthesised using the MultiScribe™ Reverse Transcriptase from 700 ng RNA. The reverse transcriptase reaction was carried out in a GeneAmp® PCR System 9600 Thermal Cycler for 10 min at 25 °C, 120 min at 37 °C and then the enzyme was deactivated for 5 min at 85 °C.

The cDNA aliquots were then utilised in qPCR reactions for *BAX and BCL-2*, with TBP, RPLP0 and TFRC used as the endogenous reference genes. PCR reactions were amplified for 40 cycles prior to which the AmpliTaq Gold® DNA polymerase was activated for 10 min at 95 °C. Each cycle consisted of a denaturing step for 15 s at 95 °C, and annealing/extension step for 1 min at 60 °C. PCR amplification was performed in a final volume of 20 μL using the Power SYBR® Green PCR Master Mix with the ABI 7500 real-time PCR machine. Primer sequences and PCR product sizes are indicated in Table 2. To confirm the absence of nonspecific amplification, PCR products were separated on 3% agarose gels, stained with ethidium bromide and images acquired with the BioRad Gel Doc® XR (Model 170-8170 Segrate, Milan, Italy). Melt curves were generated for each PCR product using the Applied Biosystems ABI 7500 software. The relative mRNA expression levels of target genes in each sample were calculated using the qbasePLUS software (Biogazelle, Zulte, Belgium). The expression stability of the reference genes was evaluated using qbasePLUS version 2.3. This software uses a pair-wise comparison model to calculate the stability of each reference gene, and selects the two or more most stable genes from a panel of reference genes for normalisation.[28] Genes are ranked based on a gene stability parameter M, where a low M value indicates high expression stability. To further indicate how stable a gene is expressed, the qbasePLUS software also calculates a coefficient of variation. PCR baseline quantification cycle (Cq) values were exported from the ABI 7500 software as an Excel file (Microsoft, Redmond, WA, USA) and imported into the qbasePLUS software. The data was analysed with the default settings and the arithmetic mean of replicates was used. Data from standard curve experiments from the ABI 7500 software, imported into the qbasePLUS software, was used to generate amplification efficiencies and standard errors that are used downstream by the qbasePLUS software to determine normalised gene expression levels. The relative quantity of each target/sample combination was scaled to the average Cq of corresponding target (scale set to untreated control in the qbasePLUS software). The relative expression of specific genes in the experiment were normalised as a ratio to the amount of the two most stably expressed reference genes according to the $-^{\Delta\Delta C}$T method of Livak and Schmittgen[29].

Table 2: Oligonucleotide sequences used for qPCR

Gene	Sequences (5'–3' direction)	Product size (bp)
BAX	F: CCT TTT CTA CTT TGC CAG CAA AC R: GAG GCC GTC CCA ACC AC	148
BCL-2	F: ATG TGT GTG GAG AGC GTC AAC C R: GCA TCC CAG CCT CCG TTA TC	122
TBP	F: TGA TGC CTT ATG GCA CTG GAC TGA R: CTG CTG CCT TTG TTG CTC TTC CAA	86
RPLP0	F: TGC AGC TGA TCA AGA CTG GAG ACA R: TCC AGG AAG CGA GAA TGC AGA GTT	178
TFRC	F: GGC ACC ATC AAG CTG CTG AAT GAA R: GTT GAT CAC GCC AGA CTT TGC TGA	133

Immunofluorescence of BAX and BCL-2

The ABCAM® double labelling procedure was followed for immunofluorescence analysis. $1x10^4$ cells were plated on cover slips 1 day before the experiment. Cells were treated with the drugs at the indicated combinations and concentrations for up to 96 h. Following treatment, cells were washed three times with 0.5% BSA in PBS, followed by fixation in 10% phosphate buffered formalin for 10 min. The fixed cells were rinsed in PBS and permeabilised with 0.05% Triton-X 100, washed and blocked with 10% normal goat serum in PBS for 60 min to eliminate non-specific binding of the secondary antibody. Cells were then incubated with polyclonal rabbit anti-human BAX (1:1000 DAKO) and monoclonal mouse anti-human BCL2 (1:100 DAKO) overnight at 4 °C in a moist chamber. Following overnight incubation, the cells were washed and incubated for 2 h with a FITC-conjugated goat anti-mouse antibody (1:500 ABCAM) and rhodamine-conjugated goat anti-rabbit antibody (1:1000 ABCAM) in the dark. Slides were rinsed, nuclei counterstained in 4',6-diamidino-2-phenylindole, dihydrochloride (DAPI) 300 nM for 5 min, rinsed and mounted with fluoromount (Sigma). Negative control groups were set up to ensure that the secondary antibodies were specific for their primary antibodies. The primary antibodies were substituted with 0.1% BSA/PBS and the normal protocol was carried out. In a second negative control, the cells were treated in the same way as the experimental slides, but the secondary antibodies were substituted with PBS. The positive controls used in this study were HeLa cells, which had been previously shown to express BAX and HepG2 cells, shown to express BCL-2.

Cells were visualised using a Zeiss Laser Scanning Confocal Microscope 780 under a Zeiss 100X oil immersion objective. Slides were kept dark once fluorescent antibodies had been added, to prevent bleaching. As fading of the fluorochromes will take place, images were analysed within the same time period. The image acquisition settings remained constant for all exposures. Images were taken and, using the ZEN 2010 (Carl Zeiss, Jena, Germany) image analysis software, the intensity of the fluorescence of each micrograph was analysed. This analysis was done by using the software to initially define the regions of interest – the nucleus and cytoplasm. The mean intensity of each region of interest from the treatment groups was then analysed with the statistics software JMP® (Version 10.0 SAS Institute Inc., Cary, NC, USA). Data are reported as mean$\pm$SEM. After verifying the normal distribution and the homogeneity of the variance using an F test ($p<0.05$), a one-way analysis of variance (where a significance level of $p<0.05$ was set) was used to compare the results.

Acridine orange staining

Fluorescence light microscopy with differential uptake of fluorescent DNA binding dyes such as AO provides a simple, rapid, and accurate method for measuring apoptosis and cell membrane integrity.[30] Acridine orange permeates all cells, making the nuclei appear green. Live cells therefore have a normal green nucleus; early apoptotic cells have a bright green nucleus with condensed or fragmented chromatin; late apoptotic cells display condensed and fragmented chromatin and membrane blebbing.[30,31] This procedure was carried out to determine whether the protease inhibitors inhibit drug-induced apoptosis in the breast cell lines. MCF-7 and MCF-10A cells were routinely grown on 22-mm square coverslips placed into 35-mm culture dishes (Costar, Cambridge MA, USA). Cells were treated with methanol and the protease inhibitor based combinations, LPV/r, for 96 h before being stimulated[32] to undergo apoptosis with 1 μM CPT (MCF-7) and 30 μM CPT (MCF-10A) for 6 h. After induction of apoptosis, cells were stained with the AO dye mix for 5 min. The dye mix was 100 μg/mL AO in PBS.[33] After AO staining, culture dishes were inverted and fixed with formaldehyde vapour for 1 min to prevent the photo-damaging effects of continuous excitation on living cells because of the photosensitising effects of most fluorescent dyes.[34] A total of 600 cells per cell line per group was counted from three independent experiments to determine the percentage of apoptotic cells.

Cell death detection

The Cell Death Detection ELISA^{plus} kit (Roche Molecular Biochemicals, Mannheim, Germany) is used to measure changes in apoptosis.[35,36] In this procedure, internucleosomal DNA fragmentation is quantitatively

assayed by antibody mediated capture and detection of cytoplasmic mononucleosome and oligonucleosome associated histone-DNA complexes. Cells were cultured (as described) in 96-well plates in duplicate. Cells were treated with 0.01% methanol (vehicle) and the protease inhibitor based combination LPV/r (in the appropriate medium) for 96 h before being stimulated[32] to undergo apoptosis with 1 μM CPT (MCF-7)[37] and 30 μM CPT (MCF-10A)[38] for 6 h.

According to the protocol described by Liu et al.[35] and Tu et al.[36], cells were centrifuged (200 *g*), resuspended in 200 μL of the lysis buffer supplied in the kit and incubated for 30 min at room temperature. Nuclei were then pelleted at 200 *g* for 10 min and 20 μL of the supernatant (cytoplasmic fraction) from the treated group, untreated group, positive control and background control were transferred into the corresponding streptavidin-coated wells for the enzyme-linked immunosorbent assay (ELISA) according to the manufacturer's standard protocol.[36] Then, 80 μL of the immunoreagent was added to each well containing the 20 μL supernatant or controls. The wells were covered with adhesive foil and incubated on a microplate shaker under gentle shaking (300 rpm) for 2 h at room temperature. The solution was then removed by gentle pipetting; wells were rinsed three times with 250 μL incubation buffer and removed. Next, 100 μL of the ABTS solution was pipetted into the wells and incubated on a plate shaker at 250 rpm until the colour development was sufficient for photometric analysis (approximately 10–20 min). Then 100 μL of the ABTS stop solution was added to each well. Absorbance was measured at 405 nm and 490 nm (reference wavelength) with a microplate reader (Anthos 2010 Model 17-550). Signals in wells containing the substrate only were subtracted as background. Data were analysed from three independent experiments. Data analysis was performed using JMP® (Version 10.0). Data are reported as mean$\pm$SEM. After verifying the normal distribution and the homogeneity of the variance using an F test ($p<0.05$), a one-way analysis of variance (significance level $p<0.05$) was used to compare the results. If there was a significant difference between the means ($p<0.05$), then a Tukey–Kramer post-hoc analysis was performed to determine which treatment groups were significantly different from each other.

Statistical analysis

Statistical analysis was performed using JMP® (Version 10.0). Data are reported as mean$\pm$SEM. After verifying the normal distribution and the homogeneity of the variance using an F test ($p<0.05$), a one-way analysis of variance (where a significance level of $p<0.05$ was set) was used to compare the results. All means were then compared using the Tukey–Kramer honestly significant different test.

Results

Effects of ARV drugs on the viability of MCF-7 and MCF 10A cells

The effects of the antiretroviral (ARV) drugs on the viability of MCF-7 and MCF-10A breast cells were investigated by the neutral red uptake assay. LPV significantly ($p<0.05$) reduced the viability of the normal breast MCF-10A cells at 24 h when compared to the vehicle control (Figure 1a). Also at 24 h, LPV significantly ($p<0.05$) reduced the percentage viability of the normal breast cells when compared to the cancer breast MCF- 7 cells (Figure 1a). With continuous exposure, the normal breast cells recovered from the initial cytotoxic effects of LPV, such that from 48 h to 96 h, there were no significant differences between the treated cells and their vehicle controls (Figure. 1a). LPV did not significantly ($p<0.05$) alter the percentage viability of the breast cancer MCF-7 cells from 24 h to 96 h when compared to their vehicle controls (Figure 1a The reduction in % viability of MCF-7 cells at 48 h was not statistically significant. RTV significantly ($p<0.05$) reduced the viability of the normal breast MCF-10A cells when compared to the vehicle control and the breast cancer MCF-7 cells at 24 h (Figure 1b). At 48 h, there was no significant ($p<0.05$) alteration in percentage viability between RTV treated and untreated in either cell line. At 72 h, RTV treated MCF-10A cells demonstrated significantly ($p<0.05$) increased viability, while at 96 h, there was no statistical significance in the difference between RTV treated and untreated MCF-10A cells (Figure 1b). RTV did not alter the percentage viability of the breast cancer MCF-7 cells from 24 h to 96 h (Figure 1b). The apparent increase in % viability of RTV treated MCF-7 cells at 24 h did not achieve statistical significance. The combination LPV/r interestingly did not significantly ($p<0.05$) alter the percentage viability of the normal (MCF-10A) and cancer (MCF-7) breast cells from 24 h to 96 h when compared to their respective vehicle controls and to each other at the respective time points (Figure 1c).

Effects of ARV drugs on BAX/BCL2 gene expression in MCF-7/MCF10A cells

The effects of the protease inhibitors on *BAX* and *BCL-2* mRNA expression in both cell lines were investigated after exposure for 24 h to 96 h (data shown for 96 h). In both cell lines treated with LPV, the differences in fold changes were not significantly ($p<0.05$) different from the untreated controls for *BAX* and *BCL-2* (Figure 2a). In both cell lines treated with RTV and LPV/r, the differences in fold changes were not significantly ($p<0.05$) different from the untreated controls for *BAX* and *BCL-2* (Figures 2b and 2c). Results were analysed with the qBASE^{PLUS} software and normalised to the expression levels of *TBP*, *TFRC* and *RPLP0*.

Effects of ARVs on BAX/BCL2 immunofluorescence in MCF-7/10A cells

Figure 3a–c is a graphical representation of immunofluorescence intensities illustrating the effects of LPV, RTV and LPV/r on expression of BAX and BCL-2 in MCF-7 and MCF-10A cell cultures after 96 h of exposure. In the untreated group of MCF-7 cells (micrographs not shown), BAX and BCL-2 proteins were expressed and co-localised in the nuclei and cytoplasm. Mean fluorescence intensities of BAX and BCL-2 in the nuclei and cytoplasm were statistically analysed (ANOVA) between the untreated/vehicle control and the treated groups (graphical representation in Figure 3a–c).There are no differences in protein staining, localisation and intensity of vehicle (0.01% methanol) treated group when compared to growth medium alone. BAX and BCL-2, localisations remain unaltered across treatment groups. The untreated group of MCF-10A cells showed the expression and localisation of BAX and BCL-2 (micrographs not shown). Mean fluorescence intensities of BAX and BCL-2 in the nuclei and cytoplasm were statistically analysed (ANOVA) between the untreated (vehicle) control and the treated groups. (graphical representation in Figure 3a–c). There were no differences in protein staining, localisation and intensity of vehicle (0.01% methanol) treated group as compared to growth medium alone. BAX and BCL-2 localisations remain unchanged across treatment groups.

Protease inhibitors and apoptosis inhibition

Numerous studies[19,20,32,39] have shown that protease inhibitors have potent anti-apoptotic effects in a variety of cell systems, the majority of which are immune cells. We assessed the anti-apoptotic effects of LPV/r to test whether these generalised effects are seen in the currently investigated human breast cell lines and to test whether they alter apoptosis via a pathway other than the *BAX/BCL-2* pathway, since they had no significant effects on *BAX* and *BCL-2* mRNA expression (Figure 2). From an evaluation of the many methods currently used to analyse apoptosis, chromatin condensation and nuclear fragmentation remain the hallmarks of apoptotic cells.[30] Because it had been suggested that as a rule, classification of cell death in a given model should always include morphological examination coupled with at least one other assay,[40] we chose the acridine orange staining for morphological evaluation and an ELISA based method for the quantitative analysis of DNA fragmentation. From previous optimisation experiments, it was determined that 1 μM and 30 μM CPT significantly induced apoptosis in the MCF-7 and MCF10-A cells respectively for 6 h. Figures 4 and 5 represent data from AO staining showing that a 96 h exposure to LPV/r did not significantly ($p<0.05$) inhibit apoptosis in both breast cell lines. Figure 6 represents data quantifying DNA fragmentation between groups. These ELISA results also confirm that LPV/r did not significantly inhibit drug-induced apoptosis in the breast MCF-7 and MCF-10A cells.

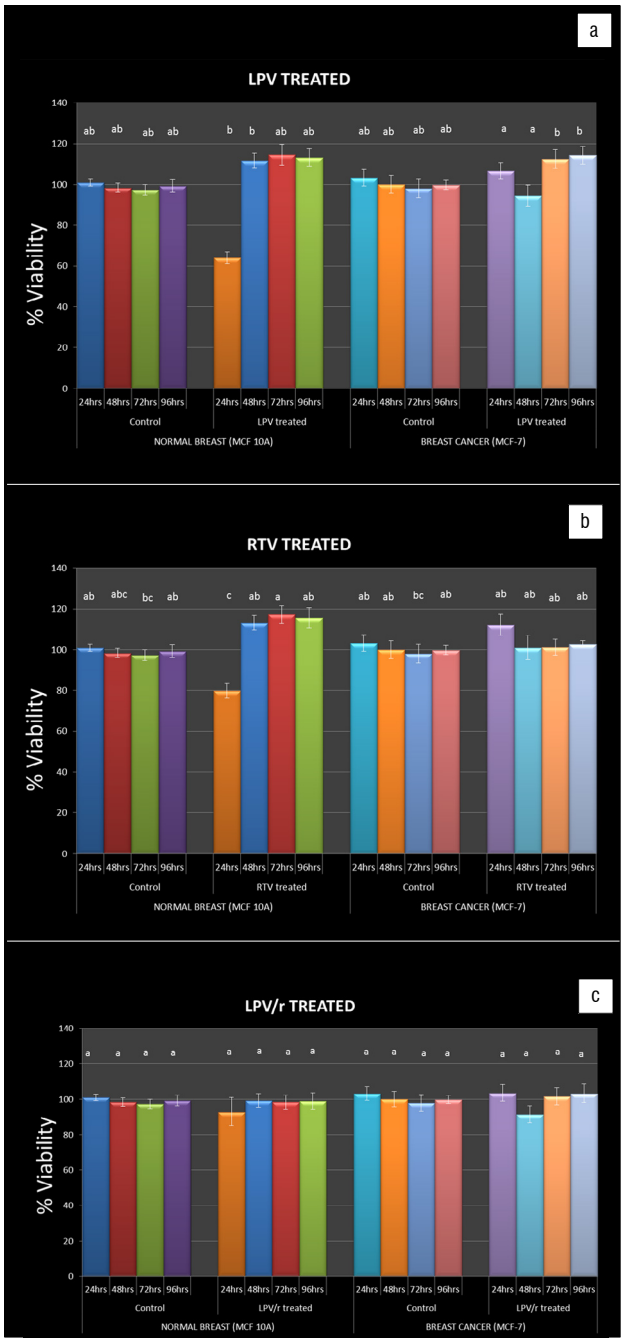


Figure 1: Effects of (a) LPV, (b) RTV and (c) LPV/r on the percentage viability of normal breast (MCF-10A) and breast cancer (MCF- 7) cells. Cells were incubated with 9.8 μg/mL LPV, 0.6 μg/mL RTV and 9.8 μg/mL LPV/0.6 μg/mL for 24–96 h. Graphs were plotted with mean+SEM percentage viability from a mean of three.

Discussion and conclusion

This study investigated the possibility that the use of protease inhibitors could be a risk factor for the initiation or development of breast cancer in patients receiving HAART. An analysis of the incidence of non-AIDS defining cancers in HIV-infected patients suggests that the incidence of NADM in these patients has significantly increased since the introduction and implementation of the combined antiretroviral therapy,[12] creating the need to effectively establish or rule out the possibility that candidate antiretroviral drugs may promote breast cancer. The breast was chosen as a model of non-AIDS-defining malignancies.

The findings reported here show that the antiretroviral drugs: lopinavir, ritonavir and their combination (all at clinically relevant concentrations, which reflect their steady-state peak plasma concentrations in patients receiving these drugs)[41,42] demonstrated some form of cytotoxicity in the breast cells, which may alter the activity of the nuclear and mitochondrial genome of cells and may cause genotoxic effects largely related to the tumour initiation processes.[43-45] LPV showed an initial cytotoxic effect on the MCF-10A cells after 24 h of exposure. Even though this cytotoxic effect disappeared after 48 h, 72 h and 96 h of exposure, this initial cytotoxicity can produce genomic rearrangements that may act as the primary step toward carcinogenesis. RTV was significantly cytotoxic to the normal breast MCF-10A cells at 24 h, when compared with the control group and the breast cancer MCF-7 cells. The cytotoxic effects of RTV on the normal breast disappeared with continuous administration but could lead to DNA damage and may trigger the processes of cancer initiation in normal breast cells. RTV demonstrated opposing effects on normal breast and breast cancer cells at 24 h. It generally appeared to promote survival in normal breast after an initial toxic effect. Interestingly, as combined in Kaletra®, did not significantly alter the percentage viability of both breast cell lines.

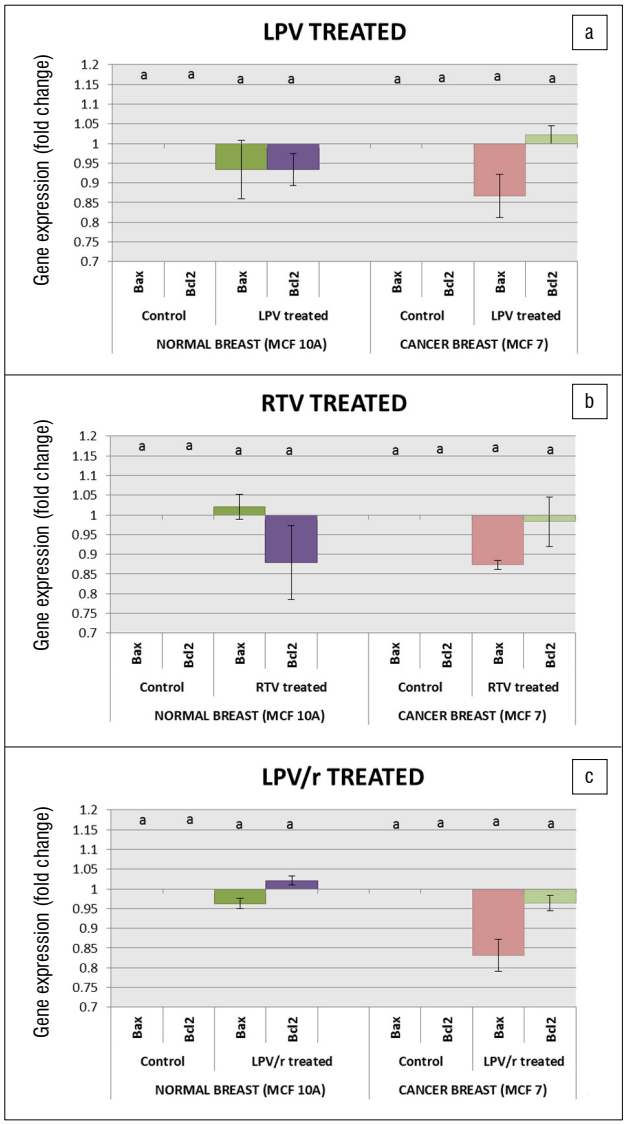


Figure 2: Effects of (a) LPV, (b) RTV and (c) LPV/r on *BAX* and *BCL-2* mRNA expression in normal breast (MCF-10A) and breast cancer (MCF-7) cells. Cells were incubated with 9.8 μg/mL LPV, 0.6 μg/mL RTV and 9.8 μg/mL LPV/0.6 μg/mL for 96 h. Data (mean±SEM) are represented as fold changes of gene expression relative to values from the vehicle-treated cells (defined as 1) and representative of three independent experiments for RNA extraction. Groups not connected by the same letter are significantly ($p<0.05$) different from each other.

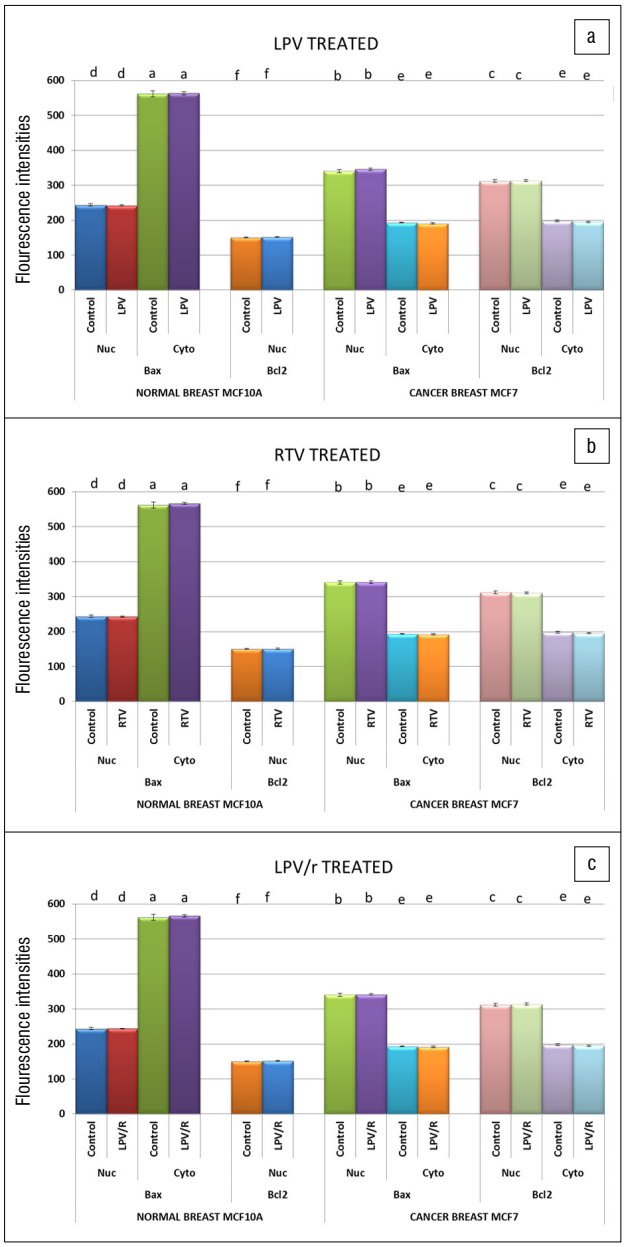


Figure 3: Effects of (a) LPV, (b) RTV and (c) LPV/r on BAX and BCL- 2 protein expression in normal breast (MCF-10A) and breast cancer (MCF-7) cells. Cells were incubated with 9.8 μg/mL LPV, 0.6 g/mL RTV and 9.8 μg/mL LPV/0.6 μg/mL for 96 h. Data (mean±SEM), are represented as fluorescence intensities of protein expression relative to values from the vehicle-treated cells and representative of 3 independent experiments for immunofluorescence staining. Groups not connected by the same letter are significantly ($p<0.05$) different from each other.

To promote cancer, the antiretroviral drugs, either individually or in combination at the clinically relevant concentrations tested were expected to up-regulate the anti-apoptotic *BCL-2* mRNA and or protein and down- regulate the pro-apoptotic *BAX* mRNA and or protein. Testing the effects of protease inhibitors (individually and in combination), our findings show that the antiretroviral drugs; lopinavir, ritonavir and Kaletra®, all at clinically relevant concentrations (which reflect their steady-state peak plasma concentrations), do not alter the mRNA expression of apoptosis related *BAX* and *BCL-2*. Neither do they alter the localisation of their proteins in the human breast cancer cell line MCF-7 and non-tumorigenic immortalised breast cell line MCF-10A. These results are consistent with the findings of other researchers. Reports by Gomez-Sucerquia et al.[46] showed that efavirenz at similar concentrations did not alter the mRNA expression of *BAX* and *BCL-2 like 1* genes following a 24 h treatment period in human Hep3B cells. Phenix et al.[32], Badley[19] and Rizza and Badley[20] reported that protease inhibitors, including lopinavir, did not alter the mRNA expression and protein synthesis of *BAX, BCL-2* and some other key apoptotic genes following a 3-day exposure in immune cells. Some other proteins and/or genes of this or another pathway, however, might be involved.

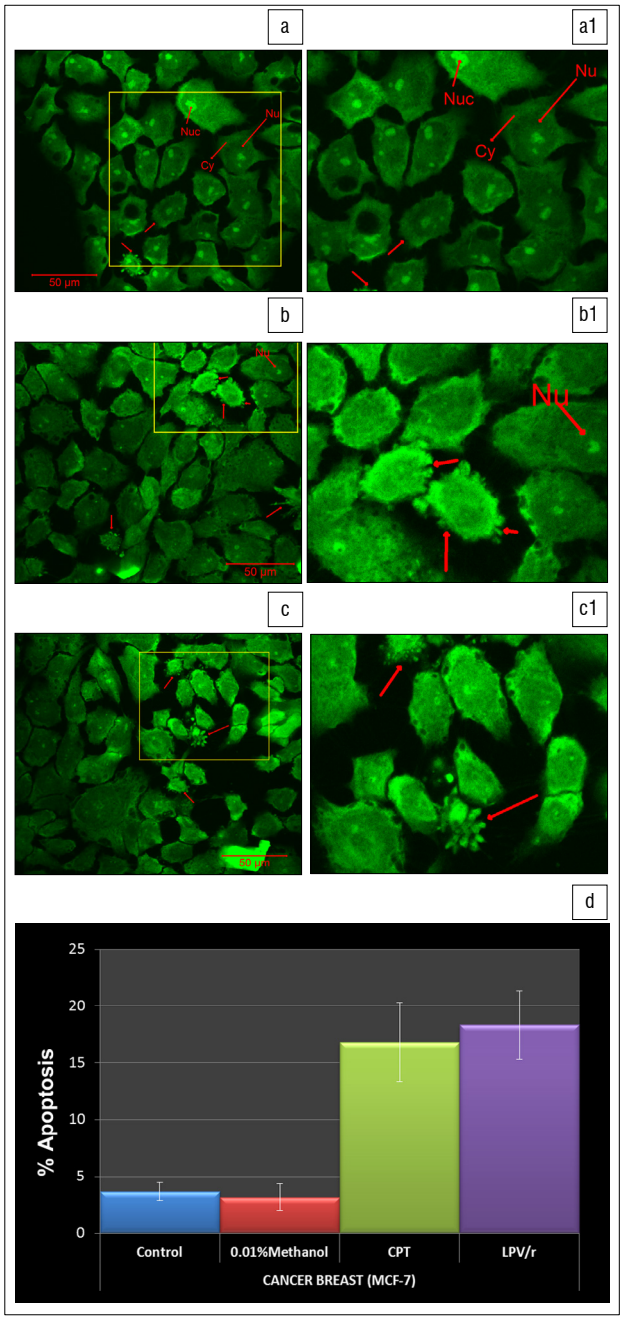


Figure 4: Effects of LPV/r on CPT mediated apoptosis in MCF-7 cells. (a) Untreated control MCF-7 cells. (b) CPT treated MCF-7 cells. MCF-7 cells were stimulated with 1 μM CPT for 6 h without prior exposure to LPV/r. (c) MCF-7 cells incubated with LPV/r prior to apoptosis induction. Apoptosis was assessed using acridine orange staining. Arrows point to apoptotic bodies and cells undergoing apoptosis with the characteristic feature of membrane blebbing. Areas within the box are magnified to the right (a1, b1, c1). (d) A total of 600 cells per group was counted from three independent experiments to determine the percentage of apoptotic cells. LPV/r had no significant inhibitory effects on CPT induced apoptosis in MCF-7 cells.

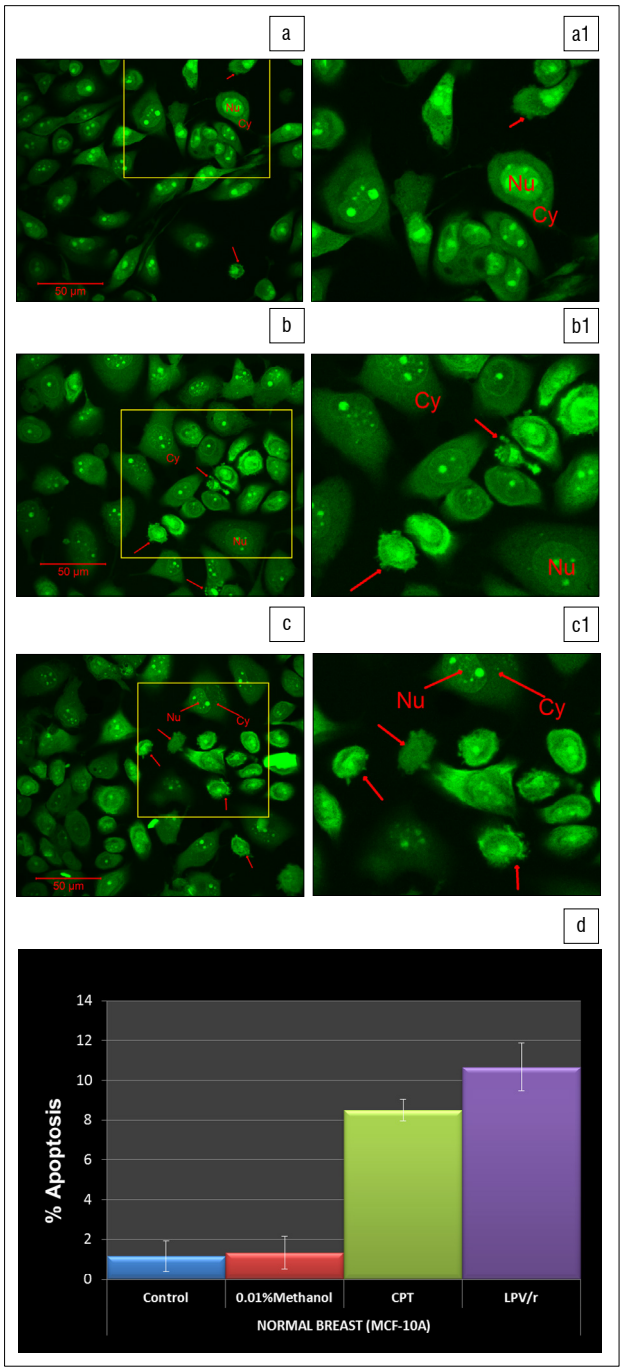


Figure 5: Effects of LPV/r on CPT mediated apoptosis in MCF-10A cells. (a) Untreated control MCF-10A cells. (b) CPT treated MCF-10A cells. MCF-10A cells were stimulated with 30 μM CPT for 6 h without prior exposure to LPV/r. (c) MCF-10A cells incubated with LPV/r prior to apoptosis induction. Apoptosis was assessed using acridine orange staining. Arrows point to apoptotic bodies and cells undergoing apoptosis with the characteristic feature of membrane blebbing. Areas within the box are magnified to the right (a1, b1, c1). (d) A total of 600 cells per group was counted from three independent experiments to determine the percentage of apoptotic cells. LPV/r had no significant inhibitory effects on CPT induced apoptosis in MCF-10A cells.

It has been shown previously that protease inhibitors have potent anti-apoptotic effects in different cellular systems, the majority of which are immune cells.[19,20,32,39] These studies reported that the protease inhibitors investigated inhibited drug induced apoptosis. This anti-apoptotic property, (exhibited by protease inhibitors at concentrations similar to those levels achieved in people receiving the drugs) is an important phase in cancer development. The anti-apoptotic effect of LPV/r was therefore assessed to test whether these generalised effects are seen in the currently investigated cell lines and to test whether they alter apoptosis via a pathway other than the *BAX*/*BCL-2* pathway, because they had no significant effects on *BAX* and *BCL-2* mRNA expression. At a dose of 1 μM CPT significantly increased the percentage of apoptotic MCF-7 cells from 3-4% (spontaneous apoptosis) to 15–18% following a 6-h exposure, while the percentage of apoptotic MCF-10A cells increased from 1–2% to 8–9% following a 6 h exposure to 30 μM CPT. Pre-incubation with LPV/r did not significantly ($p>0.05$) inhibit CPT-induced apoptosis in both breast cell lines. The cell death ELISA assessment of DNA fragmentation showed similar results. These findings conflict with the reports of Phenix et al.[32] which demonstrated that the HIV protease inhibitor nelfinavir inhibited Jurkat T-cell apoptosis induced by a variety of different stimuli, including CPT and with our previous report that pre-incubation with LPV/r for 96 h significantly inhibited the development of apoptosis in the cervical cancer HCS-2 cell line.[47] These disparities might be because different cell lines react differently to the same target molecules. The variety of responses depends on the set of receptor proteins the cell possesses, which determines the specific subset of signals it can respond to, and also depends on the intracellular machinery by which the cell integrates and interprets the signals it receives.[48] However, Phenix et al.[32] and Adefolaju et al.[47] also reported that inhibition of apoptotic death was not associated with alterations in mRNA expression of a variety of pro- and anti-apoptotic factors, and was not dependent on protein synthesis.

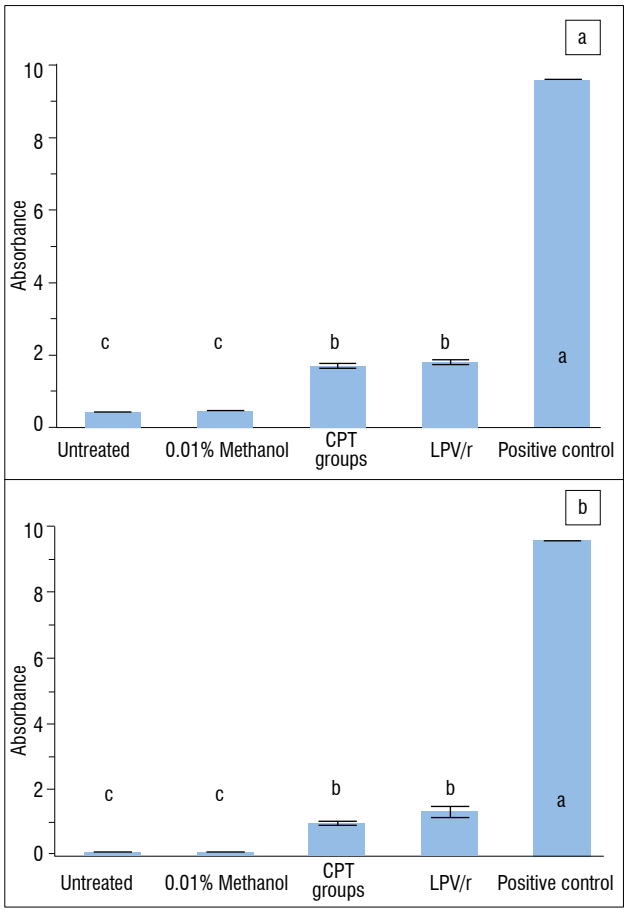


Figure 6: Effects of LPV/r on CPT mediated apoptosis in MCF-7 and MCF-10A cells evaluated by the Cell Death Detection ELISA. (a) MCF-7 and (b) MCF-10A cells were treated with methanol or LPV/r for 96 h and stimulated with CPT (1 μM and 30 μM, respectively) and apoptosis was assessed by quantifying internucleosomal DNA fragmentation. LPV/r treatment did not significantly ($p<0.05$) inhibit CPT induced apoptosis in both cell lines. Levels not connected by the same letter are significantly different from each other.

In the HAART era, it is a paradox that people living with HIV and AIDS are now at an increased risk for developing several specific non-AIDS-defining malignancies[4,49-53] including breast cancer.[12,52] Such that, despite the immune reconstitution induced by HAART, increased NADMs deaths have been reported amongst HIV patients.[54] It has also been reported that patients with NADMs often have more aggressive cancers and display more advanced stages of the disease[55,56], with breast cancer metastasising early and usually more poorly differentiated.[57] The effects of HAART on the risk for NADM have not been clearly established. Research evidence has been conflicting as to whether or which antiretroviral drugs or classes decrease, increase, or have no effect on the risk of developing NADM,[56,58] with no definite pattern emerging from such studies.

The speculation that some antiretroviral drugs or combinations may have an adverse effect on the risk of carcinogenesis among HIV patients under HAART led to this study. Our results show that these protease inhibitors do not alter the mRNA expression of apoptosis related *BAX* and *BCL-2,* neither do they alter the localisation of their proteins in the human breast cancer cell line MCF-7 and non-tumorigenic immortalised breast cell line MCF- 10A. The findings reported here also show that HIV protease inhibitors – LPV/r did not exhibit anti-apoptotic properties in MCF-7 and MCF-10A cells. If the ARVs initiate or promote cancer, these pathways are not likely to be involved, so the effects of these agents need to be investigated on other oncogenic pathways. Also, considering that this is an in-vitro study, it should be noted that the possibility of obtaining different results cannot be excluded when studied in an in vivo model, in which responses to pharmacological agents are much more complicated.

Acknowledgements

We thank Dr Clem Penny, Department of Internal Medicine, University of the Witwatersrand, Johannesburg, South Africa for providing the MCF- 7 and MCF-10A cells. This work was supported by the South African Medical Research Council.

Authors' contributions

G.A.A., K.E.T. and M.J.H. were involved with the conceptualisation of study and experimental design, standardisation and optimisation of all protocols, data acquisition, analysis and interpretation and manuscript write-up and final approval of the revised version. M.J.H was also involved with supervision of the project.

References

1. Joint United Nations Programme on HIV/AIDS (UNAIDS). Joint United Nations Programme on HIV/AIDS update on the global AIDS epidemic 2013. Geneva: UNAIDS; 2013. Available from: http://www.unaids.org/sites/default/files/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf
2. Department of Health. The South African antiretroviral treatment guidelines 2010. Republic of South Africa. Pretoria: Department of Health; 2010. Available from: http://www.uj.ac.za/EN/CorporateServices/ioha/Documentation/Documents/ART%20Guideline.pdf
3. Nachega JB, Stein DM, Lehman DA, Hlatshwayo D, Mothopeng R, Chaisson RE, et al. Adherence to antiretroviral therapy in HIV-infected adults in Soweto, South Africa. AIDS Res Hum Retroviruses. 2004;20(10):1053–1056.
4. Sikora MJ, Rae JM, Johnson MD, Desta Z. Efavirenz directly modulates the oestrogen receptor and induces breast cancer cell growth. HIV Med. 2010;11(9):603–607.
5. Maqutu D, Zewotir T, North D, Naidoo K, Grobler A. Determinants of optimal adherence over time to antiretroviral therapy amongst HIV positive adults in South Africa: A longitudinal study. AIDS Behav. 2011;15(7):1465–1474.
6. Blas-Garcia A, Apostolova N, Ballesteros D, Monleon D, Morales JM, Rocha M, et al. Inhibition of mitochondrial function by efavirenz increases lipid content in hepatic cells. Hepatology. 2010;52(1):115–125.
7. Powles T, Robinson D, Stebbing J, Shamash J, Nelson M, Gazzard B, et al. Highly active antiretroviral therapy and the incidence of non-AIDS-defining cancers in people with HIV infection. J Clin Oncol. 2009;27(6):884–890.
8. Torres SM, Walker DM, Carter MM, Cook DL Jr, McCash CL, Cordova EM, et al. Mutagenicity of zidovudine, lamivudine, and abacavir following in vitro exposure of human lymphoblastoid cells or in utero exposure of CD-1 mice to single agents or drug combinations. Environ Mol Mutagen. 2007;48(3/4):224–238.
9. Oluwole SF, Ali AO, Shafaee Z, Depaz HA. Breast cancer in women with HIV/AIDS: Report of five cases with a review of the literature. J Surg Oncol. 2005;89(1):23–27.
10. Clay PG, Taylor TA, Glaros AG, McRae M, Williams C, McCandless D, et al. "One pill, once daily": What clinicians need to know about Atriplatrade mark. Ther Clin Risk Manag. 2008;4(2):291–302.
11. Phillips AA, Justman JE. Screening HIV-infected patients for non-AIDS-defining malignancies. Curr HIV/AIDS Rep. 2009;6(2):83–92.
12. Shiels MS, Cole SR, Kirk GD, Poole C. A meta-analysis of the incidence of non-AIDS cancers in HIV-infected individuals. J Acquir Immune Defic Syndr. 2009;52(5):611–622.
13. Clifford GM, Polesel J, Rickenbach M, Dal Maso L, Keiser O, Kofler A, et al. Cancer risk in the Swiss HIV Cohort Study: Associations with immunodeficiency, smoking, and highly active antiretroviral therapy. J Natl Cancer Inst. 2005;97(6):425–432.
14. Silverberg MJ, Abrams DI. AIDS-defining and non-AIDS-defining malignancies: Cancer occurrence in the antiretroviral therapy era. Curr Opin Oncol. 2007;19(5):446–451.
15. Engels EA, Biggar RJ, Hall HI, Cross H, Crutchfield A, Finch JL, et al. Cancer risk in people infected with human immunodeficiency virus in the United States. Int J Cancer. 2008;123(1):187–94.
16. Eum KH, Lee M. Crosstalk between autophagy and apoptosis in the regulation of paclitaxel-induced cell death in v-Ha-ras-transformed fibroblasts. Mol Cell Biochem. 2011;348(1/2):61–68.
17. Llambi F, Green DR. Apoptosis and oncogenesis: Give and take in the *BCL-2* family. Curr Opin Genet Dev. 2011;21(1):12–20.
18. Sjostrom J, Bergh J. How apoptosis is regulated, and what goes wrong in cancer. BMJ. 2001;322(7301):1538–1539.
19. Badley AD. In vitro and in vivo effects of HIV protease inhibitors on apoptosis. Cell Death Differ. 2005;12:924–931.
20. Rizza SA, Badley AD. HIV protease inhibitors impact on apoptosis. Med Chem. 2008;4(1):75–79.
21. Razandi M, Alton G, Pedram A, Ghonshani S, Webb P, Levin ER. Identification of a structural determinant necessary for the localization and function of estrogen receptor alpha at the plasma membrane. Mol Cell Biol. 2003;23(5):1633–1646.
22. Pan Y, Wang L, Dai JL. Suppression of breast cancer cell growth by Na+/H+ exchanger regulatory factor 1 (NHERF1). Breast Cancer Res. 2006;8(6):R63.
23. Noor MA, Flint OP, Maa J-F, Parker RA. Effects of atazanavir/ritonavir and lopinavir/ritonavir on glucose uptake and insulin sensitivity: Demonstrable differences in vitro and clinically. AIDS. 2006;20(14):1813–1821.
24. Tong L, Phan TK, Robinson KL, Babusis D, Strab R, Bhoopathy S, et al. Effects of human immunodeficiency virus protease inhibitors on the intestinal absorption of tenofovir disoproxil fumarate in vitro. Antimicrob Agents Chemother. 2007;51(10):3498–3504.
25. Abbott Laboratories. Drug information: Norvir® [document on the Internet]. c2013 [cited 2015 Jan 15]. Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/020945s028lbl.pdf
26. Abbott Laboratories. Drug information: Kaletra® [document on the Internet]. c2010 [cited 2015 Jan 15]. Available from: http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/021226s030lbl.pdf
27. Borenfreund E, Puerner JA. Toxicity determined in vitro by morphological alterations and neutral red absorption. Toxicol Lett. 1985;24(2/3):119–124.
28. Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol. 2007;8(2):R19.
29. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods. 2001;25(4):402–408.

30. Ribble D, Goldstein NB, Norris DA, Shellman YG. A simple technique for quantifying apoptosis in 96-well plates. BMC Biotechnol. 2005;5:12.

31. Yang S, Liu M, Liang N, Zhao Q, Zhang Y, Xue W, et al. Discovery and antitumor activities of constituents from *Cyrtomium fortumei* (J.) Smith rhizomes. Chem Cent J. 2013;7:24.

32. Phenix BN, Lum JJ, Nie Z, Sanchez-Dardon J, Badley AD. Antiapoptotic mechanism of HIV protease inhibitors: Preventing mitochondrial transmembrane potential loss. Blood. 2001;98(4):1078–1085.

33. Mironova EV, Evstratova AA, Antonov SM. A fluorescence vital assay for the recognition and quantification of excitotoxic cell death by necrosis and apoptosis using confocal microscopy on neurons in culture. J Neurosci Meth. 2007;163(1):1–8.

34. Canete M, Juarranz A, Lopez-Nieva P, Alonso-Torcal C, Villanueva A, Stockert JC. Fixation and permanent mounting of fluorescent probes after vital labelling of cultured cells. Acta Histochem. 2001;103(2):117–126.

35. Liu CY, Takemasa A, Liles WC, Goodman RB, Jonas M, Rosen H, et al. Broad-spectrum caspase inhibition paradoxically augments cell death in TNF-alpha-stimulated neutrophils. Blood. 2003;101(1):295–304.

36. Tu HP, Chen YT, Chiu HC, Chin YT, Huang SM, Cheng LC, et al. Cyclosporine A enhances apoptosis in gingival keratinocytes of rats and in OECM1 cells via the mitochondrial pathway. J Periodontal Res. 2009;44(6):767–775.

37. Nieves-Neira W, Pommier Y. Apoptotic response to Camptothecin and 7-hydroxystaurosporine (UCN-01) in the 8 human breast cancer cell lines of the NCI anticancer drug screen: Multifactorial relationships with topoisomerase I, protein kinase C, Bcl-2, p53, MDM-2 and caspase pathways. Int J Cancer. 1999;82(3):396–404.

38. Rastogi S, Joshi B, Fusaro G, Chellappan S. Camptothecin induces nuclear export of prohibitin preferentially in transformed cells through a CRM-1-dependent mechanism. J Biol Chem. 2006;281(5):2951–2959.

39. Vlahakis SR, Bennett SA, Whitehead SN, Badley AD. HIV protease inhibitors modulate apoptosis signaling in vitro and in vivo. Apoptosis. 2007;12(5):969– 977.

40. Renvoize C, Biola A, Pallardy M, Breard J. Apoptosis: Identification of dying cells. Cell Biol Toxicol. 1998;14(2):111–120.

41. Apostolova N, Gomez-Sucerquia LJ, Moran A, Alvarez A, Blas-Garcia A, Esplugues JV. Enhanced oxidative stress and increased mitochondrial mass during efavirenz-induced apoptosis in human hepatic cells. Br J Pharmacol. 2010;160(8):2069–2084.

42. Bumpus NN. Efavirenz and 8-hydroxyefavirenz induce cell death via a JNK- and BimEL-dependent mechanism in primary human hepatocytes. Toxicol Appl Pharmacol. 2011;257(2):227–234.

43. Bishop AJ, Schiestl RH. Homologous recombination as a mechanism of carcinogenesis. Biochim Biophys Acta. 2001;1471(3):M109–121.

44. Olivero OA. Mechanisms of genotoxicity of nucleoside reverse transcriptase inhibitors. Environ Mol Mutagen. 2007;48(3/4):215–223.

45. Wu KM, Powley MW, Ghantous H. Timing of carcinogenicity studies and predictability of genotoxicity for tumorigenicity in anti-HIV drug development. Int J Toxicol. 2012;31(3):211–221.

46. Gomez-Sucerquia LJ, Blas-Garcia A, Marti-Cabrera M, Esplugues JV, Apostolova N. Profile of stress and toxicity gene expression in human hepatic cells treated with efavirenz. Antiviral Res. 2012;94(3):232–241.

47. Adefolaju GA, Theron KE, Hosie MJ. Effects of HIV protease, nucleoside/non-nucleoside reverse transcriptase inhibitors on Bax, Bcl-2 and apoptosis in two cervical cell lines. Biomed Pharmacother. 2014;68(2):241–251.

48. Alberts B, Johnson A, Lewis J, Raff, M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.

49. Silverberg MJ, Chao C, Leyden WA, Xu L, Tang B, Horberg MA, et al. HIV infection and the risk of cancers with and without a known infectious cause. AIDS. 2009;23(17):2337–2345.

50. Spano JP, Lanoy E, Mounier N, Katlama C, Costagliola D, Heard I. Breast cancer among HIV infected individuals from the ONCOVIH study, in France: Therapeutic implications. Eur J Cancer. 2012;48(18):3335–3341.

51. Calabresi A, Ferraresi A, Festa A, Scarcella C, Donato F, Vassallo F, et al. Incidence of AIDS-defining cancers and virus-related and non-virus-related non-AIDS-defining cancers among HIV-infected patients compared with the general population in a large health district of Northern Italy, 1999–2009. HIV Med. 2013;14(8):481–490.

52. Franzetti M, Adorni F, Parravicini C, Vergani B, Antinori S, Milazzo L, et al. Trends and predictors of non-AIDS-defining cancers in men and women with HIV infection: A single-institution retrospective study before and after the introduction of HAART. J Acquir Immune Defic Syndr. 2013;62(4):414–420.

53. Cutrell J, Bedimo R. Non-AIDS-defining cancers among HIV-infected patients. Curr HIV/AIDS Rep. 2013;10(3):207–216.

54. Zucchetto A, Suligoi B, De Paoli A, Pennazza S, Polesel J, Bruzzone S, et al. Excess mortality for non-AIDS-defining cancers among people with AIDS. Clin Infect Dis. 2010;51(9):1099–1101.

55. Grulich AE, Van Leeuwen MT, Falster MO, Vajdic CM. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: A meta-analysis. Lancet. 2007;370(9581):59–67.

56. Deeken JF, Tjen-A-Looi A, Rudek MA, Okuliar C, Young M, Little RF, et al. The rising challenge of non-AIDS-defining cancers in HIV-infected patients. Clin Infect Dis. 2012;55(9):1228–1235.

57. Gewurz BE, Dezube BJ, Pantanowitz L. HIV and the breast. AIDS Read. 2005;15(8):392–396, 399–402.

58. Burgi A, Brodine S, Wegner S, Milazzo M, Wallace MR, Spooner K, et al. Incidence and risk factors for the occurrence of non-AIDS-defining cancers among human immunodeficiency virus-infected individuals. Cancer. 2005;104(7):1505–1511.

Spatial and temporal disaggregation of anthropogenic CO_2 emissions from the City of Cape Town

AUTHORS:
Alecia Nickless[1,2]
Robert J. Scholes[2]
Ed Filby[3]

AFFILIATIONS:
[1]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa
[2]Global Change and Ecosystem Dynamics, Natural Resources and the Environment, Council for Scientific and Industrial Research, Pretoria, South Africa
[3]City of Cape Town, Air Quality Management, Cape Town, South Africa

CORRESPONDENCE TO:
Alecia Nickless

EMAIL:
alecia.nickless@phc.ox.ac.uk

POSTAL ADDRESS:
Global Change and Ecosystem Dynamics, Natural Resources and the Environment, Council for Scientific and Industrial Research, PO Box 395, Pretoria 0001, South Africa

DATES:
Received: 05 Nov. 2014
Revised: 10 Feb. 2015
Accepted: 15 Feb. 2015

KEYWORDS:
carbon dioxide inventory; emission factor; error propagation; mesoscale; interpolation

HOW TO CITE:
Nickless A, Scholes RJ, Filby E. Spatial and temporal disaggregation of anthropogenic CO_2 emissions from the City of Cape Town. S Afr J Sci. 2015;111(11/12), Art. #2014-0387, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140387

This paper describes the methodology used to spatially and temporally disaggregate carbon dioxide emission estimates for the City of Cape Town, to be used for a city-scale atmospheric inversion estimating carbon dioxide fluxes. Fossil fuel emissions were broken down into emissions from road transport, domestic emissions, industrial emissions, and airport and harbour emissions. Using spatially explicit information on vehicle counts, and an hourly scaling factor, vehicle emissions estimates were obtained for the city. Domestic emissions from fossil fuel burning were estimated from household fuel usage information and spatially disaggregated population data from the 2011 national census. Fuel usage data were used to derive industrial emissions from listed activities, which included emissions from power generation, and these were distributed spatially according to the source point locations. The emissions from the Cape Town harbour and the international airport were determined from vessel and aircraft count data, respectively. For each emission type, error estimates were determined through error propagation techniques. The total fossil fuel emission field for the city was obtained by summing the spatial layers for each emission type, accumulated for the period of interest. These results will be used in a city-scale inversion study, and this method implemented in the future for a national atmospheric inversion study.

Introduction

Anthropogenic emissions are those emissions which are the result of human activities. Performing an inventory analysis is a method of quantifying these emissions based on human activity data. The basic equation is:

$$Emissions = AD \times EF \quad \text{Equation 1}$$

where *AD* is the activity data and *EF* is the emissions factor, which converts the activity data into an emission.[1] For example, in the energy sector, in the case of carbon dioxide (CO_2) emissions, the amount of fuel consumed constitutes the activity data and the emission factor would then convert the activity data into the amount of CO_2 emitted per unit of fuel.

The Intergovernmental Panel on Climate Change (IPCC), at the invitation of the United Nations Framework Convention on Climate Change (UNFCCC) has produced a set of guidelines on how to conduct an inventory analysis for greenhouse gases, with the purpose of ensuring consistency and comparability between the greenhouse gas emissions reports of different countries. Under these guidelines, a national inventory consists of all the greenhouse gas emissions and removals which have taken place within the country's national jurisdiction. Inventory analyses are usually conducted at a national level because the activity data can easily be extracted from available national statistics.

In order to assess the magnitude of sources and sinks in a particular region within a country, a national inventory is not sufficient. The activity data need to be disaggregated between the regions which make up the country. For example, in the case of a mesoscale atmospheric inversion for CO_2, which aims to estimate fluxes based on high precision measurements of CO_2 concentrations and an atmospheric transport model, prior estimates are required for anthropogenic emissions. These estimates are required at the resolution of the source and sink regions which are used in the inversion exercise. An example of such a study is the fossil fuel emissions for the USA provided by the Vulcan project used by the Carbon Tracker inversion exercise.[2] This study was able to disaggregate the 2002 fossil fuel emissions for contiguous USA, based mainly on fuel usage data, at a 10 km $\times$ 10 km spatial resolution and a temporal resolution as high as a few hours. This project aimed to improve on its predecessor inventory which provided global spatial and temporal patterns of fossil fuel emissions, which used temporal resolutions of up to a month and spatial resolutions of one degree. The EDGAR (Emission Database for Global Atmospheric Research) is a global product on a $0.1° \times 0.1°$ grid, which calculates the total emissions of CO_2 and other species for each country, and distributes these total emissions spatially and temporally according to proxy data, such as population data or road transport network data.[3] A remote sensing based product also exists at the same spatial resolution, which calculates emissions based on night-time lights, population data, national fossil fuel data, and power plant location and statistics.[4,5] At a much more detailed level, taking into account information such as building locations and their dimensions, the Hestia project provides a bottom-up approach for quantifying fossil fuel emissions for a large city.[6] Our paper describes a bottom-up methodology approach which aims to make use of the available data for the City of Cape Town, but does so in the absence of the detailed building, road and population data which were available for Indianapolis during the Hestia project.

To determine the emissions from different source regions for a small mesoscale sub-national study, and to take advantage of hourly measurements of CO_2, it is necessary to use a method in which the data can be disaggregated into the different spatial subregions and at a time step which is congruent with the scope of the project. For a high spatial resolution study, this requires emissions inferred at high temporal scale as well, and so diurnal information on emissions from different sources is required. As explained by earlier studies,[2] data related to the consumption of fuel are lacking at these high-resolution spatiotemporal scales. In South Africa, data related to fuel consumption at individual institutions or sales at individual stations are not publically available, and

therefore special arrangements need to be made, either with individual institutions or with the reporting agency, in order to access the data.

This paper describes the methodology implemented to disaggregate anthropogenic emissions of CO_2 for a small domain, but high spatial resolution, atmospheric inversion study conducted for the City of Cape Town. At the time of the 2011 Census, the population of Cape Town was 3 740 025.[7] A report of the energy usage of the city was compiled in 2011, which calculated the energy usage per sector of Cape Town, and calculated it to be 50% from transport, 18% from residential, 16% from commercial, 14% from industrial and 1% from government.[8] But of the carbon emissions, only 27% were attributed to the transport sector as a result of the carbon intensive usage of coal for electricity generation to provide almost all of the energy to the residential and commercial sectors in South Africa. The residential and commercial sectors emit approximately 29% and 28%, respectively, of the total carbon emissions of Cape Town.

Koeberg, a nuclear power station near Cape Town and the only one in South Africa, provides 4.4% of the electricity requirements of the country.[9] It feeds electricity directly into the grid, and therefore the reduction in carbon emissions resulting from nuclear power production benefits the whole country, not just to the City of Cape Town. Carbon emitted as a result of electricity generation for the city physically, occurs where the coal-fired power stations are located, which is mainly in the northeastern parts of South Africa. In this study, we are concerned with the location of the emission sources, rather than the location of the owner of the emissions. Therefore, emissions resulting from electricity generation are all attributed to power stations where these emissions are occurring in space. We assume all emissions from the commercial sector are the result of electricity generation and so are accounted for in the power station fuel usage information and thus we do not consider the commercial sector separately.

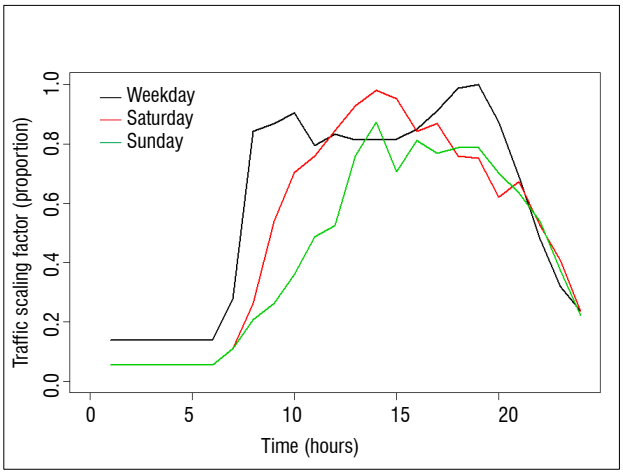
Methodology

Road transport vehicle emissions

A model describing the number of vehicles and vehicle kilometres travelled in an hour during peak hours on each section of road was obtained from an emissions report for the City of Cape Town,[10] and is based on vehicle count data. No public information is available on vehicle composition for the city, so an average emission factor was used, calculated from available vehicle types supplied by the Greenhouse Gas Protocol guidelines for emissions calculations,[11] based on the US Environmental Protection Agency's (EPA) published values, and the Department for Environment, Food and Rural Affairs' (Defra) guidelines,[12] which therefore assume an equal distribution of vehicles in each vehicle category. The average emission factor calculated was 347.01 g of CO_2 per vehicle kilometre, with a standard deviation of 239.64 g of CO_2 per km.

This emission factor converts vehicle kilometres into CO_2 emissions. The total number of vehicle kilometres travelled in a particular pixel was calculated by rasterising the line object data from the supplied shape file on vehicle kilometres provided by the City of Cape Town, so that the sum of vehicle kilometres over all lines were equal to the sum of vehicle kilometres over all pixels. The proportion of vehicle kilometres allocated to a pixel was the same as the proportion of the length of the line which occurred inside the pixel. This was performed using the *rgeos*[13] and *raster*[14] packages in R (a statistical software package).

In addition to the model describing the distribution of vehicle kilometres on Cape Town roads during peak hours, scaling factors were also provided to describe the traffic intensity at different times of the day, both over weekdays and weekends. These hourly scaling factors were used to transform the peak time weekday vehicle kilometres to match with a particular day of the week and time, so that a spatially explicit time series could be created with an hourly time step from Monday through to Sunday (Figure1).



Source: This information was obtained from the City of Cape Town city planners, estimated from vehicle count information.

Figure 1: Plot of the scaling factors used to convert peak hour traffic hours on a weekday into traffic hours at different times of the day during the week, on Saturdays and on Sundays.

Domestic emissions

To obtain the emissions from domestic fossil fuel burning for lighting, cooking and heating, data on the number of households from the 2011 census and data on the amount of residential fuel usage from the 2006 and 2009 Energy Digests[15,16] were used. The average amount of fuel usage per household was obtained by dividing the total fuel usage across the whole country by the number of households reported in the 2011 South African census (14 450 161). The average amount of fuel used per household was multiplied by the number of households in each pixel, and this value was scaled according to the proportion of fuel used for cooking, lighting and heating, where 75% of the annual heating fuel usage was assumed to take place during the winter months (March to August). It was assumed that 75% of the annual energy consumed was used for heating, 20% for cooking and 5% for lighting.

The population of Cape Town was subdivided into the different wards of the city, and these data were recorded as a shape file containing polygons of the wards and the associated population and household count. Using a similar method as for the vehicle line data, the polygon information was rasterised into pixel data so that the sum of the household counts over all the wards equalled the sum over all the pixels. The proportion of the household count that was assigned to each pixel was determined by the proportion of the polygon area located inside the pixel. This method can be extended to accommodate socio-economic data about the different wards used in our study, where the emissions from residential fossil fuel burning can be allocated based on income levels or electricity consumption because more affluent households will largely depend on electricity for heating, lighting and cooking. In order to better understand the discrepancies in household fuel usage, we have left this aspect for a future study.

Industrial emissions from listed activities

Under South Africa's Air Quality Act, industry must obtain an emissions licence to perform certain listed activities, and reporting of activity data for the purpose of emissions calculations is mandatory.[17] The dominant industrial activities listed include ceramic processes, hydrocarbon refining processes, iron and steel processes, Macadam processes for asphalt production, and waste incineration processes, as well as electricity generation at gas turbine power plants.[18] The initial approach was based on the methodology from Gurney[2] for point source industrial emissions in the USA, where CO emissions were converted into CO_2 emissions based on the ratio between CO and CO_2 emissions factors for that industry. Because of the coarseness of the reported CO emissions, we were unable to break the emissions down into different processes for which CO could be converted into CO_2 emission using the industry and process specific emission factors.

As an alternative, the reported fuel usage data for the top fuel users were converted directly into CO_2 emissions by multiplying this fuel usage date with the Defra greenhouse gas emission factors.[12] The fuel types that were considered included heavy fuel oil, coal, diesel, paraffin and fuel gas which were divided into liquid petroleum gas and refinery fuel gas. In the case of gas fuels, which were recorded in units of Nm^3, the fuel usage was first converted into kWh, and then into CO_2 emissions, where the calorific values were obtained from Rayaprolu[19] so that Defra emission factors could be used. The point data were then aggregated into the required raster format through summing the emissions from each source in a pixel. This analysis only took into account emissions from fuel combustion, not process related emissions.

Airport and harbour emissions

Emissions from aircraft are normally separated into the landing/takeoff cycle (LTO) and the cruise phase of the flight. The LTO part of flights at the Cape Town International Airport is allocated to the city's emissions, and divided evenly over the area which covers the airport. The Airports Company South Africa provides data on the number of aircraft movements, separated into domestic and international flights, for each month.[20] The IPCC guidelines provide emission factors per LTO,[21] separated into domestic and international flights. These emission factors were used to convert the monthly count of aircraft movements into CO_2 emissions.

The monthly emission was then divided equally between days, but emissions were only allocated to the hours between 06:00 and 22:00, when most of the aircraft activity takes place at the airport. The average emission factor for the domestic fleet was reported to be 2680 kg of CO_2 per LTO, and the average emission factor for the international fleet was 7900 kg of CO_2 per LTO.

The National Ports Authority of South Africa[22] publishes statistics on the harbour activity at the Cape Town Port on a monthly basis. The UK's Defra published a report on 2010 shipping emissions in the UK[23] and this report was used as a guideline to obtain estimates of the average amount of time spent by a particular vessel type in port, the average power of the main engine (ME) and auxiliary engine (AE) of each vessel type, and the emission factors for each vessel type while at berth and performing manoeuvring activities in port. The guidelines provided the equation to convert the gross tonnage of a vessel into total ME power:

$$ME\ power = 6.608 \times gross\ tonnage^{0.7033} \quad \text{Equation 2}$$

as well as the estimated proportion of power of the AE relative to the ME for a particular vessel type. The emission formula used is:

$$E_{in\ port} = T[(ME \times LF_{ME}) \times EF_{ME} + (AE \times LF_{AE}) \times EF_{AE} \quad \text{Equation 3}$$

where E is the emission, T is the time in hours spent in port, ME is the power of the main engine, AE the power of the auxiliary engine, LF is the loading factor for a particular engine and EF is the emissions factor for a particular engine. At berth and while manoeuvring, vessels are expected to operate at 20% of the maximum continuous rating for main engine operation and at 45% of the maximum continuous rating of the auxiliary engine (Table 1).

As for airport emissions, the monthly estimates were divided between each day and each hour of the month, but no assumption was made regarding when the activity took place, so emissions were allocated to all hours of the day.

The monthly emission values from the airport or harbour were allocated to a polygon describing the shape of the airport or harbour respectively. The polygons were then rasterised using the same grid as for the previous emission fields. To obtain hourly emission estimates, the total emissions for the month were divided by the number of days in the month, and then divided by 24 to get the hourly emission value. In the case of airport emissions, the daily emissions were divided by 16 instead, because it was assumed that the bulk of aircraft activity took place from 06:00 until 22:00.

Uncertainty analysis

Uncertainty estimates are required, not only to show the reliability of the estimates, but the inverse modelling approach requires prior flux estimates as well as prior error estimates for the covariance matrix of the estimated fluxes.

As no information was available on the error of the vehicle counts model, the nature of the data was used to obtain the estimate of uncertainty. The model provides the mean number of vehicle kilometres over a unit distance, and therefore it is likely that the data will follow a Poisson process, which implies that the variance of the estimate should be equal to the mean, and therefore the standard deviation equal to the square root of the mean. The CO_2 emission is the product of this count in vehicle kilometres and the average CO_2 emission factor, which has a standard deviation of 239.64 g of CO_2 per kilometre. From error propagation laws, the error in the CO_2 emission estimate will then be:

$$\delta CO_{2vehicle} = |CO_2| \times \sqrt{\left(\frac{\delta CO_2\ emission\ factor}{CO_2\ emission\ factor}\right)^2 + \left(\frac{\delta\ vehicle\ km\ count}{vehicle\ km\ count}\right)^2} \quad \text{Equation 4}$$

Table 1: Estimates, per vessel type, of fuel types, the average ratio of the power output of the auxillary engine (AE) to the main engine (ME), the average number of hours spent in port, and the CO_2 emission factors for the ME and AE

Engine type	ME fuel type	AE fuel type	Average ratio of AE/ME power per vessel	Average number of hours in port	ME CO_2 emission factor (kg CO_2 per kWh)	AE CO_2 emission factor (kg CO_2 per kWh)
Bulk carrier	RO	MGO	0.21	71.77	0.822	0.710
Container ship	RO	RO	0.22	26.50	0.822	0.745
General cargo	RO	MGO	0.33	40.01	0.822	0.710
Passenger	MDO	MDO	0.35	3.87	0.782	0.710
Ro-Ro cargo	RO	RO	0.30	14.60	0.822	0.745
Tanker	RO	MGO	0.27	35.86	0.822	0.710
Fishing	MGO	MGO	0.64	65.31	0.782	0.710
Others	MGO	MGO	0.29	53.53	0.782	0.710

RO, residual oil; MGO, marine gas oil; MDO, marine diesel oil

According to the Statistical Release of the 2011 South African census, the omission rate for the census questionnaire was approximately 15%.[7] The average fuel usage data per household was calculated by dividing the total annual amount of fuel sold by the number of households. No data were available for the difference in fuel usage between households. Therefore, to account for the missing source of uncertainty, the omission rate was elevated to 30%, double that of the omission rate, and this used as the estimate of the uncertainty in domestic emissions.

A report was published for the UK on the treatment of uncertainties of greenhouse gas emissions,[24] which provided estimates of activity data error and emission factor error under each fuel type for industrial sources. As the CO_2 emission at a particular point source is calculated as:

$$CO_2 \text{ emission} = \text{fuel usage} \times CO_2 \text{ emission factor} \qquad \text{Equation 5}$$

Error propagation techniques can be used to determine the uncertainty of the final estimate as:

$$\delta CO_{2\,\text{industry}} = |CO_2| \times \sqrt{\left(\frac{\delta \text{ activity data}}{\text{fuel usage}}\right)^2 + \left(\frac{\delta CO_2 \text{ emission factor}}{CO_2 \text{ emission factor}}\right)^2} \qquad \text{Equation 6}$$

where δ is the uncertainty value. The uncertainties provided are expressed as proportions of the amount of fuel use and of the size of the emission factor, therefore the uncertainty for the final CO_2 emission can be simplified to:

$$\delta\, CO_{2\,\text{industry}} = |CO_2| \times \sqrt{(p_{\text{activity data}})^2 + (p_{CO2\ \text{emission factor}})^2} \qquad \text{Equation 7}$$

where $p_{\text{activity data}}$ and $p_{CO2\ \text{emission factor}}$ are the proportions of error assumed for the reported fuel usage data and for the emission factor respectively for a given fuel type.

The aircraft count data are assumed to be without error, and therefore the error will be contained in the emission factors. The standard deviations of the emission factors of individual aircraft used to calculate the average emission factor for the domestic and international fleet were used to determine the uncertainty of the aircraft emissions. This was found to be 34% of the mean emission factor for the domestic fleet and 28% for the international fleet.[21]

As for the aircraft data, the counts of different ships in the harbour are assumed to be correct. Therefore, the error is assumed to lie in the emission factors for the different vessel types. From the Defra UK shipping inventory guide,[23] the assumed errors for berth and manoeuvring activities in the port are 20% and 30%, respectively. Therefore, to ensure a conservative estimate, the error is assumed to be 30% of the estimate.

Total emissions

Once the layers for each emission source type are obtained, the total emissions for a particular hour or any particular period can be obtained by summing the raster layers, where the appropriate scalar manipulations have been performed, such as multiplying the vehicle emission layer by the appropriate scaling factor for the day of the week and time of day. In order to be able to obtain the error estimates for each pixel, the uncertainties need to be expressed as variances instead of standard deviations, and then the variances for each of the source emission estimates in a pixel can be summed to obtain the total variance, which can then be converted back into a standard deviation by taking the square root of the variance.

Fossil fuel product comparison

To determine if the emissions estimated in this study are reasonable, the emissions for a weekday in March 2012 were compared with the EDGAR product. The spatial distributions of the emissions were mapped for each of the products, and the total emissions for the domain of Cape Town were calculated and compared between the two products.

Results and discussion

Road transport vehicle emissions

The rasterised vehicle emissions during peak hour traffic showed the concentration of emissions around the city centre of Cape Town and over the highway routes leading into the city from the suburban areas. Using the equation for uncertainty in the emission estimates, the pixel with the largest emission estimate of 19.74 Mg CO_2 per hour had an error estimate of 16.45 Mg CO_2 per hour. The error estimate was 83% of the emission estimate. The error in the vehicle emissions was expected to be large as there was a great deal of uncertainty in the average emission factors, with factors ranging from 100.1 g to 1034.6 g of CO_2 per kilometre.

Domestic emissions

Residential emissions from domestic fuel use were distributed over the suburban areas around Cape Town, as expected (Figures 3 and 4). Owing to the assumption that more fuel was used for heating purposes in the winter months, the maximum levels of emissions during summer were approximately half of what was consumed during the winter months (Table 2). The largest emission estimate for a pixel was 27.7 Mg CO_2 per hour during the winter months. The error of the estimate, using the assumed 30% error rate, was 8.3 Mg CO_2 per hour.

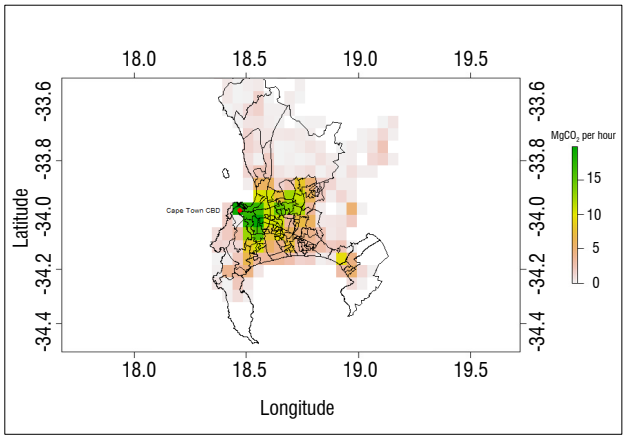


Figure 2: Map of rasterised CO_2 vehicle emissions during peak hour traffic, produced from vehicle kilometre hours and an average emission factor.

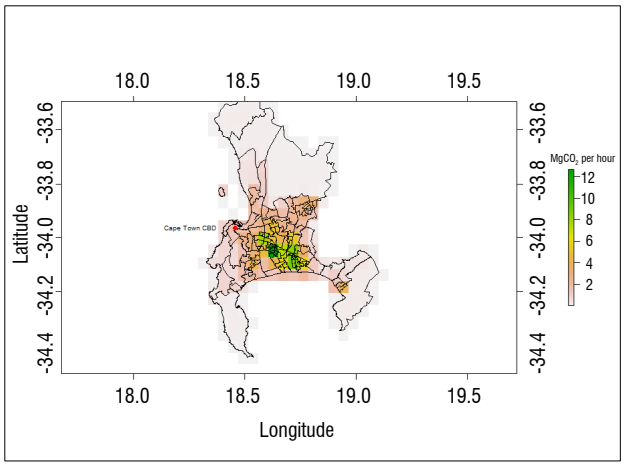


Figure 3: Map of rasterised CO_2 emissions from domestic fuel use during summer.

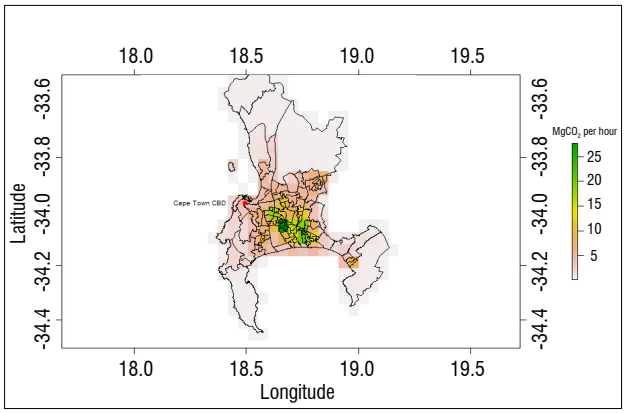


Figure 4: Map of rasterised CO_2 emissions from domestic fuel use during winter.

Industrial emissions from scheduled activities

The emissions resulting from industrial processes are distributed slightly away from the CBD towards the outskirts of the city centre, spreading outwards towards the city boundaries (Figure 5). The largest per pixel emission was 57 Mg CO_2 per hour, with an error value of 9 Mg CO_2 per hour, which is 16% of the mean value (Figure 6).

The advantage of obtaining the CO_2 emissions by using the fuel data compared to converting the CO emissions for different industries is that the error estimates are much smaller, as the fuel data have a much smaller associated error than estimated CO emissions for a particular industry. The disadvantage of the fuel data approach is that emissions from process-related activities are ignored, but would be included if the total CO emissions were converted into CO_2 emissions. Both the fuel data and the CO emission data are difficult to access and rely on accurate reporting from the industrial firms.

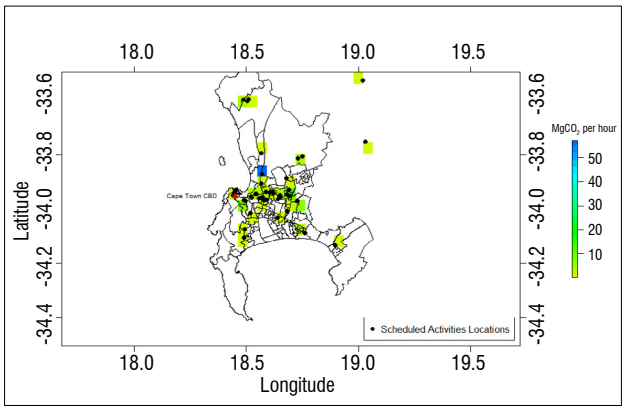


Figure 5: Map of rasterised CO_2 emissions from industrial point sources.

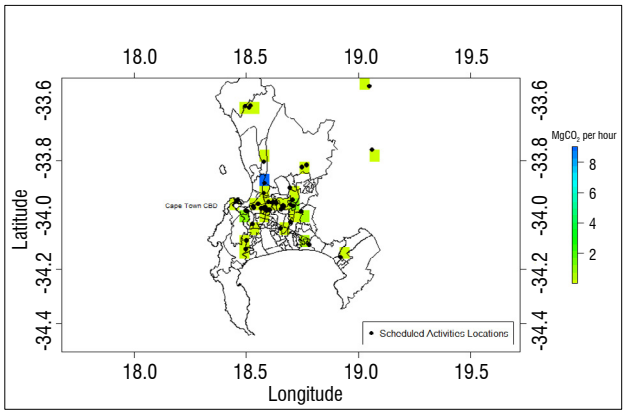


Figure 6: Uncertainties in the CO_2 emissions from industrial point sources.

Airport and harbour emissions

The emissions from aircraft at the airport are on average 10 890 Mg of CO_2 a month, with higher emissions during November to January when air traffic increases into the city (Figure 7). The average hourly emission from aircraft at the airport is 15.1 Mg of CO_2. This analysis does not include the emissions from point sources or ground vehicles at the airport and will require a count of each aircraft type arriving at the airport. The Defra guidelines for aircraft emissions supply the average amount of time each ground unit spends in operation per LTO cycle for a particular aircraft type, and these estimates could be used to determine emissions from ground vehicles at the airport.

The total emissions from shipping vessels in the Cape Town harbour, at berth and during manoeuvring procedures, are on average 4171.6 Mg of CO_2 per month (Figure 8), which is approximately 5.8 Mg of CO_2 per hour. Emissions take a dip during the mid-winter months, when the seas around Cape Town are particularly rough and storms are prevalent.

Total emissions

To demonstrate the aggregation of the different source emission layers, the total emissions estimated for a weekday in March 2012 at 18:00 were obtained. The industrial emissions and power station emissions were assumed to take place at all times, so the same layer would be used regardless of which hour was of interest. March falls into the winter period, so the winter domestic emission layer was used, and because it was a weekday at 18:00, no scaling was necessary for the vehicle emissions layer. The airport and harbour emissions for March 2012 were used. The monthly value was divided equally between all hours of the month for harbour emissions and divided evenly between all days and the hours from 06:00 to 18:00 in the case of the airport emissions. These layers were summed and the resulting emission layer was mapped (Figure 9).

Table 2: Emission factors and fuel usage figures used to estimate the amount of CO_2 emissions per household[25]

	LPG	Paraffin	Wood	Coal
Emission factor[25]	1530.1 kg CO_2 per kL	2531.9 kg CO_2 per kL	77.38 kg of CO_2 per Mg	2448.7 kg of CO_2 per Mg
Total amount used in summer per household	4.758×10^{-3} kL	9.683×10^{-3} kL	2.465×10^{-1} Mg	3.278×10^{-1} Mg
Summer emission of CO2 per household per hour	3.929×10^{-2} kg of CO_2 per hour	1.347×10^{-1} kg of CO_2 per hour	1.048×10^{-1} kg of CO_2 per hour	4.410 kg of CO_2 per hour
Total amount used in winter per household	1.047×10^{-2} kL	2.130×10^{-2} kL	5.424×10^{-1} Mg	7.212×10^{-1} Mg
Winter emission of CO2 per household per hour	8.551×10^{-2} kg of CO_2 per hour	2.931×10^{-1} kg of CO_2 per hour	2.281×10^{-1} kg of CO_2 per hour	9.598 kg of CO_2 per hour

Source: Total fuel usage figures were obtained from the 2009 Energy Digest published by the South African Department of Energy[16] and estimates of households in South Africa and in the Cape Town area were obtained from the 2011 South African Census[7].

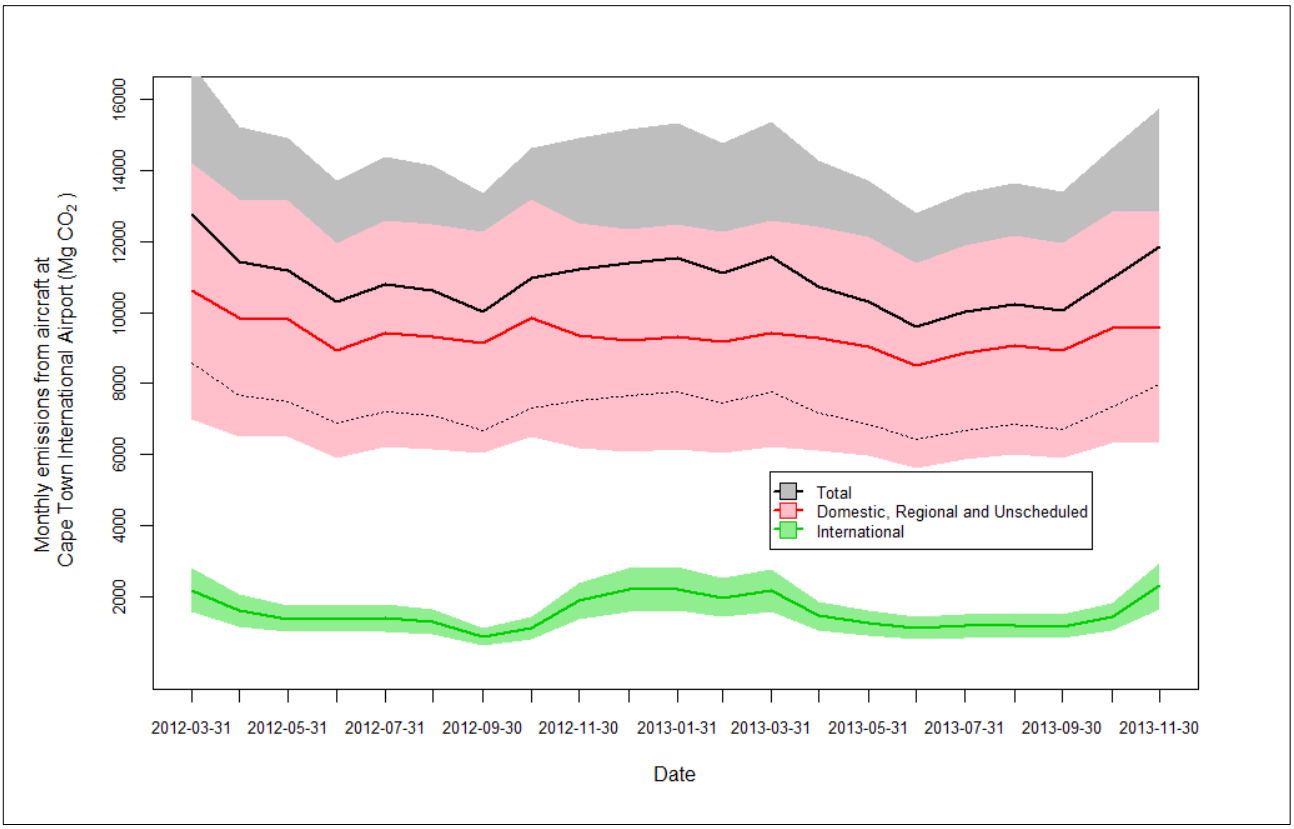
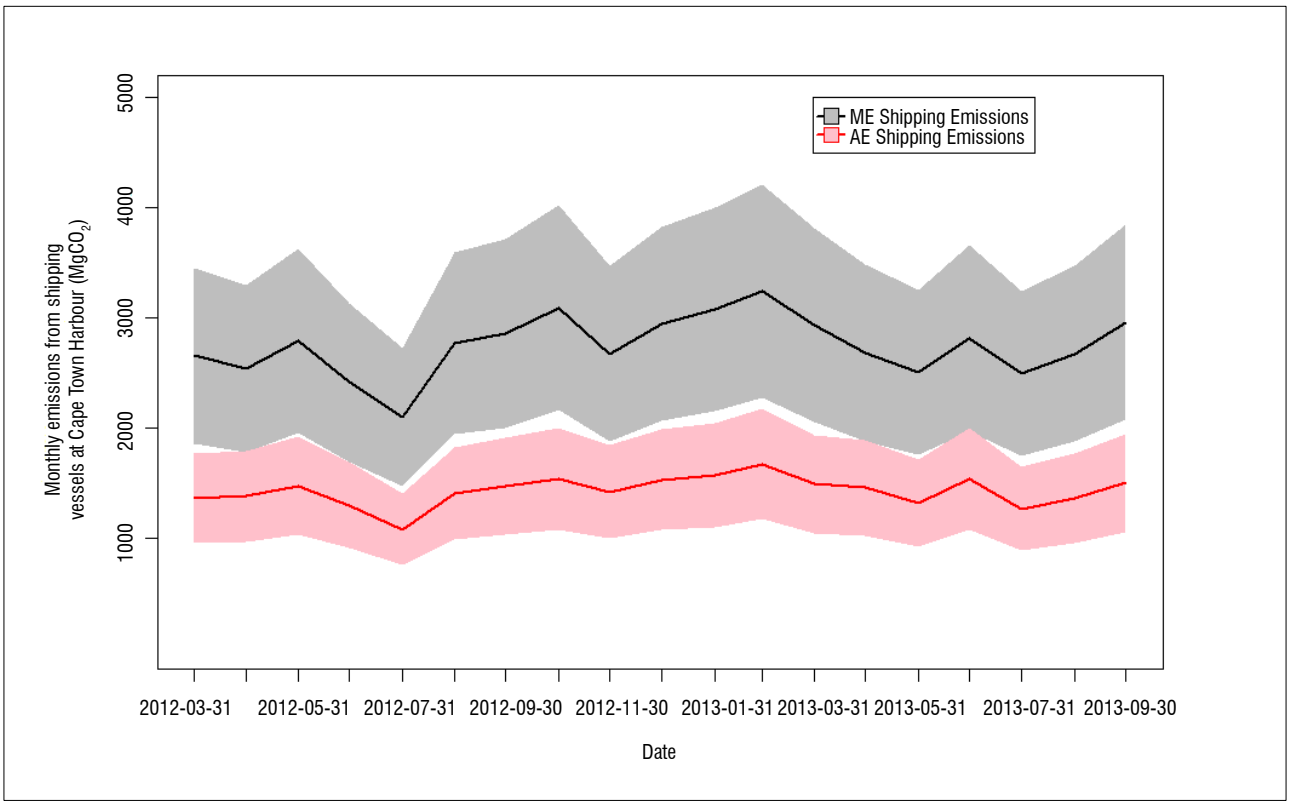


Figure 7: Monthly emissions from aircraft at Cape Town International Airport. The shaded region represents the level of uncertainty and the dashed line represents the lower limit of the total emissions from aircraft at the airport.



ME, main engine; AE, auxiliary engine

Figure 8: Monthly emissions from shipping vessels at Cape Town harbour. The shaded region represents the region of uncertainty in the emission estimates.

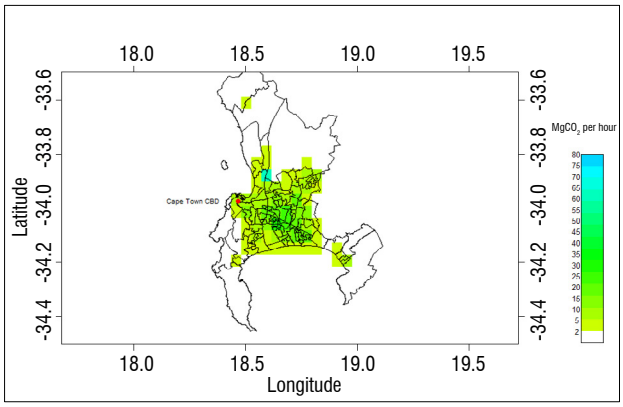


Figure 9: Aggregated CO_2 emissions from 18:00 to 19:00 during a weekday in March 2012.

Fossil fuel product comparison

A comparison with the EDGAR 0.1°×0.1° product over Cape Town revealed a similar allocation of CO_2 emissions for different areas, but with the EDGAR product providing a more smoothed product (Figure 10). The total emissions calculated in this study were strongly influenced by the large point source industrial emitters. The total emission of CO_2 during a weekday in March over the full spatial domain of Cape Town was 9 252 883 kt. Calculated from the EDGAR data, which is provided in kg CO_2 per m^2 per second, the total emission for the same area would be 7 574 559 kt. The EDGAR data are available until 2010, therefore there is a 2-year lag between the two products compared. The EDGAR data provided are also an average value for the year, and therefore it is not surprising that our study estimated a 22% higher CO_2 emission than that considered for a typical weekday in March 2012.

Discussion and conclusions

The results for the City of Cape Town show that through the use of publicly available data and reported data on activity and population levels around the city, it is possible to obtain a spatially and temporally explicit inventory of emissions. The most challenging of the sectors was the industrial sector, for which data at the required resolution are not always available in adequate detail. These fossil fuel emission estimates are essential to run an atmospheric inversion for the city to obtain improved estimates of the total CO_2 fluxes occurring in and around the city.

These estimates can be improved by obtaining detailed fleet data for vehicle, aircraft and shipping vessel movements, as the emission factors differ significantly among different fleet types. Domestic emissions estimates can be improved by using Cape Town specific surveys on the average fuel use quantities, and distinguishing these surveys between different households depending on domicile type, as those homes with better amenities are less likely to rely on solid and liquid fuels for cooking, heating and lighting. The estimates of the emissions from the power stations and industrial sources could be improved if more detailed fuel usage and specific process data were available for each facility included in the assessment. It may be useful for South Africa to publish a similar document to the one Defra publishes for the UK, with South Africa specific emission factors, to increase accuracy and so that emission estimates by different professionals can be standardised for South Africa. A comparison with a global benchmark emissions product revealed that the proposed method, making use of the currently available usage data, provides reasonable estimates of CO_2 emissions, which would be further improved if South Africa specific emission factors were used.

The results obtained through this process will provide important inputs required for an atmospheric inversion study, relying on observations from a network of CO_2 measurement equipment placed around the city. A similar approach as described in this paper will be used to disaggregate the national emissions, to provide coarser estimates of CO_2 emission from fossil fuel combustion, which will then be used in a national inversion study. The best placement of new measurement sites for the observation of CO_2 sources and sinks at a national level has already been determined through an optimal network design.[26] Improving the knowledge of the South African CO_2 budget will help to reduce the uncertainty of the global estimates of sources and sinks, as southern Africa is usually a large source of error in global inversions, because of Africa's general under-sampling of greenhouse gas concentration measurements.[27]

Acknowledgements

This work was supported by parliamentary grant funding from the Council of Scientific and Industrial Research, South Africa, project number JECOS81. We thank Karen Small of the Strategic Development Information and Geographic Information System Department of the City of Cape Town for providing the Western Cape census shape files, and Sarah Ward of the City of Cape Town for providing access to emission data collected by the City of Cape Town.

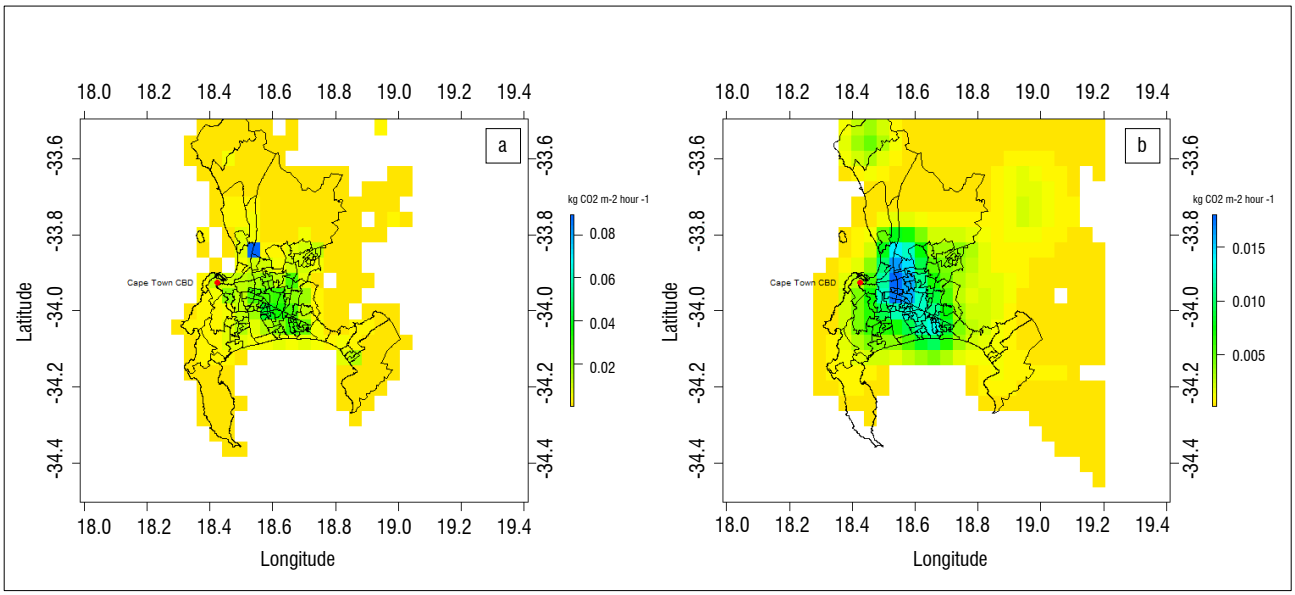


Figure 10: Comparison of (a) emissions estimated from the spatial and temporal disaggregation approach and (b) EDGAR emissions for the year 2010 over the City of Cape Town. Emissions are presented as kg CO_2 per m^2 per day.

Authors' contributions

A.N., the main author, was responsible for data acquisition, data processing, statistical analysis, model development and writing of the manuscript. R.J.S. provided scientific advice and E.F. was involved with data provision and processing and provided scientific advice.

References

1. Eggleston HS, Buendia L, Miwa K, Ngara T, Tanabe K, editors. IPCC Guidelines for National Greenhouse Gas Inventories. Tokyo: IGES; 2008.
2. Gurney KR, Mendoza DL, Zhou Y, Fischer ML, Miller CC, Geethakumar S, et al. High resolution fossil fuel combustion CO_2 emission fluxes for the United States. Environ Sci Technol. 2009;43(14):5535–5541. http://dx.doi.org/10.1021/es900806c
3. Janssens-Maenhout G, Pagliari V, Guizzardi D, Muntean M. Global emission inventories in the Emission Database for Global Atmospheric Research (EDGAR) – Manual (I). Gridding: EDGAR emissions distribution on global gridmaps. Luxembourg: European Union; 2012. p. 33.
4. Rayner PJ, Raupach MR, Paget M, Peylin P, Koffi E. A new global gridded data set of CO_2 emissions from fossil fuel combustion: Methodology and evaluation. J Geophys Res. 2010;115:D19306. http://dx.doi.org/10.1029/2009JD013439
5. Asefi-Najafabady S, Rayner PJ, Gurney KR, McRobert A, Song Y, Coltin K, et al. A multiyear, global gridded fossil fuel CO_2 emission data product: Evaluation and analysis of results. J Geophys Res Atmos. 2014;119:10213–10231. http://dx.doi.org/10.1002/2013JD021296
6. Gurney KR, Razlivanov I, Song Y, Zhou Y, Benes B, Abdul-Massih M. Quantification of fossil fuel CO_2 emissions on building/street scale for a large U.S. city. Environ Sci Technol. 2012;46(21):12194–12202. http://dx.doi.org/10.1021/es3011282
7. Statistics South Africa. Census 2011 statistical release – P0301.4. Pretoria: Statistics South Africa; 2011.
8. City of Cape Town. State of energy and energy futures report. Cape Town: City of Cape Town; 2011. Available from: http://www.capetown.gov.za/en/EnvironmentalResourceManagement/publications/Documents/State_of_Energy_+_Energy_Futures_Report_2011_revised_2012-01.pdf
9. Eskom. Generating electricity at a nuclear power station: Fact sheet revision 8 [document on the Internet]. c2013 [cited 2014 Jan 16]. Available from: http://www.eskom.co.za/AboutElectricity/FactsFigures/Documents/NU_0001NuclearEnergyBasicCycleRev6.pdf
10. Cambridge Environmental Research Consultants. Compilation of emissions inventory and preliminary air quality monitoring for Cape Town. Report no. FM865/R5/12. Cape Town: City of Cape Town; 2012.
11. Greenhouse Gas Protocol. Calculating CO_2 emissions from mobile sources [homepage on the Internet]. c2005 [cited 2014 Jan 16] Available from: http://www.ghgprotocol.org/files/ghgp/tools/co2-mobile.pdf
12. UK Department for Environment, Food and Rural Affairs (Defra). Government GHG conversion factors for company reporting: Methodology paper for emission factors. London: Crown; 2013. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/224437/pb13988-emission-factor-methodology-130719.pdf
13. Bivand R, Rundel C. Rgeos: Interface to Geometry Engine – Open Source (GEOS). R package version 0.3-2 [program on the Internet]. c2013 [cited 2014 Aug 28]. Available from: http://CRAN.R-project.org/package=rgeos
14. Hijmans RJ. Raster: Geographic data analysis and modelling. R package version 2.1-49 [program on the Internet]. c2013 [cited 2014 Aug 28]. Available from: http://CRAN.R-project.org/package=raster
15. South African Department of Minerals and Energy. Digest of South African energy statistics. Pretoria: Department of Minerals and Energy; 2006. Available from: http://www.energy.gov.za/files/media/explained/2006%20Digest%20PDF%20version.pdf
16. South African Department of Energy. Digest of South African energy statistics. Pretoria: Department of Energy; 2009. Available from: http://www.energy.gov.za/files/media/explained/2009%20Digest%20PDF%20version.pdf
17. National Environment Management: Air Quality Act, No. 39 of 2004. Government Gazette. 2005 Feb 24; vol. 476, no. 27318.
18. Eskom. Ankerling and Gourikwa gas turbine power stations fact sheet: Revision 6. Johannesburg: Generation Communication; 2013. Available from: http://www.eskom.co.za/AboutElectricity/FactsFigures/Pages/Facts_Figures.aspx
19. Rayaprolu K. Boilers: A practical reference. New York: Taylor and Francis; 2013. p. 181–183.
20. Airports Company South Africa (ACSA). Aircraft statistics [document on the Internet]. No date [cited 2014 Jan 13]. Available from: http://www.acsa.co.za/airports/cape-town-international/statistics/aircraft
21. Intergovernmental Panel on Climate Change (IPCC). Good practice guidance and uncertainty management in national greenhouse gas inventories. Montreal: IPCC; 2000. p. 93–102. Available from: http://www.ipcc-nggip.iges.or.jp/public/gp/english/
22. Transnet National Ports Authority, South Africa. Summary of cargo handled at ports of South Africa [homepage on the Internet]. c2010 [cited 2014 Jan 16]. Available from: http://www.transnetnationalportsauthority.net/DoingBusinesswithUs/Pages/Port-Statistics.aspx
23. UK Department for Environment, Food and Rural Affairs (Defra). UK ship emissions inventory. Final report. London: Crown; 2010. Available from: http://uk-air.defra.gov.uk/assets/documents/reports/cat15/1012131459_21897_Final_Report_291110.pdf
24. UK Department for Environment, Food and Rural Affairs (Defra). Treatment of uncertainties for national estimates of greenhouse gas emissions [homepage on the Internet]. c2013 [cited 2014 Mar 25]. Available from: http://uk-air.defra.gov.uk/reports/empire/naei/ipcc/uncertainty/contents.html
25. UK Department for Environment, Food and Rural Affairs (Defra). 2012 Guidelines to Defra / DECC's GHG conversion factors for company reporting. PB 13773. London: Crown; 2012. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/69554/pb13773-ghg-conversion-factors-2012.pdf
26. Nickless A, Ziehn T, Rayner PJ, Scholes RJ, Engelbrecht F. Greenhouse gas network design using backward Lagrangian particle dispersion modelling – Part 2: Sensitivity analyses and the South African test case. Atmos Chem Phys. 2015:2051–2069. http://dx.doi.org/doi:10.5194/acp-15-2051-2015
27. Denman KL, Brasseur G, Chidthaisong A, Ciais P, Cox PM, Dickinson RE, et al. Couplings between changes in the climate system and biogeochemistry. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, et al., editors. Climate change 2007: The physical science basis. Contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2007. p. 499–587.

Wildflower harvesting on the Agulhas Plain, South Africa: Challenges in a fragmented industry

AUTHORS:
Thijs Blokker[1]
David Bek[2]
Tony Binns[1]

AFFILIATIONS:
[1]Department of Geography, University of Otago, Dunedin, New Zealand
[2]Centre for Business in Society, Coventry University, Coventry, United Kingdom

CORRESPONDENCE TO:
Thijs Blokker

EMAIL:
blokkerthijs@gmail.com

POSTAL ADDRESS:
Department of Geography, University of Otago, Leith Walk, Dunedin 9017, New Zealand

DATES:
Received: 11 May 2014
Revised: 12 Nov. 2014
Accepted: 23 Feb. 2015

KEYWORDS:
Cape Floristic Region; sustainability; poverty alleviation; conservation; forum; flowers

HOW TO CITE:
Blokker T, Bek D, Binns T. Wildflower harvesting on the Agulhas Plain, South Africa: Challenges in a fragmented industry. S Afr J Sci. 2015;111(11/12), Art. #2014-0160, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140160

South Africa's Agulhas Plain is home to the Cape Floristic Region (CFR), one of the richest floras in the world and the smallest of earth's six plant kingdoms. The indigenous fynbos flora is harvested from the wild and is both exported and sold locally. The conservation value of the CFR, and the need to address deeply entrenched socio-economic disparities and high poverty levels have set a challenging context for the wildflower harvesting industry. The strong competition which exists between producers has resulted in fragmentation of the industry and a breakdown in communication. Using data gathered from interviews and meetings with a range of stakeholders, we argue that the wildflower harvesting industry needs to cooperate and improve communication levels to address the challenges collectively. Without such a collective voice, the sustainability of the industry, the CFR and the livelihoods of disadvantaged communities will be affected. The establishment of a 'Wildflower Harvesting Forum' was explored as a possible solution and is recommended as a sustainable way forward.

Introduction

The Cape Floristic Region (CFR), or Cape Floral Kingdom, located mainly in South Africa's Western Cape Province, is one of the richest biodiversity hotspots in the world (Figure 1). It is designated as one of earth's six plant kingdoms, but is the only one found within the boundaries of a single country.[1-3] However, the CFR is extremely vulnerable, and is therefore of 'high conservational priority' both within South Africa and also globally.[1,4] Four-fifths of the CFR comprises fynbos (Afrikaans for 'fine bush'), a type of vegetation of which some species, because of their durability, beauty and uniqueness, have proven popular as a harvested product, especially for the main export destinations of the UK and Europe. At the southern tip of Africa lies the Agulhas Plain where the wildflower harvesting industry constitutes an important aspect of the local economy, providing for the livelihoods of farmers, poor households and communities in areas where unemployment is as high as 80%.[4,5]

The wildflower harvesting industry on the Agulhas Plain faces complex pressures to both preserve the conservation value of the CFR, whilst also providing a much-needed means to uplift local livelihoods. A solution promoted by a local non-governmental organisation, the Flower Valley Conservation Trust (FVCT), has been to harvest fynbos in a sustainable way so that the conservation value of the CFR is not jeopardised, while at the same time allowing socio-economic gains to be made. Such practices have been promoted by the FVCT's 'Sustainable Harvesting Programme', which has received support from CapeNature's licensing system. The key premise of this programme is to promote both economic and environmental sustainability primarily through a pragmatic code of conduct and by adhering to this, gain access to the niche bouquet market in the UK, primarily Marks and Spencer. However, the industry remains loosely regulated, and there remains much work to be done before the full environmental and socio-economic benefits of sustainable harvesting are achieved. A sustainable harvesting supply chain now exists, which consists of harvesters who have signed up to the programme and have thus been able to access valuable markets, such as Marks and Spencer in the UK, which supports the tenets of the programme.[6] However, a sizeable mainstream industry continues to operate, supplying an array of less regulated markets.

To further complicate the situation, additional challenges facing the industry include a raft of environmental issues such as wildfires, climate change and invasive alien vegetation. A further significant challenge is the highly competitive nature of the industry. This competition exists at both a local and global level. Internationally, the industry is enmeshed within global value chains (GVCs), which has resulted in discrepancies between beneficiaries along the GVC. Discrepancies also exist within the local market in South Africa, and affect local competition, for which pressure within the GVC ultimately reduces prices for producers, whilst still requiring exacting product standards.

Largely as a result of the competitive pressures, the wildflower harvesting industry is very secretive and lacks effective communication. Industry stakeholders such as FVCT and the Protea Producers of South Africa (PPSA) have worked hard to improve interaction within the industry. Key stakeholders believe communication in the currently fragmented industry can be improved by establishing a forum, so that the full potential of enterprises can be unlocked through information sharing and collective bargaining. These benefits have the potential to ensure the future sustainability of the wildflower harvesting industry.

The paper is structured around the following research questions: (1) What are the key challenges facing the wildflower industry on the Agulhas Plain? (2) What institutional changes are needed to overcome these challenges? (3) What role might a wildflower harvesting forum play in furthering the interests of the industry? This paper is of interest to the scientific community as it illustrates the role that social environments play in influencing the implementation of scientifically driven conservation strategies.

The Cape Floristic Region and fynbos

The CFR covers an area of less than 90 000 m^2, of which four-fifths comprises the Cape fynbos (Figure 1).[3,7] With 8600 plant species, of which 5800 are endemic, the CFR is one of the richest ecosystems in the world.[1,2] To put the significance and uniqueness of the CFR into perspective, the rest of the African continent, some 235 times the size of the CFR, only contains three and a half times the number of species that are indigenous to the CFR.[7]

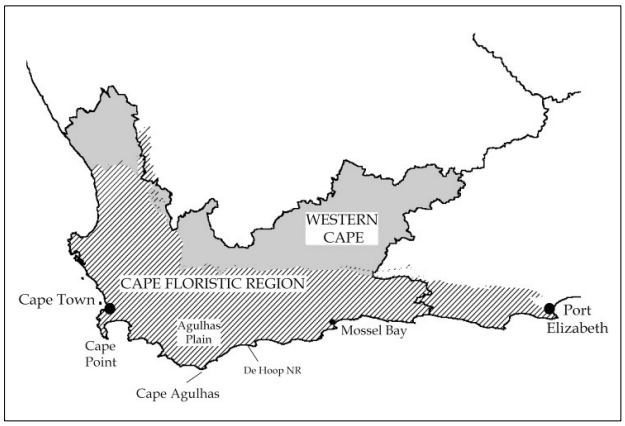


Figure 1: The Cape Floristic Region in relation to the Agulhas Plain, Western Cape Province, South Africa.[3]

The Agulhas Plain is situated at the southern tip of Africa, within the Western Cape Province (Figure 2). The Plain comprises 270 000 ha of land within the CFR,[2] and has remarkable plant diversity, with over 100 locally endemic vegetation types. As a result, the Plain is considered to be a high priority for conservation within South Africa, and indeed globally.[1]

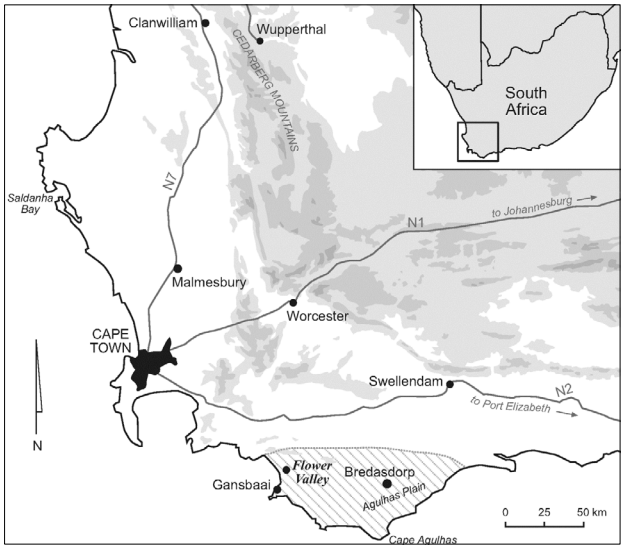


Figure 2: Map showing the location of the Agulhas Plain.[8]

Owing to South Africa's apartheid legacy, virtually all privately owned land on the Agulhas Plain is owned by white landowners. Those disadvantaged under apartheid mostly inhabit the rural settlements and small towns of the area. Many inhabitants are descendants of the KhoiKhoi people, but since the early 1990s, there has been a significant influx of mainly Xhosa people from the Eastern Cape Province.[4]

The rural areas of the Agulhas Plain are typical of rural South Africa in that they are characterised by marked spatial and social inequality. A noteworthy event occurred on 1 March 2013, when the minimum daily wage for agricultural labourers was increased by 51% from ZAR69 (USD 6.21) to ZAR105 (USD 9.46). This measure was implemented following protests that resulted in hundreds of arrests and at least three deaths.[9,10] The minimum wage increase received mixed reactions, with an understandably positive response from farmworkers, unions and charities.[10] However, many South Africans were sceptical of the possible wide-ranging negative effects, arguing that the increase is unsustainable, and will actually harm agricultural workers if rural enterprises are forced to undergo restructuring.

The wildflower harvesting industry has been in operation for over a century, with initially small amounts exported to Europe.[7] The first people credited with making a substantial living through exporting wildflowers were the people from Elim (just southwest of Bredasdorp) who, through the help of the German Moravian Church, began to export dried flowers to Germany in 1877. Since then the industry slowly developed into the multimillion rand operation it is today, with over 700 harvesters currently operating within the Western Cape.

Harvesters vary in size, from large-scale land lessees and large landowners to small-scale producers. In addition to wildflower harvesters, there are also a number of cultivators in operation. Although the cultivators were not the main focus of this research, they need to be mentioned as they are responsible for supplying the majority of the 'focal' flowers, which has meant that wildflower harvesters are increasingly focusing on fynbos 'greens'. Cape flora bouquets typically comprise a mixture of 'focal flowers' and 'greens'. Focal flowers are typically high value flowers such as proteas or pincushions, 'greens' are lower value stems which surround the focal flower. Some products are sold locally, but the majority are exported, largely to the UK and Europe, but also to the Middle East, Asia, Africa and the USA.[11]

Key stakeholders in the Agulhas Plain wildflower harvesting industry

Flower Valley Conservation Trust

Established in 1999, the FVCT focuses on utilising wild flowers commercially to promote both landscape preservation and livelihood development.[4] A number of factors have shaped how FVCT operates as an NGO today, most notably in 2004 when FVCT was split into two separate entities with the business of flower sourcing and sales assigned to a separate newly formed company, Fynsa, a separate legal entity initially operating from a pack shed on Flower Valley Farm (the birthplace of the initiative) before moving to Stanford. Meanwhile, FVCT focused on social and environmental concerns supported by donors.[6]

An agreement between the two ensured that only certified fynbos products were sourced. A link remains between the Flower Valley Farm operations and the Trust, but the finances are kept separate. This link is essential, as running a flower harvesting enterprise provides FVCT with critical insights into the challenges faced by the industry. FVCT has other important links, for example, with CapeNature, which was involved with the development of a sustainable harvesting code of practice, CapeFlora SA (known as the Protea Producers of South Africa until 2014) and the Sustainable Harvesting Committee, the latter representing the key stakeholders involved in the Sustainable Harvesting Programme.

The Sustainable Harvesting Programme is the centrepiece of FVCT's work. This programme is multi-faceted; its core objective is to ensure that wild harvesting adheres to practices that will not threaten the long-term health of the fynbos ecosystem. A Sustainable Harvesting Code of Practice provides guidelines for landowners and harvesting teams.[12] The central tenets of this Code are underpinned by FVCT's pioneering research which led to the production of two key documents – a Vulnerability Index of harvested species[13] and a Resource Base Assessment[14]. The former quantifies the extent to which individual species are at risk from extinction, whilst the latter enables landowners to assess their fynbos stocks and make informed judgements about harvesting levels.

CapeNature

CapeNature is a Western Cape provincial government organisation with the statutory responsibility for biodiversity conservation. CapeNature's vision is to establish a successful 'conservation economy' which will allow for the transformation of biodiversity conservation into a key component of local economic development.[15] CapeNature, therefore has a strong vested interest in the regulation of the wildflower harvesting industry. Harvesters, landowners and pack sheds require licences to engage in the wildflower industry. Licences are issued following an inspection to ensure that no rare or endangered species are harvested, that the resource base is sufficient, and that the *veld* (Afrikaans word for large open space or field) from which they are harvesting is the correct age. In this way, CapeNature has been a key player in the development of the Sustainable Harvesting Programme.

CapeFlora SA

CapeFlora SA is a non-profit organisation which was established as the PPSA in 2005 with the goal of:

> *identifying and addressing the strategic needs of the fynbos and protea industry with the vision to bring about a sustainable increase in the supply and demand for high value/high quality fynbos products to discerning international markets to the benefit of all role players in the South African fynbos industry.*[16]

At present, CapeFlora SA has about 90% representation from the cultivated industry, but only six or seven members from the wildflower harvesters.[16]

Situating the case study

People living in rural poverty have fewer assets, and are often very dependent on the natural resource base for their livelihoods.[17,18] This reliance has a number of associated risks for both those living in poverty and their surrounding environments. With strong linkages between socio-economic systems and ecological systems, conservation planning and poverty alleviation ideally need to be tackled as a single complex interacting system.[19] Although it has been apparent for a long time that the two issues need to be solved simultaneously, it was not really until the late 1970s that the idea of 'conservation with development' gained momentum. Indeed, the relationship between biodiversity conservation and poverty alleviation is still hotly debated in both policy forums and academia.[18,20-22] A number of sceptics believe that a win-win scenario is wishful thinking,[21,23,24] such that efforts to achieve both objectives might be seen as contradictory – a viewpoint that is frequently reflected in the characteristics and behaviour of stakeholders, power structures and policy jurisdiction, which underpin conservation and poverty alleviation.[21]

The debate has moved forward in recent years with proponents recommending a market-based approach to combined conservation and poverty alleviation programmes. Thus, sustainability is said to be more readily achieved through its three dimensions: environmental, economic and social.[25] Market-based approaches to the conservation/poverty alleviation nexus are increasingly mainstreamed by NGOs, governments, businesses and the research community.[26] However, as Fisher et al.[27] point out, despite the three pillars of sustainable development being seen as inseparable, the economic pillar has often tended to dominate. Evidence, however, suggests that in many instances, poor people benefit least from this approach to biodiversity, and the poor frequently end up bearing most of the costs.[28] Furthermore, Bell and Russell[29] question the potential of market-based mechanisms for environmental management, particularly in developing countries in which regulatory capacity is weak. This observation resonates somewhat with the 'tragedy of the commons' theory, whereby people attempt to maximise profits from a common resource when there is an economic benefit attached to the resource.[30]

In some cases, a market-based approach is associated with products which are part of global value chains (GVCs). According to Gereffi et al.[31(p.79)]:

> *Global value chain research and policy work examine the different ways in which global production and distributions systems are integrated, and the possibility for firms in developing countries to enhance their position in global markets.*

A well-documented and often critiqued aspect of GVCs relates to ethical trade issues, especially concerning labour and environmental standards.[32,33] The appalling labour conditions in some developing countries and environmental degradation have been widely publicised through the media, and in reports published by NGOs such as Action Aid, Oxfam and Greenpeace. The latter illustrate how vulnerable people in developing countries and their surrounding environments are exploited within GVCs. Although it is acknowledged that the employment created is important for livelihood improvements, the often precarious nature of the employment can reinforce societal inequalities.[34] It seems that those living in conditions of poverty will take up employment even under precarious conditions in an attempt to improve their future prospects and uplift their livelihoods. Sadly, it seems likely that exploitation will continue in developing countries whilst 'lead companies' gain more from trade under the current workings of GVCs.[34]

Gereffi et al.[31] argue that the power within the GVCs lies with the lead firms which are predominantly big enterprises, such as supermarket chains. In relation to labour conditions, Raworth[34] has noted that lead firms have the power to push costs and risks onto producers, who in turn pass them onto the weakest link in the chain, namely their employees. This puts stress on the end-line producers and employees who may have to work long hours to cater to these demands. In terms of the environment, there can be detrimental effects caused by overutilisation.

Private sector enterprises in developing countries, which make up the producer component of GVCs, are typically micro, small and medium enterprises. These enterprises in most instances contribute a significant share of employment and income opportunities to the surrounding communities within which they operate. Unfortunately, the full potential of these enterprises often remains unlocked as enterprises frequently operate in isolation, resulting in uncompetitive production patterns and ultimately an approach which is not conducive to innovation.[35]

Within the context of this investigation of wildflower harvesting, it was important to consider the impact of both formal and informal socialising processes and their effects on inter-organisational relationships. Through socialising, evidence suggests that goodwill, mutual trust and respect can develop, allowing for greater levels of cooperation and communication.[36,37] It is widely argued that communication is one of the most important elements in successful inter-firm exchange.[38] Within the highly competitive wildflower harvesting industry, Axelrod's[39] game theory seems relevant, as he observes that 'in situations where each individual has an incentive to be selfish, how can cooperation ever develop?' This observation accords with Hardin's 'tragedy of the commons', and the 'prisoner's dilemma' concept which highlights the importance of trust in building cooperative and communicative relationships.[38,40,41]

Challenges facing the Agulhas wildflower harvesting industry

In this study, we used a range of qualitative research methods during a 6-week period of field research in February and March 2013. Semi-structured interviews were undertaken with 19 key informants, who were selected based on their association with the wildflower harvesting industry and included small- to large-scale harvesters, pack shed workers, NGO representatives, government officials and botanists. There was also some participant observation, and a detailed analysis of a range of 'grey' literature which included FVCT annual reports, newspaper articles and government and NGO reports. Some of the main challenges currently facing the wildflower harvesting industry on the Agulhas Plain are discussed below.

Natural challenges of aliens and wildfire

Alien plant species are a serious problem facing the wildflower harvesting industry, with 40% of the Agulhas Plain infested to some degree.[1] Alien species displace the fynbos vegetation and exacerbate the existing water shortage problem, leading to a greater risk of wild fires. Fire is an ever-present challenge as it is a critical part of the fynbos life cycle. But if wildfires get out of control they can have devastating impacts, as was the case in 2006 when 47 000 ha of land burned, including one-third of the natural fynbos resource in the region.[42]

Thus far, both the threat of wildfire and aliens have largely been dealt with on a case by case basis, but ideally these challenges should be dealt with strategically on a long-term basis and over a large geographical area. Although some programmes have been implemented in the past to deal with aliens on both private and public land, none to date has been particularly successful, but it is hoped that the Agulhas Biodiversity Initiative Alien Clearing Programme, launched in May 2013, might prove to be more effective. Towards the end of 2012, through the Department of Environmental Affairs' Land User Incentive Scheme, ZAR18 million was awarded to the Agulhas Biodiversity Initiative to undertake alien clearing on private land within the Agulhas Plain.

Regulatory challenges

CapeNature is the main organisation involved with regulation; however, its capacity to undertake this role is limited. As Respondent 2 commented, '[CapeNature] are basically under-resourced and understaffed to deal with the processing of licences efficiently' and, as a result, opportunities for overharvesting, poaching and picking of illegal species are increased. Each of these issues has a key impact on the overall sustainability and viability of the industry.

Overharvesting, whereby insufficient seed stock is left to ensure reproductive replacement, has been found to be a serious threat to the sustainability of the industry and CFR. The largest contributing factors to overharvesting can be traced back to self-economic interests and a weak regulatory system. In an attempt to make the industry more sustainable, the Sustainable Harvesting Programme has been introduced, whilst CapeNature is also attempting to address inadequacies with the licensing system.

Poverty alleviation

The sustainable harvesting supply chain has undoubtedly improved local livelihoods by providing year-round employment rather than the part-time employment offered by many of the other industries operating on the Agulhas Plain. Every person interviewed reflected upon the importance of investing in the capacity building of staff. FVCT has a strong track record in this regard, having developed the Sustainable Harvesting Programme, the Agulhas Biodiversity Initiative Alien Clearing Programme, and the Field Monitor Programme.[12] Each of these programmes has significant capacity building and poverty alleviation components.[43] In addition, there are a number of other ways in which the capacity of staff has been enhanced. Respondent 7, for example, set up three small enterprises, each run independently, but under his mentorship. Each of these enterprises has its own *bakkie* (pickup truck) and at least seven employees. He commented, 'They can decide whether or not they pick flowers for me, they can chop wood if there are no flowers, and they can also do work for other people.' Other ways in which the capacity of staff has been enhanced is through health and safety courses, providing help with obtaining drivers' licences and supporting staff in taking various training courses.

Attempting to preserve the conservation value of the CFR while simultaneously alleviating high levels of poverty is of upmost importance within the context of the wildflower harvesting industry. There is a need to make the industry economically viable and sustainable, and the wildflower harvesting industry is making some progress towards these goals. Local poverty is being tackled through the provision of employment, whilst sustainable harvesting reduces negative ecological impacts.

Our respondents indicated that more could be done to alleviate poverty. In the past, initiatives have been put in place to provide opportunities for people from disadvantaged communities. However, Respondent 2 explained that accessing sufficient land to bring in a sustainable income is an increasing challenge, and as a result, 'we are hesitant, as we do not want to set people up for failure or give them false hope'. From a purely conservation viewpoint, the current situation involving wildflower harvesting is problematic as the resource is insufficiently protected. Although not all harvesters have acquired sustainable harvesting accreditation, the numbers seeking to do so are increasing. This is, therefore, an important step for the industry in trying to achieve a 'triple-win scenario', as a key objective of the programme is to promote economic, social and environmental sustainability.[12]

Competition within the wildflower harvesting industry

Within the wildflower harvesting industry there is intense competition to supply markets as the supply of fynbos currently exceeds demand. While there is competition between wildflower harvesters, the situation is further complicated by the increasing number of cultivators now delivering higher value focal flowers into the market. The expansion of the cultivated industry has led to the wildflower harvesting industry focusing on the gathering of fynbos 'greens'. This shift in focus has occurred because wildflowers are often damaged by heavy rainfall or sunspots, whilst the market favours carefully cultivated and unblemished flowers.

Other issues currently facing the wildflower harvesting industry include the pressure to plough up land where wild fynbos grows in favour of more profitable land uses such as farming, flower growing, vineyards and fynbos cultivation. These land uses are generally more economically viable; that is, 1 ha of cultivated proteas can generate the equivalent economic return of up to 100 ha of natural *veld*. The continuing profitability of wildflower harvesting is particularly important from a conservation point of view, to prevent large areas of natural fynbos from being ploughed up or turned over to grazing. Global economic trends have also had an impact on the industry, such that during the recent global financial crisis there was increasing competition in the markets from handmade paper and cloth 'flowers' produced in India.

Our interviews indicate that the highly competitive nature of the industry has led to much tension and secrecy as harvesters are desperate to hold onto existing markets. This context makes it very difficult for new entrants, particularly smaller harvesters, to enter the industry. As a result of this culture of rivalry, there is an absence of data relating to the industry. In contrast, the vast majority of fruit producers are members of Hortgro, and are required to submit annual statistical returns which enable industry-wide strategic planning and lobbying to take place. However, the wildflower industry is literally an unknown quantity, with only partial statistics available.

Conradie and Knoesen[43(p.3)] note with frustration that 'due to the fierce competition in the industry arising from limited market access, there is no complete producer list'. As a result, the producers are open to manipulation by other players in the supply chain. One respondent explained that, 'a major problem has always been with the 'mafia' pack sheds, they have too much power'. The pack sheds dictate the price and the required quality of fynbos received. As a result, the harvesters have become 'serious price takers', and often have no say in what their product is worth'. Therefore, they have no bargaining power. 'If I don't supply someone else will do it'. Only one harvester interviewed looked beyond the power of the pack sheds, commenting that, 'the biggest threat is the supermarkets; they dictate the price, how many stems in a bunch, what flowers they want and when they want them'.

In 2013 the UK importers made a strategic decision to place their orders through Fynbloem rather than Fynsa. Thus, the majority of exports to the UK are routed via the Fynbloem pack shed at Riviersonderend where bouquets are made before they are sent to MM Flowers Ltd in the UK which finalises the bouquet packaging and organises distribution to the retailers. Other supply chains, such as those for the dried flower market, also involve pack sheds and agents for the markets. The involvement of such intermediaries has left harvesters feeling aggrieved that they are not benefitting in an economically commensurate way. According to Respondent 16, 'the mark-up is ridiculous; some of these bunches go for ZAR150 (USD 14.93) in the UK, while here we are only getting 10 or 40 cents [ZAR0.10 or ZAR0.40] for a stem'.

Whilst we detected a power imbalance between harvesters and pack sheds, the broader context needs to be appreciated. The wildflower harvesting industry is enmeshed within Global Value Chains (GVCs), as most of the harvested product is exported to overseas markets. Figure 3 illustrates the growth in Cape Flora bouquet production between 2008 and 2011. In most circumstances, the pressure becomes greater the further down the GVC that stakeholders are situated.[34] As most of the product is exported to Europe and the UK, the lead companies that apply pressure down the chains are the supermarkets and large flower importers. These companies are introducing more exacting requirements from the harvesters in the shape of 'just in time' supplies and challenging product specifications. These requirements in some cases lead to overharvesting, with knock-on effects for the workers who have to work long hours to meet deadlines, often in difficult conditions. Despite the growing demand for Cape flora bouquets (Figure 3), suppliers report that they have been squeezed increasingly tightly because of the economic recession afflicting target markets. From the retailer perspective, market growth has been achieved by offering good value to consumers.

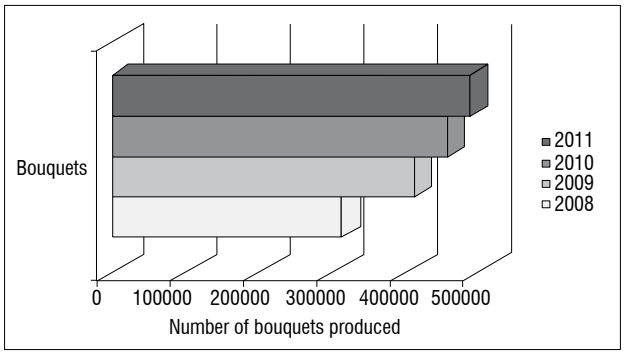


Figure 3: Production of Cape flora bouquets at Fynsa 2008–2011.[44]

Competition in the wildflower harvesting industry

As a result of this competition, it seems as if the successful companies are those that operate on a large scale. As outlined by Respondent 10: 'With wild fynbos you need a lot of land to be a supplier, you can't do it on 4000 ha, only as a side-line maybe, not as a proper business. You need at least 10 000 ha, as you must have all the different species all through the year.' The increasing focus on lower value fynbos greens perpetuates the need to harvest large quantities to make a decent profit.

Harvesting different species throughout the year necessitates access to vast areas of land in order to take advantage of economies of scale. Harvesters have incurred extra costs as a result of the 2013 rural wage increase and steadily increasing fuel prices. Those operating on a large scale are in a better position to absorb these costs into their business models, and are able to invest in plant and infrastructure development, for example, purchasing of 4x4 *bakkies* which are essential for accessing remote picking grounds and transporting the product.

From the pack shed owners' perspective, there are also advantages to sourcing from large suppliers. Pack sheds are under pressure to deliver standardised final products. Sourcing from a number of different harvesters makes achieving this standard difficult, as there can be variations in stem length and colour depending on the growing conditions in different locations. It is also logistically much easier to deal with a few large harvesters than with many smaller harvesters.

How might competition be reduced?

Respondents suggested that competition in the industry might be reduced if there was an increase in demand for the product. Respondent 14 commented: 'People are worried and say the "pie is only this big", but that is only because the demand is only so big'. Therefore, increasing demand should ease competitive pressure through allowing the 'pie' to grow, which should ultimately reduce competitive pressures. Our interviews indicate that the best way to do this is either to differentiate the product or to find new markets. With the exception of Respondent 7, who has been able to find a niche market, all the other harvesters interviewed had difficulty finding good markets. As it is difficult to differentiate the product on an individual scale, there is a need to look towards increasing the demand for fynbos for the industry as a whole. In recent years, the South African market for wildflowers has grown steadily with an increasing demand for fynbos.

In overseas markets it seems that the ethical story of poverty alleviation and conservation is a good selling point for fynbos. There is possibly some scope to sell a similar story to the local South African market, with the added value of buying a local product. However, using the ethical story will not necessarily guarantee improved sales, as a number of other products sold in South Africa and overseas also have ethical stories behind them. For example, Kenya has a long history of exporting flowers to the UK and Europe with ethical stories as a selling point.[45,46] It seems that the wildflower harvesting industry as a whole needs to become more innovative in trying to differentiate the fynbos product and thus possibly achieve an increase in demand. But in order to achieve innovation, we suggest that levels of communication must be improved within the industry.[35]

Communication within the wildflower harvesting industry

Every respondent in our field-based research spoke of a serious lack of communication within the industry and that communication is largely informal. Respondent 7 commented: 'A lot of us are pals and *braai* [barbecue] and watch rugby together on a Saturday. Come Monday, however, and we do not want each other near our sheds, so there is communication, but it is bordered off.' Whilst the larger harvesters, pack sheds, NGOs and government organisations have improved communication in recent years, the PPSA has been advocating for even better communication and has recently argued that wildflower harvesters should be part of their association. However, despite a 90% representation from the cultivated harvesters, unfortunately there are currently only six or seven wildflower harvesters involved through the Sustainable Harvesting Committee.[16]

Because of a general lack of communication, it is evident that there is serious fragmentation within the wildflower harvesting industry, largely driven by its strongly competitive nature. Stakeholders are scared to share too much information as they are afraid that it could potentially harm their businesses. There is a fear that other harvesters could outcompete them and take their harvesting lands, their innovative ideas and products, and ultimately their markets. Thus, a lack of trust is an inhibiting factor in the future development of the industry.

The first logical step to improving communication is to try to break down the barriers to communication by creating a more trusting environment. The informal communication which seems to exist should not be undervalued.[36,37] A number of respondents agreed that in order to improve more formal communication, it is essential that communication first commence at an informal level within local networks. The best approach might be to encourage communication through people who are already part of committees or organisations such as PPSA, especially people who are well known and respected within the industry. Those who are part of PPSA have explained that they have gained much from increased communication and membership of associations. As outlined by Respondent 18, 'people need to see the benefits of joining a group…if they see no real benefits for themselves then they will not see the point'. All respondents stated that communication is one of the keys to resolving some of the main challenges within the industry. A range of benefits was discussed, such as improved information sharing, but the benefit most frequently mentioned was the potential for collective bargaining.

A possible 'Wildflower Harvesting Forum'?

The potential benefits from increased information sharing, and in particular collective bargaining, are immense and it is apparent that there is an urgent need to set up some form of collective. With so many challenges, it seems the industry cannot afford to carry on with the status quo if it wants to remain sustainable and viable. In light of the evidence gained from field-based research, it seems there is a strong motivation for the establishment of a 'Wildflower Harvesting Forum'. When enterprises work largely in isolation, there is an uncompetitive production pattern which ultimately does not result in innovation.[35] Setting up a forum could be a way to enhance both innovation and sustainability within the industry.

Moving beyond merely improving communication, a forum should be able to organise and channel communication in an effective and efficient way. Such a forum could strengthen local, vertical and horizontal linkages, in addition to external linkages between harvesters and other stakeholders.[47] Through developing a forum, a 'one-stop shop' could be set up to provide a single source of contact between harvesters and various local institutions.

A forum could also improve the basis for collective bargaining, as harvesters will be in a much better position to lobby government and markets. Respondents noted that collective bargaining would enable stronger negotiation around prices from pack sheds and retailers. Additionally, more favourable conditions could also be negotiated, for example, giving more control over the products that are rejected. On an individual basis, there is little scope for bargaining, as pack sheds have the ability to simply source the product from another harvester.

However, pack sheds will have less opportunity for bargaining if they are dealing with a more powerful and unified body of harvesters. Furthermore, a forum is likely to have more power in negotiations with government concerning the development of more supportive legislation.

Ideally, a forum should develop from increased communication, co-operation and trust, building on the informal links and communication which already exist. Those leading the process need to be trusted and the benefits of cooperation emphasised. Perhaps initially, certain common challenges could be discussed such as alien clearing, fire management and the implications of the 2013 wage increase. By starting locally and informally, and discussing issues with no possible economic repercussions, levels of trust should hopefully improve, such that more sensitive economic matters could be raised. To this end, the Sustainable Harvesting Committee has proposed that regional forums, which would feed into CapeFlora SA, should be created for wild harvesters.

Conclusion

A Wildflower Harvesting Forum could be pivotal in addressing, in a cooperative and community-based manner, the many challenges facing the industry. The people who might potentially drive the establishment of a forum are probably those who are already members of key committees or organisations such as the CapeFlora SA, and therefore have well- established links.

Although harvesters are, perhaps understandably, more concerned about their immediate economic viability, a forum would act as a critical space to embed environmental and social sustainability as well as longer-term economic sustainability. In light of the importance and vulnerability of the CFR, if environmental concerns are not addressed and unsustainable harvesting occurs, the CFR will become severely degraded, resulting in the eventual collapse of the industry. In relation to the alleviation of poverty, the legacies of apartheid are still deeply embedded in society, which is most evident amongst the historically disadvantaged communities of the Agulhas Plain. Promoting sustainable harvesting and wider social ethics must be key considerations in the future development of the industry. Recent research suggests that a forum could play a much-needed role in pulling together disparate voices to promote the environmental and socio-economic credentials of the industry. Ultimately, changes need to be made to the distribution of value which may well necessitate restructuring of the value chain. For these changes to occur, industry players must collaborate and speak with a coherent voice, which can only happen if strong institutions develop at the base of the value chain. We strongly endorse the recommendation of the Sustainable Harvesting Committee to create regional forums and would urge that resources are provided to ensure that this initiative is established and that the wildflower industry develops a reputation for professionalism befitting players in an international industry.

Acknowledgements

We thank the Flower Valley Conservation Trust for all their support and the participants who contributed to this study for their time. We are further grateful for constructive comments from the anonymous referees and the editor.

Authors' contributions

T. Blokker was the primary researcher and author; and T. Binns and D.B. supervised the research and provided significant input in the writing of the paper.

References

1. Agulhas Biodiversity Initiative (ABI). Agulhas Biodiversity Initiative: Project brief. Cape Town: ABI; 2003.
2. Privett S, Lutzeyer H. Field guide to the flora of Grootbos Nature Reserve and the Walker Bay region. Cape Town: Grootbos Foundation; 2010.
3. Turpie J, Heydenrych BJ. The Cape Floral Kingdom's fynbos vegetation. In: Perrings C, Williamson M, Dalmazzone S, editors. The economics of biological invasions. Cheltenham: Edward Elgar; 2000. p. 553–556. http://dx.doi.org/10.4337/9781781008645.00018
4. Privett S. Flower Valley review for the Arcadia Fund: Final report (Confidential). Cape Town: Fynbos Ecoscapes; 2010.
5. Richardson L. Agulhas Plain fynbos monitoring project. Bredasdorp: Flower Valley Conservation Trust; 2013.
6. Nel E, Binns T, Bek D. Fair trade and alternative food networks in the internationalization of small-scale rural entrepreneurs in South Africa. In: Etemad H, editor. The process of internationalization in emerging SMEs and emerging economies. Cheltenham: Edward Elgar; 2013 p. 234–259. http://dx.doi.org/10.4337/9781781003190.00018
7. Cowling R, Richardson D. Fynbos: South Africa's unique floral kingdom. Cape Town: Fernwood Press; 1995.
8. Bek D, Binns T, Nel E. Wild flower harvesting on South Africa's Agulhas Plain: A mechanism for achieving sustainable local economic development? Sustain Dev. 2013;21(5):281–293. http://dx.doi.org/10.1002/sd.499
9. Helliker KD. NGOs and rural movements in contemporary South Africa. Soc Dynamics. 2013;39(2):317–326. http://dx.doi.org/10.1080/02533952.2013.806415
10. Smith A. South African farmworkers sacked 'in truckloads' after strike, says union. The Guardian. 2013 January 24. Available from: http://www.guardian.co.uk/world/2013/jan/24/truckloads-south-african-farmworkers-sacked/print
11. Kotze M. PPSA producer survey results. Cape Town: Hortgro Services; 2012.
12. Agulhas Biodiversity Initiative (ABI). Summary of good practice guidelines for the sustainable use of inidigenous wild fynbos resources. Cape Town: Cape Action for People and the Environment; no date.
13. Privett S, Bailey R, Raimondo D, Kirkwood D, Euston-Brown K. A vulnerability index for rare and harvested plant species on the Agulhas Plain. Bredasdorp: Flower Valley Conservation Trust; 2005.
14. Bailey R, Euston-Brown K, Privett S. Resource base assessment of wild fynbos plants on Waterford Farm, Agulhas Plain: A pilot study. Bredasdorp: Flower Valley Conservation Trust; 2007.
15. CapeNature. About CapeNature [homepage on the Internet]. c2013 [cited 2013 Jun 10]. Available from: http://www.capenature.co.za/about-us/
16. PPSA. Protea producers of South Africa [homepage on the Internet]. c2013 [cited 2013 Jun 10]. Available from: http://www.capeflora.co.za/
17. Chen S, Ravallion M. Absolute poverty measures for the developing world, 1981–2004. Proc Natl Acad Sci USA. 2007;104(43):16757–16762. http:// dx.doi.org/10.1073/pnas.0702930104
18. Salafsky N, Wollenberg E. Linking livelihoods and conservation: A conceptual framework and scale for assessing the integration of human needs and biodiversity. World Dev. 2000;28(8):1421–1438. http://dx.doi.org/10.1016/S0305-750X(00)00031-0
19. Fisher B, Christopher T. Poverty and biodiversity: Measuring the overlap of human poverty and the biodiversity hotspots. Ecol Econ. 2007;62(1):93–101. http://dx.doi.org/10.1016/j.ecolecon.2006.05.020
20. Crane W. Biodiversity conservation and land rights in South Africa: Whither the farm dwellers? Geoforum. 2006;37(6):1035–1045. http://dx.doi.org/10.1016/j.geoforum.2006.07.002
21. Adams WH, Aveling R, Brockington D, Dickson B, Elliot J, Hutton J, et al. Biodiversity conservation and the eradication of poverty. Science. 2004;306:1146–1149. http://dx.doi.org/10.1126/science.1097920
22. Roe D, Yassin Mohammed E, Porras I, Giuliani A. Linking biodiversity conservation and poverty reduction: De-polarizing the conservation-poverty debate. Conserv Lett. 2012;6(3):162–171. http://dx.doi.org/10.1111/j.1755- 263X.2012.00292.x
23. Oates JF. Myth and reality in the rain forest: How conservation strategies are failing in West Africa. Berkeley, CA: University of California Press; 1999.
24. Terborgh J. Reflections of a scientist on the world parks congress. Conserv Biol. 2004;18(3):619–620. http://dx.doi.org/10.1111/j.1523- 1739.2004.01837.x

25. Matsuoka S, Komatsu S. Social capacity development for environmental management. In: Matsuoka S, editor. Effective environmental management in developing countries: Assessing social capacity development. New York: Palgrave MacMillan; 2007. p. 23–52.

26. Bishop J, Kapila S, Hick F, Mitchell P, Vorhies F. Building biodiversity business. London: Shell International Limited and the International Conservation of Nature; 2008.

27. Fisher R, Maginnis S, Jackson W, Barrow E, Jeanrenaud S. Linking conservation and poverty reduction: Landscapes, people and power. Hoboken, NJ: Taylor and Francis; 2012.

28. Van Beukering PJ, Papyrakis EB, Bouma J, Brouwer R, editors. Nature's wealth: The economics of ecosystem services and poverty. Cambridge: Cambridge University Press; 2013. p. 1–31. http://dx.doi.org/10.1017/CBO9781139225311

29. Bell RG, Russell C. Environmental policy for developing countries. Issues Sci Technol. 2002;18(3):63–70.

30. Hardin G. The tragedy of the commons. Science. 1968;162(3859):1243–1248. http://dx.doi.org/10.1126/science.162.3859.1243

31. Gereffi G, Humphrey J, Sturgeon T. The governance of global value chains. Rev Int Polit Econ. 2005;12(1):78–104. http://dx.doi.org/10.1080/09692290500049805

32. Ponte S, Gibbon P. Quality standards, conventions and the governance of global value chains. Econ Soc. 2005;34(1):1–31. http://dx.doi.org/10.1080/0308514042000329315

33. Humphrey J, Memedovic O. Global value chains in the agrifood sector. Vienna: United Nations Industrial Development Organization; 2006. Available from: https://www.unido.org/fileadmin/user_media/Publications/Pub_free/Global_value_chains_in_the_agrifood_sector.pdf

34. Raworth K. Trading away our rights: Women working in global supply chains. Oxford: Oxfam; 2004. Available from: http://www.oxfam.org/sites/www. oxfam.org/files/rights.pdf

35. United Nations Industrial Development Organization (UNIDO). Clusters and business linkages [homepage on the Internet]. No date [cited 2013 Aug 09]. Available from: http://www.unido.org/clusters/

36. Cousins PD, Handfield RB, Lawson B, Petersen KJ. Creating supply chain relational capital: The impact of formal and informal socialization processes. J Oper Manag. 2006;24(6):851–863. http://dx.doi.org/10.1016/j.jom.2005.08.007

37. Lawson B, Petersen KJ, Cousins PD, Handfield RB. Knowledge sharing in interorganizational product development teams: The effect of formal and informal socialization mechanisms. J Prod Innovat Manag, 2009;26(2):156–172. http://dx.doi.org/10.1111/j.1540-5885.2009.00343.x

38. Prahinski C, Benton W. Supplier evaluations: Communication strategies to improve supplier performance. J Oper Manag. 2004;22(1):39–62. http:// dx. doi.org/10.1016/j.jom.2003.12.005

39. Axelrod R. The evolution of cooperation. New York: Basic Books; 1984.

40. Messner D, Meyer-Stamer J. Governance and networks. Tools to study the dynamics of clusters and global value chains. Duisburg: INEF; 2000.

41. Bolton GE, Ockenfels A. ERC: A theory of equity, reciprocity, and competition. Am Econ Rev. 2000;90(1):166–193. http://dx.doi.org/10.1257/aer.90.1.166

42. Privett S. Livelihoods out of the ashes: An opportunity analysis following the 2006 Overberg Fires. Cambridge: Fauna and Flora International; 2006.

43. Conradie B, Knoesen D. A survey of the cultivation and wild harvesting of fynbos flowers in South Africa [document on the Internet]. c2010 [cited 2013 Nov 12]. Available from: http://www.flowervalley.org.za/pdfs/PPSA%20report%20Conradie%20Knoessen.pdf

44. Bek D. End of project review for WWF Table Mountain Fund. Cape Town; 2012.

45. Hughes A. Global commodity networks, ethical trade and governmentality: Organizing business responsibility in the Kenyan cut flower industry. T I Brit Geogr. 2001;26(4):390–406. http://dx.doi.org/10.1111/1475-5661.00031

46. Barrett HR, Ilbery BW, Brown AW, Binns T. Globalization and the changing networks of food supply: The importation of fresh horticultural produce from Kenya into the UK. Trans Inst Br Geogr. 1999;24(2):159–174. http://dx.doi.org/10.1111/j.0020-2754.1999.00159.x

47. Giuliani E, Pietrobelli C, Rabellotti R. Upgrading in global value chains: Lessons from Latin American clusters. World Dev. 2005;33(4):549–573. http://dx.doi.org/10.1016/j.worlddev.2005.01.002

Incidence of non-typhoidal *Salmonella* in poultry products in the North West Province, South Africa

AUTHORS:
Roseline Y. Olobatoke[1]
Sendros D. Mulugeta[1]

AFFILIATION:
[1]Department of Animal Science, School of Agricultural Sciences, North West University, Mmabatho, South Africa

CORRESPONDENCE TO:
Roseline Olobatoke

EMAIL:
yemisirose205@yahoo.com

POSTAL ADDRESS:
Centre for Applied Radiation Science and Technology, North West University, Private Bag X2046, Mmabatho 2735, South Africa

DATES:
Received: 16 July 2014
Revised: 10 Oct. 2014
Accepted: 23 Feb. 2015

KEYWORDS:
food-borne salmonellosis; broiler; *invA*; *spvC*; fingerprinting

HOW TO CITE:
Olobatoke RY, Mulugeta SD. Incidence of non-typhoidal *Salmonella* in poultry products in the North West Province, South Africa. S Afr J Sci. 2015;111(11/12), Art. #2014-0233, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140233

This study was conducted to evaluate the incidence of non-typhoidal *Salmonella* (NTS) serotypes in raw and ready-to-eat (RTE) broiler products in the North West Province of South Africa. A total of 120 raw broiler samples, 40 samples of polonies and 20 samples of smoked viennas were obtained from retail points in major cities and towns in the province. Samples were subjected to aerobic plate count and later screened for the presence of NTS using phenotypic and genotypic techniques. The average bacterial count in raw products was 3.1×10^5 cfu/g whereas bacterial contamination of RTE products was 1.8×10^3 cfu/g. The average recovery rate of NTS species from raw broiler products was 12.5% and the serotypes identified were *S. typhimurium* (46.4%), *S. enteritidis* (30.9%) and *S. newport* (22.9%). No NTS was recovered from the RTE products. However, *S. typhimurium* was the predominant serotype in whole carcasses whereas *S. enteritidis* and *S. newport* were prevalent in chicken parts. Out of the 160 presumptive NTS isolates screened by polymerase chain reaction (PCR), 140 (87.5%) were confirmed for the presence of the *Salmonella*-specific *invA* gene. In addition, 115 (82.4%) of the confirmed isolates harboured the plasmid *spvC* gene. *Random Amplified Polymorphic DNA* (RAPD) fingerprinting of isolates using RAPD 1 and RAPD 3 primers, revealed some inter- and intra-serotype genetic diversity among isolates, suggesting varying sources of contamination. The results of this study represent the first report on the incidence and prevalent serotypes of NTS in chicken products in the North West Province of South Africa.

Introduction

Non-typhoidal salmonellosis is an important public health problem worldwide and particularly in sub-Saharan Africa where it commonly manifests as gastroenteritis and/or bloodstream infections in both children and adults.[1] The gastroenteritis form, which may sometimes be self-limiting, is commonly found in industrialised countries. Immunocompromised individuals including patients with HIV, cancer or diabetes, are at higher risk of non-typhoidal *Salmonella* (NTS) bacteraemia and often develop focal infections such as meningitis, septic arthritis, pneumonia and osteomyelitis.[2] Although more than 2500 serovars of *Salmonella enterica* have been reported, *S. typhimurium* and *S. enteritidis* are identified as the commonest causes of human infection.[3,4] A retrospective study in one hospital in the Democratic Republic of Congo from 2002 to 2006, revealed that NTS caused 59% of bacteraemia in children. *Salmonella typhimurium* and *Salmonella enteritidis* were responsible for 82% of the cases.[5] In Mozambique, NTS were reported to account for 120 cases of childhood bacteraemia per 100 000 persons/year.[6] Contaminated poultry meat and eggs, among other factors, have been implicated as vehicles of transmission for these hardy pathogens.[7] For these reasons, there have been numerous studies focusing on assessing the incidence/prevalence of NTS strains in chicken carcasses and other meat products.[7-9]

South Africa has witnessed a tremendous increase in chicken meat consumption.[10] Concurrently, the Enteric Disease Reference Unit of the National Institute for Communicable Diseases noted an increasing number of NTS isolates despite the fact that human salmonellosis cases are rarely reported. Evidence of these occurrences are the outbreaks of food-borne illnesses in Mpumalanga Province of South Africa incriminating NTS serotypes. One of the outbreaks involved the consumption of meals prepared with poultry products.[11,12] These outbreaks indicate the presence of NTS in South Africa, which may be an issue of public health concern. A few investigations have been conducted in South Africa to ascertain the contamination of chicken carcasses and ready-to-eat foods from retail stores, with various pathogenic bacteria including *Salmonella*.[13,14] However, the incidence of NTS in broiler products in the North West Province (NWP) has not been established despite the fact that the NWP is one of the provinces with the largest production and distribution of broilers in the country. The objective of this study therefore was to screen raw and ready-to-eat broiler products obtained from the NWP of South Africa for incidence of NTS contamination.

Materials and methods

Experimental design

The study was cross-sectional with sampling based on two shop types and three product groups. The sampling, which lasted for 6 months, commenced in October 2010 and ended in March 2011, corresponding with the summer months in the study area.

Sample size

Primary population size to be sampled was determined according to the following formula[15]:

$$\left(\frac{z_\alpha * SD}{L}\right)^2 \quad \text{where SD} = [p(1-p)]^{1/2} \qquad \text{Equation 1}$$

where p is the expected prevalence; L is the accepted error and Z_α is the value for normally distributed data at a confidence level α.

Based on an expected prevalence of 40% and a confidence level of 90%, a total of 57 shops were sampled, irrespective of the source of their products (i.e. either from within or outside the province). A total of 180 samples was obtained for analysis based on the purchase of a minimum of three product groups from each shop, and keeping 5% above the actual sample size to prevent the risk of sample loss during analysis (Table 1). Temperature of samples at point of purchase ranged from -20 °C to 4 °C. For all the procedures involved in the analysis, *Salmonella typhimurium* ATCC 14028 was used as a positive control strain whereas *E. coli* ATCC 25922 was used as a negative control. All culture, isolation and biochemical screening techniques were carried out using the MFHPB-20 procedure[16] with appropriate modifications.

Table 1: Number of *Salmonella* positive samples obtained from different locations

Sampling area	Number of shops visited[a]	Number of samples obtained[b]	Number of positive samples	% Positive samples
Sun City	2	6	0	0
Koster	3	8	1	12.5
Delareyville	5	16	0	0
Vryburg	4	12	0	0
Mmabatho	3	12	2	16.7
Mafikeng	10	33	6	18.2
Zeerust	6	19	0	0
Rustenburg	6	18	2	11.1
Lichtenburg	5	16	2	12.5
Klerksdorp	5	15	0	0
Potchefstroom	8	25	2	8

aThe number of shops visited was dependent on the number of shops available within a particular location. bA minimum of three samples comprising three sample types was purchased from each shop.

Determination of total bacterial count in samples

Briefly, 25 g of each sample was aseptically removed and homogenised with 225 mL of 2% buffered peptone water (BPW) in a stomacher bag (Nasco, Swedesboro, NJ, USA). Then 1 mL of the homogenate was transferred to Eppendorf tubes and serial dilutions were made using 2% BPW. An aliquot (0.1 mL) of each dilution was plated on plate count agar (Merck, SA) by spreading, and incubated at 37 °C for 24 h after which colonies were counted and recorded as cfu/g sample.

Isolation of Salmonella species from broiler carcasses and sausages

For broiler carcasses, each sample was aseptically removed from the package and transferred to a sterile plastic bag. Then 150 mL of 2% BPW was added to each bag and the mixture was shaken constantly for 2 min to obtain carcass rinse. Thereafter, 25 mL of the rinse liquid was transferred to a Whirl-pak bag (Nasco, USA) and another 75 mL of 2% BPW was added. Chicken sausages were aseptically homogenised with 225 mL of double strength (4%) BPW in a stomacher bag (Nasco, USA). Each bag was sealed and incubated at 37 °C for 18 h. Following incubation, contents of each Whirl-pak bag (Nasco, USA) was homogenised and 0.1 mL aliquot of the mixture was transferred to 10 mL of Rappaport–Vassiliadis broth (Merck, SA) for selective enrichment of *Salmonella*. The broth was incubated at 42 °C for 24 h and then kept at room temperature for 24 h.[17] Thereafter, a loopful of each broth sample was streaked on xylose lysine desoxycholate (XLD) agar (Merck, SA) and *Salmonella Shigella* agar (Merck, SA) and incubated at 37 °C for 24–36 h. Suspected colonies were further purified on XLD agar and thereafter preserved on nutrient agar (Merck, SA) slants for further analysis.

Biochemical screening of presumptive isolates

All suspect colonies from the XLD agar culture were inoculated on triple sugar iron agar (Merck, SA) slant and incubated at 37 °C for 24 h. Presumptive *Salmonella* isolates from the triple sugar iron test were further confirmed using the API-20E test (bioMerieux Inc., SA) performed according to the manufacturer's instructions.

Serological identification

Isolates showing typical *Salmonella* biochemical reactions were sero-typed by the slide agglutination test using *Salmonella* specific polyvalent antisera (Davies Diagnostics, SA) for O and H antigens. The test was performed according to the manufacturer's instructions.

Molecular characterisation of isolates

DNA extraction

Genomic and plasmid DNA were extracted using the Zymo® kit (Inqaba Biotech, SA) and alkaline lysis method,[18] respectively. The quantity and purity of all extracted DNA were estimated using a UV visible spectrophotometer (model S-22, Boeco, Germany) after which the integrity of the DNA was checked on standard submarine gel electrophoresis using 0.8% (w/v) agarose. The isolated DNA was stored at -20 °C until use.

PCR of extracted DNA

The 16S rRNA gene fragments of all isolates were amplified,[19] after which presumptive *Salmonella* isolates were screened for the presence of invasion (*invA*) and virulence (*spvC*) genes for identity confirmation.[20] All the oligonucleotide primers used for polymerase chain reactions (PCR) were obtained from Inqaba Biotech, South Africa, and details of the sequences and cycling conditions are shown in Table 2. All PCR reactions were prepared in 25 μL volumes consisting of 1 μg/μL of the template DNA, 50 pmol of each oligonucleotide primer set, 1x PCR master mix, 1U Taq DNA polymerase and nuclease free distilled water.

Amplifications were performed using a Peltier Thermal Cycler (model Dyad™ DNA Engine). A reaction blank containing all the components of the reaction mixture except the template DNA was included in each PCR procedure to check for contamination. Thereafter, 10 μL of the PCR amplicons was analysed by standard submarine gel electrophoresis using 1% (w/v) agarose at 60 V for 6 h. Lambda DNA molecular weight marker was the gene ruler used in electrophoretic analyses. Electrophoresis was conducted in a horizontal Pharmacia biotech equipment system (model Hoefer HE 99X; Amersham Pharmacia Biotech, Sweden) using 1x TAE running buffer (40 mM Tris, 1 mM EDTA and 20 mM glacial acetic acid, pH 8.0). The gels were stained in ethidium bromide (0.1 mg/mL) for 10–15 min and later visualised under UV light.[18] A Gene Genius Bio imaging system (Syngene, Synoptics; UK) was used to capture the image using GeneSnap (version 6.00.22) software. Amplicons which showed clear bands for 16S rRNA and *invA* genes were purified and sequenced at Inqaba Biotech (South Africa), for strain identification of the isolates, using the ABI 3500 XL sequencer. The search for amino acid homology of the sequence results was done using the NCBI Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/BLAST/).

RAPD-PCR

Confirmed *Salmonella* isolates identified through the expression of the *Salmonella*-specific *invA* gene were further subjected to fingerprinting by random amplified polymorphic DNA (RAPD)-PCR to evaluate the genetic diversity among the isolates. Oligonucleotides of short sequence were used[21,22] and the PCR reactions were prepared as stated above. Primer details and PCR conditions are shown in Table 2. Aliquots (10 μL) of the resulting amplicons were characterised on 1.2% agarose gel electrophoresed at 60 V for 6 h. Isolates were later scored for the presence or absence (1 or 0) of each band on agarose gel.[22]

Table 2: Primers and polymerase chain reaction (PCR) conditions used for amplification of target genes and random amplified polymorphic DNA (RAPD)-PCR

Primer	Sequence (5'-3')	Target gene	Amplicon size (bp)	PCR conditions
27F 1492R	AGAGTTTGATCCTGGCTCAG GGTTACCTTGTTACGACTT	16S ribosomal gene	1450	95 °C, 5 min; 35 cycles of 94 °C, 30 s; 61 °C, 30 s; 72 °C, 1 min; extension at 72 °C, 5 min
InvF invR	CGCGGCCCGATTTTCTCTGGA AATGCGGGGATCTGGGCGACAAG	*invA*	321	94 °C, 90 s; 30 cycles of 94 °C, 45 s; 60 °C, 45 s; 72 °C, 90 s; extension at 72 °C, 3 min
VirF VirR	GGGGCGGAAATACCATCTACA GCGCCCAGGCTAACACG	*spvC*	392	94 °C, 90 s; 30 cycles of 94 °C, 45 s; 60 °C, 45 s; 72 °C, 90 s; extension at 72 °C, 3 min
RAPD 1	AGCGTCACTC			94 °C, 5 min; 42 cycles of 94 °C, 1 min; 27 °C, 45 s; 72 °C, 90 s; extension at 72 °C, 5 min
RAPD 2	GCGGAAATAG			94 °C, 60 s; 40 cycles of 94 °C, 1 min; 36 °C, 1 min; 72 °C, 2 min; extension at 72 °C, 10 min
RAPD 3	AACGCGCAAC			94 °C, 90 s; 40 cycles of 94 °C, 1 min; 36 °C, 1 min 72 °C, 2 min; extension at 72 °C, 10 min

Data analysis

To obtain the cfu/g of sample for total bacterial count, colonies on plate count agar were counted and compared with the dilution factor.[23] All *Salmonella* populations were transformed to base 10 logarithms before analysis. Recovery rate of *Salmonella* isolates was calculated using a previously described equation.[24] Means in recovery rates between locations, shop and product types were compared using the Student's *t*-test of SPSS (SPSS 10.0 for windows, SPSS Chicago, IL, USA). Differences were considered significant at $p \leq 0.05$.

$$\text{Recovery rate} = \frac{\text{No of positive samples}}{\text{Total no of samples}} \times 100 \qquad \text{Equation 2}$$

For cluster analysis, the distance matrices, means and standard deviations were calculated using Statistica and the result was used to construct a phylogenetic relationship among isolates by the neighbour-joining method of Saitou and Nei[25].

Results and discussion

Total bacterial count of samples

On average, total bacterial count of butchery samples (2.2×10^5 cfu/g) was higher than that of supermarket samples (1.4×10^4 cfu/g). Similarly, raw chicken parts were more contaminated (7.6×10^5 cfu/g) than whole chicken carcasses (2.5×10^5 cfu/g). RTE products were the least contaminated (1.8×10^3 cfu/g) perhaps because the products had been subjected to further processing. However, the level of microbial contamination of all analysed products was within the recommended range for some developed countries such as United Kingdom (10^4–10^5 cfu/g) and Australia (10^6 cfu/g).[26] Sources of contamination of raw poultry products vary and could originate from the live birds, processing procedures or from the environment.[27] The higher microbial contamination rate of butchery samples and raw chicken portions when compared with supermarket products might be an indication of poor hygienic conditions in the processing environment. This suggests a

lack of strict hygiene control measures during product processing and could have a public health implication on the consumers as poultry are usually contaminated with a relatively high frequency of pathogenic bacteria. Indeed, raw poultry products have been reported in quite a number of human food poisoning cases, particularly following handling, undercooking or mishandling of the cooked products.[28,29] Furthermore, high level of product contamination, such as recorded in some samples in the current study, could facilitate product spoilage, particularly when organisms such as *Pseudomonas* are involved.

Incidence of non-typhoidal Salmonella *in broiler products*

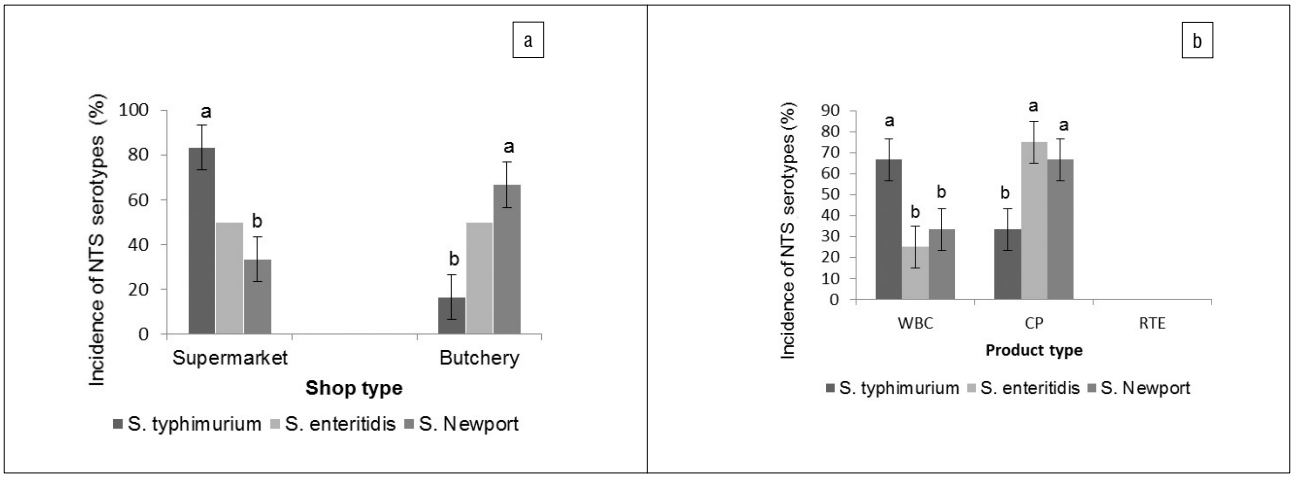
Out of 11 locations sampled, *Salmonella* was recovered from products obtained from 6 (54.5%) of the locations (Table 1). The difference in *Salmonella* isolation rates between positive locations was significant ($p=0.009$) with Mafikeng having the highest rate. Mafikeng is a major town with about the highest number of retail shops in the province. As a result, the number of shops visited and samples obtained from Mafikeng in this study were more than other locations and this could be responsible for the difference in *Salmonella* isolation rates between locations. *Salmonella* spp. was not recovered from RTE products whereas average recovery rate in raw broiler products was 12.5%. This rate is lower than the 19.5% reported in the Gauteng Province of South Africa[13], 20% in the USA[30], 30% in Canada[31], 39.7% in Mexico[32] and 60% in Portugal[8]. Other countries such as Zambia[33], Saudi Arabia[34], Turkey[7] and Sudan[9], however, recorded lower incidences of NTS (4.7%, 5.92%, 8% and 9.2%, respectively) in broiler products than noted in the current study. Incidence of *Salmonella* contamination was higher in whole chicken carcasses (30%) than chicken parts (20%). This was contrary to the observations of Chaiba et al.[35] and Moussa et al.[34] who noted higher incidences of NTS in chicken cuts in Morroco and Sudan respectively. The presence and distribution of NTS serovars vary from region to region depending on the sampling plan and detection limits of the methodology employed.[36]

Butchery products were more contaminated with NTS (32%) than supermarket products (18%). This result, though contrary to the report by Yang et al.[37], was in agreement with the findings of Chaiba et al.[35] and could be the result of better hygienic standards maintained in supermarkets than butcheries. Frozen products had a higher NTS recovery rate (10.4%) than fresh products (8.8%). This may suggest initial high contamination of the frozen products with NTS. Although previous studies identified storage temperature as an important risk factor for pathogen growth and survival, the result of the current study suggests that freezing temperature range was not sufficient to inactivate the pathogens. It has been documented that freezing may not be regarded as a means of destroying food borne microorganisms. This is because low freezing temperatures of about –20 °C are less harmful to the pathogens than the median temperature ranges.[38]

Molecular characterisation of non-typhoidal Salmonella isolates

A total of 160 presumptive *Salmonella* isolates obtained from analysed broiler products were screened for the presence of *invA* and *spvC* genes, following the positive amplification of the 16S rRNA fragment. The *invA* gene was detected in 140 (87.5%) of the 160 screened isolates. Sequencing of the 16S rRNA amplicons of the remaining 20 isolates showed that they belonged to three other coliform species namely *Klebsiella*, *Escherichia* and *Serratia*. These could be the contaminants that showed biochemical and serological reactions similar to *Salmonella*. The *invA* gene, located on the Salmonella pathogenicity island 1 (SPI 1), is highly conserved in *Salmonella* species and encodes a type III secretion system (TTSS) that exports proteins in response to bacterial contact with epithelial cells.[39] The *invA* gene operon is essential for full virulence in *Salmonella*.[40] The high detection of *invA* gene in the current study may be an indication of the potential pathogenicity of the isolated *Salmonella* strains and could be a cause for public health concern. Torpdahl et al.[40] and Khan et al.[41] similarly detected *invA* gene in all *Salmonella* isolates recovered from poultry products and orange juice, respectively. Analysis of the sequenced isolates in the current study revealed three NTS serotypes, namely, *S. typhimurium* (46.4%; $n=65$), *S. enteritidis* (30.7%; $n=43$) and *S. newport* (22.9%; $n=32$). The sequences of the three serotypes were deposited in GenBank with accession numbers JX859913, KC683709 and JX859912, respectively. The rate of isolation of the different NTS serotypes was in agreement with the reports from previous research.[35,36] All the identified serotypes were isolated from both supermarket and butchery samples as well as whole carcasses and chicken parts at different rates (Figure 1). This shows that the isolated *Salmonella* serotypes may be circulating within the environment and possible evidence is the fact that some of the products were contaminated with multiple NTS serotypes, which may also indicate varying sources of product contamination.

Out of the 140 isolates confirmed for the presence of the *invA* gene, 115 (82.4%) isolates were found to harbour the *spvC* genes representing 71.9% of all the screened isolates. The *spvC* gene in *Salmonella* spp. interacts with the host immune system and is responsible for an increased growth rate in host cells.[41] The prevalence of the *spvC* gene fragment was higher in isolates from butchery samples (71.4%) than supermarket samples (25.9%) and isolates from chicken parts (57.1%) than whole broiler carcasses (42.9%). The detection rate of the *spvC* gene in the current study agreed with the findings of Bolton et al.[42] and Khan et al.[20] who detected the gene fragment in 97% and 88% of *Salmonella typhimurium* DT104 strains, respectively. However, Khan et al.[41] did not detect *spvC* gene in any of the *Salmonella* serotypes subjected to multiplex PCR. The recovery rate of *S. typhimurium* in this study was notably higher than that of other serotypes (*S. enteritidis* and *S. newport*), as opposed to the



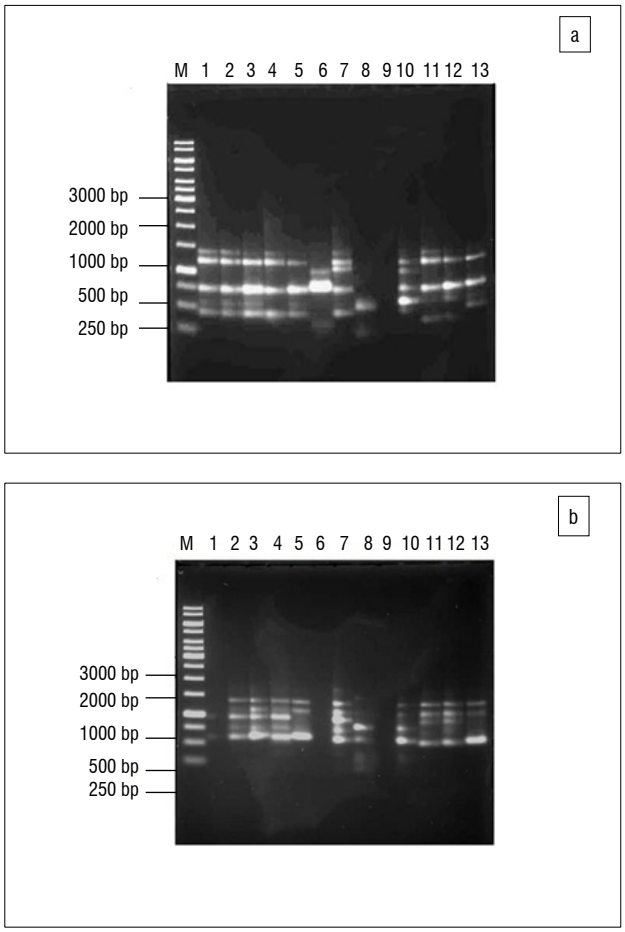
WBC, whole broiler carcasses; CP, chicken parts; RTE, ready-to-eat. Bars with different labels are significantly different ($p<0.05$).

Figure 1: Percentage incidence of non-typhoidal *Salmonella* (NTS) serotypes by (a) shop type and (b) product type.

report of Khan et al.[41] who did not recover *S. typhimurium* isolates in the juice samples analysed. Thus the disparity in the detection of the *spvC* gene fragment could possibly be related to the prevalence of the gene in *S. typhimurium* serotype.

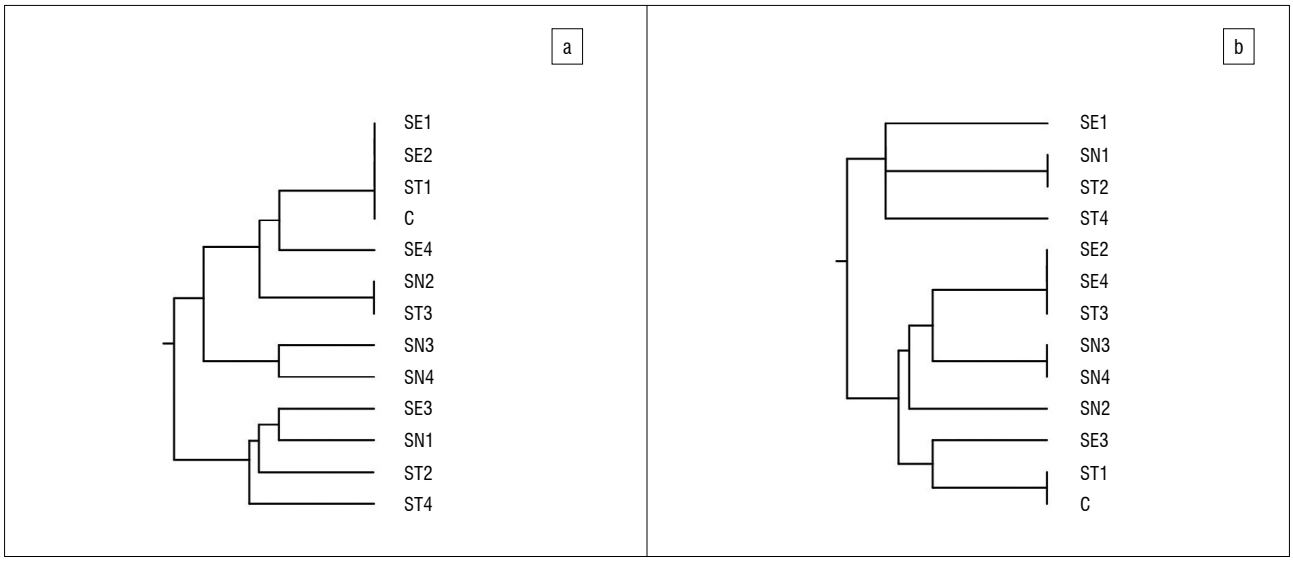
Analysis of RAPD-PCR

The profiles generated by the RAPD-PCR were composed of 1–5 bands ranging between 0.25 kb and 1.4 kb. Both RAPD 1 and RAPD 3 primers each generated four banding profiles for *S. typhimurium,* three for *S. newport* and two profiles for *S. enteritidis* (Figure 2a and 2b). One strain of *S. newport* (SN1), isolated from chicken parts obtained from the butchery, showed no bands with both RAPD 1 and RAPD 3 primers. Similarly, a strain of *S. typhimurium* (ST2) isolated from supermarket whole broiler carcass, showed three bands with RAPD 1 but no bands with RAPD 3 primer. However, both strains produced one band of about 0.45 kb as did most of the isolates with RAPD 2 primer although the primer had very poor discriminatory power and was therefore not included in the analysis. Cluster analysis of the RAPD 1 profiles revealed three different RAPD types for *S. newport* and *S. enteritidis* whereas *S. typhimurium* was classified into four types (Figure 3a). RAPD 3 primer on the other hand, grouped S. *typhimurium* into four, *S. newport* into three and *S. enteritidis* into two RAPD types (Figure 3b). Although a few isolates of the same serotype showed either similar or same banding patterns, some level of heterogeneity was generally noticed in the isolates, irrespective of the sources (i.e. geographical location, shop or product type). This may again indicate varying sources of product contamination (e.g. live birds, water, environment or product handler). Furthermore, changes in bacterial genome may occur in the same serovar through plasmid acquisition/genetic mutations without manifesting phenotype alterations.[43] Interestingly, a particular strain of *S. typhimurium*, isolated from a Mafikeng product namely chicken drumstick, was closely clustered with the positive control strain by both RAPD 1 and RAPD 3 primers. This attests to the efficiency of the *Salmonella* detection method employed in the current study. In spite of the heterogeneity manifested by the isolates however, it may still be concluded that most of the isolates have the same clonal origin.



Lane M, DNA marker (1 kb); Lane 1, SE1; Lane 2, SE2; Lane 3, SE3; Lane 4, SE4; Lane 5, ST1; Lane 6, ST2; Lane 7, ST3; Lane 8, ST4; Lane 9, SN1; Lane 10, SN2; Lane 11, SN3; Lane 12, SN4; Lane 13, S. typhimurium *ATCC 14208 (control).*

Figure 2: Agarose gel (1.2%) electrophoresis showing DNA fingerprinting of non-typhoidal *Salmonella* serotypes generated using (a) random amplified polymorphic DNA (RAPD 1) and (b) RAPD 3 primers.



SE, S. enteritidis*; ST,* S. typhimurium*; SN, S.* newport*; C, control* (S. typhimurium *ATCC 14208*).

All serovars within a serotype showing the same band type were grouped as a sub-serotype (e.g. ST1).

Figure 3: Dendrogram generated based on the RAPD analysis of non-typhoidal *Salmonella* serotypes using (a) random amplified polymorphic DNA (RAPD) 1and (b) RAPD 3 primers.

Conclusion

The results of this study represent the first report on the incidence and prevalent serotypes of NTS in chicken products in the NWP. It is, however, worthy to note that the three NTS serotypes isolated in this study were also the same serotypes identified in 49 NTS cases reported from the NWP to GERM-SA in 2011.[44] This indicates that raw chicken products may serve as major vehicles contributing to foodborne salmonellosis in the NWP. Furthermore, multiple contamination of one sample by different serotypes, and the genetic diversity among isolates of the same or different serotypes (as expressed by RAPD fingerprinting) suggest varying sources of product contamination. This is a cause for public health concern and calls for adequate monitoring and establishment of effective strategies to control contamination along the production/supply chain, in order to safeguard public health. The fact that NTS was not recovered from polonies and viennas (RTE products) in the current study may indicate effectiveness of employed RTE processing techniques in eliminating *Salmonella* contaminants from raw products. This report should provide confidence and reassurance to both RTE producers and consumers.

Acknowledgement

We thank the management of the North-West University, South Africa, for providing the financial and material resources for this work.

Authors' contributions

R.O. was responsible for the project design and execution and for writing the manuscript. S.M. was the project leader and was responsible for the data analysis.

References

1. Morpeth SC, Ramadhani HO, Crump JA. Invasive non-typhi *Salmonella* disease in Africa. Clin Infec Dis. 2009;49:606–611. http://dx.doi.org/10.1086/603553
2. Hohmann EL. Nontyphoidal salmonellosis. Clin Infect Dis. 2001;32:263–269. http://dx.doi.org/10.1086/318457
3. Humphrey TJ. Public-health aspects of *Salmonella* infections. In: Wray C, Wray A, editors. *Salmonella* in domestic animals. Wallingford, England: CABI Publishing; 2000. p. 245–263. http://dx.doi.org/10.1079/9780851992617.0245
4. Herikstad H, Motarjemi Y, Tauxe RV. *Salmonella* surveillance: A global survey of public health serotyping. Epidemiol Infect. 2002;129:1–8. http://dx.doi.org/10.1017/S0950268802006842
5. Vandenberg O, Nyarukweba DZ, Ndeba PM, Hendriksen RS, Barzilay EJ, Schirvel C, et al. Microbiological and clinical features of *Salmonella* species isolated from bacteremic children in Eastern Democratic Republic of Congo. Pediatr Infect Dis J. 2010;29:504–510. http://dx.doi.org/10.1097/inf.0b013e3181cd615a
6. Sigauque B, Roca A, Mandomando I, Morais L, Quinto L, Sacarlal J, et al. Community-acquired bacteraemia among children admitted to a rural hospital in Mozambique. Pediatr Infect Dis J. 2009;28:108–113. http://dx.doi.org/10.1097/INF.0b013e318187a87d
7. Dogru AK, Ayaz ND, Gencay YE. Serotype identification and antimicrobial resistance profiles of *Salmonella* spp. isolated from chicken carcasses. Trop Anim Health Prod. 2010;42:893–897. http://dx.doi.org/10.1007/s11250-009-9504-7
8. Antunes P, Réu C, Sousa JC, Peixe L, Pestana N. Incidence of *Salmonella* from poultry products and their susceptibility to antimicrobial agents. Int J Food Microbiol. 2003;82:97–103. http://dx.doi.org/10.1016/S0168-1605(02)00251-9
9. El Hussein AA, Mayha M, Elmadiena N, Elsaid SM, Mohammed AMS, Muckle CA, et al. Prevalence of *Salmonella enterica* subspecies *enterica* serovars in Khartoum State, Sudan. Res J Microbiol. 2010;5:966–973. http://dx.doi.org/10.3923/jm.2010.966.973
10. Mons G. Economy and imports pose biggest threats to African poultry industry. World Poult. 2010;26:6–7.
11. Smith AM, Gouws A, Hoyland G, Sooka A, Keddy KH. Outbreaks of food-borne disease – A common occurrence but rarely reported. S Afr Med J. 2007;97:1272.
12. National Institute for Communicable Diseases (NICD). *Salmonella Virchow* foodborne illness outbreak. Communicable Diseases Communique. Johannesburg: NICD; 2010. p. 3.
13. Van Nierop W, Dusé AG, Marais E, Aithma N, Thothobolo N, Kassel M, et al. Contamination of chicken carcasses in Gauteng, South Africa by *Salmonella*, *Listeria monocytogenes* and *Campylobacter*. Int J Food Microbiol. 2005;99:1–6. http://dx.doi.org/10.1016/j.ijfoodmicro.2004.06.009
14. Christison CA, Lindsay D, Von Holy A. Microbiological survey of ready-to-eat foods and associated preparation surfaces in retail delicatessens, Johannesburg, South Africa. Food Control. 2008;19:727–733. http://dx.doi.org/10.1016/j.foodcont.2007.07.004
15. Magnani R. Sampling guide. Food and Nutrition Technical Assistance Project (FANTA) Washington DC: Academy for Educational Development; 1997. p. 8–22.
16. Reid A. Isolation and identification of *Salmonella* from food and environmental samples. Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Press; 2009. p. 1–14. Available from: http://doc.mbalib.com/view/d2b3d2bfed24ef386714adab0de8c57c.html
17. Waltman WD. Methods for the cultural isolation of *Salmonella*. In: Wray C, Wray A, editors. *Salmonella* in domestic animals. Wallingford: Cabi Publishing; 2000. p. 355–372. http://dx.doi.org/10.1079/9780851992617.0355
18. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning. 2nd ed. Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Press; 1989.
19. Muyzer G, Teske A, Wirsen CO, Jannasch HW. Phylogenetic relationships of *Thiomicrospira* species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. Arch Microbiol. 1995;164:165–172. http://dx.doi.org/10.1007/BF02529967
20. Khan MAH, Chowdhoury AH, Mosaddik MA, Shajahan M. Multidrug resistant gene (S) harboring on a 20 Kb plasmid in *Salmonella typhi* that causes typhoid-enteric fever. Pak J Biol Sci. 2000;3:911–914. http://dx.doi.org/10.3923/pjbs.2000.911.914
21. Quintaes BR, Leal NC, Reis EM, Hofer E. Optimization of randomly amplified polymorphic DNA-polymerase chain reaction for molecular typing of *Salmonella enterica* serovar Typhi. Rev Soc Bras Med Trop. 2004;37:143–147. http://dx.doi.org/10.1590/S0037-86822004000200006
22. Habtamu-Taddele M, Rathore R, Dhama K, Agarwal RK. Epidemiological characterization of *Salmonella gallinarum* isolates of poultry origin in India, employing two PCR based typing methods of RAPD-PCR and PCR-RFLP. Asian J Anim Vet Adv. 2011;6:1037–1051. http://dx.doi.org/10.3923/ajava.2011.1037.1051
23. McLandsborough LA. Food microbiology laboratory. Boca Raton, FL: CRC Press; 2004.
24. Charimba G, Hugo CJ, Hugo A. The growth, survival and thermal inactivation of *Escherichia coli* O157:H7 in a traditional South African sausage. Meat Sci. 2010;85:89–95. http://dx.doi.org/10.1016/j.meatsci.2009.12.010
25. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4:406–425.
26. Dahal N. Prevalence and antimicrobial resistance of Salmonella in imported chicken carcasses in Bhutan [MSc dissertation]. Chiang Mai/Berlin: Chiang Mai University/ Freie University; 2007.
27. Cohen N, Ennaji H, Bouchrif B, Hassar M, Karib H. Comparative study of microbiological quality of raw poultry meat at various seasons and for different slaughtering processes in Casablanca (Morocco). J Appl Poult Res. 2007;16:502–508. http://dx.doi.org/10.3382/japr.2006-00061
28. Geornaras I, De Jesus A, Van Zyl E, Von Holy A. Microbiological survey of a South African poultry processing plant. J Basic Microbiol. 1995;35:73–82. http://dx.doi.org/10.1002/jobm.3620350204
29. Mead GC. Microbiological quality of poultry meat: A review. Braz J Poult Sci. 2004;6:135–142. http://dx.doi.org/10.1590/S1516-635X2004000300001
30. White DG, Zhao S, Sudler R, Ayers S, Friedman S, Chen S, et al. The isolation of antibiotic-resistant *Salmonella* from retail ground meats. New Engl J Med. 2001;345:1147–1154. http://dx.doi.org/10.1056/NEJMoa010315
31. Bohaychuk VM, Gensler GE, King RK, Manninen KI, Sorensen O, Wu JT, et al. Occurrence of pathogens in raw and ready-to-eat meat and poultry products collected from the retail marketplace in Edmonton, Alberta, Canada. J Food Protec. 2006;69:2176–2182.

32. Zaidi MB, McDermott PF, Fedorka-Cray P, Leon V, Canche C, Hubert SK, et al. Nontyphoidal *Salmonella* from human clinical cases, asymptomatic children, and raw retail meats in Yucatan, Mexico. Clin Infec Dis. 2006;42:21–28. http://dx.doi.org/10.1086/498508

33. Hang'ombe BM, Sharma RN, Skjerve E, Tuchili LM. Occurrence of *Salmonella* enteritidis in pooled table eggs and market-ready chicken carcasses in Zambia. Avian Dis. 1999;43:597–599. http://dx.doi.org/10.2307/1592662

34. Moussa IM, Gassem MA, Al-Doss AA, Mahmoud WAS, Abdel-Mawgood AL. Using molecular techniques for rapid detection of *Salmonella* serovars in frozen chicken and chicken products collected from Riyadh, Saudi Arabia. Afr J Biotechnol. 2010;9:612–619.

35. Chaiba A, Fouzia RF, Abdelkader C, Rachida SB, Mouloud Z. Occurrence of *Salmonella* in chicken carcasses and giblets in Meknès-Morocco. Pak J Nutr. 2008;7:231–233. http://dx.doi.org/10.3923/pjn.2008.231.233

36. Dominguez C, Gomez I, Zumalacarregui J. Prevalence of *Salmonella* and *Campylobacter* in retail chicken meat in Spain. Int J Food Microbiol. 2002;72:165–168. http://dx.doi.org/10.1016/S0168-1605(01)00638-9

37. Yang B, Xi M, Wang X, Cui S, Yue T, Hao H, et al. Prevalence of *Salmonella* on raw poultry at retail markets in China. J Food Prot. 2011;74:1724–1728. http://dx.doi.org/10.4315/0362-028X.JFP-11-215

38. Jay JM, Loessner ML, Golden DA. Protection of foods with low temperature. In: Jay JM, editor. Modern food microbiology, food science text series. 7th ed. New York: Springer Science and Business Media Inc.; 2005. p. 395–409.

39. Galán JE, Ginocchio C, Costeas P. Molecular and functional characterization of the *Salmonella typhimurium* invasion gene *invA*: Homology of *InvA* to members of a new protein family. J Bacteriol. 1992;17:4338–4349.

40. Torpdahl M, Skov MN, Sandvang D, Baggesen DL. Genotypic characterization of *Salmonella* by multilocus sequence typing, pulsed-field gel electrophoresis and amplified fragment length polymorphism. J Microbiol Methods. 2005;63:173–184. http://dx.doi.org/10.1016/j.mimet.2005.03.006

41. Khan AA, Melvin CD, Dagdag EB. Identification and molecular characterization of *Salmonella* spp. from unpasteurized orange juices and identification of new serotype *Salmonella* strain *S. enterica* serovar Tempe. Food Microbiol. 2007;24:539–543. http://dx.doi.org/10.1016/j.fm.2006.09.002

42. Bolton LF, Kelley LC, Lee MD, Fedorka-Cray PJ, Maurer JJ. Detection of multidrug-resistant *Salmonella enterica* serotype *Typhimurium* DT104 based on a gene, which confers cross-resistance to florfenicol and chloramphenicol. J Clin Microbiol. 1999;37:1348–1351.

43. Baggesen DL, Christensen J. Distribution of *Salmonella enterica* serotypes and phage types in Danish pig herds. In: Bech-Nielsen S, Nielsen JP, editors. Proceedings of the Second International Symposium on Epidemiology and Control of *Salmonella* in Pork; 1997 Aug 20–22; Copenhagen, Denmark. Copenhagen: Federation of Danish Pig Producers and Slaughterhouses; 1997. p. 107–109.

44. Crowther-Gibson P, Govender N, Keddy K, Perovic O, Quan V, Von Gottberg A. Non-typhoidal *Salmonella enterica* (NTS). In: Germs-SA annual report. Johannesburg: NICD; 2011. p. 12–14

Measurement of innovation in South Africa: An analysis of survey metrics and recommendations

AUTHOR:
Sibusiso T. Manzini[1]

AFFILIATION:
[1]GreenMatter, Johannesburg, South Africa

CORRESPONDENCE TO:
Sibusiso Manzini

EMAIL:
sibusisomanzini@yahoo.com

POSTAL ADDRESS:
PO Box 2752, Pretoria 0001, South Africa

DATES:
Received: 15 June 2014
Revised: 11 Nov. 2014
Accepted: 04 Mar. 2015

KEYWORDS:
NSI measurement; indicators; innovation survey; non-technological innovation; innovation metrics

HOW TO CITE:
Manzini ST. Measurement of Innovation in South Africa: An analysis of survey metrics and recommendations. S Afr J Sci. 2015;111(11/12), Art. #2014-0163, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140163

The National System of Innovation (NSI) is an important construct in South Africa's policy discourse as illustrated in key national planning initiatives, such as the National Development Plan. The country's capacity to innovate is linked to the prospects for industrial development leading to social and economic growth. Proper measurement of innovation activity is therefore crucial for policymaking. In this study, a constructive analytical critique of the innovation surveys that are conducted in South Africa is presented, the case for broadening current perspectives of innovation in the national policy discourse is reinforced, the significance of a broad perspective of innovation is demonstrated and new metrics for use in the measurement of the performance of the NSI are proposed. Current NSI survey instruments lack definition of non-technological innovation. They emphasise inputs rather than outputs, lack regional and sectoral analyses, give limited attention to innovation diffusion and are susceptible to respondent interpretation. Furthermore, there are gaps regarding the wider conditions of innovation and system linkages and learning. In order to obtain a comprehensive assessment of innovation in South Africa, there is a need to sharpen the metrics for measuring non-technological innovation and to define, account for and accurately measure the 'hidden' innovations that drive the realisation of value in management, the arts, public service and society in general. The new proposed indicators, which are mostly focused on innovation outputs, can be used as a basis for plugging the gaps identified in the existing surveys.

Introduction

Various empirical studies conducted across the world in the past half-century suggest a high correlation between innovation, on one hand, and industrial competitiveness and economic growth, on the other.[1,2] Through economic growth and its direct social impacts, innovation can be linked to improved quality of life as measured, for instance, by the Human Development Index.[3] For this reason, innovation has become an important part of public policy in many countries, including South Africa.[3] For the purpose of this discussion, innovation is:

> *a process of generating, acquiring and applying knowledge for economically and socially beneficial purposes and takes place through efficient unfolding of various learning processes, rather than being determined by the mastery of science and technological knowledge.*[4]

In this broad sense, innovation can be regarded as the deployment of new value to society through the exercise of human ingenuity in any sphere of activity.

It is commonly accepted that innovation processes germinate and develop within what is referred to as innovation systems.[5] These include private and public organisations and other actors that connect in various ways to bring together the technical, commercial and financial competencies and inputs required for innovation. It is on these systems that government innovation policies focus.[6] Comprehensive country reviews on science, technology and innovation demonstrate that innovation is an important policy construct in South Africa.[7,8]

The National Development Plan is the latest government programme of action to recognise the role of innovation in economic development.[9] This plan, apart from pre-existing theoretical and practical justifications, provides a new impetus for a review of the National System of Innovation (NSI) indicators.

The purpose of this paper is to (1) present a constructive analytical critique of the innovation surveys that are conducted in South Africa; (2) reinforce the case for broadening current perspectives of innovation in the national policy discourse; (3) demonstrate the implications of a broad perspective of innovation; and accordingly, (4) propose new metrics for use in the measurement of the performance of the NSI.

This report is based on a qualitative study, conducted using an adaptation of the Delphi method, which was carried out in the following sequence:

1. Meeting of an expert panel to define the problem and map out the process and parameters
2. Analysing the national research and development (R&D) and innovation surveys that are currently carried out in South Africa
3. Critically assessing the metrics, methodologies and outcomes of the current national surveys that pertain to innovation
4. Proposing, based on the above analysis, a broader framework and appropriate metrics for conducting innovation surveys
5. Consolidating inputs from the expert panel

The expert panel provided inputs on an iterative basis, throughout the above stages, after which the final narrative was consolidated.

Measurement of performance of the National System of Innovation

Given the importance that is attached to innovation for a country's economic development, efforts are made in many countries to measure it. The results are used to inform government policy and funding for innovation. South Africa's National Advisory Council on Innovation (NACI) expresses the rationale for regular innovation surveys as follows:

> *Sound measurement of innovation is crucial in policy formulation and implementation, to monitoring spending in this regard, assessing the contribution of innovation to achieving social and economic objectives. Reporting on the measurement of innovation serves to legitimise public intervention by enhancing public accountability.*[10]

Two major surveys are currently used to measure the performance of South Africa's NSI. These are the research and development (R&D) survey and the innovation survey. Both of these surveys are conducted by the Centre for Science, Technology and Innovation Indicators of the Human Sciences Research Council and commissioned by the Department of Science and Technology (DST).

DST is the national government department that is currently responsible for innovation policy in South Africa. R&D surveys have been conducted on an annual basis since the establishment of DST in 2002. The innovation survey is a more recent endeavour with its second report, which spans the period 2005–2008, only published in 2008.

The key indicators used in conducting the R&D and innovation surveys[11,12] are shown in Table 1. The analysis and recommendations made in this report take into account the metrics, methodologies and outcomes of the two surveys mentioned above.

Gaps and shortcomings in current NSI surveys

Poor definition of non-technological innovation

The key indicators in the R&D survey have a clear focus on research and experimental development, which are some of the important input processes towards innovation. The R&D survey provides useful data both on the input and output metrics.

However, an analysis of the key indicators used in the innovation survey raises a concern that they lack clarity and focus. Under organisational innovation, the survey looked into the following dimensions:

- knowledge management systems to better use or exchange information
- major changes to the organisation of work
- external relations with other firms or public institutions[12]

The above indicators provide ample room for confusion because they don't make explicit the innovative aspects of the organisational change that result in value addition. Many organisational changes that are in line with the above may not be innovative at all. For instance, changes in the organisation of work bear no innovation import if they don't deliver value to the clients or the organisation. Furthermore, the organisational indicators outlined above may eschew other forms of non-technological innovations. These are outlined in detail elsewhere in this report.

Emphasis on inputs rather than outputs

The current surveys tend to focus on the inputs rather than the outputs of innovation. The R&D survey and the innovation survey provide fairly comprehensive sets of input factors, such as the level of public resources invested in innovation promotion. However, there are, by comparison, much fewer measures that address the social and economic outputs of innovation. A consequence of this imbalance is the inability of the surveys to provide definitive indications as to whether the investments that are being committed to innovation promotion are yielding the desired results.

Lack of regional and sectoral analysis

By definition, national surveys are meant to provide a country perspective on innovation. However, if the purpose is to provide a meaningful basis for economic and industrial policy, there is a need to provide a provincial or regional analysis in order to take account of local innovation and production systems. If, for an example, a particular trend emerges dominantly in the Western Cape, that result should not conceal the attributes of a Limpopo-based industrial sector which may be experiencing a diametrically opposite phenomenon owing to variant regional dynamics. A regional analysis is important for policy making in South Africa as there are strong geographical differences and inequalities. This level of analysis would enable better understanding of 'specific knowledge spill-overs that occur around certain firms, industries or institutions unique to that space'.[13] The results of the surveys should be reported in a way that would enable provincial and local governments to derive policy implications for their respective regional economies. Furthermore, a geographical analysis of innovation is crucial to achieve a better understanding of innovation at the national level.[14]

Table 1: Key research and development (R&D) and innovation indicators used in current National System of Innovation surveys

R&D survey indicators	Innovation survey indicators
• Gross domestic expenditure on R&D (GERD)	• Rate of innovation
• Gross domestic product (GDP) at market prices	• Characteristics of enterprises covered by the survey
• GERD as a percentage of GDP	• Types of innovations
• Civil GERD as a percentage of GDP	• Product (goods or services) innovation
• Total R&D personnel (FTE)	• Process innovation
• Total researchers (FTE)	• Innovation activities and expenditures
• Total researchers per 1000 total employment (FTE)	• Financial support for innovation activities
• Total R&D personnel per 1000 total employment (FTE)	• Sources of information and cooperation for innovation activities
• Total researchers (headcount)	• Cooperation partners for innovation activities
• Female researchers as a percentage of total researchers	• Effects of innovation
	• Factors hampering innovation activities
	• Intellectual property rights

FTE, full-time equivalent

Similarly, the survey data should provide a sectoral analysis that is based on the different industrial sectors of the economy. This level of analysis is useful for bringing out into the open trends and anomalies that may otherwise be concealed in aggregate figures.

The sectoral analysis would provide the key basis for a more nuanced, fact-based industrial policy. Sectoral innovation[15] is an important conceptual framework for innovation activity that provides a multidimensional, integrated and dynamic view. The descriptions of methodologies in the reports of the surveys conducted by the Centre for Science, Technology and Innovation Indicators provide some provincial and sectoral data but these are not reflected in the published innovation survey reports. For this reason, the innovation survey data cannot be readily accessed by policy makers to inform plans and programmes in specific regions and sectors.

Limited attention to innovation diffusion

Like all other countries, South Africa's social and economic development depends on both local and imported innovations. The impact of those innovations is dependent on the extent of their diffusion across society and industry. Knowledge outputs such as patents that emanate from firms and research institutions might indicate the performance of the individual institutions. However, it is only through implementation that impact can be realised. A more realistic measure of the impact of those outputs on the economy is the extent to which they are dispersed in the relevant industrial and social sectors.

In view of this, it is necessary to measure the extent of diffusion of innovations within the NSI. The current surveys tend to concentrate on measuring technological development through the indicators that look at the different types of innovation outputs within an enterprise. These measures should be supplemented with indicators that provide an estimate of the extent of diffusion of the acquired or produced innovations.

A critical point that needs to be taken cognisance of is the strategic positioning of the country's innovation agenda. South Africa's capacity to be the prime mover at the cutting edge of innovation is limited to only a few areas of knowledge. In most cases, the country is well equipped to focus on attracting and adapting global knowledge to address local needs and conditions. Measures in this regard should then track the level of diffusion of the innovations through various entrepreneurial activities in the economy. The key issue therefore is to strike the right balance between using or attracting existing knowledge and innovations, adapting them to local contexts, on one hand, and pursuing focused research, including on frontier technology when appropriate, on the other.[6]

Respondent interpretation

Both the R&D and innovation surveys are based on the Organisation for Economic Cooperation and Development's framework and methodology.[16] On the methodological aspect of the survey, it is standard procedure that questionnaires are mailed to the respondents. Proper attention is given to providing the necessary information in the instructions and telephonic support, if required.

The questionnaires also provide definitions of the various indicators that are being measured. However, in spite of all these efforts, the questions still leave too much room for the respondents to answer some key questions based on their own misconceptions. For example, when reporting, there may be variances among respondents as to the meaning of 'significant improvements' in relation to products or services.

The wider conditions for innovation

Each institution and organisation within the NSI does not exist in isolation. All the components are embedded in a broader social and economic environment that is subject to both local and international influences. An organisation's capacity to innovate is not only dependent on its internal competencies and circumstances, it is also dependent on the dynamic influences of the factors in the broader environment. Both the R&D and innovation surveys address the outputs and outcomes that were obtained within individual firms and institutions. They do not provide data about the broader conditions for innovation in the country. Government policy has to concern itself with the aggregate social and economic factors that allow or limit innovation. It is essential to formulate evidence-based advice on the general environment in which innovation occurs.

With respect to the wider conditions of innovation, relevant data are available in various forms and reports that are produced by different public institutions such as Statistics SA, the Department of Trade and Industry (dti), the South African Revenue Service (SARS) and the South African Reserve Bank, among others. There is no need to proliferate indicators by instituting additional metrics when useful data are already available. In-depth analysis of relevant existing data, taking into account the R&D and innovation survey reports, should provide advice on the broader conditions of innovation.

Linkages and learning

The NSI is a nationwide network of diverse policies, institutions and organisations that work together in various ways to promote innovation. In essence, the NSI is more about the connections and linkages between its elements than it is about the individual entities.[17] In order to measure the strength of the NSI and the time-series progression of its innovation capacity, it is necessary to focus on the linkages.

The target basis of measurement in current surveys is the individual organisation. The various data sets are then pooled into aggregates, according to the various metrics, to provide an aggregate national picture. It is not always possible to develop an accurate measure of innovation for a system in this way. A dynamic system, such as the NSI, is never a sum of its parts.[17] There is a need to develop indicators that measure the quantity, quality and efficiency of the linkages and networks between the various elements of the NSI. This will contribute towards making better sense of the NSI as a dynamic whole.

An aspect that is closely linked to the discussion of networks and linkages is that of learning. The specific focus here is on technological learning. Marcelle[18] provides a detailed description of how the technological capability building (TCB) approach can be applied at firm level in order to promote organisational learning and capability building. The TCB incorporates the various technical and non-technical components that are essential for sustained learning. The TCB approach can be adapted, using appropriate indicators, to measure the rate at which the NSI is progressing with respect to innovation.

Measurement of non-technological innovation

Traditionally, surveys of South Africa's NSI tended to focus on science and technology based innovation. The consequence of this was the exclusion of other forms of innovation that could also be contributing significantly to the country's economic and social development. The Organisation for Economic Cooperation and Development[19] reports that in many countries there are firms that introduce new products and services without performing any R&D. It is reasonable to expect that the same holds true for South Africa. The recent efforts to broaden the definition of innovation in the innovation survey are therefore appropriate.

A study of the economic sectors in Table 2 suggests that the sectors that are major contributors to R&D, such as agriculture, mining and manufacturing, are relatively small.[20] These sectors also add up to a small percentage of South Africa's total productive economy. Sectors such as finance, real estate, business and government services are the more dominant components of the country's gross domestic product. The latter sectors do not innovate through R&D or in the manner that is anticipated by the existing innovation indicators.

South Africa's economy is replete with examples of innovations that do not follow the traditional R&D-based pathway. The innovations that led to the emergence of Capitec as a force in South Africa's banking sector may not have been captured by the current surveys. The introduction of new methods of registering for and paying out social services to the recipients that have been introduced by the South African Social Services Agency, is another example of an innovation that might fall through the cracks and left unaccounted for by the current innovation surveys. The same could be said of the methods that have been implemented by the South African Revenue Service (SARS) to optimise the recovery of taxes.

During the development of the cellular telephone industry, Vodacom introduced several non-technological innovations, some of which were world leaders.[21] In the biodiversity sector, various innovations are being implemented to protect species from environmental and fabricated threats. It is possible that the current surveys fail to capture these innovations owing to the metrics definitions.

Table 2: South Africa's economic sectors

Seasonally adjusted and annualised quarterly value added (Q4 2011)		
Industry	**Value added (ZAR billion, 2005 prices)**	**% Contribution to total GDP**
Agriculture, forestry and fishing	40.528	2
Construction	58.599	3
Electricity, gas and water	34.734	2
Finance, real estate and business services	404.974	21
General government services	262.627	14
Manufacturing	289.294	15
Mining and quarrying	96.817	5
Personal services	104.676	5
Taxes less subsidies on products	205.653	11
Transport, storage and communication	174.621	9
Wholesale and retail trade, hotels and restaurants	239.367	13
GDP at market prices	1 911.890	100

In order to provide a more comprehensive analysis of the level of innovation within South Africa's NSI, it is recommended that surveys should expand their range of indicators to include other aspects of innovation. These aspects include organisational innovation, soft innovation[22] and social innovation.

Organisational innovation

Although organisational innovation is included in the current surveys, there has not been a comprehensive analysis of its nature and contribution to South Africa's NSI. Under the current definition, any organisational restructuring process, including downsizing, could pass as organisational innovation. Case studies that reflect success stories of novel approaches would form a useful supplement to the reports in order to recognise and promote this type of innovation.

Soft innovation

Soft innovation[22] is the type of innovation that takes place across all sectors of the economy. The arts and other creative industries, such as film and television, are driven by soft or 'artistic' innovation. These industries are major contributors to the country's formal and informal economy. Within the traditional R&D-based sectors, such as industrial manufacturing, soft innovation adds a lot of value in the form of product design, packaging and other aesthetic value additions. Soft innovation is likely to be missed by the current innovation metrics because it does not always constitute a new or significantly improved product. Yet, according to studies conducted in the United Kingdom,[22] this type of innovation may add significant economic value to products and services. In this regard, a proper analysis of the contribution of the Design Institute, and similar initiatives, needs to be conducted.

The work conducted by the National Endowment for Science, Technology and the Arts (NESTA)[22] with respect to soft innovation makes the following important points: Soft innovation is a concept that reflects aesthetic changes; soft innovation and technological innovation are interrelated; only soft innovations with high market share are considered significant; non-traditional metrics are needed to measure soft innovation; there are high rates of soft innovation in the creative industries; soft innovation is significant outside the creative industries too; missing soft innovation gives a biased account of total innovation activity; intellectual property rights are an important area for policy; sub-optimal levels of soft innovation may justify government intervention; the commercial benefits of soft innovation may be high; government policy must embrace all innovation activities, not just technological or scientific.

The above findings concerning the role of soft innovation in the context of overall innovation activity within the economy hold true for South Africa's NSI.

Social innovation

In addition to the types of innovation discussed above, there is a further more encompassing form of innovation referred to in this report as social innovation. Social innovation is a value-adding outcome that emanates from a variety of ways that involve interactions between people. The above innovation types tend to focus on products and services. However, a deep analysis of South Africa's economic activity suggests that value can be created through the quality of human–to-human contact. While the nature of social innovation is difficult to define and measure, its social and economic impact is undoubted. It should therefore be reflected in the key metrics that are used to measure innovation activity within the NSI.

Figure 1 illustrates a comprehensive perspective of innovation that includes the types that are discussed in this report. It emphasises that the different types of innovation are not discreet but overlap and often complement one another in value creation.

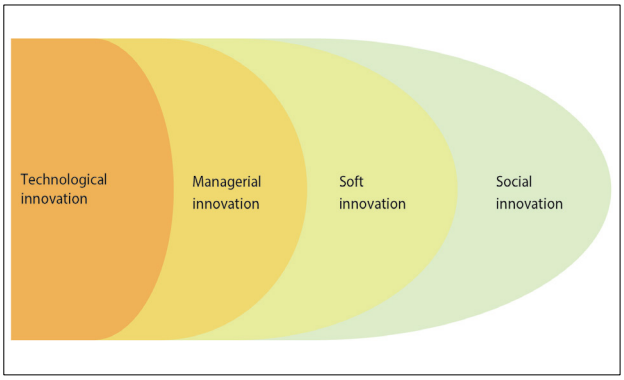


Figure 1: An illustration of the overlapping types of innovation.

In order to account for the full spectrum of the types of innovation that are active in the NSI, it is recommended that consideration be given to the following:

- All the above types of innovation must form a component of indicators of innovation within the NSI, in the longer term.
- Further work needs to be done in order to formulate rigorous definitions of soft and social innovation to build on the proposals presented in this work.
- There is a need to develop consensus around the meaning and application of social innovation in the South African context.
- Studies should be conducted to explore the dynamics of the interplay, if there is any, between technological and non-technological innovation.
- Appropriate indicators should be identified and used to measure the baseline and subsequent development trends for all types of innovation.
- The NSI indicators should adopt flexible and appropriate ways to report on innovation, including using non-quantitative reports, as necessary.

Proposed new indicators

A key message is that current surveys of the NSI do not cover the full spectrum of innovation activities that are important for South Africa's social and economic well-being. The above discussion outlines the various forms of innovation that have been missed by the R&D-based surveys that are currently in place. This report proposes a broader view of innovation that encompasses knowledge domains beyond science, engineering and technology. Furthermore, there is a need to focus more strongly on the innovation metrics on the demand-side of the economy, i.e. those that measure the actual uptake of knowledge-intensive factors, rather than merely measuring supply.

Four principles guide the approach adopted by this report in formulating its proposals:

- *Parsimony* – For practical reasons, it is absolutely important to limit the number of indicators. By their very nature, indicators are, at best, proxies that are used to draw credible conclusions about the behaviour of the NSI. For this reason, as few indicators as possible are selected to explain as much as possible about the NSI.
- *Complementarity* – The aim of this work is not to introduce a new regime of metrics of innovation. Instead, it is to identify gaps in the existing surveys and propose ways to fill them. The proposed indicators are meant to complement the R&D and innovation surveys. Similarly, the use of other appropriate national and international surveys that collect data that are applicable to innovation measurement are recommended.
- *Optimisation* – There are strong indications that where relevant data are available (see examples in Table 3), they are not always subjected to sufficient analysis in order to provide better understanding of innovation activities in the country. This report therefore calls for more diligence in the analysis of innovation-relevant data from all credible sources.
- *Output focus* – The new indicators that are proposed below aim to shift the focus of measurement towards innovation outputs. Input measurements are important to inform policy, for example, with regard to investment decisions. However, as discussed above, a healthy balance needs to be struck between input and output indicators.

Some of the indicators that are proposed below are drawn from a pilot study that was commissioned by the Cooperation Framework on Innovation Systems between Finland and South Africa (Cofisa)[23] while others are new conceptions that emerged from this study.

The proposed new indicators are organised around the five themes that are at the centre of innovation activity. These are:

- knowledge demand indicators
- knowledge mobilisation indicators
- knowledge application indicators
- knowledge flow indicators
- social impact indicators

Knowledge demand indicators

The focus here is on the actual knowledge that drives the economy. The indicators that are chosen address directly the attractors of skills in business and knowledge centres. These are the knowledge and technology intensity of manufactured goods, and the contribution of knowledge to the manufacturing trade balance.

Knowledge and technology intensity of manufactured goods

'Exports of knowledge intensive goods' is a meaningful indicator of the innovation capacity of a country. Such goods form an important component of the economy in that they provide local needs, thus substituting imports, and are able to compete successfully in the global market.

To measure the level of knowledge and technology input in the export products, international guidelines are used. The data are collected by SARS and reported by the dti. For further details, see Cofisa[23].

Contribution to the manufacturing trade balance

The sectoral trade balance indicates the structural strengths and weaknesses of the different sectors of an economy. It is measured relative to the performance of total manufacturing trade. A sectoral analysis of the trade balance is an incisive measure in that it can focus on priority industries. Their respective performances can be measured relative to overall manufacturing data, to indicate comparative performance, irrespective of whether aggregate manufacturing is growing or falling. The data are collected by SARS and reported by the dti.[23]

Knowledge mobilisation indicators

Since the 1996 White Paper on Science and Technology, there has been a recognition that the growth of innovation in South Africa will depend on the continuous improvement of skills and education levels in the country. These indicators address not just the education levels in society but also the skills that are accessible for deployment in the economy. The key indicator here is participation in lifelong learning.

Participation in lifelong learning

The South African government recognises the importance of lifelong learning as a means towards social and economic development. This is evidenced by the provisions for work-based learning in the Skills Development Act and various other supporting institutions and instruments, such as the Sector Education and Training Authorities (SETAs) and the Skills Development Levy. Many employees use the tertiary education institutions to further their skills. Participation in lifelong learning is a good indicator of a knowledge economy that demands and promotes continuous knowledge and skills acquisition.

The annual General Household Survey that is conducted by Statistics South Africa can supply the baseline data required for individuals 20 years old and above. Employer-based reports, such as the skills development plans, as well as the reports of SETAs and other training authorities, can provide annual data that can be used to establish a baseline and thereafter track trends.

Knowledge application indicators

Current indicators tend to focus on the supply side of knowledge outputs. The indicators that are proposed under this heading complement this by looking at the knowledge outputs that are actually absorbed into the economy. These include the licensing of patents, and entrepreneurship

Licensing of patents

Patent output is used in current surveys as a proxy for knowledge intensity. It is proposed here that the licensing of patents and other intellectual property for industrial application is perhaps the more appropriate indicator. This is because it indicates the intellectual property that is actually engaged productively in the economy. This is in stark contrast to patents that are filed and never put to practical use.

Current R&D and innovation surveys can be used to collect data on patenting licensing. The patents involved here are those that are developed exclusively or jointly by South Africans. The data can be compared to the licensing information that can be supplied by the relevant patent offices.

Entrepreneurship (company registrations)

Enterprise creation is an important indicator of economic activity. A vibrant and growing knowledge economy will be reflected in the registration of new business ventures. Alternatively, the company de-registrations can, on aggregate, hint at a reduction of economic activity.

It is proposed that the registration of new companies with the Companies and Intellectual Property Commission, as a percentage of registered companies, can be used as an indicator of entrepreneurship and knowledge intensity. The data can be obtained from this commission.

This may require further development of registration categories in order to adequately inform innovation activity.

Knowledge flow indicators

The strength of the NSI is dependent on the extent to which knowledge flows and is converted into practical use within the networks of institutions and organisations. The diffusion of knowledge and innovation enables the widening of impacts to society. The indicator that provides an alternative approach to measure this aspect is innovation networks.

Innovation networks

The strength of an innovation system is dependent on the quality, quantity and efficiency of the flow of key information within the elements that make up the NSI. A disconnected system has a poor chance of developing. However, a vibrant and well-connected system, wherein all the components interdependently share and cooperate, is likely to result in an aggregate systemic increase in innovation intensity.

This indicator is an important missing link in the current surveys. Appropriate measures need to indicate the rate of change of connectivity within all the key players in the NSI. This could be done by determining the number and extent of partnerships, cooperative projects or joint ventures with other entities in the NSI. The degree of communication and networking could be measured through determining the number and sizes of knowledge-sharing networks an organisation is participating in. A higher weight should be given to partnerships and networks that involve different types of institutions. Regional innovation forums could be tasked with documenting local activity and providing indications of participation. This data could be obtained by adding the necessary questions in the current innovation survey.

Social impact indicators

An important argument made in this report is the need to extend the measurement of innovation activity to include areas of activity beyond the traditional R&D sectors. This category could also accommodate measures of innovation activity in the public sector,[24] as they are developed. In addition, there is a need to determine, with better accuracy, the social impact of innovation. The social impact indicators discussed are social cohesion, social impact innovations and innovations in the public sector.

Social cohesion

Social cohesion, particularly in a South African context, is difficult to define or measure. However, it is reasonable to accept that, depending on how it is used, an increase in innovation, and the benefits that accrue as a result, will manifest in social cohesion. This could be in the form of a more inclusive society, with low rates of income inequality, poverty and other social maladies and divides. Social cohesion, therefore, can be regarded as both an instrument and a goal for innovation policy.

Social impact of innovations

The ultimate measure of innovation impact is the rate of social development that accrues to society. Some innovations may provide direct social impact even when they may not be commercial successes, such as those that pertain to improved public service delivery. The innovations that have an economic impact may result in social benefits through improved wealth. Measures of social impact, such as the Human Development Index, may serve as an indicator of the knowledge intensity of the country. This indicator is currently being used although the focus is mostly on the ranking attained by the country, for instance in the indexes of the World Bank and the Institute for Management Development (IMD). While the

Table 3: Summary of proposed new innovation indicators

Category	Proposed indicator	Input/output	Primary/ secondary	Data source
Knowledge demand	Knowledge and technology intensity of manufactured goods	Output	Primary	The dti; SARS
	Contribution to the manufacturing trade balance	Output	Primary	The dti; SARS
	Foreign student population in higher education	Output	Secondary	Higher Education Management Information System (HEMIS)
	Employment of tertiary level graduates	Output	Secondary	Stats SA
Knowledge mobilisation	Participation in lifelong learning	Input	Primary	Stats SA; SETAs; company HR data
	Education system resources	Input	Secondary	Education Management Information System (EMIS)
	Access to ICTs	Input	Secondary	Stats SA
Knowledge application	Licensing of patents	Output	Primary	Innovation surveys
	Entrepreneurship	Output	Primary	CIPC
	Economic impact of innovations	Output	Secondary	The dti; SARS; Stats SA
Knowledge flows	Foreign direct investment networks	Input	Secondary	SARS; Stats SA
	Innovation networks	Input and output	Primary	Adapted Innovation Survey
	International flows of human resources	Output	Secondary	DHA; company HR data
Social impact	Social cohesion	Input and output	Primary	Stats SA; other existing sources
	Social impact of innovations	Output	Primary	World Bank; IMD
	Innovation in the public sector	Input and output	Secondary	CPSI; Innovation Survey

comparisons may provide useful insights, the focus should be more on how and why the total score changes over time. This data should be complemented by identifying the quality and quantity of the innovations that impacted society. Broader social surveys among communities and individuals should elicit this data.

Innovation in the public sector

Government plays an important role in society in creating enabling policy environments for social and economic development as well as in rendering essential services. Government departments, other public institutions at national, provincial and local levels, have to adapt to increasing demands and strive to attain higher levels of efficiency. The public service should be encouraged to innovate in order to meet these challenges. The measurement of overall innovation activity in the country should take into account innovation in the public sector.

The key indicators of public sector innovation should include various innovation activities, as defined elsewhere in this report, as well as their contribution towards service delivery. The current innovation survey can be adapted, taking into account the proposals made here, in order to measure public sector innovation. This work could be done in association with the Centre for Public Service Innovation (CPSI) in order to provide a comprehensive view of innovation and its impact in the country.

Primary versus secondary indicators

In keeping with the need to keep the number of indicators to the minimum, the proposed indicators are divided into primary and secondary indicators (Table 3). The primary indicators will provide key data that will fill the gaps that exist in the current R&D and innovation surveys, as discussed in this report. The secondary indicators are regarded as important metrics to provide supporting data for a more comprehensive understanding of innovation activity in the country. It is recommended that a further process should unfold to refine this distinction.

The primary indicators that are proposed for immediate adoption are:

- knowledge and technology intensity of manufactured goods
- contribution to the manufacturing trade balance
- participation in lifelong learning
- licensing of patents
- entrepreneurship
- innovation networks
- social cohesion
- social impact of innovation

Conclusion

The R&D and innovation surveys that are currently conducted in South Africa provide necessary and important information about the growth and development of elements of the NSI. The data that they produce could be optimised through more rigorous analysis. Regional and sectoral analyses would provide important nuances that carry useful import for public policy at national, provincial and local government levels. The international benchmarks that are applied in the metrics and methodologies that are followed to conduct the studies permit useful comparisons to be made to continuously stretch development targets.

However, in order to get a comprehensive assessment of innovation in South Africa, there is a need to sharpen the metrics for measuring non-technological (e.g. the social impact indicators) innovation. The practice of innovation in the South African economy takes place well beyond the confines of science and technology, which have dominated the innovation policy discourse since the advent of the NSI. There is a need to define, account for and accurately measure the 'hidden' innovations that drive the realisation of value in management, the arts, public service and society in general.

The current innovation indicators need to be supplemented in keeping with the broader view of innovation presented here. The new proposed indicators, which are mostly focused on innovation outputs, can be used as a basis for plugging the gaps in existing surveys.

Acknowledgements

I thank Dr Neville Commins, Dr Gillian Marcelle and Dr Takalani Rambau for making invaluable contributions to the study and the report; ASSAf for coordinating the expert group; NACI for supporting the study; and the individuals who formed the expert group that participated in the study.

References

1. Cameron G. Innovation and economic growth. CEPDP 277. London: Centre for Economic Performance, London School of Economics and Political Science; 1996.
2. Juma C, Yee-Cheong L. Innovation: Applying knowledge in development. London: Earthscan; 2005.
3. World Economic Forum (WEF). The global competitiveness report: 2010–2011. Geneva: WEF; 2010.
4. Marcelle GM, Nkhumise L, Vawda S. Making innovation and science relevant for poor communities: The case of a water management project in South Africa. Paper presented at International workshop: New models of innovation for development; 2013 July 4–5; Manchester, UK.
5. Lundvall B, editor. National innovation systems: Towards a theory of innovation and interactive learning. London: Pinter Publishers; 1992.
6. World Bank. Innovation policy: A guide for developing countries. Washington, DC: World Bank; 2010. http://dx.doi.org/10.1596/978-0-8213-8269-1
7. Organisation for Economic Cooperation and Development (OECD). Review of South Africa's innovation policy. Paris: OECD; 2007.
8. Department of Science and Technology (DST). Ministerial review on science, technology and innovation. Pretoria: DST; 2012.
9. National Planning Commission. National development plan 2030: Our future – make it work. Pretoria: National Planning Commission; 2011.
10. National Advisory Council on Innovation (NACI). South African science and technology indicators. Pretoria: NACI; 2010.
11. Department of Science and Technology (DST). National survey of research and experimental development (2009/2010 fiscal year): Main results. Pretoria: DST; 2013.
12. Human Sciences Research Council (HSRC). South African innovation survey: Main results 2008. Pretoria: HSRC; 2011.
13. Cunningham S. The fundamentals of innovation system promotion for development practitioners. Mesopartner Monograph 5. 3rd ed. Pretoria: Mesopartner; 2012. Available from: http://www.mesopartner.com/fileadmin/user_files/books_monographs/MM05_Ed_3_edited_The_fundamentals_of_innovation_system_diagnosis_and_promotion_Comic_with_front___back_page.pdf
14. Asheim BT, Gertler MS. The geography of innovation: Regional systems of innovation. In: Fagerberg J, Mowery DC, Nelson RR, editors. The Oxford handbook of innovation. Oxford: Oxford University Press; 2005. p. 291–317.
15. Malerba F. Sectoral systems of innovation and production. Res Policy. 2002;31(2):247–264.
16. OECD/Eurostat/European Union. Proposed guidelines for collecting and interpreting technological innovation data: Oslo manual. Paris: The Measurement of Scientific and Technological Activities, OECD Publishing; 1997. http://dx.doi.org/10.1787/9789264192263-en
17. Manzini ST. South Africa's national system of innovation: A systems, chaos and complexity perspective. Lambert: Herstellung; 2012.
18. Marcelle GM. Technological learning: A strategic imperative for firms in the developing world. Cheltenham: Edward Elgar Publishers; 2004. http://dx.doi.org/10.4337/9781845426910
19. Statistics South Africa (Stats SA). Gross domestic product: First quarter 2013. Statistical release P0441. Pretoria: Stats SA; 2013. Available from: http://www.statssa.gov.za/publications/P0441/P04411stQuarter2013.pdf
20. Knott-Craig A, Afonso E. Second is nothing. Johannesburg: Pan Macmillan SA; 2009.

21. National Endowment for Science, Technology and the Arts (NESTA). Soft innovation: Toward a more complete picture of innovation change. Research report. London: NESTA; 2009.

22. Cooperation Framework on Innovation Systems between Finland and South Africa (Cofisa). Southern African science, technology and innovation scoreboard 2008. San Francisco, CA: Cofisa; 2008. http://dx.doi.org/10.1016/j.urology.2008.08.258

23. National Endowment for Science, Technology and the Arts (NESTA). Innovation in public sector organisations: A pilot survey for measuring innovation across the public sector. Index report. London: NESTA; 2011.

24. National Endowment for Science, Technology and the Arts (NESTA). Innovation in public sector organisations: A pilot survey for measuring innovation across the public sector. Index report. London: NESTA; 2011.

Paper-based smart microfluidics for education and low-cost diagnostics

AUTHORS:
Suzanne Smith[1]
Klariska Moodley[1]
Ureshnie Govender[1]
Hao Chen[1]
Louis Fourie[1]
Sibusiso Ngwenya[1]
Shavon Kumar[1]
Phumlani Mjwana[1]
Hastings Cele[1]
Mesuli B. Mbanjwa[1]
Suretha Potgieter[1]
Trudi-Heleen Joubert[1]
Kevin Land[1]

AFFILIATION:
[1]Materials Science and Manufacturing, Council for Scientific and Industrial Research, Pretoria, South Africa

CORRESPONDENCE TO:
Suzanne Smith

EMAIL:
ssmith@csir.co.za

POSTAL ADDRESS:
Materials Science and Manufacturing, Council for Scientific and Industrial Research, PO Box 395, Pretoria 0001, South Africa

DATES:
Received: 13 Oct. 2014
Revised: 02 Mar. 2015
Accepted: 06 Mar. 2015

KEYWORDS:
diagnostics; printed electronics; biosensors; micro paper-based analytical devices; μPADs

HOW TO CITE:
Smith S, Moodley K, Govender U, Chen H, Fourie L, Ngwenya S, et al. Paper-based smart microfluidics for education and low-cost diagnostics. S Afr J Sci. 2015;111(11/12), Art. #2014-0358, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20140358

Current centralised healthcare models pose many challenges, particularly for developing countries such as South Africa, where travel and time costs make it difficult for patients to seek healthcare, even when urgently needed. To address this issue, point-of-care (PoC) tests, which are performed at or near the site of clinical care, have gained popularity and are actively being developed. Microfluidic systems, in which small volumes of fluids can be processed, provide an ideal platform on which to develop PoC diagnostic solutions. Specifically, the emerging field of paper-based microfluidics, with advantages such as low-cost, disposability and minimal external equipment requirements, provides unique opportunities for addressing healthcare issues in developing countries. This work explores the field of paper-based microfluidics, with step-by-step instructions on the design, manufacture and testing processes to realise paper-based devices towards diagnostic applications. Paper-based microfluidic and electronic components are presented, as well as the integration of these components to provide smart paper-based devices. This serves as an educational tool, enabling both beginners and experts in the field to fast-track development of unique paper-based solutions towards PoC diagnostics, with emphasis on the South African context, where both the need for and impact of these solutions are great.

Introduction

The need to move away from a centralised healthcare system to one which is more home/patient based is becoming more urgent.[1] In developing countries, patients have to travel long distances to obtain medical attention and are often required to make return visits to either receive test results or medication, as well as to monitor their response to treatment. However, with the cost of healthcare and transportation being higher than the average household can afford, patients only seek professional healthcare as a last resort, by which time their illness may have progressed to dangerous levels. Furthermore, patients seldom make follow-up visits once they receive short-term relief from their symptoms, resulting in increased risk of spreading disease. Similarly for environmental monitoring, the risk of the disease spreading is increased by the typically low rate of monitoring, which is affected by the high costs and time delays associated with sampling, sample transportation and specialised testing.[2-5]

Point-of-care (PoC) tests are laboratory diagnostic assays performed at or near the site where clinical care is delivered and may therefore be used in both a first world laboratory and the most remote corner of the developing world. In the latter case, these tests have the ability to overcome many of the disadvantages of centralised testing facilities because they are low-cost, provide faster test results, and negate the need for samples/patients to be transported over long distances. As a result, PoC tests promote regular patient and/or environmental sample testing, enabling more effective health monitoring across large geographical regions and populations. This early warning system serves to significantly reduce the impact of disease.

Microfluidics is a field in which devices made from materials such as silicon, glass or polymers are used to automate standard desktop laboratory processes on a micro scale. In these devices, small volumes of fluids are manipulated and transported through micron-sized channels, and the various processes that occur on board the device are monitored using microscopes and computer software. Microfluidic devices – also called lab-on-chip devices – have been identified as a key technology for the development of PoC diagnostic tools and have found application in the fields of blood chemistry, immunoassay, nucleic-acid amplification testing and flow cytometry.[1] However, because of their requirement for supporting equipment, as well as their inability to be mass-produced at low-cost, their use in low resource settings has proven difficult.[6]

The World Health Organization (WHO) has introduced the ASSURED criteria around which diagnostic tools should be designed. The criteria states that diagnostic tests should be Affordable, Sensitive, Specific, User-friendly, Rapid and Robust, Equipment-free and Deliverable to end-users.[7] To overcome the limitations of microfluidics according to the WHO ASSURED criteria, Harvard University introduced the field of paper-based microfluidics, where the silicon or glass substrate is replaced with a paper substrate.[8]

Microfluidic paper-based analytical devices (μPADs)[9] – or microfluidics 2.0[10] – have received substantial attention in the recent past as an attractive solution for implementing low-cost diagnostics at the point-of-care, because the technology has a number of desirable features, both from a technical and a social impact perspective. Coupled with the use of a low-cost paper substrate, paper-based microfluidic devices enable sample flow through capillary action, negating the need for pumps or electricity. The use of microscopes can also be eliminated by employing colour change reactions identifiable by eye, a method preferred at the PoC as it requires minimal user interpretation. Paper provides the ideal background contrast for evaluating colour changes. Moreover, reagents can be stored on paper in desiccated form which helps to prolong the shelf life of the devices. This allows them to be delivered in bulk to remote testing locations. Paper is thin and lightweight, so the devices are highly portable and easy to transport.

Although paper-based microfluidic devices have become analogous with common paper tests like the pH test strip and home based lateral flow pregnancy tests,[11] these devices also offer improvements to lateral flow technology by enabling multiplexed analyses, without making the test results ambiguous. Various applications for paper-based microfluidic devices, as well as various manufacturing techniques, have been reported in several reviews.[11-13] By employing less specialised paper types for manufacture and a more efficient high throughput fabrication process,

the overall cost of the tests are reduced.[14] Wax printing manufacturing[14,15] has become the preferred choice of manufacture because of its simplicity and ability to enable mass production at low cost. Three-dimensional (3D) paper microfluidics were introduced by Martinez et al.[16], wherein the unique multiplexing ability of these devices was demonstrated. Later, Liu and Crooks[17] improved the fabrication method by using the ancient Japanese technique of folding (origami) to stack together the various layers of the 3D device.

Furthermore, combinations of microfluidics, micro-electronics and micro-optics provide remarkable opportunities for advancing the biosensor industry, for example, the Clearblue Digital Pregnancy Test[18] is a disposable lateral flow test powered by a small button-cell battery and performs an optical measurement, while displaying semi-quantitative results on an LCD screen. Printed electronics[19] have also received significant recent attention for the ability to enhance diagnostic device functionality without compromising the attractive features of paper-based microfluidics.

Combining smart biosensors such as these with the simplicity, portability, disposability and low-cost multiplex analysis of microfluidic paper test devices, allows for the rapid high sensitivity analysis and real-time monitoring required from PoC diagnostics. Moreover, integrating all the functions on a single paper substrate would enable a quantitative analysis result without requiring an external reader – an objective which may be fulfilled by leveraging the rapid expansion of printed electronics technology.

The digital and additive nature of printing processes, the range of materials that can be formulated as inks, and the possibility to cater to various production scales, ranging from prototype to large-scale, are some of the key factors that enable the deployment of printing processes in electronics fabrication. Printing is accessible to a broader audience, providing an energy-saving, environmentally-friendly alternative that can dramatically reduce the volume of raw materials consumed and waste generated when compared to conventional production methods. Printed electronics is still an emerging industry, with several key components at research level,[19,20-23] but it is envisioned that cost-effective, autonomous, portable, environmentally friendly, and disposable printed biosensor systems for telecommunications connected PoC testing will become a reality, providing sensitive and quantitative multiplexed results.

Compliance with many of the WHO ASSURED criteria positions the paper-based electronic microfluidics platform as a front-runner in achieving the goal of rapid, inexpensive PoC diagnostics. In addition, South Africa is ideally located to understand first-hand the challenges of implementing effective diagnostic devices for addressing healthcare issues in developing countries, enabling South Africans to be at the forefront of developing effective paper-based diagnostic solutions.

Although there have been many publications focusing on specific manufacturing methods and applications of paper-based microfluidics, to our knowledge, there are currently no papers that have been published with the specific goal of assisting new users in the field. Papers have been published where the concept of paper-based fluidics has been used for educational purposes,[24] but full workflows have not been given.

This paper presents the basic steps for getting started with integrating paper-based fluidics with electronic functions for people with little to no experience in microfluidics or manufacturing. Step-by-step guidelines for the fabrication of an origami 3D paper-based microfluidic device using the wax printing technique, as well as the addition of electronic features are provided. The detailed protocols in this paper allow first-time users to fast-track their development work, whether it is for the creation of life-saving diagnostics or for the creation of educational tools. The aim is to evoke interest and insight into paper-based electronics-enhanced microfluidics in people from a variety of technical backgrounds, enabling the generation of creative, multi-disciplinary solutions applicable to health and the environmental challenges of the developing world. As an immediate application, the article introduces a paper-based device which may be utilised for educational projects, advancing the principles of paper microfluidics and relevant manufacturing techniques. An important intention is to create awareness and promote conversations that will further the applications of this field, and augment their impact on the world.

Component development

In order to get started with paper-based microfluidics, individual components are first developed and then integrated into a functional device. The sections below take the reader through this process in a step-by-step manner.

Origami three-dimensional fluidic network

A 3D paper microfluidic device was designed and implemented using paper folding or origami principles.[17,25] Initially, the size of the 3D microfluidic device must be determined. It should be sized to fit comfortably into the user's hand, and ensure that test result signals are easily visualised with the naked eye. For this demonstration, a 50 mm x 50 mm device was used. For wax printing, the distance between adjacent lines in the design should be no smaller than 2 mm. This spacing allows for wax spreading during melting, and prevents the channels from becoming clogged and consequently preventing fluid flow.

The number of layers required in a 3D device will depend on the functionality required and the complexity of the detection mechanism. For example, should there be a need to only add reagents after the sample is loaded onto the device, then a layer must be incorporated into the device to serve this purpose. However, we have found it easier to either dry these reagents at the test result signal position itself (the final layer of the device), or mix them directly into the sample (should this be acceptable for the chemical reaction). For the purposes of demonstration, the design of a three-layer device is shown in Figure 1. In this device, a single sample can be used to perform three sets of analyses, demonstrating that 3D paper-based microfluidic devices are capable of performing multiplexed tests.

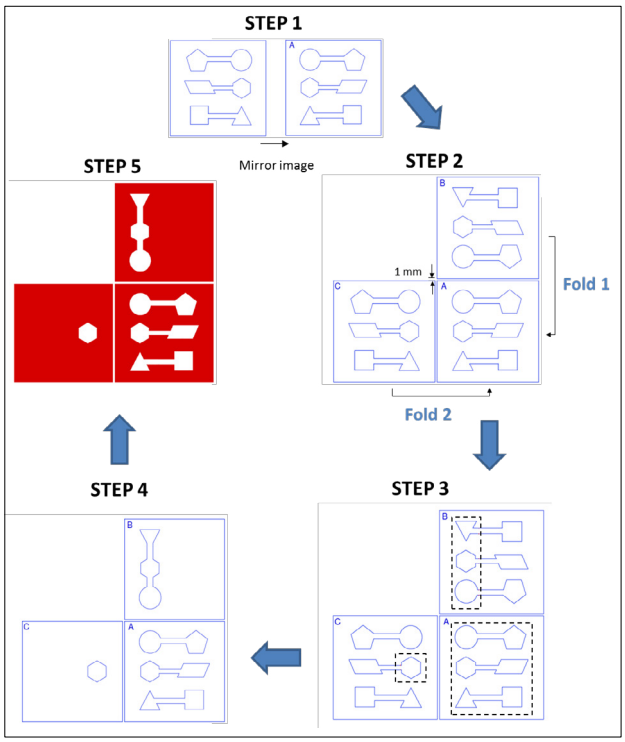


Figure 1: Illustration of design steps carried out to create an origami 3D paper fluidic device. STEP 1 shows the paper device design outline, which is mirrored to create layer A. STEP 2 illustrates further mirroring of the design to create two extra layers, B and C, for a three-layered paper device. STEP 3 highlights the vertical fluidic flow pathways required through the device (dotted lines) for the fluid to be introduced to the top layer, C, flow through to the middle layer, B, and onto the bottom layer, A. By deleting the design sections that fall outside the dotted lines, the final design is realised, as shown in STEP 4. The design is then shaded in the areas where the wax will be printed, leaving the fluidic flow pathways open, as shown in STEP 5.

Devices are designed using a standard computer aided design program. The design procedure for a 3D origami-type microfluidic network is illustrated in Figure 1, showing how layers are folded from a 2D design into a 3D stack, while performing the steps that follow.

Step one

The shape of the test result signal(s) must be determined. Result signals can be words (YES/NO), icons (such as a tick or cross), or any other shape deemed relevant. Once the test result signal layout design is created, a mirror image of it must be used as a template for the design steps, as illustrated in STEP 1 in Figure 1. This is because once assembled, the user will turn over the device in order to read the result. In mirroring the design around itself, the original design remains centralised and will form the bottom layer of the 3D device after folding. This bottom layer of the device is depicted as layer A in Figure 1, and contains three test result signal shapes – the pentagon, the parallelogram and square – which will be viewed by the user after running the test. The other shapes and connection lines are part of the flow path, and will be discussed in further steps.

Step two

The number of mirrored images that should be created depends on the number of layers needed in the device. Because three layers are being used for this demonstration, the designed bottom layer is mirrored twice, as shown in STEP 2 of Figure 1. Mirrored images should be spaced 1 mm from each other,[26] creating inactive spaces where origami folds can be made during the assembly of the device.

Step three

The vertical fluidic connections must be determined by verifying how the sample fluid will flow from the sample inlet at the top of the device to test result signal(s) at the bottom of the device. In Figure 1, layer C forms the uppermost sample introduction layer, layer B is the intermediate layer, and layer A is the bottom test result layer. The fluidic path design for this demonstration is enclosed in dashed lines of STEP 3 in Figure 1. As a result of the horizontally positioned device layers being in direct contact with each other after folding, the sample moves vertically downward from layer C, through to layer B, and down to layer A. Because layer B will be folded down directly onto layer A, fluid will move vertically downward between these two layers through the enclosed intermediate shapes – the circle, the hexagon, and the triangle. Within the dashed line, the sample will be loaded from the top into the enclosed hexagonal region on layer C. Layer C will be folded down directly onto layer B, so fluid will move vertically downward from the enclosed hexagonal region on layer C into the same shape enclosed region of layer B. Deleting the areas of the design not enclosed by dashed lines on layers B and C completes this step.

Step four

The horizontal fluidic connections must be incorporated into the design. STEP 4 in Figure 1 shows the creation of the horizontal flow paths or channels required on layer B. Most of the enclosed intermediate regions of layer B – the circle, the hexagon, and the triangle – become filled with sample as a result of horizontal capillary fluid movement. Also, verify the fluidic connections between the sample inlet on the intermediate shapes of layer A – the circle, the hexagon and the triangle – and the test result signal shapes, the pentagon, the parallelogram and square.

Step five

The design is finalised by shading the areas of the device not contributing to the fluid flow path, which represent the areas where hydrophobic wax will be printed. The wax will help to contain the fluid flow to only where it is required. The design as shown in STEP 5 of Figure 1 is now ready for printing, melting and assembly.

For printing, a Xerox Colorqube 8870DN (Xerox, Bytes Technology Group, Johannesburg, South Africa) solid ink printer was used, as shown in Figure 2a. To date, wax printing has been the technique most commonly used for the manufacture of paper-based microfluidic devices, although various other techniques have also been used to create paper-based microfluidic devices using wax. Lu et al.[15] used wax pens to draw on the desired hydrophobic patterns, Dungchai et al.[27] used a wax screen printing technique and Songjaroen et al.[28] created devices using a wax dipping technique. While all these methods require minimal investment to perform, they suffer from poor reproducibility and are not well suited for mass-production. Using a solid ink printer has therefore proven most efficient as it provides better design resolution, is less labour intensive and provides a higher manufacturing throughput. Solid ink printers function by heating up blocks of wax (considered the ink cartridges in these printers) and depositing the molten wax onto the paper. The wax components used in this work were Cartridge-Free ColorQube inks, compatible with the 8570/8870 Xerox ColorQube printer range (Xerox, by Bytes Technology Group, Johannesburg, South Africa). The available colours are magenta, yellow, black and blue, which can be combined in different ratios to produce any colour required. Once printed, the ink cools rapidly, and is then ready for use. Chromatography paper (Whatman Chromatography paper no. 1, Merck, Johannesburg, South Africa) has been the substrate of choice for creating paper-based microfluidic devices.[9] These papers are low cost, widely available and suitable for most chemical assays.

After the design is printed, it is melted on a hot plate at a temperature of 255 °C for 1 min as depicted in Figure 2b. These parameters were found to be optimal in a separate study performed by Govindasamy et al.[29] During melting, the wax layers on the paper surface are melted through the cross section or thickness of the paper. The wax barriers then serve as hydrophobic walls that guide the movement of fluids

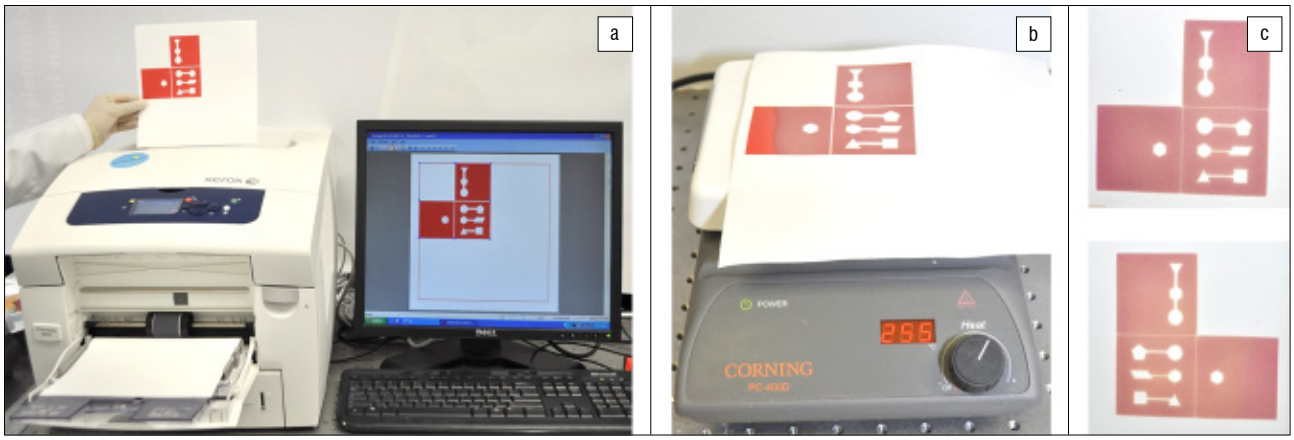


Figure 2: (a) Printing of wax device design onto chromatography paper from a computer aided design drawing using a wax printer connected to a computer, followed by (b) melting of the wax through the paper using a hot plate at 255 °C. (c) The front and back of the finished device after the melting stage. Complete infiltration of the wax through the paper device is observed.

along the paper. The formation of a dark image on the reverse side of the paper indicates efficient melting. Figure 2c shows both the front and back of the device after successful wax melting. Once melted, the device is cooled down for 30 s and then cut out.

Three-dimensional paper-based microfluidic designs require effective sealing between the layers of paper to ensure that fluid does not leak and that there is good contact between the fluidic connections, thereby ensuring the overall functionality of the device. Clamping of the paper layers can be used to achieve this,[26] or alternatively, double-sided adhesive layers can be used to stick the layers together.[16] Lewis et al.[30] also demonstrated that spray adhesive can efficiently seal these devices and ensure good contact between the device layers. This technique is simple to use, low-cost and enables high-throughput production. This method utilises adhesive layers, with the required materials and equipment readily available. Using the paper device design as a template, the fluid flow path cut-out (i.e. the sections of the paper design not covered with wax) can be designed using a computer aided design program as shown by the grey shaded areas in Figure 3a. The adhesive layers are manufactured using pressure sensitive adhesive sheeting (Flexmount DFM 200 clear V-95 150 Poly H-9 V-95 400 H-9, Flexcon, Synchron, Johannesburg, South Africa), which is cut according to the specific design on a vinyl cutter plotter (Roland CAMM-1 GX-24, Telpro Management, Johannesburg, South Africa). The adhesive in the sections for fluid flow is removed using tweezers as shown in Figure 3b. The larger individual layer adhesive blocks can be cut out using scissors.

Figure 4 illustrates the application of adhesive layers to the device to enable the 3D paper device to be assembled. In the case of the design shown in Figure 1, the adhesive is applied onto layers C and A before folding. Layer B is first folded onto layer A. A roller is used to apply pressure on this fold to ensure the adhesive securely binds these two layers together. Next, layer C is folded onto layer B. Pressure is applied once again and thereafter the device is ready for use. This ensures good sealing between layers B and C, and between layers A and B, respectively. The difference in the channel dimensions is negligible for lines that are designed to have widths greater than 1 mm. Tests carried out showed that lines with a design width of 1 mm resulted in printed lines that were 13% wider, while lines with design widths of 800 μm and 500 μm gave printed lines approximately 25% wider. The small difference of the channel dimensions after wax melting enables the adhesive cut-outs to align well with the wax designs. The sealing effect of the adhesive on the wax devices was found to be problematic at times, with small amounts of fluids leaking between layers of the device. Different types of adhesives could be explored to optimise the sealing between layers. The assembled device is now ready for use.

Signal reactions

The end result of biological/chemical diagnostic tests is dependent on sensitive reactions to produce a signal. The reaction that takes place requires an analyte of interest (antigen, substrate, etc.) to make contact with the complementary detection agent (antibody, enzyme, etc.). The presence of the analyte in the sample solution (e.g. blood or urine) followed by contact between the analyte and detection agent is observed as a user-friendly result output (colorimetric, digital readout, etc.). A well-known example is a home-based pregnancy test, which relies on a paper-based lateral flow platform[12] (Figure 5).

Although popular, a drawback to lateral flow tests as observed in Figure 5, is that reactions are restricted to a single analyte and therefore a single result. Paper-based microfluidics employ the same principles as lateral flow tests, but can make use of hydrophobic channelling to create multiplexed assays.[31] This section describes a colour reaction for a pH test whereby more than one sample is assayed on the same platform (multiplexing) using paper-based microfluidics.

Laboratory and household materials were selected for use as analytes and red cabbage dye was used as the pH detection agent. Red cabbage dye contains a natural pH detection compound called anthocyanin, which causes a highly acidic analyte (low pH) to turn red, and a basic analyte to turn light blue or yellow.

In preparation for the test, anthocyanin was extracted from red cabbage into distilled water. To achieve this, cabbage leaves were finely chopped, placed in the extraction solution, and heated in a microwave at a high heat setting for 1 min intervals. It is important to cover the solution to

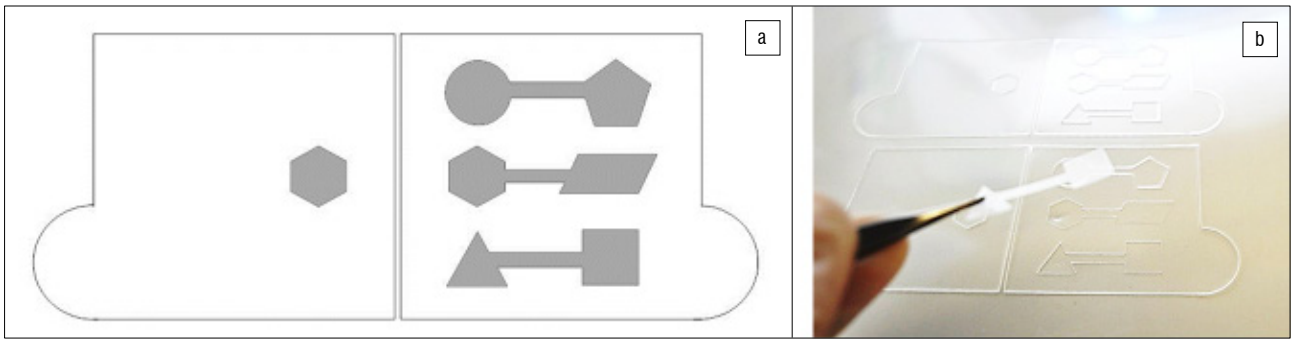


Figure 3: (a) Computer aided design drawing of adhesive layers used for the assembly and sealing of the layers of the design in Figure 1. The shaded areas indicate the fluidic flow pathways and illustrate the adhesive sections that must be removed. (b) The adhesive layer designs are manufactured using a vinyl cutter and the sections for fluid flow are removed using tweezers.

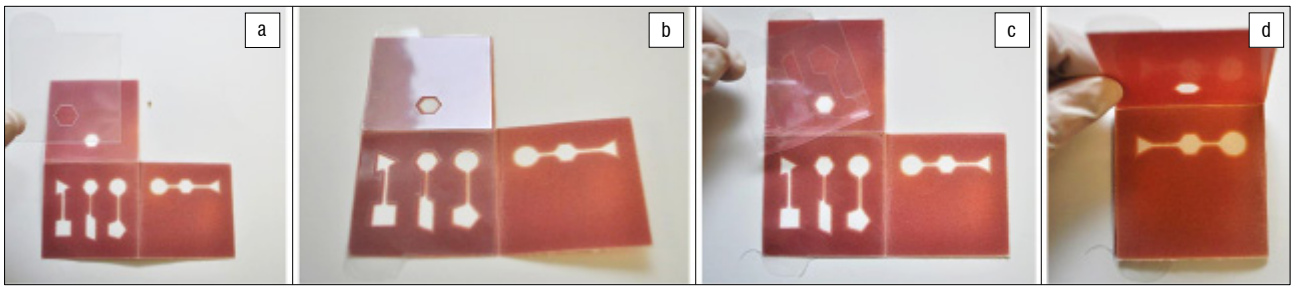


Figure 4: Example of device assembly using adhesive layers: (a, b) the layers are applied to the wax device and (c) the top protective layer is then removed before (d) folding and sealing of the layers of the 3D device.

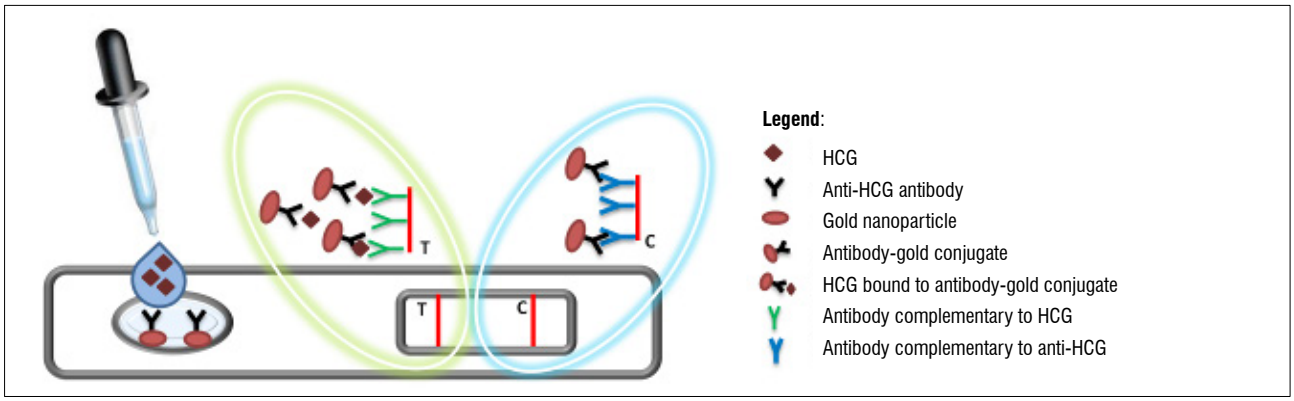


Figure 5: Illustration of a lateral flow pregnancy test. Urine is placed in the sample introduction window onto a sample pad. If the hormone human chorionic gonadotropin (HCG) is present in urine, it binds to a complementary anti-HCG antibody which is attached to a gold particle. The HCG-antibody-gold conjugate flows to meet a second antibody which is complementary to HCG, resulting in the stripe at the test line (T). Antibody-gold conjugates that do not bind to HCG flow to a third complementary antibody, resulting in a stripe at the control line (C). A positive pregnancy result occurs when both the test and control lines are observed, a negative result occurs when the control line is observed alone.

prevent evaporation and to stir the solution after each heating interval. It was noticed that the solution developed a purple colour after heating over a few minutes, while in contrast, the cabbage leaves lost colour. The solution was cooled before use.

The next stage involved spotting and drying of the analyte onto chromatography paper with wax channelling. Various analytes found in a laboratory (e.g. NaOH, HCl) and household materials (e.g. lemon juice and vinegar) were dried onto the channelled paper (layer A of the design in Figure 1) using a range of drying times (seconds, hours, overnight) and methods (air drying, oven drying, heat gun). The analytes were then tested for colour reactions using the anthocyanin solution, which was introduced at the sample inlet of the paper design (layer C). It was established that overnight air drying produced the most brilliant and distinct colours at the different pH values.

The analytes were also assayed in the presence of NaCl. The paper-based microfluidics platform in this paper also made use of electronics for an output and NaCl created the conductive salt bridge for the electronics. The analytes were air dried overnight in the presence and absence of NaCl. It was found that the range of analytes dried with NaCl did not have the expected colour reactions as were observed in the absence of NaCl. For this reason, NaCl was dried separately from the analyte in an area of the test that would not interfere with the pH reactions.

The volume of anthocyanin solution was also tested to identify the correct amount of solution required for flow through the channels and layers of paper. From this step, it was established that 120 μL of solution pipetted slowly onto the inlet on the first layer of paper was sufficient for the solution to flow through to the last layer and react with the analytes of interest to produce a visual result.

As demonstrated in Figure 6, one can track the path of this paper-based microfluidic tool, whereby the flow of the anthocyanin solution is directed by wax channelling towards three analytes housed in separate compartments. Within each compartment, the reagent makes contact with different analytes of varying pH, resulting in a range of colour reactions. In this demonstration, Milk of Magnesia turned green, alkaline solution turned yellow and ammonia turned blue. This is a multiplexing tool with a qualitative output that can also be used for more complex reactions. A quantitative output can be achieved by means of a printed electronics transducer or a handheld reader that converts the result of the chemical reaction to a measurable value. It should be noted that a higher concentration of colour is observed at the edges of the detection zones as opposed to the centre of the detection zones, also known as the coffee-ring effect, and is an important consideration in paper-based microfluidics as discussed by Deegan et al[32].

In summary, each step in testing the reaction on paper has highlighted the importance of understanding the nature of the reaction and how it will take place under various conditions on a paper-based microfluidic platform. This also creates an awareness of the effect of chemically altering the paper to manipulate the flow of sample on the paper. It is imperative that the optimised reactions on the paper-based microfluidics platform produce sensitive, accurate, specific results to meet the needs of the target market.

Sample delivery

Sample introduction to the microfluidic device can be achieved manually by using a pipette, or in an automated fashion, for example, via blister packs, which provide a simple and effective mechanism for reagent storage and release. Blister packs have become increasingly popular in microfluidic devices geared towards point-of-care solutions. The limited research on blister packs has focused on numerical modelling[33] and sealing of the blister[34]. Manufacture of blister pack devices often requires specialised materials and equipment, but sample delivery can also be undertaken using simple and cost-effective methods, as demonstrated here. Sample delivery for the demonstration device was achieved by designing and implementing a low cost fabrication method to produce re-usable fluidic 'buttons' using polydimethylsiloxane (PDMS) to mimic the functionality and advantages of blister packs. The 'button', which was designed and manufactured in-house, allows for easy and low-cost introduction of fluids for the test. Similar design and manufacturing methods have been applied in-house for low-cost microfluidic devices.[35] The 'button' was realised by assembling three layers of PDMS sheets: a blank top layer, a middle layer with a round opening which functions as a reservoir, and a bottom layer with a hole which forms the outlet.

The properties of the sample fluid, such as viscosity and surface tension, can affect the sample volume required for each origami microfluidic device. The volume of sample required also depends on the paper thickness (and hence its wicking ability), the length and width of the fluid flow channels on the device, the size of fluidic connection points, and most importantly, the size of the test result signal shapes. As a result, the amount of sample required on a paper test is dependent on its design. It is therefore recommended that once a device is designed, it is tested with fluid in order to gauge the sample volume requirements. Ideally, the sample inlet should be designed to contain the total volume of sample required to perform the test. This will prevent the need for the user to continually re-load sample fluid, while at the same time not wasting high volumes of fluid.

To manufacture the 'button', PDMS was cast in a mould to form a flat sheet with a thickness of 3–5 mm and then cut into blocks of 15 mm x 15 mm. The reservoir of the middle layer was created using a punch with a diameter of 10 mm, which had a volume adequate for the 120 μl dye samples for the demonstration device. A 1 mm diameter punch was used to create the outlet in the bottom layer. The dimension of the outlet hole was suitable to prevent sample

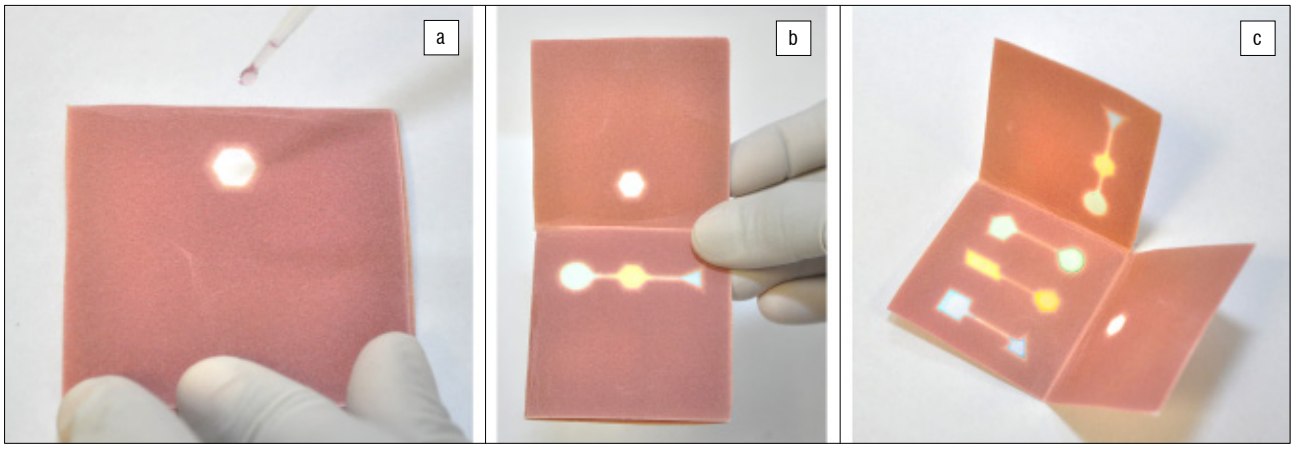


Figure 6: A demonstration of the reaction following the flow of the cabbage dye in distilled water on the paper-based microfluidic platform: (a) the solution is pipetted into the sample introduction shape; (b) layer C of the device is opened to show that the dye flows into the intermediate shapes on layer B, resulting in staining of the intermediate shapes; (c) after opening all the layers, one observes the effect of the multiplexing reaction in which the dye stains Milk of Magnesia green, alkaline solution yellow and ammonia blue in the test result shapes.

leaking out, while a typical finger press ejected a satisfactory sample volume onto the wicking paper substrate. The three PDMS blocks were then sealed with liquid PDMS or glue to achieve the final assembled button device. Reagents can be preloaded into the button using a syringe and needle by injecting the fluid into the 1 mm outlet hole. The design and manufacturing process of the sample delivery buttons is shown in Figure 7.

Electronics

The promise of printed electronics integrated on paper has attracted significant international attention. An appropriate commencement point for low-cost smart healthcare products in need of considerable processing resources is the hybrid approach of combining rigid electronic components with printed electronics interconnects. A reasonable progression from printed interconnect wiring alone, would be the printing of electrodes[36] and communication antennas[23], to printed biorecognition and transducer materials[19,21], as well as dielectric and semiconductor materials, followed by passive components such as resistive heating elements[19], and then active devices[36], transistor circuits and memories, heading eventually to integrating a power source[20] and display[22,36], all potentially contained in a 3D-printed mechanical assembly.

The heterogeneous integration approach of having rigid electronic components with flexible printed interconnects is appropriate for PoC healthcare products. In this approach, standard components such as microprocessors, sensor chips and passive components are embedded into paper-based circuits, particularly by using surface mount components. The interconnect wiring network is generated with conductive ink and colloidal nanosilver is by far the most ubiquitous in paper electronics. Laboratory devices have been fabricated by direct writing with either a syringe or a roller-ball pen filled with nanosilver ink[36], as well as by spraying conductive ink through a stencil,[19] while several digital printing technologies are used, most notably screen printing[21] and inkjet printing[23] for small volumes and prototyping, and high volume roll-to-roll printing[23] for commercial devices. Where equipment is not available, manual painting of electronic pathways using conductive paint provides an effective solution for rapidly developing simple electronics on paper devices. In the demonstration examples here, a nanosilver conductive paint was used to hand-draw conductive tracks.

A number of electronic components useful for bioanalysis have been investigated for integration onto paper-based microfluidic devices. A switching element – either a manual button or a chemical reaction region – may be utilised to trigger electronic functions and can most simply be implemented by two conductive pads or electrodes which can be electrically connected in a controlled manner. Components such as light emitting diodes (LEDs) may be employed to display the result of an analysis. The electronic components used for the examples in this paper are off-the-shelf surface mount LEDs and 1 kΩ current limiting resistors where required. These are secured into position on the paper device using glue, and the nanosilver conductive paint is applied to the terminals of the components to complete the circuit.

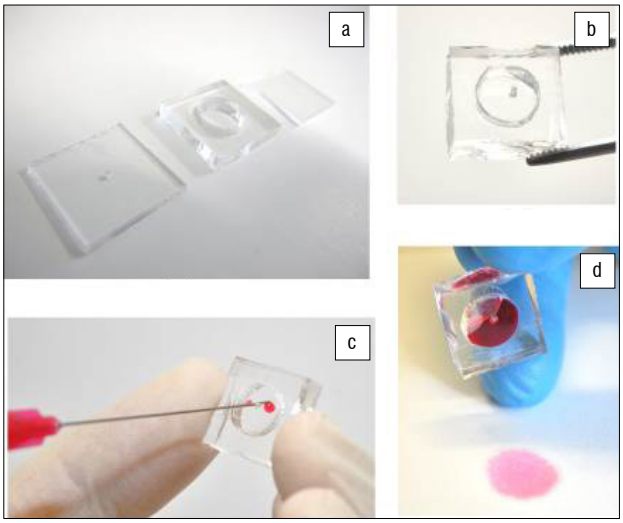


Figure 7: Sample delivery 'button' design and manufacture: (a) three polydimethylsiloxane (PDMS) blocks are layered with a 1 mm punched hole as outlet, a 10 mm punched hole as reservoir, and a blank cover; (b) assembling the layers with liquid PDMS produces a sealed 'button' device; (c) preloading the 'button' by injecting a reagent using a syringe and needle tip; (d) sample spot delivered onto paper by pressing on the 'button'.

Figure 8 illustrates the design and implementation of a simple paper electronic manual button function with an LED power indicator. A compact CR1220 3V watch battery was used to power the electronics, with an additional paper flap included on the paper-based device to house the battery. The LED was illuminated when the power source made contact and the paper power button was pressed. The paper device was designed using a similar principle to the origami microfluidic device in Figure 1, and consists of three layers which can be folded to produce the assembled device. Referring to the design in Figure 8a, at fold 1, the middle layer folds down onto the bottom layer. The top layer then folds over onto the middle layer at fold 2. The grey lines are printed on the paper as guidelines for hand-

painting the conductive traces, while the circles indicate where pads must be painted. The LED is connected at the oval dashed shape and the battery is placed at the circular dashed circle. Figure 8b depicts the assembly of the folded device, while the operational device is presented in Figure 8c.

When integrating electronics with wax-printed microfluidic paper-based devices, it is desirable to create circuitry over the wax, but without any visible guidelines as to the placement of conductive tracks and components. Adhesive sheeting, which is cut according to the specific design on a vinyl cutter plotter (Roland CAMM-1 GX-24, Telpro Management, Johannesburg, South Africa), can be utilised to create a stencil which can be placed on top of the wax layers to serve as guidelines for the layout of the electronic circuits.

Figure 9 illustrates the use of a stencil design to assist in the positioning of the conductive ink pathways and surface mount components that make up the electronics of the device. The device functions in the same way, and uses the same basic design as the device in Figure 8. The adhesive stencil is stuck lightly over the wax printed device as shown in Figure 9a, and the conductive paint can be applied in the open lines of the stencil directly onto the wax resulting in the device in Figure 9b. Once the conductive paint is dry, the adhesive stencil can easily be removed from the wax device as seen in Figure 9c. The electronic components can then be secured in the relevant places in the circuit and connected with conductive paint to produce the completed device in Figure 9d.

Integrated device example

An integrated electronics-supported paper-based microfluidic device with a specific application could be developed using the individual components as described in the previous sections. In each case, components could be developed using simple techniques, requiring little to no specialised equipment. In this example, a three-dimensional paper-based device was created using wax barriers. Surface mount electronic components and a compact watch battery, along with manually applied conductive ink tracks were implemented to illuminate an LED once the paper device was assembled and all layers made contact. Furthermore, a second LED was illuminated to indicate that the test was complete. The collage shown in Figure 10 visually summarises the fully assembled and operational device.

Using the procedures discussed in the origami 3D fluidic network section, the three-layer microfluidic device was created. A rectangular sample inlet port is on the top layer, and sample fluid passes through four circular intermediate fluidic connection points in order to reach the test result signal zones on the bottom layer. The positions of the circular connection points ensure that sample enters the upper end of each alphabetical character in the 'C-S-I-R' test result symbol. Sample fluid therefore moves towards the lower end of each letter as the test progresses. Once each letter zone is completely filled with fluid, the test is deemed complete. The microfluidic design was printed on a wax printer and then placed on a 255 °C hot plate for 1 min, as described in the origami 3D fluidic network section. To construct the origami device, layer B was folded onto layer C, after which layer A was folded onto layer B.

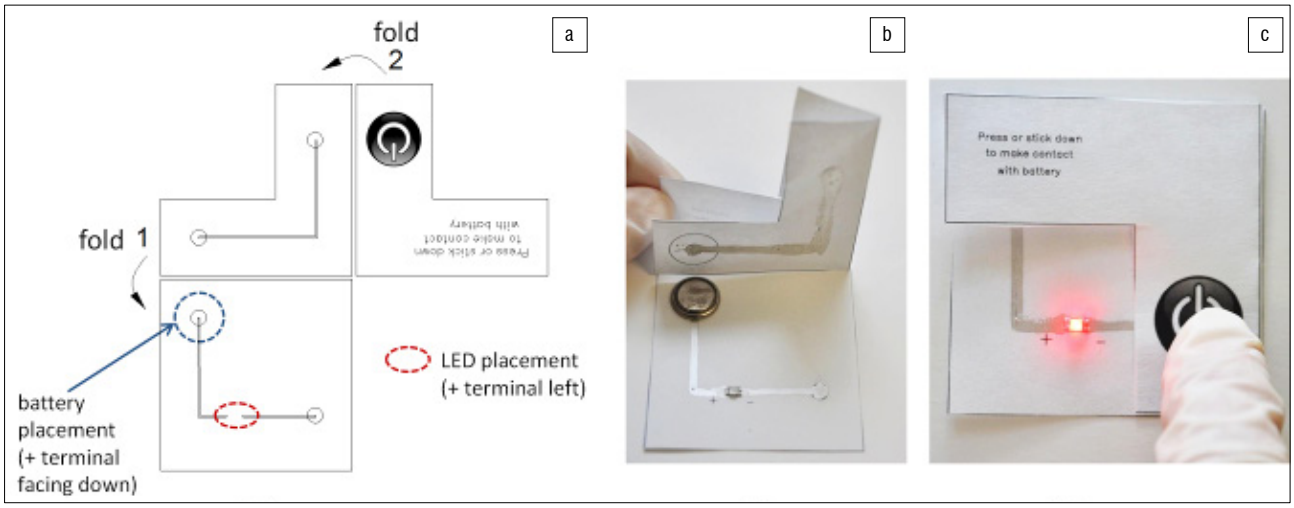


Figure 8: Manual manufacture of a simple paper-based electronic device on standard printing paper: (a) layout design of a manual button and LED power indicator. The three-layer design houses a battery and LED on the bottom layer, onto which the middle layer is folded (fold 1), followed by the folding of the top layer containing the paper button. (b) Device assembly and folding is illustrated, with electronic pathways painted in silver and electronic components mounted. (c) The complete functional device is illustrated; the LED illuminates when the paper button is pressed.

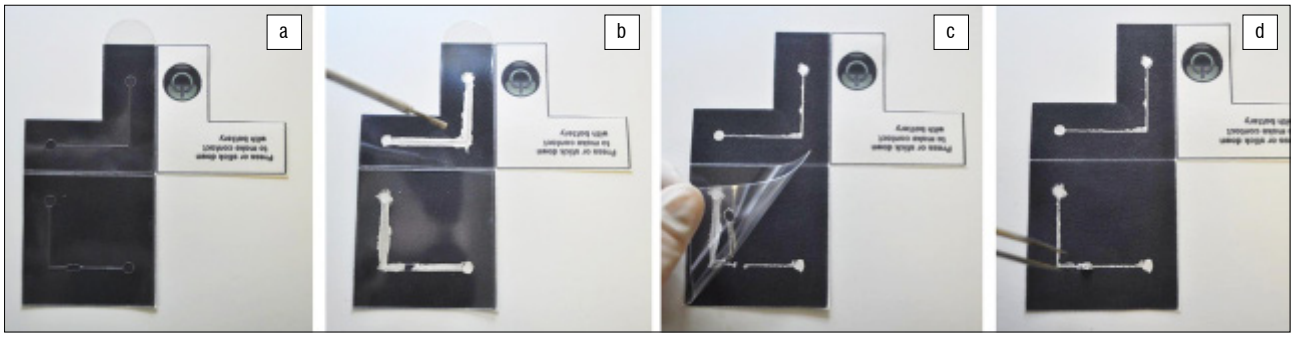


Figure 9: Manual manufacture of a simple electronic device integrated on wax-printed paper: (a) adhesive layer stencils are stuck onto the wax printed areas; (b) application of the conductive nanosilver paint within the adhesive stencil areas; (c) removal of the stencil; (d) mounting of electronic components for the complete device.

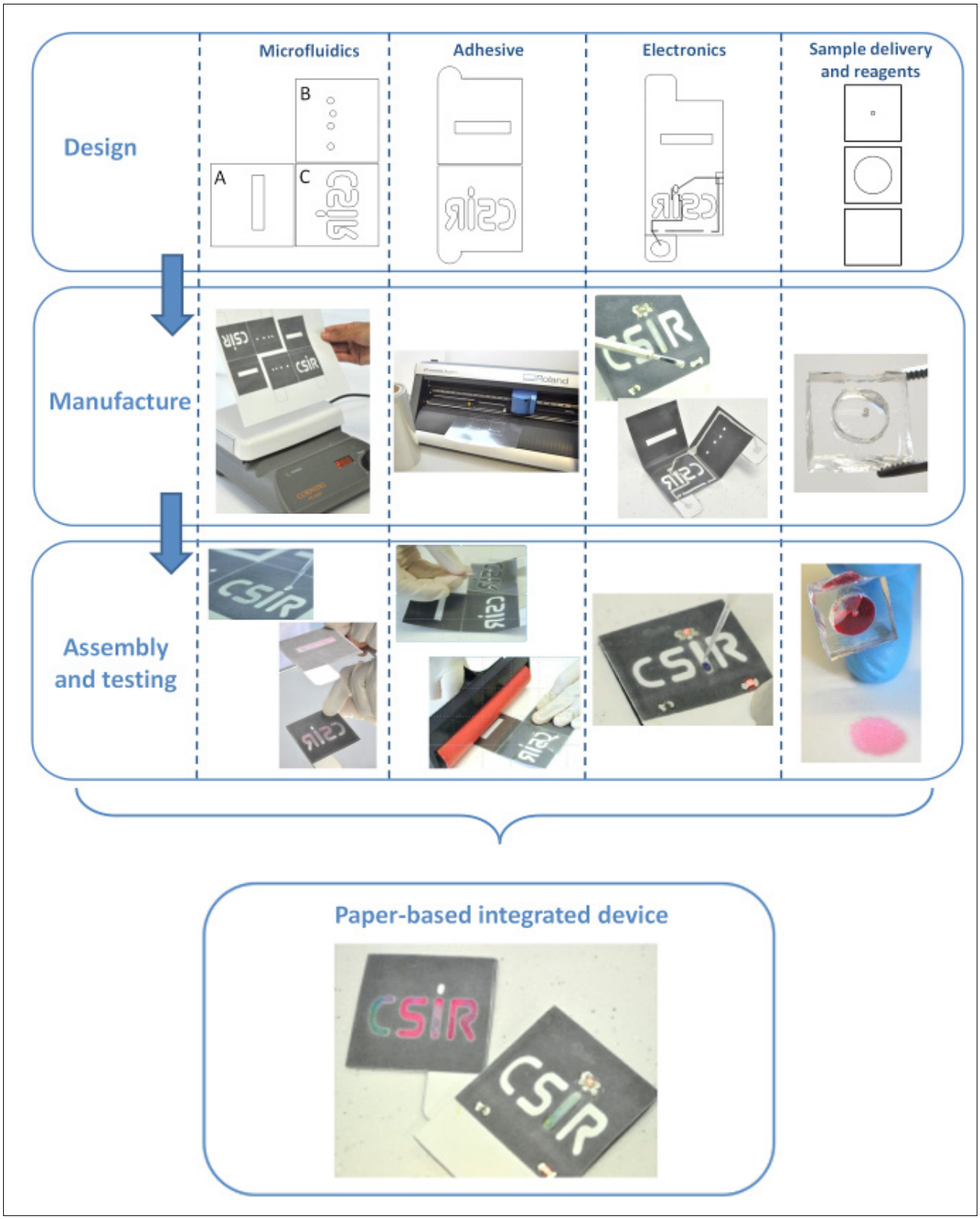


Figure 10: Collage showing the design, manufacturing, assembly and testing steps of the fully operational integrated electronics-enhanced microfluidic origami device. Each step encompasses microfluidic, adhesive, electronic and sample delivery components. The design step consists of the computer aided designs for the microfluidic paper device design (illustrated in Figure 1), the adhesive layer design (shown in Figure 3), as well as the electronic and sample introduction 'button' designs. The manufacturing step includes wax printing of the paper-based microfluidic device and melting of the wax using a hotplate (shown in Figure 2), vinyl cutting of the adhesive layers, painting of conductive tracks and mounting of electronic components such as LEDs and batteries, as well as polydimethylsiloxane casting of layers to create the sample delivery button. The assembly and testing step includes depositing reagents onto the paper device and the addition of adhesive layers to the paper device, which can then be folded and sealed. A sample can then be introduced to the device using the sample delivery 'button' and resulting colour changes and/or LED illuminations can be observed when reactions take place on the paper, as illustrated by the integrated paper-based devices.

The pH-dependent colour change reactions discussed in the signal reactions section were used as a model multiplexed chemical test. The cabbage dye was introduced via the fluidic button discussed in the sample delivery section of this paper, via the inlet port at the top of the device. After separation into four fluidic paths in the intermediate layer, the sample flowed to the test result section where four individual colour change reactions occurred in each of the 'C-S-I-R' letters.

To indicate completion of the test to the user, an LED at the dot of the 'I' lights up once fluid has travelled through the device and along the length of the letter 'I'. A solution of NaCl was dried between the sensing electrodes to ensure electrical conductivity. Stencil designs for manual application of conductive ink were created as a guideline for the positioning of conductive tracks on each device layer, according to the methods discussed in the electronics section of this paper. The circuit interconnection network was constructed on the inside of the folded device, with the electronic components visible on the outside. The conductive tracks were painted onto the wax-printed paper-based device by hand and surface mount LEDs and resistors were secured onto the paper device using glue. Through small pinholes in the paper device, conductive paint was applied to ensure contact between the electronic components and the conductive tracks on the reverse side of the paper device.

The wax-printed microfluidic paper device with electronic tracks and mounted electronic components, as well as an inserted battery, was assembled using adhesive layers by means of the approach discussed in the origami 3D fluidic network section of this paper. An LED lights up to indicate that electrical contact between the layers has been achieved and the device is powered.

To test this device, cabbage dye was loaded into the sample inlet. The large inlet zone allows for an excess volume of fluid to be contained in the sample inlet, and the user does not have to repeatedly load dye as it is drawn into the device. It is also important to ensure that the bottom layer of the device (containing the CSIR test signal) is not placed onto a hydrophilic surface while a test is being performed. Hydrophilic surfaces may exert additional capillary forces and draw the fluid out of the device before it can complete the reactions. It is better to keep the device suspended in air in order to avoid these effects. However, in the future it may be beneficial to use a transparent, hydrophobic backing on such devices.

Conclusion

The fundamental manufacturing steps for starting out with paper-based microfluidics and electronics – and the integration of these on a single origami device – were expounded in this paper. A demonstration electronics-enhanced microfluidic device was developed with emphasis on simple manufacturing methods. The demonstration device lends itself to educational activities affording insight into paper-based electronics-enhanced microfluidics. Additionally, the uncomplicated construction of these devices enables wide experimentation with paper-based microfluidics and electronics in various formal or informal environments. The latter facilitates community engagement which, along with market research, will help tailor devices towards customer need in low resource communities and industry.

Educational activities, such as the 2014 'Bring-a-girl-child-to-work day'[37], in which high-school students were exposed to the paper-based device development processes discussed here, cultivates an interest in this technology. Stemming from this, innovative ideas and solutions are likely to emerge, and can easily and rapidly be prototyped using existing infrastructure.

FabLabs,[38] with the many people utilising their facilities, offers an opportunity to test these devices, from a production, manufacturing as well as from a usability perspective. Valuable inputs can be gained for further product improvement.

As paper-based electronics-enhanced microfluidic solutions journey through the research, development and innovation chain to commercial success, dedicated production equipment will become necessary. But as the infrastructure required to support a production facility is relatively small, a viable model is distributed production, where manufacturing can occur geographically close to the target market, tailored to local needs in a way that is neither practical nor economical using mass production. In this way, paper-based electronics-enhanced microfluidic devices can provide support for micro-sized entrepreneurial manufacturing enterprises.

It is clear that following on from the immediate educational use of the developed demonstration tool, there is significant potential for paper-based electronics-enhanced microfluidics in sensitive and cost effective diagnostic applications in health, agriculture and environmental monitoring, particularly for low-resourced areas.

Acknowledgements

We gratefully acknowledge funding made available through the Council for Scientific and Industrial Research (CSIR) parliamentary grant.

Authors' contributions

All authors contributed to the technical work of component development and realisation of the functional integrated device. K.M. led the paper-based microfluidic design and manufacture, U.G. managed the development and testing of the reactions, J.C. was responsible for the sample delivery component design and manufacture, and S.S. headed up the electronic component development. T.J. and K.L. provided guidance and conceptual contributions towards the component and integrated device development. T.J, K.M., S.S., U.G., J.C. and K.L. compiled the manuscript.

References

1. Yager P, Edwards T, Fu E, Helton K, Nelson K, Tam MR, et al. Microfluidic diagnostic technologies for global public health. Nature. 2006;442(7101):412–418. http://dx.doi.org/10.1038/nature05064
2. Gous N, Scott L, Potgieter J, Ntabeni L, Enslin S, Newman R, et al. Feasibility of performing multiple point of care testing for HIV anti-retroviral treatment initiation and monitoring from multiple or single fingersticks. PLoS One. 2013;8(12):e85265. http://dx.doi.org/10.1371/journal.pone.0085265
3. Losina E, Bassett IV, Giddy J, Chetty S, Regan S, Walensky RP, et al. The 'ART' of linkage: Pre-treatment loss to care after HIV diagnosis at two PEPFAR sites in Durban, South Africa. PLoS One. 2010;5(3):e9538. http://dx.doi.org/10.1371/journal.pone.0009538
4. Corbett EL, Marston B, Churchyard GJ, De Cock KM. Tuberculosis in sub-Saharan Africa: Opportunities, challenges, and change in the era of antiretroviral treatment. The Lancet. 2006;367(9514):926–937. http://dx.doi.org/10.1016/S0140-6736(06)68383-9
5. Patton JC, Akkers E, Coovadia AH, Meyers TM, Stevens WS, Sherman GG. Evaluation of dried whole blood spots obtained by heel or finger stick as an alternative to venous blood for diagnosis of human immunodeficiency virus type 1 infection in vertically exposed infants in the routine diagnostic laboratory. Clin Vaccine Immunol. 2007;14(2):201–203. http://dx.doi.org/10.1128/CVI.00223-06
6. Chin CD, Linder V, Sia SK. Commercialization of microfluidic point-of-care diagnostic devices. Lab Chip. 2012;12:2118–2134. http://dx.doi.org/10.1039/c2lc21204h
7. Mabey D, Peeling RW, Ustianowski A, Perkins MD. Diagnostics for the developing world. Nat Rev Microbiol. 2004;2(3):231–240. http://dx.doi.org/10.1038/nrmicro841
8. Martinez AW, Phillips ST, Butte MJ, Whitesides GM. Patterned paper as a platform for inexpensive, low-volume, portable bioassays. Angew Chem Int Ed. 2007;46(8):1318–1320. http://dx.doi.org/10.1002/anie.200603817
9. Martinez AW, Phillips ST, Whitesides GM, Carrilho E. Diagnostics for the developing world: Microfluidic paper-based analytical devices. Anal Chem. 2010;82(1):3–10. http://dx.doi.org/10.1021/ac9013989
10. Microfluidics 2.0 [homepage on the Internet]. c2010 [cited 2014 Oct 09]. Available from: http://www.mf20.org/
11. Yetisen AK, Akram MS, Lowe CR. Paper-based microfluidic point-of-care diagnostic devices. Lab Chip. 2013;13(12):2210–2251. http://dx.doi.org/10.1039/c3lc50169h
12. Liana DD, Raguse B, Gooding JJ, Chow E. Recent advances in paper-based sensors. Sensors. 2012;12(9):11505–11526. http://dx.doi.org/10.3390/s120911505

13. Li X, Ballerini DR, Shen W. A perspective on paper-based microfluidics: Current status and future trends. Biomicrofluidics. 2012;6(1):011301. http://dx.doi.org/10.1063/1.3687398

14. Carrilho E, Martinez AW, Whitesides GM. Understanding wax printing: A simple micropatterning process for paper-based microfluidics. Anal Chem. 2009;81(16):7091–7095. http://dx.doi.org/10.1021/ac901071p

15. Lu Y, Shi W, Jiang L, Qin J, Lin B. Rapid prototyping of paper-based microfluidics with wax for low-cost, portable bioassay. Electrophoresis. 2009;30(9):1497–1500. http://dx.doi.org/10.1002/elps.200800563

16. Martinez AW, Phillips ST, Whitesides GM. Three-dimensional microfluidic devices fabricated in layered paper and tape. Proc Natl Acad Sci USA. 2008;105(50):19606–19611. http://dx.doi.org/10.1073/pnas.0810903105

17. Liu H, Crooks RM. Three-dimensional paper microfluidic devices assembled using the principles of origami. J Am Chem Soc. 2011;133(44):17564–17566. http://dx.doi.org/10.1021/ja2071779

18. Clearblue Digital Pregnancy Test [homepage on the Internet]. c2014. Available from: http://za.clearblue.com/pregnancy-tests/digital-with-weeks-indicator

19. Maxwell EJ, Mazzeo AD, Whitesides GM. Paper-based electroanalytical devices for accessible diagnostic testing. MRS Bull. 2013;38(04):309–314. http://dx.doi.org/10.1557/mrs.2013.56

20. Esquivel JP, Del Campo FJ, Gómez de la Fuente, JL, Rojas S, Sabaté N. Microfluidic fuel cells on paper: Meeting the power needs of next generation lateral flow devices. Energ Environ Sci. 2014;7(5):1744–1749. http://dx.doi.org/10.1039/c3ee44044c

21. Wu Y, Xue P, Hui KM, Kang Y. A paper-based microfluidic electrochemical immunodevice integrated with amplification-by-polymerization for the ultrasensitive multiplexed detection of cancer biomarkers. Biosens Bioelectron. 2014;52:180–187. http://dx.doi.org/10.1016/j.bios.2013.08.039

22. Kawahara J, Andersson Ersman P, Nilsson D, Katoh K, Nakata Y, Sandberg M, et al. Flexible active matrix addressed displays manufactured by printing and coating techniques. J Polym Sci Phys. 2013;51(4):265–271. http://dx.doi.org/10.1002/polb.23213

23. Vena A, Sydanheimo L, Tentzeris M, Ukkonen L. A fully inkjet-printed wireless and chipless sensor for CO_2 and temperature detection. IEEE Sens J. 2014;(99):1–12. http://dx.doi.org/10.1109/jsen.2014.2336838

24. Ravgiala RR, Weisburd S, Sleeper R, Martinez A, Rozkiewicz D, Whitesides GM, et al. Using paper-based diagnostics with high school students to model forensic investigation and colorimetric analysis. J Chem Educ. 2014;91(1):107–111. http://dx.doi.org/10.1021/ed300261a

25. Govindarajan AV, Ramachandran S, Vigil GD, Yager P, Böhringer KF. A low cost point-of-care viscous sample preparation device for molecular diagnosis in the developing world: An example of microfluidic origami. Lab Chip. 2012;12(1):174–181. http://dx.doi.org/10.1039/C1LC20622B

26. Ge L, Wag S, Song X, Ge S, Yu J. 3D origami-based multifunction-integrated immunodevice: Low-cost and multiplexed sandwich chemiluminescence immunoassay on microfluidic paper-based analytical device. Lab Chip. 2012;12(17):3150–3158. http://dx.doi.org/10.1039/c2lc40325k

27. Dungchai W, Chailapakul O, Henry CS. A low-cost, simple, and rapid fabrication method for paper-based microfluidics using wax screen-printing. Analyst. 2011;136(1):77–82. http://dx.doi.org/10.1039/C0AN00406E

28. Songjaroen T, Dungchai W, Chailapakul O, Laiwattanapaisal W. Novel, simple and low-cost alternative method for fabrication of paper-based microfluidics by wax dipping. Talanta. 2011;85(5):2587–2593. http://dx.doi.org/10.1016/j.talanta.2011.08.024

29. Govindasamy K, Potgieter S, Land K, Muzenda E. Fabrication of paper based microfluidic devices. Proceedings of the World Congress on Engineering 2012 Vol III WCE; 2012 July 4–6; London, UK. London: International Association of Engineers; 2012.

30. Lewis GG, DiTucci MJ, Baker MS, Phillips ST. High throughput method for prototyping three-dimensional, paper-based microfluidic devices. Lab Chip. 2012;12(15):2630–2633. http://dx.doi.org/10.1039/c2lc40331e

31. Wang J, Monton MRN, Zhang X, Filipe CD, Pelton R, Brennan JD. Hydrophobic sol–gel channel patterning strategies for paper-based microfluidics. Lab Chip. 2014;14(4):691–695. http://dx.doi.org/10.1039/C3LC51313K

32. Deegan RD, Bakajin O, Dupont TF, Huber G, Nagel ST, Witten TA. Capillary flow as the cause of ring stains from dried liquid drops. Nature. 1997;389(6653):827–829. http://dx.doi.org/10.1038/39827

33. Selvakumar S, Linares R, Oppenheimer A, Anthony B. Variation analysis of flow rate delivered using a blister pump. Proc SPIE. 2012;8251. http://dx.doi.org/10.1117/12.907502

34. Inamdar T, Anthony BW. Characterizing fluidic seals for on-board reagent delivery. Proc SPIE. 2013;8615. http://dx.doi.org/10.1117/12.2006257

35. Land KJ, Mbanjwa MB, Govindasamy K, Korvink JG. Low cost fabrication and assembly process for re-usable 3D polydimethylsiloxane (PDMS) microfluidic networks. Biomicrofluidics. 2011;5:036502. http://dx.doi.org/10.1063/1.3641859

36. Ahn J, Je JH. Stretchable electronics: Materials, architectures and integrations. J Phys D Appl Phys. 2012;45(10):103001. http://dx.doi.org/10.1088/0022-3727/45/10/103001

37. Take a girl child to work day [page on the Internet]. c2014 [cited 2014 Oct 09]. Available from: http://www.cellc.co.za/girlchild

38. Fab Foundation [homepage on the Internet]. c2014 [cited 2014 Oct 09]. Available from: http://www.fabfoundation.org/

Keurbooms Estuary floods and sedimentation

AUTHOR:
Eckart H. Schumann[1]

AFFILIATION:
[1]Department of Geosciences, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

CORRESPONDENCE TO:
Eckart Schumann

EMAIL:
eckarts@mweb.co.za

POSTAL ADDRESS:
26 Marine Gardens, Glengarry Crescent, Humewood 6013, South Africa

DATES:
Received: 30 Oct. 2014
Revised: 26 Feb. 2015
Accepted: 09 Mar. 2015

KEYWORDS:
estuary morphology; barrier dune; mouth migration; river floods; ocean waves

HOW TO CITE:
Schumann EH. Keurbooms Estuary floods and sedimentation. S Afr J Sci. 2015;111(11/12), #Art. #2014-0380, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140380

The Keurbooms Estuary at Plettenberg Bay lies on a wave-dominated, microtidal coast. It has a dune-topped sandy barrier, or barrier dune, almost 4 km long, with a narrow back-barrier lagoon connected to its source rivers, the Keurbooms and Bitou. The estuary exits to the sea through this barrier dune, and it is the geomorphology and mouth position in relation to floods, which is the subject of this paper. Measurements of rainfall, water level, waves and high- and low-tide water lines were used to analyse the mouth variability over the years 2006–2012. Two major floods occurred during this time, with the first in November 2007 eroding away more than 500 000 m^3 of sediment. The new mouth was established at the Lookout Rocks limit – the first time since 1915. The second flood occurred in July 2012 and opened up a new mouth about 1 km to the north-east; high waves also affected the position of the breach. The mouth has a tendency to migrate southwards against the longshore drift, but at any stage this movement can be augmented or reversed. The effectiveness of floods in breaching a new mouth through the barrier dune depends on the flood size and the nature of the exit channel in the back-barrier lagoon. Other factors such as ocean waves, sea level, vegetative state of the dune and duration of the flood are also important and can determine where the breach occurs, and if the new mouth will dominate the old mouth.

Introduction

The Keurbooms Estuary is situated in Plettenberg Bay on the wave-dominated, microtidal south coast of South Africa. The southern boundary of the estuary is formed by the Lookout Rocks of the Cape Supergroup, while the study area consists of the Holocene tide-dominated sedimentary deposits extending to the bridge where the N2 national road crosses the estuary. The situation is depicted in Figure 1.

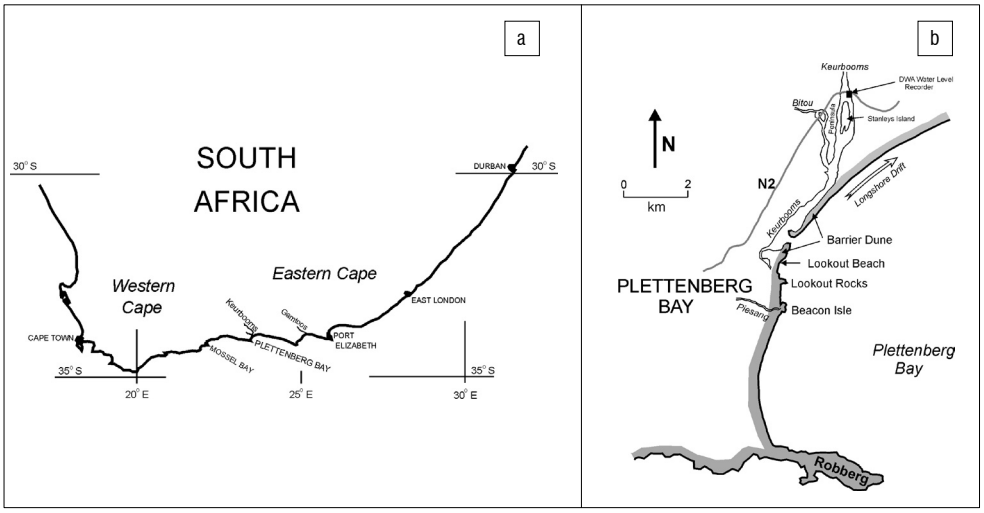


Figure 1: (a) The position of Plettenberg Bay on the south coast of South Africa. (b) Plettenberg Bay, showing the Keurbooms Estuary, Robberg Peninsula and other sites mentioned in the text.

The estuary is formed by the confluence of two rivers, namely the Keurbooms River itself, and its tributary the Bitou River. The two estuaries combine to form a narrow back-barrier lagoon behind a dune-topped sandy barrier which extends almost 4 km – here termed a *barrier dune*. The tidal inlet or mouth is not anchored in any substrate resistant to erosion, and it exits through this barrier dune. It is the variations in the mouth position in response to floods and ocean forcing which form the primary subject of this paper.

Historical records show the mouth situated at various positions along the barrier dune. Storrar[1] shows maps drawn by JC Frederici in 1789 and Lieutenant W McPherson Rice in 1797 which both depict the mouth positioned approximately in the middle of the barrier dune.

A later map of the region drawn by the Royal Navy Hydrographic Office in 1867 shows the mouth situated at the Lookout Rocks. A divisional map of Knysna dated 1890 confirms this position, while the Colonial Surveyor-General in 1900 depicted the mouth nearly 1 km north of the Lookout Rocks.

Reddering[2] reports that up to 1915, the mouth was at the Lookout Rocks, though it is not clear when it changed from the position depicted in 1900. In 1915, a severe flood breached the barrier dune at its northern end; this new mouth was big enough to dominate the previously-existing southern mouth which subsequently closed.

Aerial photographs became available in the 1930s and were used to trace further developments. In particular, those taken by the Chief Directorate: Surveys and Mapping (Trig Survey) provide a sequence, though not at regular

intervals. The first of these aerial photographs dates from 1936 and is shown in Figure 2.

Source: Chief Directorate: Surveys and Mapping.

Figure 2: An aerial photograph of the Keurbooms Estuary in 1936.

At the time, the mouth was still situated at the northern end of the barrier dune, while a complex system of sandbanks is evident on the estuary side. The barrier dune extending southwards had a sparse collection of vegetation, while a prominent overwash fan is indicated. Farther south, the wide beach/barrier dune seems void of vegetation, with marine sediments extending some distance into the back-barrier lagoon. On the western side there appears to be a number of channels with zostera-covered sandbars.

Fromme[3] analysed the mouth migrations using a set of 14 aerial photographs over a period of 44 years. The first two photographs from 1936 (Figure 2) and 1942 show the mouth still in the northerly position, while all the other 12 photographs located it in the centre of the barrier dune, some 2 km north of the Lookout Rocks. The mouth position over the 19 years from 1961 to 1980 varied over a distance of around 500 m.

Reddering[2,4] and Duvenhage and Morant[5] concluded that there was a general migration of the mouth southwards to the Lookout Rocks. It was furthermore proposed that the position of the Keurbooms Estuary mouth goes through long-term cycles, starting with a breach of the barrier dune in the north.[2,4,5] It then slowly migrates south-westward to the Lookout Rocks, with temporary reversals of this migration caused by the wave climate. The cycle repeats when a major flood breaches the barrier dune in the north again. Nonetheless, in the available historical record of over 200 years, it was only in 1915 that the mouth was situated in the extreme north.

More recently, the barrier dune was vegetated from the Lookout Rocks side. It appeared as if this vegetation would inhibit further erosion of the barrier dune[3], though in 2003, Illenberger[6] concluded that the mouth would reach the Lookout Rocks in 6–12 years.

In August 2006, a large flood removed about 200 m of the remaining barrier dune, in the process destroying a toilet/shower building. A year later, in November 2007, a massive flood removed what was left of the barrier dune, and a new mouth of the Keurbooms Estuary was formed adjacent to the Lookout Rocks. Some infrastructure, including two toilets and change-room facilities were destroyed, with rocks on the southern boundary of the estuary preventing further damage.

In July 2012, another flood breached the low barrier dune about 1 km north of the Lookout Rocks. The estuary continued to function with two mouths, but gradually the mouth at the Lookout Rocks silted up and in 2014, the estuary had only one mouth.

This paper details the changes in the geomorphology of the estuary mouth region over the last two floods from 2006 to 2012. It is proposed that the manner in which the estuary develops is dependent on a number of physical processes, and that the changes that are caused by, for instance a flood, depend on the developmental stage of the estuary at that time.

Keurbooms estuary dynamics

The catchments of the Keurbooms and Bitou Rivers are shown in Figure 3 and have areas of 859 km^2 and 237 km^2, respectively.[2] The mean annual run-off of both catchments was estimated to lie in the range of 70 x 10^6 m^3 to 160 x 10^6 m^3 by Duvenhage and Morant.[5] On the other hand, Fromme[3] calculated a median discharge of 8.29 x 10^6 m^3 per month over a period of 56 years. These values correspond to an average inflow between 2.2 m^3/s and 5 m^3/s, although these values include occasional floods, which means that the daily flow from the two rivers is probably considerably less.

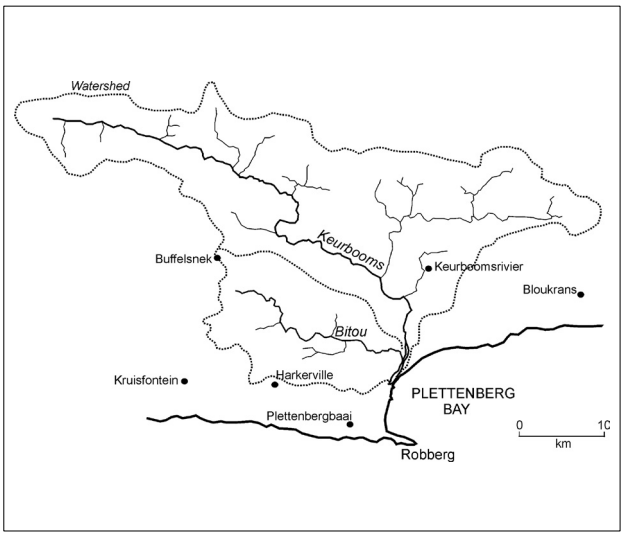


Figure 3: Catchment areas of the Keurbooms and Bitou Rivers.[2] The rainfall stations used in the analysis are shown.

Reddering[2] calculated the tidal prism, or volume of water entering and exiting the estuary over a tidal cycle, to be 1.8 x 10^6 m^3. This tidal prism is relatively large and is probably the reason why the Keurbooms has a permanently open inlet.[7,8]

A fresh water inflow of 5 m^3/s amounts to about 6% of the tidal prism volume over half a tidal cycle, though this volume will vary from spring to neap tide. These values indicate that it is probable that freshwater plays little part in the estuarine dynamics of the back-barrier lagoon, apart from the flood events.

Waves approaching a coastline at an angle will move available sediment along the coast in a longshore drift dependent on the wave height, the

beach slope and the angle at which the waves approach the beach.[9] With the dominant wave direction from the south/southwest[10] it means that the average longshore sand transport along the South African south coast is predominantly towards the west.

Potential sand transport calculations made for the Cape Recife/Algoa Bay area to the east found total transport volumes varied from around 100 000 m^3/annum to 400 000 m^3/annum[11], while in the Mossel Bay region to the west, values between 320 000 m^3/annum and 500 000 m^3/annum were determined[12]. The actual alongshore sediment transport rate is dependent on the availability of sand, but it can be expected that similar volumes will occur here.

Martin and Flemming[12] report the presence of a massive spit of unconsolidated sediment extending out from the end of the Robberg peninsula (Figure 1b). While this spit may indicate a deposition zone, longer period swell will move sediment in deep water past such capes,[13] so that Robberg should not be seen as a barrier to longshore sediment transport. Martin and Flemming[12] also found extensive deposits of Holocene sediments across Plettenberg Bay.

Plettenberg Bay conforms to a log-spiral or half-heart bay shape, with a resistant headland – the Robberg Peninsula – and a gently curving bay behind it. Wave energy can be expected to increase progressing around the bay away from the protection of Robberg until at the farther end where it is open to the Southern Ocean. The shape and characteristics of the beaches can also be expected to reflect these changes and Flemming[14] reports fine sand in the shelter of such a headland and coarser sand in the more exposed areas. This distribution conforms to the sand distribution found by Fromme[3]. The longshore drift of sand is therefore in a north/northeasterly direction in Plettenberg Bay.

As discussed, the migration of the Keurbooms Estuary mouth appears to be in a southward direction, i.e. counter to the longshore drift. This migration is a comparatively rare situation[4], with the tidal passing of the sand being over-efficient[9].

The manner in which sand progresses past the Keurbooms Estuary mouth has been discussed by Reddering[4], while Fitzgerald et al.[15] describe nine different bypassing models, although these are more specific to ebb-tidal systems. The processes occurring in the mouth are complex, with separate flood and ebb channels being formed, flood and ebb tidal deltas developing and being shaped by wave action, while at the same time the longshore drift continuously adds to a longshore drift delta extending northwards from the southern side of the mouth.

It is apparent in all these scenarios that the final movement of sand across the mouth of the estuary occurs as an event, where breaching occurs of a spit or of the longshore drift delta, allowing the ebb channel in particular a more direct route to the sea. The cut-off section of the longshore drift delta is left on the northern side of the mouth and melds onto the ebb-tidal delta, with the sand continuing onwards in the longshore drift.

High waves are also important in their direct effect on the estuary, causing a wave set up of the background sea level.[16] Over periods of days, coastal trapped waves can affect sea level by more than 0.5 m,[17] while the peak of a spring tide will add to background sea level. Zhang et al.[18] documented such an event where the sea level in the Gamtoos Estuary was increased by about 1.15 m, resulting in large scale inundation.

High waves and elevated sea level conditions can result in overwash of the barrier dune, where waves are large enough to go right over the dune and into the back-barrier lagoon beyond. These waves can erode a channel through the aeolian dune ridge and spill the sand into the lagoon in the form of a lobate washover fan. Reddering[2] estimated that between 1915 and 1942, a total of 325 000 m^3 of sand entered the southern part of the back barrier lagoon by barrier overwash.

Almost all South African estuaries have a subtidal sandy barrier that extends across the seaward section of the mouth.[8] This barrier is a wave-built sand ridge, distorting the tidal signal that enters the estuary. In particular, the tidal amplitude within the estuary is decreased relative to the adjacent ocean, and in most estuaries, there is also a tidal lag as well as an asymmetric tidal signal, with a shorter flood tide and a longer ebb tide. To enable the same volume of water to flow out on the ebb as flowed in on the flood means that ebb currents are generally slower than flood currents.[19]

Sand is moved in and out of the estuary by these currents. Bedload transport is roughly proportional to the cube of the shear velocity, and it means that small increases in current speed lead to marked increases in sediment transport.[20] With increasing shear, the sand starts bouncing, or saltating, along the bed, finally being put into suspension and carried in the body of the fluid.

The tidal asymmetry means that more sand is carried into the estuary on the flood tide than is taken out again on the subsequent ebb tide. This sediment transport is enhanced by waves, and it means that there is a net ingress of sand into the estuary. Moreover, current speeds increase in the narrow mouth, and once clear, the speeds decrease again and sediment in suspension settles. The settling velocity is dependent on the sand grain size, and flow which is slowing down will carry finer grains further: This process leads to a natural sorting where finer sands are found higher up in estuaries and such sand forms complex flood tidal deltas within the estuary. Correspondingly, ebb currents also speed up in the narrow mouth and then deposit their sediments when they slow down after exiting into the adjacent sea. In a high wave energy environment such ebb-tidal deltas are poorly developed.[21]

Data

The aim of the analysis is to investigate changes in the geomorphology of the Keurbooms Estuary, in particular during flood events. For this, rainfall data as well as water levels in the sea and estuary, wave heights in the ocean and estuarine morphology are available. Details are given below.

Rainfall and floods

Rainfall data from various stations surrounding the Keurbooms and Bitou rivers were obtained from the South African Weather Services (SAWS) for the period 1990–2013. Daily rainfall was measured at 08:00, although at all stations there were days when no measurements were taken.

Stations were selected on the basis of being relevant to the two river catchments, i.e. they needed to be close enough, and positioned on the southern side of the Tsitsikamma Mountain range. On this basis, the six stations shown in Figure 3 were selected to represent rainfall in the catchments. Mean values for every day over all six stations were calculated, the advantage being that missing values from any one station would not impact too greatly on the final value.

Water levels and waves

Sea level is measured by the South African Naval Hydrographer at Mossel Bay (Figure 1a) relative to Chart Datum, now defined as the Lowest Astronomical Tide. Land Levelling Datum is often referred to as mean sea level (msl) by surveyors, and for this analysis a value of 0.933 m was subtracted from the measured values to obtain mean sea level.

The site of the Department of Water Affairs (DWA) water level recorder is at the N2 bridge over the Keurbooms Estuary (Figure 1b). Hourly values are given in metres, although the actual values have not been adjusted to chart datum. Lack of adjustment means there was no opportunity to make a direct comparison with the sea level measurements.

It proved difficult to obtain detailed wave data for use in this analysis. As such, wave estimates from voluntary observing ships (VOS) data were obtained from the South African Data Centre for Oceanography (SADCO). The observers on such ships can be assumed to estimate the significant wave height, and thus their reports can be taken to represent waves and swell in the representative areas; only swell height is considered, with no consideration of wave direction.

The area from which VOS wave data was used extends from 34°S to 36°S, and 22°E to 25°E. It was found that the distribution of data varied considerably, and a number of days did not have any data at all. Nonetheless, only the relative changes in wave heights were used

here, and when there was data, significant changes in wave height could be ascertained.

Global Positioning System surveys and imagery

Global Positioning System (GPS) surveys were carried out when possible and represent a rough measure of the morphology of the Keurbooms Estuary at that time. They were executed by running/walking along the water's edge on the beaches and in the estuaries while marking out the route with a GPS; waypoints were taken at suitable locations. In general, the distances covered along the water's edge varied between about 3 km and 10 km, and could be completed in about 1–1.5 h. Both the low tide and high tide surveys were started on the ocean side of the estuary at about the predicted low and high water times. With the known lag of 1–2 h in the estuary,[2] it meant that the surveys in the estuary were finished during the corresponding estuarine low and high water levels.

Where possible, the surveys were done at low and high tides, and more specifically at spring low and high tides. These two surveys could then give some indication of the intertidal sediment distribution.

There is a certain amount of subjectivity involved in this process, particularly on the beach face, because the waves present at the time would have influenced the assessment of what constitutes the water's edge. Nonetheless, the technique gives a good representation of the estuary's sedimentary morphology, primarily because the changes recorded have been massive and small deviations do not affect the conclusions in any substantial manner.

The surveys up to and including July 2009 were undertaken using a Garmin 12 GPS (Garmin International, Kansas City, KS, USA). This instrument did not record positions continuously, and waypoints had to be taken at selected positions. From December 2009, a Garmin 60CSx GPS was (Garmin International) used, and was set to record positions every 3 s on the way around the water's edge, while some additional waypoints were still taken.

It proved difficult at times to include surveys on the northern side of the barrier dune as this involved getting across the mouth. At times, when high tide surveys could not be completed, estimates were made of the high water positions from the location of wrack and small scarps on the beach.

The accuracy of the GPS instruments is given as better than 10 m. Fixed reference points (shower, pumphouse, rock) were marked as waypoints for each of the surveys, and all GPS measurements on that day were then referred to this position. Conversion scales for latitude and longitude were obtained from 1:50 000 South African Naval Hydrographer (SAN) charts and programs were written to plot the results onto scale maps of 1 cm:100 m. These were transferred into Coreldraw, where the various features could be depicted. Google Earth maps on the same scale were used to include all the adjacent features such as roads, buildings and bush, while Google Earth images taken on 07 August 2005, 10 November 2010, 20 August 2011 and 30 August 2013 were also used to add to the survey results.

Results

Daily rainfall over the years 2006–2012 is shown in Figure 4. Intense rainfall events of more than 100 mm over 3 or 4 days are circled. Figure 5 shows the corresponding water levels at the N2 gauge in the Keurbooms Estuary. A standard symmetrical cosine–Lanczos filter with 97 weights and a quarter power point at 0.035 cph was used to filter these measurements, thereby effectively eliminating periods shorter than about 30 h, and in particular removing the tidal signals.

The correspondence between the intense rainfall events and the elevated water levels in the Keurbooms Estuary are clearly evident. In particular, the floods of 2006, 2007 and 2012 appear prominently, with the major flood of 2007 actually breaking the recording gauge. Nonetheless, there are a number of smaller events which stand out as potentially significant.

The flood of 2007

As indicated, the flood of August 2006, took away about 200 m of the barrier dune at the back of Lookout Beach, including dense bush and dune vegetation. After this flood, the normal estuarine sedimentary processes continued, and the general geomorphology for October 2007 is depicted in Figure 6. This figure also shows the positions of various sites mentioned in the text.

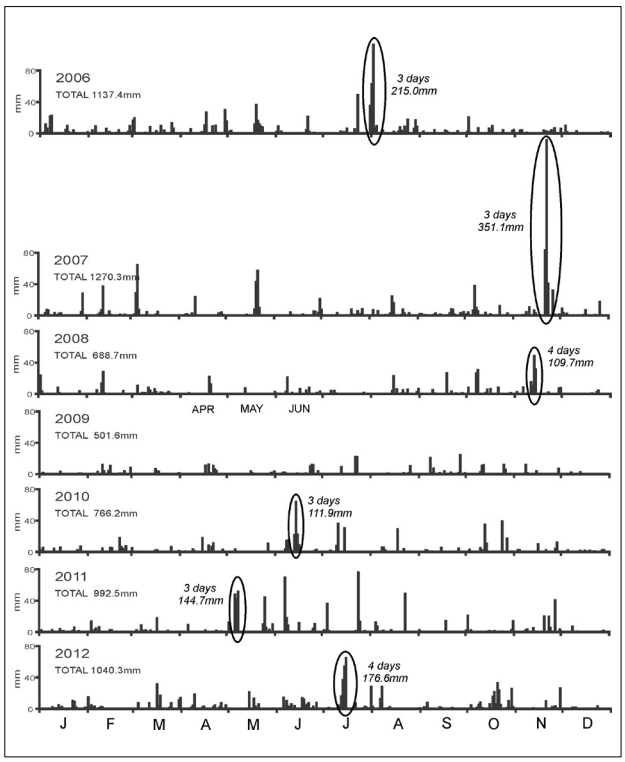


Figure 4: Daily rainfall from 2006 to 2012. Intense rainfall events of more than 100 mm over 3 or 4 days are circled.

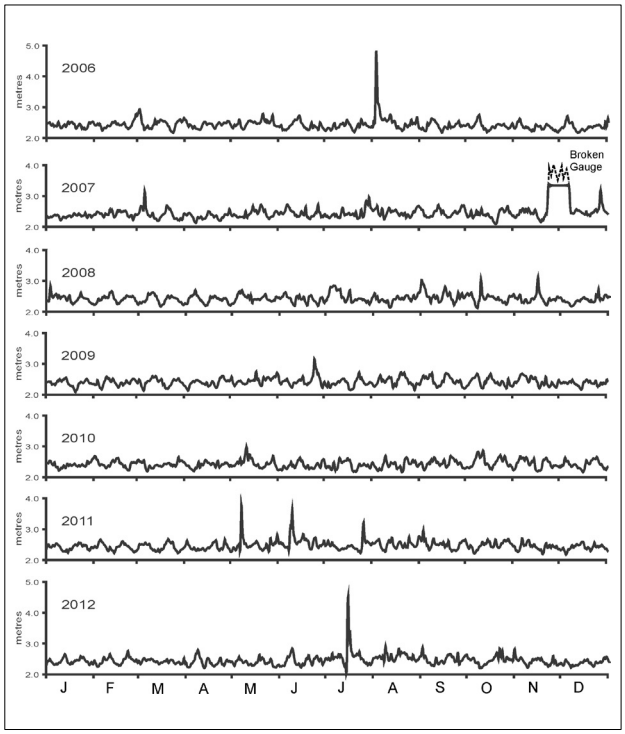


Figure 5: Filtered water levels measured at the Department of Water Affairs N2 water level recorder from 2006 to 2012.

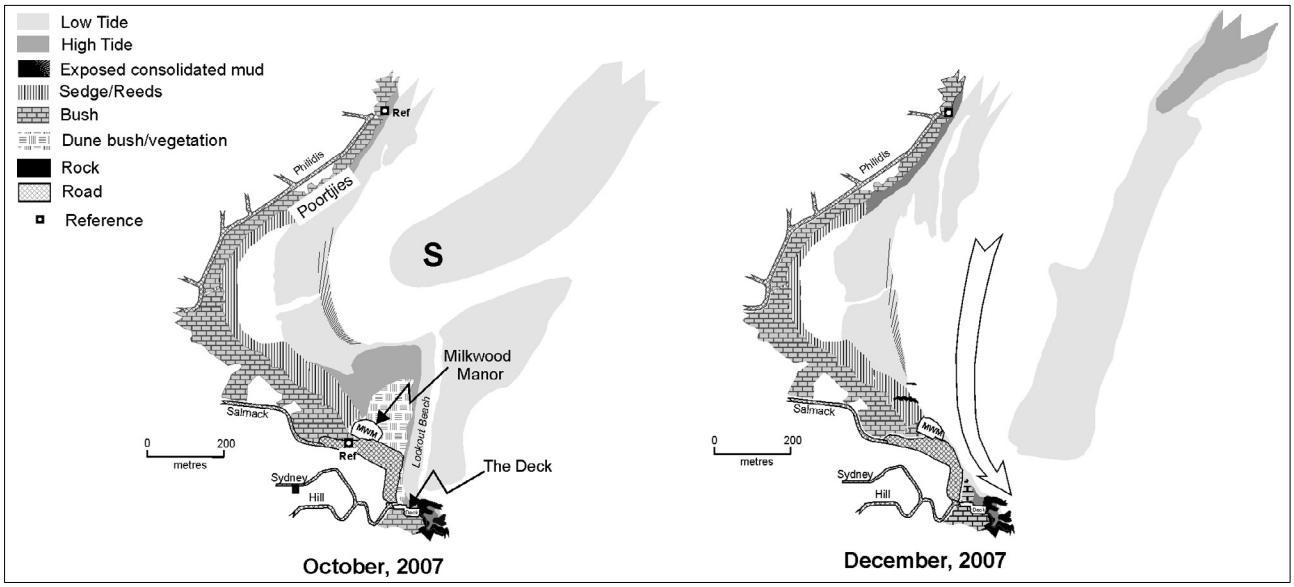


Figure 6: Sediment structure maps in the Keurbooms Estuary on the given dates. The legend gives the structures depicted, while the October 2007 map identifies sites mentioned in the text. The arrow depicts the main floodwater flow.

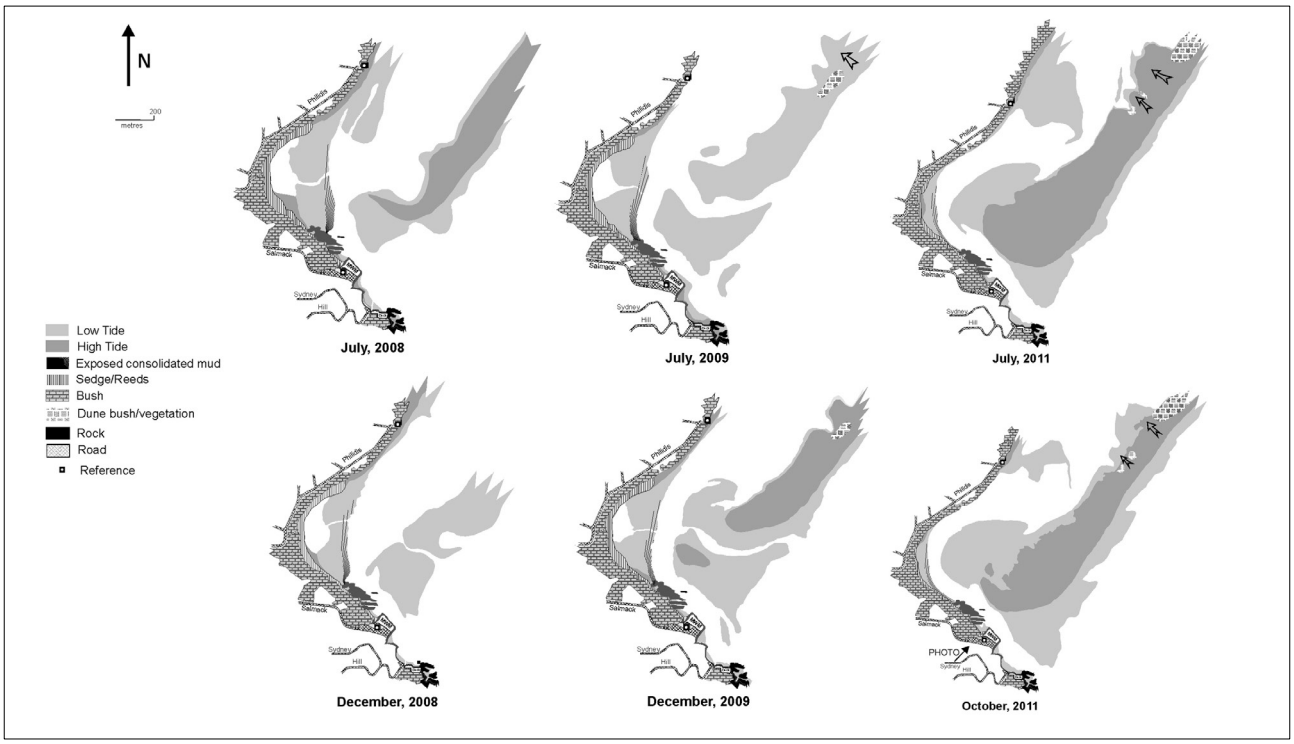


Figure 7: Sediment structures in the Keurbooms Estuary on the given dates. The arrows on the barrier dune indicate washover points.

Flows at a corner have slower currents on the inner boundary leading to deposition of sediment, while maximum currents on the outer boundary lead to erosion and possible scour holes.[22] The spit S formed such an inner boundary for both the flood tidal flow turning north-eastwards and the ebb tidal flow turning to exit through the mouth, and consequently accreted, extending south-westwards. Wave accretion also took place at the inlet beach face, because accretion is the only process that deposits sand above the water level[4].

On the other hand, the outer boundary eroded very slowly, as it consisted of consolidated fluviogenic mud and sand that had been subjected to extensive bioturbation. This outer boundary would have originated from the depositions over the years, from when the barrier dune was breached in the northeast in about 1915.[2]

At the mouth of the estuary, the longshore drift had formed an extensive longshore drift delta, with a low tide runnel in front of Lookout beach. This meant that a very convoluted channel served as the exit to the sea, particularly at the end of the ebb tide outflow.

The flood waters in November 2007, were too great to be constrained by the narrow, convoluted ebb channel and flowed directly out to sea, in the process removing the last bit of the barrier dune forming Lookout Beach. The Cape Supergroup rocks at the southwestern boundary of the estuary were exposed and were the primary reason that the estuary could not erode this bank any further.

Cooper[23] found that in such tide-dominated estuaries, floods typically erode accumulated sediment and deposit some of it as an ephemeral delta. This is what happened here, and Figure 6 shows that in December

a new subtidal barrier dune had formed with its tip some 400–500 m seawards of the previous position.

The area of sand forming the Lookout barrier dune and longshore drift delta was some 15.3 ha. In places amongst the dune vegetation, the dune was more than 6 m high, so the volume of sand removed in the flood was probably in excess of 500 000 m^3. This sand excludes any of the sand which formed the spit S, as some of that would have been washed out to form the new subtidal barrier dune.

Estuary development up to October 2011

After the massive flood in November 2007, Figure 7 shows the changes over the next 4 years. It is expected in such tide-dominated estuaries that reworking of the barrier dune by wave action should move it back more or less to its previous orientation within a relatively short period,[23] and this movement is evident by July 2008. It had also accumulated subaerial sand.

The owner of Milkwood Manor surrounded the guesthouse/restaurant with a structure of rocks, limiting further damage. However, the beach section to The Deck and the base of the Lookout Rocks was no longer protected from ocean waves and suffered severe erosion. In particular, a large part of the car park was removed and the toilet/shower facilities collapsed into the sea.

By December 2008, the ebb flow had continued eroding the consolidated muddy western boundary, while much of the shallow sandbank in front of Poortjies had been removed. Of particular interest are the changes to the barrier dune in that a new channel had developed about 200 m to the northeast of the main mouth channel. Moreover, the section of barrier dune shown was submerged at high tide, and even though the water levels at the time of measurement were some 200 mm higher than in July, the indications are that additional erosion had occurred; there were also several washover points further north.

Inspection of rainfall records in Figure 4 shows that a total of 109.7 mm of rain fell over the 4 days 11–14 November 2008, but there was only a small corresponding peak in the DWA water level record in Figure 5. Similarly, the 59 mm of rain that fell on 7 and 8 October also only caused a minor increase in water level.

The reasons for the changes can then probably be found in Figure 8. The VOS data showed that very high wave conditions occurred around 1 September, with wave heights greater than 8 m. At the time there was also considerable damage reported from other coastal areas along the south coast. Such storm waves would have eroded the barrier dune, with extensive washover areas occurring. One such washover was probably sufficiently large to form a channel, and once formed, water would have selectively exited along this channel. Moreover, the waves penetrating over the barrier dune would have been responsible for the erosion of the sand delta at Poortjies.

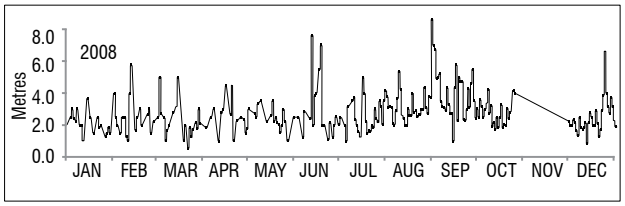


Figure 8: Daily significant wave heights from voluntary observing ships data for 2008. The dots indicate the values of wave height, joined by the thin line. There are numerous days when there were no data.

In July 2009, both the new secondary mouth and the Lookout Rocks mouth were operating, although the latter was substantially deeper. Note that the high tide survey did not extend to the barrier dune, although there was a small area exposed on the separate island, and the north-eastern barrier dune was also exposed some distance south-westward.

The input of sediment into the system was very apparent at the Lookout Rocks mouth, and a beach had built up between The Deck and Milkwood Manor, even at high tide. At Milkwood Manor, the sand at low tide extended well away from the protective ring of rocks. An ebb-tidal/longshore drift delta was evident to the east of the mouth.

By December 2009, some closing up of the secondary mouth had occurred, but substantial flows were still using this channel. Moreover, erosion of the west bank of the estuary had continued, with the low tide backwater also reduced in area. Erosion of the sandbanks to the north of The Deck had again occurred, and there was no beach available at high tide.

A low tide survey a year later in December 2010 (not shown here) confirmed that the secondary mouth had closed up and it is not evident at all in July 2011. Substantial changes had occurred in the flow patterns, with a large low tide sandbank extending out from the western bank at the toilet/changeroom reference point. As happened after the 2006 flood, the barrier dune had extended westward at its southern end as a result of being on the inner boundary of the flow, while more erosion had occurred at the western bank in the vicinity of the remnant backwater. This erosion led to increasingly convoluted flows. Extensive overwash fans were evident on the barrier dune, while clumps of dune bush had also become established in the north-east. Dune pioneers were also evident over much of the rest of the barrier dune exposed at high tide.

It should be noted that there were several minor floods in 2011 (Figure 5), which caused an appreciable increase in the water level at the N2 bridge – up to 1.5 m. However, it is apparent that the convoluted exit channel could cope with the additional flow, and there were no significant changes in the estuary's morphology.

The processes continued until October 2011. The sandbank extending from the toilet/changeroom reference on the western bank maintained its position, while the sandbank on the eastern side of the mouth channel continued accreting. However, of concern was the erosion of the western bank occurring in the Poortjies area, where even at low tide, in some places there was no beach in front of the bushes. Slow erosion continued on the western bank, but with the consolidated mud and vegetation resisting the continual flow.

The flood of July 2012

The next appreciable flood to cause major changes in the Keurbooms Estuary occurred over the four days 12–15 July 2012, when 176.6 mm of rain was recorded (Figure 4). The amount of rain is not much more than the 144.7 mm which fell over 6–8 May 2011; however, the water level at the N2 bridge increased by about 2.5 m, and this rise had a markedly different effect on the estuary.

The rainfall, sea level at Mossel Bay, as well as the water level at the N2 gauge are depicted in Figure 9, starting on 12 July and ending eight days later when conditions had returned to normal. In addition, Figure 10 shows that high wave conditions were experienced at the time, with peak wave heights close to 8 m being reported. Inspection of Figure 7 shows that in July and October 2011, overwash points were observed on the barrier dune south of the more heavily vegetated section. These points were evident, both from a lowered dune profile where waves had penetrated over the dune, and the washover fans where sand had subsequently been deposited in the estuary. These points were particularly vulnerable to breaching by the floodwaters and Figure 9 shows that the water level in the estuary rose rapidly from about midday on 14 July in response to the continuing rain from 12 July onwards. Such a damming up of water behind the barrier dune would have been caused by the convoluted exit channel past Poortjies being too small and obstructive to accommodate the volume of water flowing into the estuary.

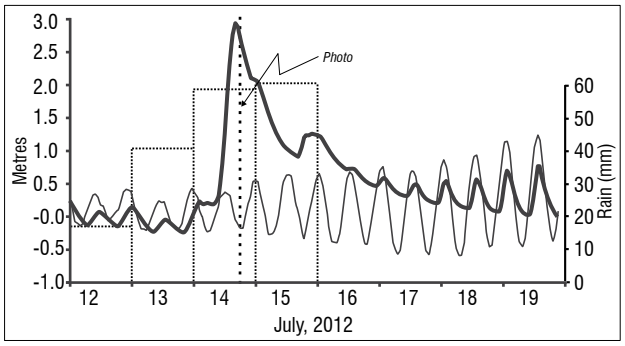


Figure 9: Hourly sea level data from Mossel Bay (light line), hourly water level data from the N2 gauge (dark line), and daily rainfall records (dotted line). The time at which the photo in Figure 11 was taken is shown.

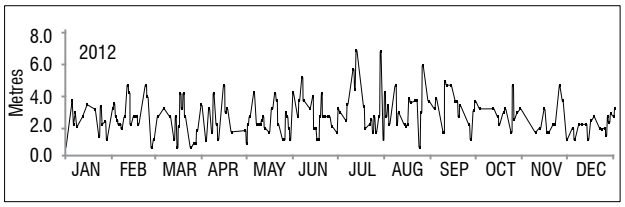


Figure 10: Daily significant wave heights from voluntary observing ships data for 2012.

The neap tide on 12 July and the ebb tide early on 14 July allowed more water to flow out thereby checking the rising water levels for a short time. The following flood tide served to limit this outflow, and the water level in the estuary rose rapidly. The high waves at the time would also have caused a wave set up thereby further blocking up the water in the estuary.

The flood water level reached its maximum and then from about 5pm on 14 July there was a rapid decrease. The reason for this decrease can be seen in the photograph in Figure 11, taken just after the peak water level had been reached (Figure 9).

It is apparent that a combination of the high water levels in the estuary, as well as probable overwashing from the high sea waves caused a breach in the barrier dune at the position indicated. This was an old overwash position observed in both July and October.

Figure 11: Photo taken at the time of the breach of the barrier dune (see Figure 9). The photo was taken from the square on Sydney Street depicted in Figure 7, October 2011.

At the same time, the mass of water going over the southern part of the barrier dune and flowing directly to a position just northwest of Milkwood Manor can be seen. Here the rocks deflected the water to the mouth. The floodwaters washed out numerous items from the estuary, included the boat shown in Figure 11.

Once the breach had been made, the dammed up waters started flowing out rapidly, leading to the abrupt drop in water levels. The new channel provided a much shorter and more direct exit to the sea and once started, would have been quickly scoured out further. Figure 9 shows that it was still raining hard at the time, and in fact, 15 July recorded a higher rainfall than 14 July. Nonetheless, the new channel could accommodate the additional flow and was being widened and deepened in the process. The water level continued falling, though the rate of decrease slowed down until late on 15 July when it had dropped some 2 m from its peak level. Then there was a slight increase again, possibly when floodwaters from another tributary reached the estuary, and thereafter the decrease in water level continued. It was only on 17 July that tidal fluctuations again reached the gauge, and on 19 July, the situation had returned to normal.

The surveys carried out on 17, 18 and 19 July are depicted in Figure 12 and show the new estuary configuration. The new mouth was situated about 1 km from the Lookout Rocks and was accommodating a substantial volume of the tidal exchanges. The remaining barrier dune was still intact, but was now an island between the two mouths. The sand delta extending from the reference point on the western bank of the estuary was still largely intact. Most of the floodwaters exited through the new mouth and thus did not pass and erode this delta.

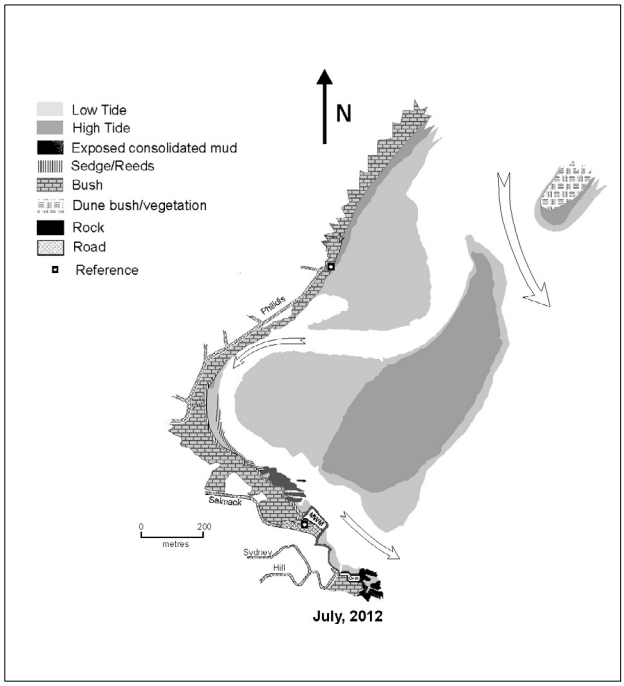


Figure 12: Keurbooms Estuary sediment structures after the breach of the new mouth in the northeast. Ebb current flows are indicated.

At this stage it was not clear which mouth would dominate. However, a survey in December 2012 revealed that substantial sandbanks had developed off the Lookout Rocks, extending across the old mouth. A survey in February 2013 showed that a low tide sandbank extended right across the mouth, and it was clear that this mouth would soon be closed completely. The closing happened by the time of a survey in July 2013.

Discussion and conclusion

This analysis followed the geomorphological changes in the Keurbooms Estuary over a period during which two substantial flood events occurred. The results confirm earlier conclusions[4,5] on the sedimentation patterns in the estuary mouth area, in particular the substantial longshore drift and the development of flood-tidal and longshore drift deltas, and smaller ebb-tidal deltas. As already discussed, the southwards migration tendency of the mouth is effected when the longshore drift delta extends too far northwards and is then breached by the ebb-tidal flow to provide

a more direct channel to the sea. The slug of sand at the end of the delta has then bypassed the mouth[15] and continues northward, with wave swash coalescing it onto the beach face.

However, the long barrier dune as well as the associated narrow back-barrier lagoon give the Keurbooms Estuary unique characteristics, and the manner in which waves and floods impact on the estuarine morphology also depends on the back-barrier channels. Thus the development of a south-west trending spit on the northern barrier dune forces the main channel westwards and southwards, in the process eroding the estuary banks and causing a more convoluted exit channel. This development is clearly evident in the sequence shown in Figure 7.

The impact of a flood will depend on the volume of floodwaters, as well as the resistance to flow of the convoluted backwater channel. Figures 4 and 5 show that there were at least two reasonable floods in 2011 which caused the water levels at the N2 gauge to rise around 1.5 m, however, there appeared to be little impact on the estuary. The flood of July 2012 raised the N2 gauge water level by about 2.5 m, and the impact on the estuary was massive. As already described, this impact was probably because the back channels had become more convoluted and the flood waters breached the barrier dune to find a more direct route to the sea. The high waves and overwash fan assisted in finding a suitable point to breach the barrier dune.

The convoluted back-barrier channel was also evident before the 2007 flood, although the volume of floodwaters at that time ensured that a new direct exit channel would be formed. This channel encompassed an extensive area of the southern part of the estuary.

If a new mouth is breached through the barrier dune, there is no guarantee that it will become dominant: this depends to a large measure on the duration of flow after breaching. The volume of water which exits will determine how well the new mouth is scoured, and whether it will then allow an easier route than the old mouth for tidal flows after the floodwaters have abated.

It appears that extreme wave events, such as those of September 2008, can also cause breaching of the barrier dune. The channel evident in December 2008 (Figure 7) remained operative for more than a year before it was eventually closed when the northward extending longshore delta became too long.

The effect of increasing vegetation on the barrier dune will also determine whether a breach can occur. Such vegetation will trap increasing amounts of wind-blown sand, causing the dune to grow both vertically and horizontally, making it more of a barrier to breaching.

Considering all these factors, it is clear that no definite predictions can be made about the effect of a flood, or indeed if and where breaching will occur. The available historical record from 1789 appears to show that the mouth was situated in the middle or southern sections of its range for most of the time, and that the breach in the north in 1915 was an unusual occurrence. Nonetheless, there is a tendency for the mouth to migrate southwards, although at any stage, breaching of the barrier dune could again occur at any other position. It all depends on the conditions at the time.

Acknowledgements

The provision of relevant data by the following authorities is gratefully acknowledged: the South African Naval Hydrographer for sea level data at Mossel Bay; the Department of Water Affairs for water levels in the Keurbooms Estuary; the South African Weather Service for rainfall data; the South African Data Centre for Oceanography for voluntary observing ships data, and the Chief Directorate: Surveys and Mapping for the 1936 aerial photographs. The work formed part of the Department of Environment Affairs – Nelson Mandela Metropolitan University (NMMU) Memorandum of Understanding, and the support of both the DEA and NMMU is gratefully acknowledged. One referee is thanked for valuable insights to improve the paper.

References

1. Storrar P. Portrait of Plettenberg Bay. Cape Town: Purnell; 1978.
2. Reddering JSV. The sedimentology of the Keurbooms Estuary [MSc dissertation]. Port Elizabeth: University of Port Elizabeth; 1981.
3. Fromme GAW. The dynamics of the Keurbooms-Bitou Estuary. CSIR Report T/SEA 8511. Stellenbosch: CSIR; 1985.
4. Reddering JSV. An inlet sequence produced by migration of a small microtidal inlet against longshore drift: The Keurbooms Inlet, South Africa. Sedimentology. 1983:30;201–218. http://dx.doi.org/10.1111/j.1365-3091.1983.tb00665.x
5. Duvenhage IR, Morant PD. Report no. 31: Keurbooms/Bitou System (CMS 19) and Piesang (CMS 18). In: Heydorn AEF, Grindley, JR, editors. Estuaries of the Cape. Part 2: Synopses of available information on individual systems. CSIR Research Report No. 430. Stellenbosch: CSIR; 1984 p. 64
6. Illenberger WKI. Management of Keurbooms Estuary sediment dynamics. Report for the Plettenberg Bay Municipality. Plettenberg Bay: Illenberger and Associates; 2003.
7. Reddering JSV, Schumann EH. Sedimentation: Causes and processes. In: Schumann EH, editor. Towards the management of marine sedimentation in South African estuaries with special reference to the Eastern Cape. WRC Report no.: 1109/1/03. Pretoria: Water Research Commission; 2003.
8. Cooper JAG. Geomorphological variability among microtidal estuaries from the wave-dominated South African coast. Geomorphology. 2001:40:99–122. http://dx.doi.org/10.1016/S0169-555X(01)00039-3
9. Carter RWG. Coastal Environments: An introduction to the physical, ecological and cultural systems of coastlines. London: Academic Press; 1988. http://dx.doi.org/10.1016/B978-0-08-050214-4.50007-3
10. Rossouw J. Review of existing wave data, wave climate and design waves for South African and South West African (Namibian) coastal waters. CSIR Report T/SEA 8401; Stellenbosch: CSIR; 1984.
11. Watermeyer Halcrow and Partners. The causes and prevention of periodic beach erosion of the southern beaches of Port Elizabeth. Part 1. Port Elizabeth: City Engineers Department; 1986.
12. Martin AK, Flemming BW. The Holocene sediment wedge off the south and east coasts of South Africa. In: Knight RJ, McClean JR, editors. Shelf sand and sandstones, Memoir II. Calgary: Canadian Society of Petroleum Geologists; 1986. p. 27–44.
13. Silvester R, Hsu JRC. Coastal stabilization: Innovative concepts. Englewood Cliffs, NJ: Prentice-Hall; 1993.
14. Flemming BW. Geology, morphology and sedimentology of estuaries and coasts. In: Wolansky E, McClusky DS, editors. Treatise on estuarine and coastal science. Waltham: Academic Press; 2011. p. 7–38. http://dx.doi.org/10.1016/B978-0-12-374711-2.00302-8
15. Fitzgerald DM, Kraus NC, Hands, EB. Natural mechanisms of sediment bypassing at tidal inlets. ERDC/CHL CHETN-IV-30, Vicksburg, MS: US Army Engineer Research and Development Center; 2000.
16. Raubenheimer B, Guza RT, Elgar S. Field observations of wave-driven setdown and setup. J Geophys Res-Oceans. 2001;106:4629–4638. http://dx.doi.org/10.1029/2000JC000572
17. Schumann EH, Brink KH. Coastal trapped waves off the coast of South Africa: Generation, propagation and current structures. J Phys Oceanogr. 1990;20:1206–1218. http://dx.doi.org/10.1175/1520-0485(1990)020<1206:CTWOTC>2.0.CO;2
18. Zhang P, Schumann EH, Shone RW. Effects of a storm surge coincident with a high spring tide in the Gamtoos Estuary, May 1992. S Afr J Sci. 1995;91:412–414.
19. Schumann EH, Largier JL, Slinger JH. Estuarine hydrodynamics. In: Allanson BR, Baird D, editors. Estuaries of South Africa. Cambridge: Cambridge University Press; 1999. p. 27–52. http://dx.doi.org/10.1017/CBO9780511525490.003

20. Open University. Waves, tides and shallow-water processes. Oxford: Pergamon Press; 1989.

21. Cooper JAG, Wright CI, Mason TR. Geomorphology and sedimentology. In: Allanson BR, Baird D, editors. Estuaries of South Africa. Cambridge: Cambridge University Press; 1999. p. 5–26. http://dx.doi.org/10.1017/CBO9780511525490.002

22. Dyer KR. Estuaries: A physical introduction. Chichester: John Wiley and Sons; 1997.

23. Cooper JAG. The role of extreme floods in estuary-coastal behaviour: Contrasts between river-dominated and tide-dominated microtidal estuaries. Sediment Geol. 2002;150:123–137. http://dx.doi.org/10.1016/S0037-0738(01)00271-8

Co-supervision in postgraduate training: Ensuring the right hand knows what the left hand is doing

AUTHORS:
Elly S. Grossman[1]
Nigel J. Crowther[2,3]

AFFILIATIONS:
[1]Faculty of Health Science Research Office, University of the Witwatersrand, Johannesburg, South Africa
[2]Department of Chemical Pathology, University of the Witwatersrand, Johannesburg, South Africa
[3]Department of Chemical Pathology, National Health Laboratory Service, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Elly Grossman

EMAIL:
grossmane@gmail.com

POSTAL ADDRESS:
PO Box 14, Cannon Rocks 6186, South Africa

DATES:
Received: 01 Sep. 2014
Revised: 18 Dec. 2014
Accepted: 18 Apr. 2015

KEYWORDS:
research higher degrees; novice supervision; cross-discipline; health science; contracts

HOW TO CITE:
Grossman ES, Crowther NJ. Co-supervision in post graduate training: Ensuring the right hand knows what the left hand is doing. S Afr J Sci. 2015;111(11/12), Art. #2014-0305, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140305

South Africa is experiencing a steep rise in postgraduate candidature and a backlog in research training and supervision. Co-supervision is a means to address such challenges. This study investigated how co-supervision could effectively and efficiently be implemented within a Faculty of Health Sciences. Supervisors and postgraduates brainstormed co-supervisory practice to identify: (1) the reasons for co-supervision, (2) what co-supervisors should discuss to facilitate their interactions and (3) how best to initiate the novice supervisor into supervisory practice. Co-supervisors are formally appointed for different reasons and all co-supervisory activities should be directed towards meeting the purpose of that appointment. Points to consider in facilitating a co-supervisor memorandum of understanding and novice supervisor training were discussed. Our findings provide suggestions to develop accountable co-supervisory practices, enhance novice supervisor training and to design discipline-specific best practice policy at institutional level to enable a common understanding of co-supervisory roles and responsibilities. Threats to effective co-supervision identified were the implications of co-supervision in staff promotion, inequitable workload recognition and no official acknowledgement of informal supervisory activities. Unless these issues are addressed, the full potential of co-supervision will remain unrealised. Supervision pedagogy and research teaching is a sophisticated skill worthy of professionalisation.

Introduction

South African academic institutions are struggling to rapidly establish a knowledgeable, qualified supervisory cohort to efficiently and effectively cope with the influx of expected postgraduates. Diminishing supervisory skills consequent to the ageing, experienced supervisory cohort have been well discussed[1] The growing need for supervisors, together with the desire for cross-disciplinary research to maximise innovation for future socio-economic benefits,[2] means that co-supervision will become the norm as opposed to the traditional apprentice-type supervision[1] to best address the proposed increase in research graduates as per the National Development Plan (NDP): Vision for 2030[3] and the Health Professionals Council of South Africa's (HPCSA)[4] recent policy. Co-supervision, whether cross-disciplinary or to augment core supervision, brings with it realities which could impact on the smooth functioning of the process. Differences in inter-supervisory expectations, varying departmental norms of how supervision is undertaken, divergent understandings of supervisory tasks and, to tap into the potential supervisory capacity of doctoral graduates outside of tertiary educational institutes, an entire neophyte cohort with no supervisory experience at all could give rise to any number of inconsistent approaches. Although the Higher Education Qualifications Framework requirements for doctoral degree candidates include: 'A graduate must be able to supervise and evaluate the research of others in the area of specialisation concerned'[5], in reality, a newly qualified PhD or master's graduate would find their first process of supervision daunting unless supported by an experienced co-supervisor and mentor.

South African understanding of co-supervision

We investigated how co-supervision could effectively and efficiently be implemented within the Faculty of Health Sciences at the University of the Witwatersrand (Wits). Co-supervision (also called joint or dual supervision) is used to support 33–70% of all South African PhD candidates[1] and can take various forms. On the one hand, a co-supervisor forms part of a formally appointed committee or panel where one postgraduate is supervised by between three and five qualified academics (commonly referred to as the American system). The committee members are selected on the basis of their expertise within the research field being investigated. The size of the supervisory contribution will depend on the portion of the research project which falls within the scope of each particular committee member. A principal or main supervisor, considered to be the head of the supervisory team, is charged with directing the research project and takes overall responsibility for the postgraduate. On the other hand, the more common approach is to use two formally appointed academic co-supervisors to take responsibility for the educational path of the postgraduate. At Wits, the roles of the supervisor and co-supervisor are described in the Senate Standing Orders on Higher Degrees as follows:

> *The Supervisor' is the person who is principally responsible for the supervision of the student and is responsible for 50% or more of the supervision. 'The Co-supervisor' is the person who is responsible for more than 10% and 50% or less of the supervision of the student.*[6]

In reality, this hierarchy and task allocation is not always followed. Obviously, there may be more than one 'co-supervisor' with percentage supervision assigned to each accordingly.

At the Faculty of Health Sciences at Wits, co-supervision is voluntary and not enforced as encountered elsewhere.[7] Co-supervisors are usually formally appointed at the research proposal stage or less frequently at another stage along the research process when it is clear that additional expertise is required. The initiative to appoint a co-supervisor is usually done by the primary supervisor and follows standard university procedures, although elsewhere, doctoral students themselves initiate the process.[8]

Apart from the formal co-supervisory structures, informal supervision of postgraduates also occurs, with differing degrees of support roles and institutional recognition. Informal supervision was found to supplement the formal supervisory arrangement for the majority of PhD students studied by Ives and Rowley[9] who concluded that such informal supervision served as a way of meeting all the *students'* research needs. Wingfield[10], on the other hand, suggested that informal co-supervision serves as a way to supplement the *supervisors'* needs by using the co-supervision potential (official or otherwise) of co-workers, postdoctoral students and senior postgraduate students to assist in the training of other postgraduates. In this way, Wingfield believed, academics are able to supervise more students than the potential ideal maximum of three postgraduate students per year.[10] Finally, the University of Edinburgh endorses the importance of informal supervisors. Their 'Code of Practice for Supervisors and Research Students'[11] reads:

> *In many research programmes other staff members will be involved in an informal advisory capacity, especially if specialised equipment is to be used. It is the duty of the principal supervisor to ensure that these informal advisors are prepared both to undertake this work and to take responsibility for matters of instruction and safety.*

The informal supervisor carries no obligations to meet deadlines or responsibilities for completing the research, might be acknowledged by the student for contributing to the completion of the degree, is seldom credited for the success of the student and never receives recognition from either the Faculty or Postgraduate Office for their inputs made to the work.

It can be seen from the above that co-supervision can take many forms and is subject to diverse regulatory practices. However, there is no working description as to what co-supervision encompasses or the range of activities within which such supervision could be organised, in the South African context or elsewhere. Furthermore, while the transference of research and related skills is core to supervision, anecdotally, South African supervisors appear to undertake far more administrative and procedural responsibilities than their counterparts elsewhere. Co-supervision is poorly covered in the literature[12,13] and is usually dealt with in general studies of supervisory practices. In many cases, reference to co-supervision is peppered with conflicting phrases such as 'co-supervision [is] believed to bring great benefit to both students and supervisors'[14] and 'the idea of co-supervision tends to present more problems and challenges than solutions in graduate research supervision'[15] with little to support either assertion[16,17]. For this reason, there is much need to explore both the supervisor and postgraduate experience of co-supervision to unpack and clarify co-supervision roles.[7,18]

Our initiative was therefore to 'get smarter' about co-supervision and investigate how this practice could be more effectively and efficiently implemented, given the steep rise in postgraduate candidature and the ill-equipped potential supervisor cohort. Our point of departure was that if there is no overall scheme of co-supervisory activities, there will be a lack of skills transfer and diffusion of responsibility, giving rise to problematic situations such as conflicting advice, lack of documentation, imbalance of workload, duplication of activities and at worst, no supervision at all.[13,19]

Methodology

Permission was granted by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (ethics clearance certificate M130120) to undertake this study. Two Supervisor Support Group focus groups were organised during February and April 2012 and co-supervision brainstormed, from both the co-supervisors and postgraduate perspective. (The Supervisor Support Group was started in the Wits Faculty of Health Sciences in 2007 to provide an informal forum for experienced and novice supervisors to exchange ideas, listen to presentations by guest speakers, discuss university higher degree policies and share interesting publications and news items. At the time of writing, there were about 160 participants from mainly the Health Sciences, with others from the rest of campus also joining the meetings).

Approximately 60 people attended the meetings and contributed to discussions. There was further email input from those who could not attend on the days or from those who chose to give a private contribution. Three topics were central to our thinking: (1) identifying the reasons for co-supervision within the Faculty of Health Sciences, (2) ascertaining what discussion points could form a 'memorandum of understanding' between co-supervisors and (3) identifying how best to initiate the novice supervisor into supervisory practice. From each topic, thrusts were developed to formulate questions and stimulate discussion around what the group identified as desired key outcomes of each activity, paths whereby the outcomes could be addressed and finally bridges and barriers within the system which could hamper or help co-supervisory practice. Crucial to our brainstorming was an understanding of best practice procedures for co-supervision and contextualisation of such best practice within the Faculty of Health Sciences. Extensive notes were taken by both authors during the debates and written contributions were scrutinised. Responses to targeted questions were listed and themes which emerged from the debates identified. Draft documents were circulated to Supervisor Support Group members, participating postgraduates, others who had contributed their thoughts as well as heads of schools and the deanery for comment. The extensive literature on supervisor practice was consulted to further develop our thoughts on co-supervision and how this practice could be enhanced within an academic environment.

Results and discussion

The guidelines that arose from these activities were grouped according to our three discussion topics. Suggestions on co-supervisory 'best practice' and novice supervisor mentoring are given in Boxes 1 and 2, respectively.

Circumstances under which formal co-supervision is considered

Co-supervision is usually considered at various stages of the supervisory process, commonly (1) during research protocol or proposal preparation and (2) during the course of the research project prior to examination. Appointment of a co-supervisor post-examination does occur and addresses an important gap in the progress of some candidates to graduation. Eleven circumstances relating to co-supervision were encountered in the Faculty of Health Sciences:

1. *Expert* advice which can include specialist subject knowledge, administrative, bureaucratic or procedural knowhow. This reason is the most frequent given for co-supervision.[12,13,16,19,20]

2. *Load sharing* by division of labour is also a commonly cited reason for co-supervision.[7,19] A recent publication[21] linking academic staff burnout to rising numbers of postgraduates predicts an associated escalation in co-supervision.

3. *Local vs distant* (or guardian) co-supervision[7,22] occurs when postgraduates are registered at the home faculty but work outside in industry or beyond South Africa's borders. Co-supervisors are appointed at both sites to support the postgraduate and to ensure continuity and oversight. Local co-supervision is also called upon to ensure that departmental requirements are met when the subject supervisor is outside of the student's registered home department. The roles of distant and local co-supervisors are set to increase with the upsurge of joint and double degree programmes to meet developments in the internationalisation of higher education.[23]

4. An academic mobility '*safety net'* to counteract the consequences of imminent supervisor retirement, sabbatical or re-location.[7,9,12,13,24-26] This reason for co-supervision is set to rise as South African universities shift from full-time to part-time academic staff on short-term contracts.[27-29] In this way, South Africa is following the international trend towards increased proportions of sessional staff on short-term contracts[30], which has the potential to impact on

supervisor continuity and postgraduate completion times.[25] Rather unfairly, it is the postgraduate who bears the consequences of university short-term staffing policy viz. uncertainties regarding supervisor appointment, a vacuum when the supervisor is lost and the shift to realign their research to meet the new supervisor's vision. A 'safety net' co-supervisory arrangement will provide an essential line of continuity and support in such cases.

5. A university *regulatory* co-supervisor to compensate for the large number of clinical specialist academic staff who lack a formal research qualification. Most South African medical schools permit clinical specialists to supervise postgraduates if they have an independent, published research record. However, it is obligatory to appoint a 'regulatory' co-supervisor who has the degree equivalent or above that which is being supervised, to comply with the Council for Higher Education policy[31], a requirement similarly reported in other health sectors.[32] The university regulatory co-supervisor is generally not an expert in the clinically oriented research topic, but will actively co-supervise the postgraduate in aspects of research approaches, higher degree format and layout, and oversee the administrative processes required for degree completion.
6. The training of *novice* supervisors by experienced supervisors.[12,13,33]
7. *Pragmatism.* 'I always have a co-supervisor so that my postgraduate has someone to go to when they are fed up with me.' An insightful comment from an esteemed supervisor and researcher, sensitive to the emotional dimension of the supervisory process.[34]
8. *Financial incentives* are an increasingly important consideration because of the new university funding formula in which research monies in the form of subsidies and incentives are assigned to supervisors following successful completion of higher degrees.[1,35] Such incentives can increase the co-supervisory pool but can also encourage freeloaders.[24]
9. *Policing* roles have become necessary to keep tabs on errant supervisors, often the excellent researcher with a productive publishing record who attracts postgraduates to their laboratory for perceived benefits of exposure to cutting-edge research. Unfortunately, supervision responsibilities are lacking and the 'policeperson' appointed from the outset is able to preside over completion of the postgraduate process and ensure university obligations are met. A 'policeperson' usually oversees the process, not the research per se, and could be regarded as a formally appointed 'critical friend'.[13,36]
10. When the supervisory track derails for whatever reason, a *minder* could be called upon to forestall a total breakdown in the relationship. This co-supervisor is often a senior academic who acts as a buffer between the supervisor and postgraduate but never takes over the supervisory role. The minder and policeperson roles appear similar, but the minder is appointed when a problem arises during the supervisory process. A policeperson is there from the start to contain an inevitable problem.
11. *Rescue* co-supervisors are formally appointed to oversee the substantial re-working of the thesis following examination. Often the amount of re-working is extensive and well beyond the ability of the supervisor. In our experience, the problems encountered in the thesis are the result of inexperience of the supervisor or negligence on the part of the postgraduate or both. A 'rescue' co-supervisor works with the postgraduate and supervisor to correct the thesis and prepare it for re-examination. Co-supervision in this case is regarded as a formative exercise for the supervisor and student and constructive engagement occurs with both parties to ensure that the 'rescue' is a positive learning experience.

The co-supervisory circumstances or roles which we have identified should not be regarded as a complete listing of all possible interventions: in other faculties in other countries, other co-supervisory experiences will occur, which require categorisation. Whatever the case, it is recommended that the purpose of the co-supervision should be clearly identified at the outset, and supervisory roles explicitly directed and executed to support those purposes efficiently, optimally and without duplication. Noteworthy from the above is that the latter three co-supervisory roles, i.e. the policing, minder and rescue roles, have not previously been described. The co-supervisors in these instances may reflect South African supervisory realities as highlighted in the PhD study by ASSAf[1] viz. increasing post-graduate numbers, ill-prepared students, high undergraduate teaching loads and administrative responsibilities together with staff having academic requirements for supervision, but lacking the experience, aptitude, interest or time to do the job. Even dedicated supervisors bewail their inability to supervise optimally because of university pressures to chase and raise third stream research funding.

Supervision is crucial to doctoral students' successful completion of their theses[16,37] and this applies to co-supervisory practices as well. Our emphasis in the following section is on developing a mutual agreement as to the roles of each co-supervisor with the idea of developing a memorandum of understanding between co-supervisors, similar to the postgraduate-supervisor contract, rather than the deeper considerations of tertiary education supervision pedagogy.

Co-supervision activities and outcomes

Box 1 is largely self-explanatory. In summary, co-supervision needs to be actively managed with all co-supervisors starting on common ground with periodic, built-in reviewing activities flagged within the programme on an ongoing basis. They need to agree on expectations of the research project, the postgraduate and their own roles and responsibilities, as well as those of the postgraduate. It would be preferable to include the postgraduate in some or all of the co-supervisory discussions, but that can be determined by the circumstance. Ongoing progress is reliant on keeping co-supervisory relationships open and transparent, regularly reviewing the postgraduate's advancement; meeting administrative requirements; monitoring project development and fulfilling co-supervisory roles. All these require an enabling environment, the right frame of mind and a constructive attitude towards diversity, flexibility, willingness to learn and recognition that if things go wrong, it is the postgraduate who suffers.

Divergent interpretations of the work needed for the various research degrees offered by the Faculty are not uncommon.[32] Divergent inter-pretations, together with differing research and supervisor experience levels, and understandings of what supervision, let alone co-supervision entail, led one of the younger participants to exclaim, 'Are we supposed to do all that?' when viewing the multiplicity of tasks identified in the two boxes. Delany[38] warns of the pitfalls when presenting supervisory practice as a series of lists, as we have done, as this implies a fragmented approach to what is in reality a rich array of dynamic activities and multifarious phenomena. However, he acknowledges that effective supervision requires 'a salient need for a program of coherent empirical validation'. Similarly, Lahenius et al.[8] concluded that co-supervision in the engineering disciplines would benefit from specific rules and responsibilities.

The current study was undertaken to provide guidelines to establish and support the co-supervisory needs within a Health Sciences Faculty environment, guidelines that we believe would apply to many other academic institutions. We further contend that listing supervisory duties, as we have done, is essential to convey the scope and variety of tasks expected for successful co-supervision, be it intra-departmental, cross-disciplinary or for the training of novices. A poor understanding of the range of tasks expected, without explicit discussion of the scope of each, can impact negatively on the nature of the relationship between co-supervisors and theirs with the postgraduate. No matter which reason forms the basis for co-supervision, it is imperative that a memorandum of understanding be drawn up between co-supervisors as a matter of course, just as the student-supervisor 'contract' is utilised in many tertiary academic institutions. Such a document is increasingly appropriate in an academic climate where universities must demonstrate to the tax-payer accountability, quality assurance and quantifiable training outcomes which are likely to facilitate successful postgraduate student completion.

Box 1: What co-supervisors should clarify in advance

1. Common aims of the research project
2. Common goals for dealing with the postgraduate student
3. Formal requirements of the higher degree to be supervised
4. The division of work (to be reviewed or renegotiated annually or biannually)
5. Informal co-supervision policy
6. Time commitments and availability
7. Roles and contributions of each co-supervisor towards the project and postgraduate with regard to:
   a. The purpose of the co-supervision, i.e. distinct roles that complement each other without duplication of effort or diffusion of responsibility
   b. Which aspects of the process will be dealt with by whom
   c. Who will take final responsibility for the content, bureaucratic process, meeting administrative deadlines etc. Will there be a 'principal' supervisor who has the final say on all matters?
   d. Commitment to the postgraduate student and to each other
   e. Expectations of the postgraduate and of each other
8. Roles regarding facilitating the project:
   a. How much help does each co-supervisor think is reasonable to give a postgraduate, e.g. the amount of technical, statistical or editing assistance permitted?
   b. How long will a postgraduate be allowed to struggle before a co- supervisor offers assistance?
   c. Who will oversee the development of research skills, methodological knowhow, technical skills, information retrieval etc.?
9. Style of supervision
   a. Will it be 'hands on', 'hands off'; according to an existing university 'supervisor–student contract'; as the need arises; an acknowledged educational approach; or to each his own?
   b. Are there specific aspects of the process that are thought to be critical and need to be understood by the other supervisors. For example:
      - Industrial or funding partnership interventions
      - Embargo on data or confidentiality clauses
      - Informal co-supervisor policies
      - Intellectual property
10. How the drafts of the research report/thesis will be handled
    a. What is a reasonable turnaround time on drafts?
    b. How many drafts is it reasonable for a member of a supervisory team to read?
    c. Will both co-supervisors read each draft or only portions of the draft?
    d. Will the same draft be read consecutively by each co-supervisor so they can see the others corrections or will each supervisor receive a copy of the same draft to comment on?
    e. What are the arrangements for the final draft regarding reading, correcting, standards to be met and overall quality of the final thesis prior to all co-supervisors signing the document off for examination?
    f. What sort of assistance should the postgraduate be given in terms of writing?
11. Meetings
    a. Will there be an agreed format for keeping records of meetings?
    b. Will both co-supervisors, together, meet with the student or have independent meetings with the student?
    c. Who will organise meetings? How often? Where?
    d. Will meetings be minuted and if so by whom, to what level of detail and will the notes be copied to both co-supervisors and the postgraduate?
    e. How often will co-supervisors meet on their own to discuss the postgraduate's progress?
12. Workload percentages
    a. What percentage split is accorded a 'primary vs secondary' or 'senior vs junior' co-supervisor?
    b. If each co-supervisor is responsible for different aspects of supervision, how is each aspect recognised and split?
    c. Financial compensation split (incentive) should be clarified upfront and where necessary, agree with university policy
    d. Workload percentages should be revised from time to time
    e. The line of supervision should be clearly defined (if required)
13. How conflicting advice to the postgraduate will be handled
    a. If views differ, does each advise the student of their recommendation and let the postgraduate reach an independent decision or should the co-supervisors try to reach a consensus first and present a united approach?
    b. Will the principal co-supervisor make the final decision?
14. Communication
    a. Will all emails be copied to co-supervisors and the postgraduate?
    b. Will the postgraduate send emails to both co-supervisors, or just to one co-supervisor? If so, which one?
    c. How will the long absence of a co-supervisor be managed?
    d. What vacation leave or study leave will be granted to the postgraduate?
    e. What notice of leave should be given and how far in advance and should it be given to all concerned?
    f. What will be the best way to access one another, especially if off-campus?
    g. Have a clear contact procedure for all when crises emerge
15. University requirements and policies
    a. Who is responsible for ensuring the candidate achieves milestones and meets the administrative requirements in a timely manner?
    b. How will the postgraduate be kept up to date with policies and procedures?
    c. Whose role is it to ensure that the student knows, and follows, policies related to research ethics, plagiarism etc.?
    d. Whose role is it to take the student through the ethics application process and sign-off?
    e. Who will take care of the postgraduate administration file, progress reports and records of the supervision?
16. Provision of funds and facilities for the postgraduate's research project. Who will be responsible for:
    a. Overseeing the payments and budgeting of the project
    b. Arranging access to basic resources such as bench space, computer etc.
    c. Fieldwork and conference funding
    d. Communicating with the funder as regards regular reports, dissemination of funds, etc.
17. Publishing and conference attendance expectations:
    a. Supervisors' expectations with regard to the postgraduate publishing and presenting
    b. Authorship on papers – who will appear as authors and how will the order of authors be determined?
    c. What contribution is reasonable without being a co-author or alternatively, the size of the contribution required to be named as a co-author?
    d. Who will take charge of conference organisation, registration, costs and presentation preparation?
18. The process for discussing concerns and conflicts:
    a. About the candidate
    b. With and between co-supervisors
    c. If the postgraduate goes to one co-supervisor with a problem about another supervisor, how will this be handled?
19. Integration of the postgraduate. Who will:
    a. Introduce the postgraduate to the department and staff members
    b. Take the student on a campus tour
    c. Ensure peer group assimilation and common room access
    d. Ensure participation in seminars, journal clubs, meetings and academic social events
    e. Obtain personal, family background and contact information

Box 2: Mentoring of novice supervisors: Training and support

The mentor should:

1. Establish goals for the novice supervisor
   a. Annual plan with milestones
2. Assist the novice in engagement with the student
   a. Encourage the novice to lead discussions during supervisions
3. Provide referencing and library assistance
4. Provide networking opportunities for the novice supervisor within the academic environment
5. Critique drafts:
   a. At first, the experienced supervisor should forward drafts with comments to the novice for their input. A meeting between the two supervisors should be arranged before meeting with the student to discuss the feedback and present a common approach to the postgraduate.
   b. With time, the process should be reversed with the novice providing initial comments.
6. Provide personal feedback on their co-supervisory performance
7. Clarify the novice supervisor's role with respect to the experienced supervisor's role as per Box 1
8. Encourage the novice supervisor to attend supervision courses
9. Ensure the novice has knowledge of the university administrative and policy requirements for higher degrees. The novice should keep copies of all updated documents to hand such as:
   a. University Standing Orders and Policies dealing with postgraduates and supervisors
   b. Faculty rule books and handbooks
   c. Faculty Standing Orders and documents from the Faculty Graduate Studies Office
   d. Ethics requirements and procedures
   e. Protocol and examination procedures
   f. Official 'supervisor–student contract'
   g. Details of student support services and international student policies
10. Keep watch to ensure the postgraduate, under instruction from the novice, complies with rules, especially concerning submission dates for assessors for the protocol, ethics clearance, annual registration, dates for fee payments, length of registration requiring motivation for an extension and submission for examination
11. Encourage the novice to have an independent research and publishing career
12. Monitor time management of both the novice supervisor and the supervisory/research process. The novice has probably realised from their own experience that completion of a higher degree inevitably takes longer than anticipated
13. Encourage active acquisition of new research skills and extend knowledge outside the novice supervisor's specialist research field
14. Assist the novice in establishing a supervision style preference, ranging from very hands-on and directive to very hands-off and non-directive. In addition, ensure the novice is well acquainted with different supervisory approaches, is able to critically evaluate educational assumptions and has a good understanding of when and which supervisory roles to optimally apply to fully meet the needs of individual research students at any given time
15. Ensure the novice is aware of funding and grant opportunities and procedures

Mentoring of novice supervisors

While there are many academic articles which focus on the role of the supervisor in the intellectual development of the postgraduate, there is an absence of similar texts on the development of the novice supervisor by the experienced supervisor. Although such a mentoring activity seems straightforward, a one-on-one relationship such as this can be better and more effectively served if an explicit agenda is prepared to chart tasks and procedures which the novice needs to know in order to be effective in her/his early supervisions. Our concerns are pragmatic rather than driven by the values, beliefs and concepts of supervision *per se* and we focus on tasks which will ultimately set the novice up for their supervisory careers. In Box 2 we suggest ways in which the novice supervisor may be eased into their supervisory role prior to embarking on independent supervisor practice. The list is not definitive, but can form the basis for interaction between the novice and experienced supervisor in most situations.

As a point of departure, the following was generally agreed upon as being good novice supervisor mentoring:

- A novice supervisor should not supervise on their own, a decision which is in line with other codes of practice.[11]
- Even with a co-supervisor, a novice supervisor with a master's degree should start by supervising honours students. Those with a PhD should start with a master's student before negotiating a PhD supervisory role (assuming the newly qualified PhD has had some experience overseeing and supervising honours projects).
- The novice supervisor should begin by overseeing a project within their field of research and methodological expertise.
- The first postgraduate assigned to the novice supervisor should be of reasonable standard and the novice should preferably have some say in the selection of the postgraduate.

The 'see one, do one' of novice supervisor development is insufficient[39] and this sentiment was echoed in our debates. There was consensus that the novice should undertake at least three co-supervisions before going solo and if possible, be mentored by three different supervisors, preferably with related, but differing, research interests. Thereby the novice would be exposed to different supervisor styles with the added benefit of widening their research horizons with different research topics and methodologies. This would permit the novice supervisor to acquire new research skills and provide further impetus towards developing an independent research and publishing career as described above. As to who should co-supervise with a novice supervisor, one respondent, a potential novice supervisor, remarked:

> *It would be nice when a novice is co-supervised by a willing hard working mentor with a good track record. Now who is going to make a list of good and bad mentors? What criteria are best to be used? Publications, complaints or comments from others?*

Buttery et al.[24] warns that students are unimpressed with their status as a postgraduate 'guinea pig' when they perceive their supervisor is an 'on the job' trainee. A more structured intervention when training novice supervisors, as we advocate, has the added spin-off of reassuring the supervised postgraduate that experienced oversight is present and makes them more amenable to the process. Contrary to ordinary co-supervision practice where co-supervisors can be called in at any stage of the research, Stack[40] feels that co-supervision with a novice should involve joint supervisory consultations with postgraduate students from the very start of the research process. The novice should play an active role in the choice of topic, designing of the research, carrying out the fieldwork and data analysis and giving feedback on written drafts submitted by the student. Bringing the novice supervisor in at a later stage is inadvisable.

It happens all too often that supervision activity is so directed towards the postgraduate that the interaction between experienced and novice supervisor gets overlooked. It is paramount that the novice and experienced supervisor have meetings separate from the postgraduate so that supervision mentoring can take place apart from postgraduate mentoring. During such sessions, administrative and procedural aspects

of supervision can be conveniently covered (see point 9 in Box 2). In South Africa, the focus of the registered postgraduate is on the research and the student is seldom party to the 'behind the scenes' administrative aspects of the degree. Consequently, when they oversee postgraduates, they are surprised to find this as an expected supervisory responsibility. Thus we recommend that a programme of continuous feedback be built into the mentoring co-supervision calendar and is best done at the direction of the experienced supervisor,[40,41] who offers suggestions in this regard to the novice to maintain morale.

Most novice supervisors base their supervisory practice on their own experiences as a research student, whether good or bad.[7,33,34] Novices are often perplexed to find that utilising strategies which worked for them is not necessarily appropriate for their supervisees.[41] Thus, novices should be aware of the variety of supervision styles available[36,42-46] and the outcomes of each. Formal training is often used to expose the novice to the entire gamut of supervisory approaches and to critically evaluate each one's assumed value. However, Turner[41] cautions that for practical purposes, formal training is insufficient for many doctoral supervision dilemmas. A further challenge is the current predominance of an overtly Western approach to learning which does not take into account the more didactic approaches found in other parts of the world.[47] This challenge has implications for South African research co-supervisors working with divergent cultural groups and the increasing numbers of international students[48] accessing South African tertiary institutions.[1,28]

The assumption that a research degree is adequate for effective supervision is damning.[1] Shannon[49] reflects: 'It would be very rare for a person inactive in research to be even a barely adequate supervisor in mathematics or science.' Such sentiments are reflected in other disciplines too.[12,15,24,32] Indeed, Grossman et al.[50] recommend that independent supervision can only be satisfactorily accomplished once the supervisor has authored or co-authored four accredited research publications. Thus, the experienced supervisor should encourage and assist the novice in developing an independent research and publishing career during the novice co-supervision experience, where postgraduate research training can occur in tandem with novice research mentoring.

Co-supervisory threats and systems

Although there was agreement that co-supervision is beneficial to postgraduates, supervisors and university throughputs alike, some threats to healthy and productive co-supervision were raised of which three generated a great deal of discussion.

The first threat was noted by the following response:

> *I am now trying to go through the motions of applying for promotion; I get a sense that it does count against my application having all my students co-supervised. It has not been said in so many words, but the gist of it is there.*

This perception was substantiated by participants who encountered similar experiences when seeking promotion or during job interviews. Co-supervision was seen as militating promotion and advancement prospects and therefore to be avoided. Tellingly, University policy is silent on the merits of co-supervision versus sole supervision in staff promotion and the literature appears equally silent on the matter.

Secondly, there was considerable debate about the practicalities of implementing point 4 of Box 1. A typical stance is illustrated by the following:

> *I've also had several cases where the bulk of the workload of supervision and all the nitty-gritty technical stuff gets dumped on me because the others tend to disappear. ... in the end the other person gets all the kudos for the work that I do.*

Our situation is not unique. Buttery et al.[24] urges equitable reflections of effort in the workload agreement, stressing that requisite expert input may vary throughout the thesis lifecycle. Workloads and financial compensation of co-supervisors have been implicated by Spooner-Lane et al.[13] as being subject to power play and institutional duplicity, and incidents of such were recounted by our group members. Universities should heed Shannon's[49] warning: 'If there is no workload recognition for the supervisory role then it may not be done well or it may be avoided except for the most dedicated.'

Finally, the substantial inputs expected of 'experts' to give of their time as informal advisors without formal recognition also raised ire during discussion. A postgraduate describes his predicament which necessitated him seeking informal supervision:

> *My first draft of the thesis (intro and materials and methods) has been looked at by Mrs R, whilst the draft sent to Supervisor 1 is still on his desk since ... [eight months prior] and has not been looked at once. If I would need to guess, Mrs R and [you] will be the only ones I can effectively rely on ... for outside help as Supervisor 1 generally has made poor attempts in this area, with Supervisor 2 only in it for the ride.*

Informal postgraduate assistance has occurred because participants felt 'sorry' for the student or else did not wish to be perceived as churlish or 'un-collegial' when refusing 'informal' supervisory requests from colleagues. However, informal supervisors expressed resentment at the lack of recognition for their inputs and unhappiness about claims made on their time, similar to those described by Spooner-Lane et al.[13] Participants wanted a formal university policy or guideline to regularise 'informal' co-supervision so that their inputs could be officially recognised. This demand is not unreasonable, considering the curriculum vitae orientated job market, the new university funding formula subsidy base and expanding network of informal supervisors needed for increased higher degree throughputs. Indeed, Deuchar[36] goes so far as to predict a future escalating reliance on such advisors to meet the current 'neo-liberal focus on efficiency, economies of scale and the image of students as customers with rights' in higher education performativity.

These three outcomes were completely unexpected and appeared to be at odds with the university desires of improved quality and quantity of postgraduate throughput. Unless suitable changes are made to university regulations to positively address the role of co-supervision in employee advancement, equitable workload recognition and regularising informal supervisory activities, the full potential of co-supervision will remain unrealised.

In the final analysis, our results show that one of the ironies of the 'secret garden'[51] is that, in hiding supervisor-student practices, it has also obscured the enormous amount of indirect work and tasks associated with student supervision. This secretive practice has dumbed down the role and range of professional skills, administrative rigour and vast institutional knowledge associated with supervision to such an extent that supervision is poorly regarded by many non-supervising academics and non-academic faculty. Furthermore, supervision appears to be largely considered by university bureaucrats as an 'add-on' to lecturing tasks. There is no notion that the supervisor performs a complex variety of tasks and roles,[44] many only remotely related to monitoring and improving postgraduate performance,[42] and of an increasing awareness that supervision pedagogy and research teaching is a sophisticated skill.[48,49] The latter was enlarged upon during our discussions, concluding with the view that the time has come to create the niche of postgraduate supervision within university employment categories and to give it a professional ranking much like lecturer, tutor, specialist, scientist and other academic categories.

Conclusion

Our findings can provide a starting point for the development of co-supervisory practices, enhancement of training of novice supervisors, and the design of cross-discipline 'best practice' policy at an institutional level. The fleshing out of the bare bones of our Boxes 1 and 2 could form the basis of a co-supervisor memorandum of understanding and we suggest discussions should occur within departments to meet specific needs of local conditions and academic traditions. Additionally, these findings might also stimulate topics for reflection and discussion among educators and

administrators in a higher education climate, which increasingly requires evidence of the organisation and administration of the postgraduate research degree as a tangible verification of quality assurance.

Acknowledgements

We acknowledge the members of the Supervisor Support Group who contributed toward the discussion on postgraduate student co-supervision.

Authors' contributions

E.S.G. conceived the project; and both authors participated in the design, conceptual contributions and completion of the work. E.S.G. wrote the manuscript with inputs from N.J.C.

References

1. Academy of Science of South Africa (ASSAf). The PhD study. Pretoria: ASSAf; 2010.
2. Van Jaarsveld A. Research capabilities reverse SA's brain drain. Mail and Guardian. 2014 July 11–17; Supplement: SA Research Chairs Initiative; p. 1.
3. National Planning Commission. National Development Plan: Vision for 2030. [document on the Internet]. c2011 [cited 2015 Oct 21]. Available from: http://www.poa.gov.za/news/Documents/NPC%20National%20Development%20Plan%20Vision%202030%20-lo-res.pdf
4. Subcommittee for Postgraduate Education and Training (Medical), Health Professionals Council of South Africa (HPCSA). Guideline on the minimum conditions of service, education and training of registrars and subspecialist trainees in South Africa. Pretoria: HPCSA; 2010.
5. Higher Education Act No. 101 of 1997 (as amended as per section 3 Government Gazette). The Higher Education Qualifications Framework. Government Gazette. 2007 October 05.
6. University of the Witwatersrand. Senate standing orders on higher degrees [document on the Internet]. c2013 [updated 2013 Jan 26; cited 2013 Jul 07]. Available from: https://www.wits.ac.za/files/6tbeh_380959001365373856.pdf
7. Lee A. Supervision teams: Making them work. London: Society for Research into Higher Education; 2008.
8. Lahenius K, Ikävalko H. Joint supervision practices in doctoral education – A student experience. J Fur High Educ. 2012;36:1–20. http://dx.doi.org/10.1080/0309877X.2012.706805
9. Ives G, Rowley G. Supervisor selection or allocation and continuity of supervision: Ph.D. students' progress and outcomes. Stud High Educ. 2005;30:535–555. http://dx.doi.org/10.1080/03075070500249161
10. Wingfield B. How much time does it take to supervise a PhD student? S Afr J Sci. 2012;108(11–12), Art. #1454, 2 pages. http://dx.doi.org/10.4102/sajs.v108i11/12.1454
11. University of Edinburgh. Code of practice for supervisors and research students [document on the Internet]. c2007 [cited 2015 Mar 30]. Available from: http:// http://www.docs.sasg.ed.ac.uk/academicservices/codes/copsupervisorsresearchstudents.pdf
12. Pole C. Joint supervision and the PhD: Safety net or panacea? Assess Eval High Educ. 1998;23:259–271. http://dx.doi.org/10.1080/0260293980230303
13. Spooner-Lane RS, Henderson DJ, Price RA, Hill G. Practice to theory: Co-supervision stories. Int J Res Sup. 2007;1:39–51.
14. Martens E, Suri H. A model for the evaluation of postgraduate supervision. In: Santhanam E, editor. Student feedback on teaching: Reflections and projections. Refereed proceedings of teaching evaluation forum; 2000 August 28-29; Perth, Australia. Perth: Evaluation of Teaching Unit, Organisational and Staff Development Services; 2001. p. 69–80. Available from: http://www.csd.uwa.edu.au/spot/forum/forum_monograph.pdf
15. Wadesango N, Machingambi S. Post graduate students' experiences with research supervisors. J Sociology Soc Anth. 2011;2:31–37.
16. Gill P, Burnard P. The student/supervisor relationship in the PhD/doctoral process. Br J Nurs. 2008;17:668–671. http://dx.doi.org/10.12968/bjon.2008.17.10.29484
17. Watts JH. Team supervision of the doctorate: Managing roles, relationships and contradictions. Teach High Educ. 2010;15:335–339. http://dx.doi.org/10.1080/13562511003740908
18. Abiddin NZ, Ismail A, Ismail A. Effective supervisory approach in enhancing postgraduate research studies. IJHSS. 2011;1:206–217.
19. Taylor S, Beasley N. A handbook for doctoral supervisors. London: Routledge; 2005. http://dx.doi.org/10.4324/9780203415740
20. Sze D. Effective student-centred PhD supervision from a social constructivist viewpoint. Development program for research higher degree supervision. Sydney: University of Sydney, Institute for Teaching and Learning; 2008. Available from: http://www.itl.usyd.edu.au/supervision/old_site/casestudies/files/sze.pdf
21. Watts J, Robertson N. Burnout in university teaching staff: A systematic literature review. Educ Res. 2011;53:33–50. http://dx.doi.org/10.1080/00131881.2011.552235
22. Wisker G, Robinson G, Shacham M. Postgraduate research success: Communities of practice involving cohorts, guardian supervisors and online communities. Innov Educ Teach Int. 2007;44:301–320. http://dx.doi.org/10.1080/14703290701486720
23. Knight J. Joint and double degree programmes: Vexing questions and issues. London: The Observatory on Borderless Higher Education; 2008. Available from: http://www.eahep.org/web/images/Malaysia/joint_and_double_degree_programmes_vexing_questions_and_issues_september_2008.pdf
24. Buttery EA, Richter EM, Filho WL. An overview of the elements that influence efficiency in postgraduate supervisory practice arrangements. Int J Educ Manage. 2005;19:7–16. http://dx.doi.org/10.1108/09513540510574920
25. Wisker G, Robinson G. Doctoral orphans: Nurturing and supporting the success of postgraduates who have lost their supervisors. In: Kiley M, editor. Quality in postgraduate research: Educating researchers for the 21st century. Proceedings of the 2010 Quality in Postgraduate Research Conference; 2010 April 13–15; Adelaide, Australia. Canberra: The Centre for Educational Development and Academic Methods, The Australian National University; 2010. p. 59–63. Available from: http://qpr.edu.au/2010/qpr_2010.pdf
26. Taylor SE. Changes in doctoral education: Implications for the professionality of supervisors in developing early career researchers. IJRD. 2012;3:118–138. http://dx.doi.org/10.1108/17597511311316973
27. Hamilton S. Varsities buckle under cash crunch. Mail and Guardian. 2013 June 14–20; p. 47–48.
28. VitalStats. Public higher education 2011. Pretoria: Council on Higher Education; 2014.
29. Department of Higher Education and Training (DHET). Report of the ministerial committee for the review of the funding of universities. Pretoria: DHET; 2013.
30. Jackson D, Peters K, Andrew S, Salamonson Y, Halcomb EJ. If you haven't got a PhD, you're not going to get a job: The PhD as a hurdle to continued academic employment in nursing. Nurs Educ Today. 2011;31:340–344. http://dx.doi.org/10.1016/j.nedt.2010.07.002
31. Higher Education Quality Committee. Criteria for programme accreditation. Pretoria: Higher Education Quality Committee, Council on Higher Education;2004.
32. Thompson DR, Kirkman S, Watson R, Stewart S. Improved research supervision in nursing. Nurs Educ Today. 2005;25:283–290. http://dx.doi.org/10.1016/j.nedt.2005.01.011
33. Hill, G. Research supervisor capability development for neophyte research supervisors. In: Kiley M, editor. Quality in postgraduate research: Educating researchers for the 21st century. Proceedings of the 2010 Quality in Postgraduate Research Conference; 2010 April 13–15; Adelaide, Australia. Canberra: The Centre for Educational Development and Academic Methods, The Australian National University; 2010 p. 159–162. Available from: http://qpr.edu.au/2010/qpr_2010.pdf
34. Pearson M, Brew A. Research training and supervision development. Stud High Educ. 2002;27:135–150. http://dx.doi.org/10.1080/03075070220119986c
35. Woodiwiss A. Publication subsidies: Challenges and dilemmas facing South African researchers. Cardiovasc J Afric. 2012;23:421–427.
36. Deuchar R. Facilitator, director or critical friend? Contradiction and congruence in doctoral supervision styles. Teach High Educ. 2008;13:489–500. http://dx.doi.org/10.1080/13562510802193905

37. Zhao F. Postgraduate research management: A process of knowledge management [article on the Internet]. c2001 [cited 2009 Jan 27]. Available from: http://www.providersedge.com/docs/km_articles/Postgrad_Research_Supervision_-_A_Process_of_KM.pdf

38. Delany D. A review of the literature on effective PhD supervision. Centre for Academic Practice and Student Learning. London: Trinity College; 2008. Available from: http://www.tcd.ie/CAPSL/assets/doc/Effective_Supervision_Literature_Review.doc

39. Lee A. Doctoral supervision: At the boundary of research and teaching. In: Fanghanel J, Colet NR, Bernstein D, editors. London scholarship of teaching and learning: 7th International Conference Proceedings Volume 4. London: City University; 2008. p. 108–116.

40. Stack EM. Reflections on the supervision of postgraduate research in accounting departments. Southern African Accounting Association Conference; 2008 June 25–28; Johannesburg, South Africa. Available from http://eprints.ru.ac.za/1077/1/Stack_SAAAPaper10Supervision.pdf

41. Turner G. I didn't really appreciate how hard work it would be! – New supervisors' experiences of doctoral supervision. Society for Research into Higher Education Annual Research Conference; 2010 December 14–16; Newport, UK. Available from: http://www.srhe.ac.uk/conference2010/abstracts/0104.pdf

42. Abiddin NZ, Hassan A, Ahmad AR. Research student supervision: An approach to good supervisory practice. Open Educ J. 2009;2:11–16. http://dx.doi.org/10.2174/1874920800902010011

43. Nulty D, Kiley M, Meyers N. Promoting and recognising excellence in the supervision of research students: An evidence-based framework. Assess Eval High Educ. 2009;34:693–707. http://dx.doi.org/10.1080/02602930802474193

44. Edwards B. Postgraduate supervision: Is having a PhD enough? Australian Association for Research in Education Conference. Problematic Futures: Educational Research in an Era of Uncertainty; 2002 December 1–5; Brisbane, Australia. Canberra: Australian Association for Research in Education; 2002. p. 1–15. Available from: http://www.aare.edu.au/data/publications/2002/edw02382.pdf

45. Freire P. The banking concept of education. Pedagogy of the oppressed. New York: Continuum. Available from: https://libcom.org/files/Freire-PedagogyoftheOppressed.pdf

46. Vilkinas T. The PhD process: The supervisor as manager. Educ Train. 2002;44:129–137. http://dx.doi.org/10.1108/00400910210424337

47. Evans C, Stevenson K. The experience of international nursing students studying for a PhD in the UK: A qualitative study. BMC Nurs. 2011;10:11. http://dx.doi.org/10.1186/1472-6955-10-11

48. McCallin A, Nayar S. Postgraduate supervision: A critical review of current practice. Teach High Educ. 2012;17:63–74. http://dx.doi.org/10.1080/13562517.2011.590979

49. Shannon AG. Research degree supervision: More mentor than master. Aus Univ Rev. 1995;38:12–15.

50. Grossman ES, Cleaton-Jones PE. What becomes of dental research trainees once they leave the Dental Research Institute? An analysis over 53 years. Eur J Dent Educ. 2008;12:69–74. http://dx.doi.org/10.1111/j.1600-0579.2008.00471.x

51. Park C. Redefining the doctorate. York: The Higher Education Academy; 2007. Available from: http://learning.ox.ac.uk/rsv.php?page=326

In vitro sensitivity testing of *Cladobotryum mycophilum* to carbendazim and prochloraz manganese

AUTHORS:
Alinesi Chakwiya[1,2]
Elna J. van der Linde[3]
Lise Korsten[1]

AFFILIATIONS:
[1]Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria, South Africa
[2]Plant Sciences Research Centre, National Institute for Scientific and Industrial Research, Kitwe, Zambia
[3]Biosystematics Division, Agricultural Research Council, Pretoria, South Africa

CORRESPONDENCE TO:
Lise Korsten

EMAIL:
Lise.Korsten@up.ac.za

POSTAL ADDRESS:
Department of Microbiology and Plant Pathology, University of Pretoria, Private Bag X20, Hatfield 0028, Pretoria, South Africa

DATES:
Received: 21 Nov. 2014
Revised: 16 Feb. 2015
Accepted: 24 Mar. 2015

KEYWORDS:
Agaricus bisporus; cobweb disease; fungicide effectiveness; fungicide resistance; fungicolous fungi

HOW TO CITE:
Chakwiya A, Van der Linde EJ, Korsten L. In vitro sensitivity testing of *Cladobotryum mycophilum* to carbendazim and prochloraz manganese. S Afr J Sci. 2015;111(11/12), Art. #2014-0408, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140408

Limited information of fungicide efficacy on cultivated mushrooms and resistance development potential is available. Minor crop industries in general have a smaller arsenal of protectants to rely on and the likelihood of resistance build-up is of greater concern. This study focused on *Cladobotryum mycophilum*'s sensitivity to carbendazim and prochloraz manganese following recent reports on decreased efficacy of both fungicides. The median effective dose (ED_{50}) values for carbendazim ranged between 0.02 mg/L and 4.31 mg/L with 60% of the South African isolates being moderately resistant. The highest resistance factor for carbendazim was 215. Prochloraz manganese ED_{50} values varied from 0.00001 mg/L to 0.55 mg/L. A significant difference in mean ED_{50} values for both fungicides tested was observed. Using cluster analysis, no discrimination of isolates previously exposed and unexposed to prochloraz manganese was observed. A wide range of differences in ED_{50} values indicated moderate resistance to carbendazim and high sensitivity to prochloraz manganese among isolates under investigation. Discriminant analysis indicated significant differences between clusters contributed by one or a few variables. This study provided evidence that prochloraz manganese remains highly fungitoxic to *C. mycophilum.* However, prochloraz manganese is to be used in a disease management strategy in combination with strict farm hygiene management strategies to retain product efficacy and ensure crop protection.

Introduction

Cobweb disease, caused by *Cladobotryum dendroides* (teleomorph *Hypomyces rosellus*) and *Cladobotryum mycophilum* (teleomorph *Hypomyces odoratus*), is one of the three major fungal diseases of the mushroom *Agaricus bisporus* (Lange) Sing.[1] *Cladobotryum mycophilum* is associated with cobweb disease in Australia, the mainland of Europe, the United States of America (USA) and South Africa. *C. mycophilum* is also the dominant species in the British Isles, where it is reportedly resistant to benzimidazoles.[2,3] *Hypomyces aurantius,* a teleomorph of *Cladobotryum varium* Nees, is another pathogen of mushrooms[4], whose importance as well, as that of *C. multiseptatum* de Hoog, is not fully understood[2].

Until the mid1990s, cobweb disease had not been a major problem in mushroom production in most countries. However, the disease has since become an important obstacle to viable mushroom production in countries such as Australia[2], the UK and Serbia[6]. It is reported that the disease caused crop losses of up to 40% in 1994/1995 in the Irish and British mushroom industries as a result of serious spotting and early crop termination at the peak of the epidemic.[5] According to Potočnik et al.[6] and Back et al.[7,8], cobweb disease was first reported on *A. bisporus* in Serbia and Korea in 2007 and 2009, respectively. Sporadic cases of cobweb disease have been reported on several mushroom farms in South Africa during the summer months.

Management of mushroom fungal diseases relies mainly on strict farm hygiene in combination with the use of fungicides.[2] Although fungicides are used in the control of cobweb disease, there are certain inherent challenges including the fact that both the host and pathogen are fungi.[9,10] In addition, very few studies on fungicide efficacy on cultivated mushrooms have been conducted and because of its minor crop status, only a few fungicides are currently registered for commercial use.[11,12] For the South African mushroom industry, only prochloraz manganese, prochloraz zinc[2] and thiabendazole[13] are registered for use on mushrooms. Grogan[14] proposed that the mushroom industry's heavy reliance on only two fungicides is a concern for sustainable production. Factors leading to resistance include (1) inherent risk factors relating to fungus biology and fungicide chemistry and (2) management risk factors relating to usage of fungicides in crop management in the wider sense. The inherent factors serve to assess the basic resistance risk for a fungicide/fungus combination in a given area and cannot be controlled. Management of risk factors can be controlled as they are in the hands of farmers, agriculture officials and distributors of fungicides.[15]

Benomyl, the first site-specific fungicide to be used in the mushroom industry, has been available from the late 1960s and provides good control of all fungal diseases.[16] Resistance development in *Lecanicillium fungicola* Zare & Gams (syn. *Verticillium fungicola* (Preuss) Hassebrauk and *H. rosellus* was reported in 1974 and 1993, respectively.[17,18] A few years later, resistance of *Cladobotryum mycophilum* to benzimidazole fungicides was reported in the UK.[19]

Prochloraz manganese, one of the demethylation inhibitors (DMIs), was introduced to the mushroom industry in the late 1980s to control wet bubble, dry bubble and cobweb diseases.[1,20] This fungicide effectively controlled mushroom diseases but over time its effectiveness against *Lecanicillium fungicola* decreased. Since then, increased loss of sensitivity of the pathogen to the fungicide has been reported worldwide.[21] Prochloraz manganese was the most effective fungicide in the prevention of cobweb disease,[1] but in recent years, it has no longer been able to control the spotting symptoms caused by the disease.[14]

Monitoring of fungicide sensitivity within pathogen populations over time is important to ensure more effective management of the disease as well as retention of efficacy of the chemical.[1,21] Distribution of sensitivity to DMI fungicides among individual strains in an unexposed pathogen population varies from highly sensitive to considerably less sensitive phenotypes. Therefore, exposure to a given DMI fungicide will provide selection pressure favouring the more resistant isolates in the population.[20,22,23] Progressive reduction in DMI efficacy results

in a practical resistance once the frequency of phenotypes prevents commercially acceptable disease control levels to be achieved under standard usage regimes.[23] The fact that resistance has developed in *L. fungicola*[21,24] and many other pathogens of field crops,[25-27] combined with the fact that most *Cladobotryum* spp. produce abundant conidia and have repeated infection cycles, may suggest that *Cladobotryum* spp. might also be at risk.[28]

We investigated the sensitivity of 32 *Cladobotryum* isolates to prochloraz manganese in order to determine a possible population shift in South African isolates in comparison with isolates obtained from the USA which have not been exposed to this fungicide. The isolates were obtained over a period of time and preserved in a central collection. Carbendazim, although not registered for use on mushrooms in South Africa, but occasionally used, was included in this study to determine its potential fungi toxicity against these isolates.

Materials and methods

Origin and collection of fungal isolates

In total, 32 isolates of *Cladobotryum* spp. were obtained from mushroom farms in South Africa as well as from research institutes around the world (Table 1). Isolates collected from local mushroom farms were either sampled from diseased mushrooms (five) or collected as infected casing (seven), while those from research institutions around the world were received as cultures on agar slants. For each isolate, a single spore culture was prepared for use in this study and preserved in sterile distilled water and in 10% glycerol at -75 °C for short- and medium-term storage for future reference.

Confirmation of isolate identity

To confirm identity of isolates, conidia morphology was examined under a phase contrast microscope (Zeiss, Germany). Colony characteristics were observed on malt extract agar (MEA) (Merck, Johannesburg, South Africa) and potato dextrose agar (PDA) (Merck) upon receipt at the laboratory. Methods for pathogenicity tests were conducted by inoculating a pathogen spore suspension on the pilei of freshly harvested mushrooms according to the Serbian report by Potočnik et al.[29] Molecular analysis was conducted to confirm the identity of all isolates through conventional polymerase chain reactions (PCR) using the partial internal transcribed spacer (ITS) region (ITS1 and ITS4) according to Tamm and Põldmaa[30]. The DNA sequence was then blasted using the NCBI GenBank database and results were recorded.

Fungicides

Prochloraz manganese (Octave 500 g/kg WP, Bayer Crop Science, Johannesburg, South Africa) and carbendazim (Bendazid 500 g/L EC, Plaaskem, Johannesburg, South Africa) were provided by mushroom farmers. To prepare stock solutions, prochloraz manganese was dissolved in sterile distilled water and then added to molten MEA cooled to approximately 50 °C at the desired test concentration. For carbendazim, the stock solution was added to MEA before sterilisation. For both fungicides, the MEA was poured into 90-mm sterile plastic Petri dishes and allowed to solidify before use.

Fungicide sensitivity test

Mycelial inhibition was assessed by transferring a 5-mm diameter disc taken from the edges of an actively growing single colony culture and placing it inverted in the centre of a plate. To assess the initial in vitro response of mycelial growth to different fungicide concentrations, all isolates were exposed to a concentration range of 1–100 mg/L for both products tested. From these observations, the isolates were grouped into two categories with reference to carbendazim concentration: low concentration (0.1–0.75 mg/L) and high concentration (5–20 mg/L). A series of MEA plates was subsequently prepared with concentrations of 0.0, 0.1, 0.2, 0.35, 0.5 and 0.75 mg/L carbendazim representing the low concentration category and concentrations of 5.0, 10.0, 15.0, 17.0 and 20.0 mg/L carbendazim representing the high concentration category. Concentrations used for the in vitro sensitivity test to prochloraz manganese were 0.0, 1.0, 2.0, 3.5, 4.5 and 5.0 mg/L. For both fungicide tests, each isolate was tested in triplicate per fungicide concentration and incubated at 25 °C for 96 h. All the media and the fungicide stocks used were prepared as a single batch and used within 24 h per trial. Perpendicular colony diameters were measured every day starting at 48 h of incubation. The experiment was repeated twice.

Table 1: Identity and collection details of isolates used in the study

Isolate code	Year isolated	Species identity (molecular)	Source of origin
IR22	Unknown	*C. mycophilum*	Ireland
IR23	Unknown	*C. dendroides*	Ireland
IR24	Unknown	*C. mycophilum*	Ireland
AU1	2007	*C. mycophilum*	Australia
AU17	2007	*C. mycophilum*	Australia
NZ25	2007	*C. mycophilum*	New Zealand
US20	2003	*C. mycophilum*	USA
US21	2003	*C. mycophilum*	USA
US34	2000	*C. mycophilum*	USA
US35	2000	*C. mycophilum*	USA
US36	2000	*C. mycophilum*	USA
US37	2000	*C. mycophilum*	USA
US40	2000	*C. mycophilum*	USA
US41	2003	*C. mycophilum*	USA
US45	2008	*C. mycophilum*	USA
US46	2008	*C. mycophilum*	USA
US16	2008	*C. mycophilum*	USA
US18	2008	*C. mycophilum*	USA
US19	2008	*C. mycophilum*	USA
SA29	2010	*C. mycophilum*	South Africa
SA28	2010	*C. mycophilum*	South Africa
SA3	2010	*C. mycophilum*	South Africa
SA26	2010	*C. mycophilum*	South Africa
SA30	2011	*C. mycophilum*	South Africa
SA31	2011	*C. mycophilum*	South Africa
SA32	2011	*C. mycophilum*	South Africa
SA3	2010	*C. mycophilum*	South Africa
SA4	2010	*C. mycophilum*	South Africa
SA27	2011	*C. mycophilum*	South Africa
SA5	2011	*C. mycophilum*	South Africa
SA12	2010	*C. mycophilum*	South Africa
NE33	2011	*C. mycophilum*	Netherlands

Statistical analysis

The mean effective concentration value (ED_{50}) for each isolate at 96 h was calculated whenever possible by the regressing percentage inhibition

against fungicide concentration. The ED_{50} values of some isolates could not be determined because their growth was too slow. The mean colony diameter for each treatment was expressed as percentage growth inhibition using a SAS general linear model procedure and the means were compared using a *t*-test. Multivariate analysis (Statistica version 10.0) was used to cluster and discriminate between and within groups among isolates tested.

Results

Confirmation of isolate identity and pathogenicity

Conidia of all isolates were 1–3 septate and mostly obvoid in shape except for isolate SA4 which had a few curved conidia. Conidia sizes ranged from 15 to 34 μm long. On PDA, all isolates developed a yellowish colony colour after four days' incubation. On MEA, three distinct colony morphologies were observed in *C. mycophilum*. Colonies with a 'cotton-fluffy' texture were formed by US34, US35, US36, US37 and US41, unlike the more compact or 'felt-like' texture of US40 and IR23 (only *H. rosellus*), an ATCC isolate of Irish origin. The rest of the isolates were 'cotton-like' – neither 'fluffy' nor 'compacted'. Most isolates' colonies were initially yellowish to pinkish, turning purplish after more than 96 h of growth. Blasting of ITS sequence against GenBank gave a 98–99% similarity to *C. mycophilum* for all isolates and IR23 was confirmed as *H. rosellus* (Table 1). These results confirmed the isolates' identity as *Cladobotryum* species.

Pathogenicity of all isolates was confirmed by the development of disease symptoms (cobweb) on the mushroom pilei. The majority (85%) of isolates resulted in cobweb symptom development within 48 h while others, such as *H. rosellus* (IR23) and some *C. mycophilum* (SA3, SA4, SA27 and IR22), were less virulent and took between 72 and 96 h to cause symptoms.

Fungicide bioassays

Most isolates under investigation demonstrated an ability to grow in the presence of bendazid at low concentrations. The ED_{50} values of bendazid against *C. mycophilum* isolates varied between 0.02 mg/L and 4.31 mg/L. The ED_{50} value of a few isolates could not be determined because they were too low. The highest resistance factor (RF) for bendazid was 215.5 mg/L for isolate US21. Isolate SA29 had the highest ED_{50} and RF values although values were not significantly different when compared with other isolates from South Africa.

All isolates tested were highly sensitive to prochloraz manganese and ED_{50} values ranged from 0.00001–0.55 mg/L. Isolates IR23, SA13 and US40 had ED_{50} values less than 0.00001 mg/L and could thus not be determined. Isolate SA12 had the highest RF of 55 000. Only four isolates (AU1, US36, US41 and SA27) had RFs of less than 100. There were significant differences among the isolates under examination. However, none of the isolates showed resistance to prochloraz manganese when their ED_{50} values were compared.

Cluster analysis

Cluster analysis of the isolates based on their response to fungicide treatment tests resulted in 100% classification of six, four and three clusters under prochloraz manganese and carbendazim low and high concentration categories, respectively. Clustering within each treatment was analysed for significant differences using discriminate analysis. For all three treatments, there were significant differences between but not within clusters. Not all variables (test concentrations) contributed to the variations observed.

Cluster analyses of prochloraz manganese treatment results grouped the 32 isolates tested into four clusters (Figure 1). Clustering was done a priori. The largest cluster (IV) included 12 isolates with an average ED_{50} value of 0.33 mg/L. The isolate with the highest ED_{50} value was SA12 (0.55 mg/L) followed by US18 (0.54 mg/L), with RF values of

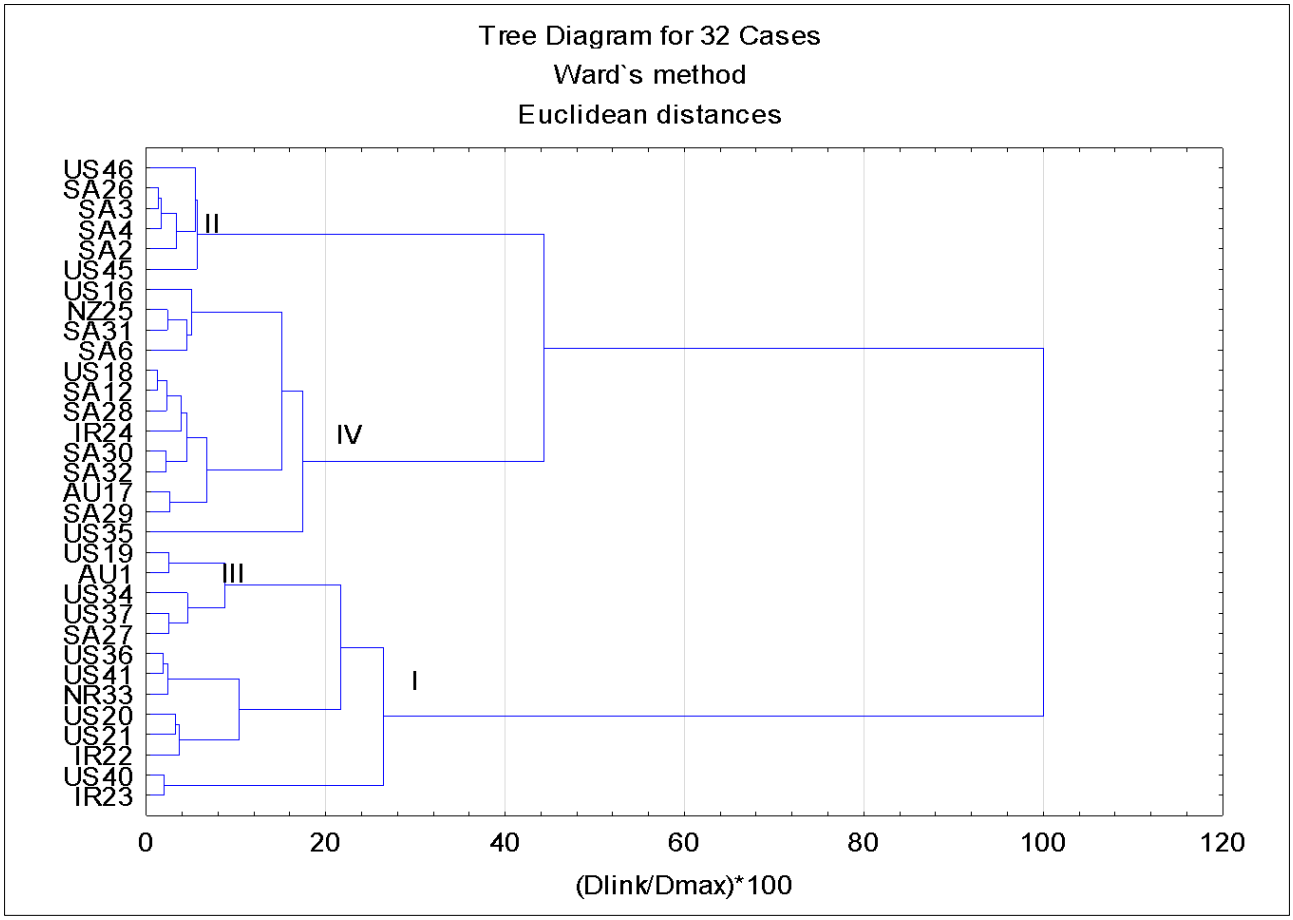


Figure 1: Dendogram showing clustering of *Cladobotryum mycophilum* isolates into four groups based on prochloraz manganese in vitro sensitivity screening.

55 000 and 54 000, respectively (Table 2). There were no significant differences among the members of this group. The range in ED_{50} values was between 0.05–0.55 mg/L. South African isolates made up 50% of this cluster, followed by 25% USA; the rest of the isolates were from Australia, Ireland and New Zealand.

Cluster I was the second largest, represented by eight members (IR22, IR23, NR33, US20, US21, US36, US40 and US41) from Ireland, Netherlands and the USA with a mean ED_{50} value of 0.047 mg/L. The highest ED_{50} value in the cluster was 0.13 mg/L for IR22, an Irish isolate. Cluster II comprised 71.4% isolates from South Africa (SA26, SA2, SA3, SA4 and SA6) and 28.6% from the USA (US45 and US46). The average ED_{50} value for Cluster II was 0.23 mg/L with a range of 0.06–0.4 mg/L. The ED_{50} values of some of the individual members of this group were not significantly different from each other despite the wide range of ED_{50} values observed (Table 2.). In Cluster III, 60% of the isolates were from the USA, 20% were from Australia and the rest were from South Africa. Cluster III had an average ED_{50} value of 0.074 mg/L with a range of 0.0002–0.13 mg/L. The highest ED_{50} value in the group was for isolates from the USA (US19 and US34). The lowest ED_{50} value was from SA27.

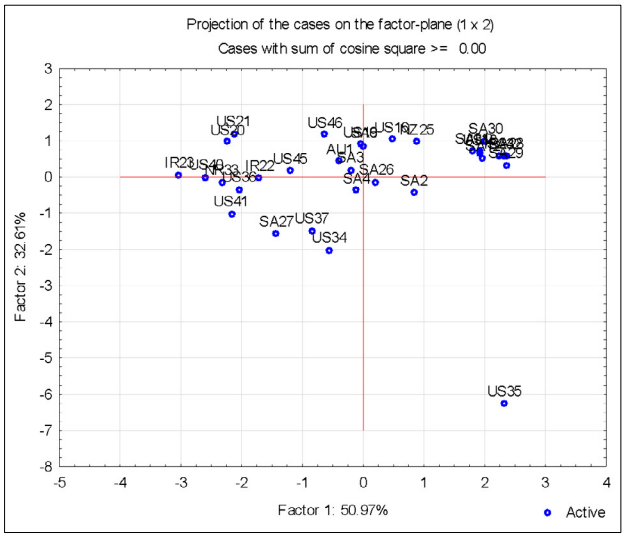


Figure 2: Projection of *Cladobotryum mycophilum* isolates following exposure to prochloraz manganese in vitro sensitivity screening.

The projection of isolates exposed to prochloraz manganese on the plane were for Factors 1 and 2. In Figure 2, eight of the South African isolates are clustered in the higher region of the plane. This shows a tendency of those isolates to tolerate higher concentrations of prochloraz manganese than the rest of the isolates under study.

Cluster analysis for the carbendazim fungicide in vitro test was carried out separately for the low concentrations (ranging from 0.0–0.75 mg/L) and the high concentrations (ranging from 0.0–20.0 mg/L). For easier reference, these are called Analysis 1 and Analysis 2. For Analysis 1, four clusters were formed a priori, with the largest (Cluster III) consisting of six isolates (Figure 3). Four of the isolates in Cluster III were from South Africa and one each was from Ireland and the USA. The average ED_{50} value for Cluster III was 0.058 mg/L. Isolate US41 was significantly different from the rest of the group. Cluster I had an average ED_{50} value of less than 0.02 mg/L, which was the lowest of all clusters formed in Analysis 1. Three isolates (US45, IR23 and US34) that were included in this lower concentration category had EC_{50} values below 0.02 mg/L, the lowest determined value in the category. Cluster II consisted of four US isolates (US35, US36, US37 and US40) and one South African isolate (SA6). Isolates US40 and US37 were significantly different from each other but not from the rest of the cluster. Overall, ED_{50} values for the isolates exposed to the low carbendazim concentrations ranged from 2.5–10.5 mg/L (Table 3).

For Analysis 2 (carbendazim high concentration), three clusters were also determined a priori (Figure 4). Cluster II with nine isolates had the most members and had an average ED_{50} value of 3.0–4.31 mg/L.

Table 2: ED_{50} values indicating the level of sensitivity of *Cladobotryum mycophilum* to prochloraz manganese chloride complex, an imidazole fungicide

Isolate code	†Mean ED_{50}	Median ED_{50}	ED_{50} range	††Resistance factor
SA12	0.55[a]	0.51	0.0076–1.121	55000
US18	0.54[ab]	0.65	0.002–1.005	54000
SA32	0.52[abc]	0.59	0.002–0.983	52000
SA28	0.49[abcd]	0.46	0.001–1.043	49000
IR24	0.43[abcdef]	0.22	0.00003–1.172	43000
US45	0.40[abcdefg]	0.35	0.008–0.883	40000
US46	0.38[abcdefg]	0.18	0.002–1.534	38000
AU17	0.33[abcdefg]	0.32	0.0013–0.666	33000
NZ25	0.16[abcdefg]	0.0085	0.0018–0.727	16000
SA30	0.34[abcdefg]	0.0143	0.002–1.273	34000
SA31	0.15[abcdefg]	0.0087	0.002–0.614	15000
US35	0.31[abcdefg]	0.17	0.0049–0.827	31000
SA6	0.15[abcdefg]	0.18	0.002–0.329	15000
US19	0.13[bcdefg]	0.02	0.00001–0.475	13000
US37	0.13[bcdefg]	0.00013	0.0001–0.536	13000
IR22	0.13[bcdefg]	0.14	0.0002–0.234	13000
US34	0.11[cdefg]	0.00055	0.003–0.342	11000
SA26	0.09[defg]	0.07	0.00001–0.0140	9000
SA3	0.08[defg]	0.07	0.013–0.185	8000
SA4	0.06[defg]	0.07	0.008–0.111	6000
SA2	0.05[efg]	0.04	0.198–0.103	50000
US20	0.03[fg]	0.012	0.00006–0.153	3000
AU1	0.0007[g]	0.0006	0.00002–0.002	70
US36	0.00004[g]	0.00003	0.00001–0.00008	4
US41	0.00001[g]	0.00001	0.00001–0.00003	3
US21	0.007[g]	0.005	0.001–0.017	700
SA27	0.0002[g]	0.002	0.0001–0.0002	20
NE33	0.008[g]	0.008	0.00001–0.0014	800
SA5	0.002[g]	0.002	0.0002–0.002	200
IR23	nd	nd	nd	nd
SA13	nd	nd	nd	nd

ED_{50} is the fungicide concentration that inhibits 50% of fungal growth.

†Means followed by different letters are significantly different.

††Expressed as the ratio of the highest ED_{50} to the lowest ED_{50} of isolates tested.

nd, not determined.

The two isolates (US21 and SA32) with the highest overall RF values (215.5 mg/L and 207.5 mg/L respectively) belonged to Cluster II. There was no significant difference in the ED_{50} values of the isolates in Cluster II. Cluster I consisted of all US isolates: US16, US19, US20 and US24. Apart from US16, with an ED_{50} value of 0.58 mg/L, the ED_{50} values of

the isolates could not be determined, but they were all lower than that for US16. The average ED_{50} value for this cluster was 0.14 mg/L. Cluster III had only two isolates, AU17 and SA30, for which the ED_{50} values was 2.03 mg/L and 3.95 mg/L, respectively, with an average of 3.0 mg/L.

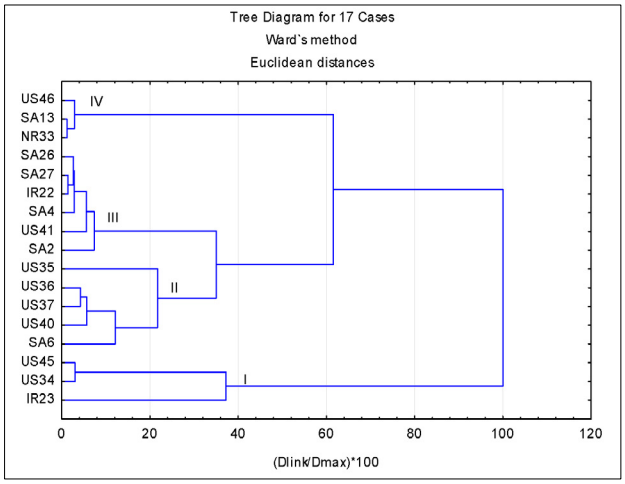


Figure 3: Dendogram showing clustering of *Cladobotryum mycophilum* isolates into four groups after carbendazim (low concentration) in vitro sensitivity screening.

Discussion and conclusion

A total of 32 *Cladobotryum* isolates was assessed for their sensitivity to prochloraz manganese and carbendazim (Figure 5). Isolates from the USA, where prochloraz manganese is not registered for use in the mushroom industry,[31] were used as a baseline reference culture for this study. All isolates were sensitive to both carbendazim (0.02–4.31 mg/L) and prochloraz manganese (0.00001–0.55 mg/L). The highest RFs observed with bendazid and prochloraz manganese were 215 (US210) and 5 5000 (SA13), respectively. According to the criteria used by Delp et al.[32] and Gouot[33] to define resistance (in which an RF value of more than 2 is considered resistant), both isolates would have been considered highly resistant to the tested fungicides. The RF provides an indication of the potential for resistance development in a given population. According to the definition by Gea et al.[9], the higher the RF the greater the difference in sensitivity between the least sensitive individuals in a population relative to the mean sensitivity of the individuals in the entire population and therefore, the greater the potential for resistance to develop.

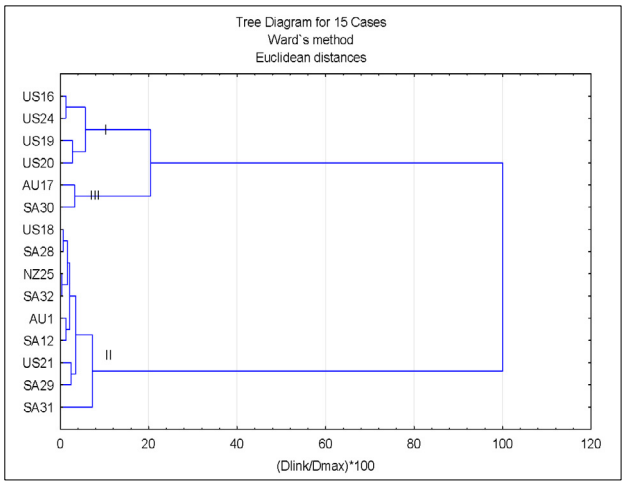


Figure 4: Clustering of *Cladobotryum mycophilum* isolates into four groups after carbendazim high concentration under in vitro treatment.

However, a higher resistance factor value does not imply that resistance will be rapidly expressed in the isolates under study because the ED_{50} values obtained were very low. In this instance, RF values were much higher for both fungicides tested, indicating the likelihood of resistance development for the isolates under investigation.[32] ED_{50} and RF values are used as an index to determine resistance or sensitivity of a fungal pathogen to fungicide treatment.[18] According to Potočnik et al.[17] and

Table 3: ED_{50} values indicating the level of sensitivity of *Cladobotryum mycophilum* to carbendazim, a benzimidazole fungicide

Carbendazim low concentration group (mg/L)					Carbendazim high concentration group (mg/L)				
Isolate code	**†Mean ED_{50}**	**Median ED_{50}**	**ED_{50} range**	**††Resistance factor**	**Isolate code**	**ED_{50}**	**Median ED_{50}**	**ED_{50} range**	**††Resistance factor**
US34	0.21^{a}	0.20	0.142–0.288	10.5	US16	0.58^{d}	0.58	0.00–0.58	29
US41	0.17^{ab}	0.01	0.088–0.300	8.5	SA30	3.95^{e}	3.18	2.84–6.59	197.5
US36	0.16^{ab}	0.16	0.158–0.162	8.0	SA31	4.13^{e}	3.84	0.86–6.48	206.5
US35	0.15^{ab}	0.15	0.113–0.189	7.5	SA29	4.15^{e}	4.64	2.99–5.03	207.5
SA6	0.16^{ab}	0.16	0.099–0.178	8.0	US21	4.31^{e}	4.44	2.79–5.99	215.5
US40	0.12^{b}	0.14	0.088–0.149	6.0	NZ25	3.75^{e}	4.21	1.54–4.98	187.5
US46	0.02^{c}	0.02	0.0001–0.021	1	US18	3.88^{e}	3.75	2.33–5.42	194
SA26	0.05^{c}	0.05	0.0312–0.756	2.5	AU1	3.00^{ef}	3.52	0.36–4.65	150
SA27	0.05^{c}	0.05	0.023–0.086	2.5	SA12	3.46^{ef}	3.45	2.16–4.91	173
IR22	0.03^{c}	0.03	0.019–0.043	1.5	SA32	3.67^{ef}	4.32	1.45–5.21	183.5
SA4	0.05^{c}	0.04	0.021–0.075	2.5	SA28	3.70^{ef}	3.76	1.12–5.14	185
AU17	2.03^{fg}	1.94	1.05-3.10	101.5	SA5	nd	nd	nd	nd

ED_{50} is the fungicide concentration that inhibits 50% of fungal growth.

†Means followed by different letters are significantly different.

††Expressed as the ratio of the highest ED_{50} to the lowest ED_{50} of isolates tested.

nd, not determined.

Gea et al.[21] a wide range in ED_{50} values indicates a wider range of sensitivity for isolates under investigation. Both Potočnik et al.[17] and Gea et al.[21] reported that an increase in the range of sensitivity increased the risk of individuals within a population becoming less sensitive to the test fungicide. The gradual selection pressure exerted on the pathogen population can therefore contribute to triggering resistance build-up in *C. mycophilum.* In this study, resistance was not observed in any of the isolates challenged with the two fungicides. However, production practices, such as strict farm hygiene to reduce the incidence of disease, should be maintained to minimise the potential risk of build-up of pathogen resistance.[21]

Among the isolates from the USA, the lowest and highest RFs were 1 and 54 000 with ED_{50} values of 0.00001 mg/L and 0.54 mg/L, respectively. This observation corresponds with the report by Erickson et al.[22] indicating that sensitivity to a DMI fungicide among individuals not yet exposed to such a fungicide varies from highly sensitive to more resistant phenotypes. However, in this study, no isolate had an ED_{50} value of less than 5 mg/L or more than 50 mg/L to render them, respectively, moderately or highly resistant according to the criteria set by Russell et al.[34] cited by Gea et al.[24]

The 12 isolates studied from the South African mushroom industry were highly sensitive to prochloraz manganese (ED_{50}=0.0002–0.55 mg/L) and there were significant differences among the isolates under study. However, none of the isolates was resistant to prochloraz manganese. Prochloraz manganese was therefore more toxic than carbendazim to all isolates investigated in the study.

According to Bonnen et al.[31], the presence of moderately resistant isolates prevalent prior to the benomyl era indicates the existence of a certain level of resistant isolates within the natural population. A number of studies has been conducted to determine sensitivity of mushroom pathogens to prochloraz and benomyl (carbendazim). In the mid-1970s, British researchers conducted a survey in the UK of *Cladobotryum* spp. sensitivity to benomyl.[18] These researchers showed that none of the isolates grew at a benomyl concentration of 5 mg/L. Benomyl has since been replaced by carbendazim; a primary breakdown fungitoxic product of benomyl.[35] In their cropping experiment, Fletcher et al.[1] observed that benomyl, thiabendazole and prochloraz manganese controlled *Cladobotryum* spp. satisfactorily. In the Netherlands, it was reported that prochloraz manganese complex was effective against *Hypomyces rosellus*, *Lecanicillium fungicola*, *Mycogone perniciosa* (Magnus) in the cultivation of *A. bisporus and A. bitorquis* (Quel.) Sacc.[20]

In the 1994/1995 outbreak of cobweb disease in the UK, it was reported that *Cladobotryum* spp. isolates were inhibited at a 10 mg/L concentration of carbendazim.[35] Resistance of *Cladobotryum* spp. to a benzimidazole fungicide was reported in Ireland in 1992. Further investigations by McKay et al.[19] confirmed that these isolates were resistant to the benzimidazole fungicides benomyl and carbendazim with an ED_{50} of up to 10 mg/L.[19] Grogan and Gaze[36] also observed that the majority of the *Cladobotryum* spp. isolated in the UK was strongly resistant to thiabendazole (ED_{50}>200 mg/L) and weakly resistant to carbendazim (3.5–4.4 mg/L).[35]

Among the isolates in our study, only a third of the isolates were moderately resistant and none were highly resistant. In accordance with the definition by Gea et al.[21], an ED_{50} of 0–5 mg/L indicates sensitivity, 5–50 mg/L moderate resistance and >50 mg/L resistance – most isolates were sensitive to carbendazim (ED_{50}=0.002–2 mg/L). Of all isolates moderately resistant to carbendazim, 60% were isolates from the South African mushroom farms. Based on Gouot's standard,[33] isolate response to carbendazim exposure (ED_{50}=0.05–4.15 mg/L) ranged from extremely sensitive to moderately resistant among the South African isolates.

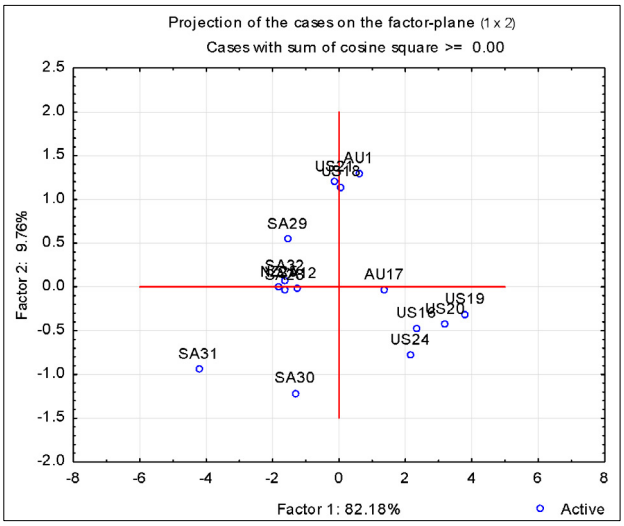


Figure 5: Projection of *Cladobotryum mycophilum* isolates following exposure to carbendazim (high concentration) after in vitro sensitivity screening. Isolates from South Africa did not form defined clusters. Only the four isolates from the USA (US16, US19, US20 and US24) grouped together.

Research findings by Grogan et al.[36] showed that isolates in the UK exhibited a range of reactions to prochloraz manganese with ED_{50} values of 0.19–7.80 mg/L. Within their study population, the duo stated that a total of 75% were weakly resistant (1.1–3.7 mg/L) to prochloraz manganese. In another report, Potočnik et al.[29] showed that Serbian isolates of *Cladobotryum dendroides* were highly sensitive, with ED_{50} values ranging from 0.01 mg/L to 0.09 mg/L for prochloraz manganese and 0.24–2.92 mg/L for carbendazim.[29] In our investigation, none of the isolates showed reduced sensitivity to prochloraz manganese and the chemical remained highly toxic to the isolates under investigation (ED_{50}=0.00001–0.55 mg/L).

From our findings, clustering of isolates using a multivariate analysis did not discriminate between isolates never exposed to prochloraz manganese and those that had been exposed. The earliest (2000) collected isolates received from the USA did not form one cluster, nor was such clustering obtained from the South African isolates. This finding is probably in accordance to observations made by McGath[23] and Erickson et al.[22] who proposed that in a population of isolates that have never been exposed to a DMI fungicide, there is a continuously varying range from highly sensitive to highly resistant phenotypes. However, in our study, no isolate had an ED_{50} value of 5 mg/L or more; therefore, no isolate investigated was highly resistant.

Even though the risk of resistance development in the *Cladobotryum* spp. under investigation seemed low, continuous fungicide usage selects for pathogen resistance; thus farm hygiene measures should be managed to reduce incidence of the disease and minimise the risk of build-up of resistance.[26] Emergence of benzimidazole resistance reduces the value of these fungicides available for disease control. In addition, a lack of effective alternatives makes it important to manage fungicides with a holistic strategy to retain efficacy. Regular monitoring therefore becomes even more important.[14] In conclusion, mushroom growers can no longer rely exclusively on prochloraz manganese as a primary management tool for fungal diseases.[21]

Acknowledgements

We are grateful to Frikkie Calitz and Liesl Morey (Biometry Unit, Agricultural Research Council, South Africa) for their statistical advice. We acknowledge the South African Mushroom Association (SAMFA) research team and the University of Pretoria for supporting the research. We also express our gratitude to Sida through the Organisation of Women in Science for the Developing World (OWSD) for a postgraduate fellowship.

Authors' contributions

L.K. was the project leader and promoter and was involved in the experimental design. E.J.v.d.L. was the co-promoter and was responsible for the experimental set-up and for arranging the statistical analyses. Both promoters were also responsible for editing the manuscript. A.C. was responsible for performing most of the experimental work and writing the manuscript.

References

1. Fletcher JT, Hims MJ, Hall PJ. The control of bubble disease and cobweb disease of mushrooms with prochloraz. Plant Pathol. 1983;32:123–131. http://dx.doi.org/10.1111/j.1365-3059.1983.tb01310.x
2. Fletcher JT, Gaze RH. Mushroom pest and disease control – A colour handbook. London: Manson Publishing Ltd; 2008.
3. McKay GJ, Egan D, Morris E, Scott C, Brown AE. Genetic and morphological characterization of *Cladobotryum* species causing cobweb disease of mushrooms. Appl Environ Microbiol. 1999;65(2):606–610.
4. Eicker A, Van Greuning M. Fungi in the cultivation of the *Agaricus bisporus* – An updated list of species. In: Van Greuning M, Van Griensvenled, editors. Genetics and breeding of *Agaricus*. Proceedings of the first International Seminar on Mushroom Science; 1991 May 14–17; Wageningen, the Netherlands. Wageningen: Centre for Agriculture Publishing and Documentation; 1991. p. 89–96.
5. Adie B, Grogan H, Archer S, Mills P. Temporal and spatial dispersal of *Cladobotryum* conidia in the controlled environment of a mushroom growing room. Appl Environ Microbiol. 2006;72:7212–7217. http://dx.doi.org/10.1128/AEM.01369-06
6. Potočnik I, Recanovick E, Milijašević S, Todorović B, Stepanović M. Morphological and pathogenic characteristics of the fungus *Cladobotryum dendroides*, the causative agent of cobweb disease of the cultivated mushroom *Agaricus bisporus* in Serbia. Pestic Phytomed. 2008;23:175–181. http://dx.doi.org/10.2298/PIF0803175P
7. Back C-G, Kim Y-H, Jo W-S, Chung H, Jung H-Y. Cobweb disease on *Agaricus bisporus* caused by *Cladobotryum mycophilum* in Korea. J Gen Plant Pathol. 2010;76:232–235. http://dx.doi.org/10.1007/s10327-010-0236-3
8. Back C-G, Lee C-Y, Seo G-S, Jung H-C. Characterisation of species of *Cladobotryum* which cause cobweb disease of edible mushrooms grown in Korea. Mycobiology. 2012;40(3):189–194. http://dx.doi.org/10.5941/MYCO.2012.40.3.189
9. Gea FT, Tello JC, Navarro M-J. Efficacy and effects on yield of different fungicides for the control of wet bubble disease of mushroom caused by the mycoparasite *Mycogone perniciosa*. Crop Prot. 2010;29:1021–1025. http://dx.doi.org/10.1016/j.cropro.2010.06.006
10. Mehrparvar M, Goltapeh ME, Safaei N. Resistance of Iranian *Lecanicillium fungicola* to benzimidazole and ergosterol demethylation inhibition fungicides. J Agricult Sci Technol. 2013;15:389–395.
11. Chrysay-Tokousbalides M, Kastanias MA, Philippoussis A, Diamantopoulou P. Selective fungitoxicity of famoxadone, tebuconazole and trifloxystrobin between *Verticillium fungicola* and *Agaricus bisporus.* Crop Prot. 2007;26:469–475. http://dx.doi.org/10.1016/j.cropro.2006.02.016
12. Stoddart H, Garthwaite DG, Thomas MR. Pesticide usage survey report 197. Mushroom crops in Great Britain 2003. Department of Environment, Food and Rural Affairs and Scottish Executive Environment and Rural Affairs Department. London: Defra; 2004.
13. Association of Veterinary and Crop Associations of South Africa (AVCASA). Fungicides [document on the Internet]. c2015 [cited 2015 Oct 29]. Available from: http://www.croplife.co.za/Portals/13/Croplife_Documents/Fungicides.pdf
14. Grogan HM. Fungicide control of mushroom cobweb disease caused by *Cladobotryum* strains with different benzimidazole resistance profiles. Pest Manag Sci. 2006;62:153–161. http://dx.doi.org/10.1002/ps.1133
15. Staub T, Sozzi D. Fungicide resistance. Plant Dis. 1984;68(12):1026–1031. http://dx.doi.org/10.1094/PD-69-1026
16. Samuels GJ, Johnston PR. Benomyl and the *Verticillium* diseases of cultivated mushrooms. New Zeal J Agr Res. 1980;23:155–157. http://dx.doi.org/10.1080/00288233.1980.10417861
17. Potočnik I, Milijaešvić S, Rekanović E, Todorović B, Stepanović M. Sensitivity of *Cladobotryum* spp., a pathogen of the button mushroom (*Agaricus bisporus*), to some fungicides. Pestic Phytomed. 2007;22:233–240.
18. Fletcher JT, Yarham DJ. The incidence of Benomyl tolerance in *Verticillium fungicola*, *Mycogone perniciosa* and *Hypomyces rosellus* in mushroom crops. Ann Appl Biol. 1976;84:343–353. http://dx.doi.org/10.1111/j.1744-7348.1976.tb01777.x
19. McKay GJ, Egan D, Morris E, Brown AE. Identification of benzimidazole resistance in *Cladobotryum dendroides* using a PCR-based method. Mycol Res. 1998;102:671–676. http://dx.doi.org/10.1017/S095375629700542X
20. Van Zaayen A, Van Adrichem JCJ. Prochloraz for control of fungal pathogens of cultivated mushrooms. Neth J Plant Pathol. 1982;88:203–213. http://dx.doi.org/10.1007/BF02140883
21. Gea FT, Navarro MJ, Tello JC. Reduced sensitivity of mushroom pathogen *Verticillium fungicola* to prochloraz-manganese in vitro. Mycol Res. 2005;109:741–745. http://dx.doi.org/10.1017/S095375620500242X
22. Erickson EO, Wilcox W. Distribution of sensitivities to three sterol demethylation inhibitor fungicides among populations of *Uncinula necator* sensitive and resistant to triadimefon. Phytopathology. 1997;87:784–791. http://dx.doi.org/10.1094/PHYTO.1997.87.8.784
23. McGrath MT. Fungicide resistance in cucurbit powdery mildew: Experiences and challenges. Plant Dis. 2001;85:236–245. http://dx.doi.org/10.1094/PDIS.2001.85.3.236
24. Gea FJ, Tello JC, Honrubia M. In vitro sensitivity of *Verticillium fungicola* to selected fungicides. Mycopathologia. 1996;136:133–137. http://dx.doi.org/10.1007/BF00438918
25. Albertini C, Gredt M, Lerous P. Polymorphism of 14α-demethylase gene (CYP51) in cereal eyespot fungi *Tapesia acuformis* and *Tapesia yallundae*. Eur J Plant Pathol. 2003;109:117–128. http://dx.doi.org/10.1023/A:1022584822191
26. Dyer PS, Hansen J, Delaney A, Lucas JA. Genetic control of resistance to the sterol 14α-demethylase inhibitor fungicide prochloraz in the cereal eyespot pathogen *Tapesia yallundae*. Appl Environ Microbiol. 2000;66(11):4599–4604. http://dx.doi.org/10.1128/AEM.66.11.4599-4604.2000
27. Guarnaccia V, Aiello D, Giancarlo P. Emergence of prochloraz-resistant population of *Calonectria pauciramosa* and *Calonectria polizzi* in ornamental nurseries in Southern Italy. Plant Dis. 2014;98(3):344–350. http://dx.doi.org/10.1094/PDIS-04-13-0425-RE
28. Dekker J. The fungicide resistance problem: Will it grow worse? Bulletin OEPP/EPPO Bulletin. 1985;15:337–344. http://dx.doi.org/10.1111/j.1365-2338.1985.tb00238.x
29. Potočnik I, Vukojević J, Stajić M, Rekanović E, Milijaešvić S, Todorović B, et al. In vitro toxicity of selected fungicides from the groups benzimidazole and demethylation inhibitors to *Cladobotryum dendroides* and *Agaricus bisporus*. J Environ Sci Health B. 2009;44:364–370. http://dx.doi.org/10.1080/03601230902801059
30. Tamm H, Põldmaa K. Diversity, host associations, and phylogeography of temperate aurofusarin-producing *Hypomyces*/*Cladobotryum* including causal agents of cobweb disease of cultivated mushrooms. Fungal Biol. 2013;117:348–367. http://dx.doi.org/10.1016/j.funbio.2013.03.005
31. Bonnen A, Hopkins C. Fungicide resistance and population variation in *Verticillium fungicola*, a pathogen of the button mushroom, *Agaricus bisporus.* Mycol Res. 1997;101:89–96. http://dx.doi.org/10.1017/S0953756296002237
32. Delp CJ, Dekker J. Fungicide resistance: Definitions and use of terms. Bulletin OEPP/EPPO Bulletin. 1985;15:333–335. http://dx.doi.org/10.1111/j.1365-2338.1985.tb00237.x
33. Gouot J-M. Characteristics and population dynamics of *Botrytis cinerea* and other pathogens resistant to dicarboximides. In: Delp CJ, editor. Fungicide resistance in North America. St Paul, MN: APS Press; 1988. p. 53–55.
34. Gea FJ, Tello JC, Honrubia M. In vitro sensitivity of *Verticillium fungicola* to selected fungicides. Mycopathologia. 1996;136:133–137. http://dx.doi.org/10.1007/BF00438918
35. Hassall KA. The biochemistry and uses of pesticides. 2nd revised ed. New York: VCH Weinheim; 1990.
36. Grogan HM, Gaze RH. Fungicide resistance among *Cladobotryum* spp. – Causal agents of cobweb disease of the edible mushroom *Agaricus bisporus*. Mycol Res. 2000;104:741–745. http://dx.doi.org/10.1017/S0953756299001197

Potential human impact on the environmental central niche of the chacma baboon

AUTHORS:
Olivia M.L. Stone[1]
Shawn W. Laffan[1]
Darren Curnoe[1]
Andy I.R. Herries[2]

AFFILIATIONS:
[1]School of Biological Earth and Environmental Science, University of New South Wales, Sydney, Australia
[2]Department of Archaeology and History, College of Arts, Social Science and Commerce, La Trobe University, Melbourne, Australia

CORRESPONDENCE TO:
Olivia Stone

EMAIL:
oliviamlstone@gmail.com

POSTAL ADDRESS:
School of Biological Earth and Environmental Science, University of New South Wales, High St. Kensington, Sydney, New South Wales 2052, Australia

DATES:
Received: 12 Aug. 2014
Revised: 29 Jan. 2015
Accepted: 08 Apr. 2015

KEYWORDS:
Papio ursinus; habitat distribution; GIS; core habitat; anthropogenic impact

HOW TO CITE:
Stone OML, Laffan SW, Curnoe D, Herries AIR. Potential human impact on the environmental central niche of the chacma baboon. S Afr J Sci. 2015;111(11/12), Art. #2014-0279, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140279

We assessed the human impact on regions identified as the environmental central niche for the chacma baboon (*Papio ursinus*) across southern Africa. This central niche is the area within an animal's natural range that is most insulated from changes to the environmental variables that influence that animal's distribution. We used an environmental envelope model constructed with geographic information system software to predict the geographic extent of the central niche. The predicted chacma baboon central niche was 389 000 km^2, with substantial overlap with human settlements in several countries. Of note is that although Botswana contains nearly 60 000 km^2 of predicted central niche, the International Union for the Conservation of Nature chacma baboon distribution map implies that much of this area is uninhabited by baboons. A regional assessment of the province of KwaZulu-Natal (South Africa) suggests more than 95% of its central niche is uninhabited. Additionally, the very limited and likely disturbed central niche area in Lesotho coupled with the unknown status of chacma baboons within Lesotho warrants further attention. Overall, it appears likely that significant proportions of the predicted central niche in southern Africa are currently uninhabited by the chacma baboon. These uninhabited areas correspond with areas of high human population density and anthropogenic land alteration. The remaining central niche areas that are still inhabited are potentially key areas for conservation and are important for ensuring the sustainability of future populations. However, these areas may be undergoing degradation whilst also becoming more inaccessible to baboons, thus increasing the difficulty of conservation efforts. This preliminary assessment highlights the urgent need for detailed assessments at a finer scale.

Introduction

Human population expansion has left many other animal species with decreased available habitat or has forced them into direct competition with humans. Primates in particular are greatly affected by anthropogenic land alteration and land use conflict. As 48% of primate species are listed as 'vulnerable', 'endangered' or 'critically endangered' by the International Union for the Conservation of Nature (IUCN), including 37% of all African primates[1], habitat fragmentation can be extremely detrimental to population size and dynamics and metapopulation survival[2]. Often these same animals are in direct competition with humans, become known as pest species and suffer considerable persecution.[3] In many cases these persecuted species are those less charismatic fauna that are believed to be abundant. The chacma baboon (*Papio ursinus*) is one such animal.

Data on the relationship between a species and its habitat can provide fundamental information about distribution patterns, migration, population size and density. These data can be crucial for informing conservation efforts. Obtaining data about areas most valuable to a species will allow assessments of ecological requirements and help target geographical regions crucial to species sustainability. These types of data are undefined for the chacma baboon, a species with a supposed broad distribution across much of southern Africa.[4] Although the chacma baboon is classified as being of 'least concern',[5,6] further investigation into chacma baboon distribution is warranted by the recent recognition that the distribution within at least one province (KwaZulu-Natal, South Africa) is now highly fragmented owing to anthropogenic land alteration, resulting in declining and fragmented populations.[7] Until recently, there were only qualitative published predictions detailing the current distribution of the chacma baboon,[8-10] although this situation is changing, with a quantitative predicted habitat distribution prior to anthropogenic land disturbance now available.[11]

Previous studies refer to potentially important areas as prime or core habitat, but there is no consensus on the use of these terms. Rather there are several definitions, ranging from areas buffered from human disturbance, used in landscape assessments[12] to any area of habitat large enough to support at least one animal of the target species, as applied in giant panda (*Ailuropoda melanoleuca*) habitat predictions.[13] Core habitat has been identified as the area most frequently utilised, which can vary from feeding and ranging in howler monkeys (*Alouatta palliata*)[14] to den and rendezvous locations in grey wolves (*Canis lupus*)[15]. Alternatively, core habitat has been described as the area where the population can successfully survive and reproduce with the inclusion of intervening lands for dispersal.[16] Core or prime habitat could also refer to habitat within an individual animal's home range or to areas identified within the total range of a species. The chacma baboon has no species-specific core habitat definition and available core habitat definitions are unsuitable. However, as the chacma baboon occupies a range of environments throughout southern Africa[9,10,17,-18], and as environmental variables are most influential over a continental extent[19], modelling with environmental variables becomes the primary option. Consequently, we proposed and investigated the environmental central niche.

We describe the central niche as that within an animal's fundamental niche (where the fundamental niche is the full environmental range, including land and resources, with the potential to be inhabited, in the absence of limiting factors such as species competition). The central niche is land that is insulated and protected from the environmental extremes that influence the animal's distribution. Such areas are likely to place less stress on the animals, offer more favourable habitat and thus increase species survival. This statement is based on the assumption that individuals tend to live within the most advantageous habitat available to them. The central niche can contain a variety of different habitats and does not limit the total species range.

The aim of this study was to identify the chacma baboon's central niche and to compare it with human occupation and, by association, land disturbance. We used climatic data to predict the central niche of the chacma baboon.

Methods

Central niche areas are the set of geographical locations that fall within the central part of the environmental range for each of the environmental variables known to influence an animal's distribution. These ranges are determined using species observation records. The central 50% is identified as those locations which fall within the inter-quartile range of a set of environmental variables. Therefore, we estimate that approximately 50% of our known baboon populations are living within these limits. Using the central part of the range allows for a buffer (25%) to counter for outliers and less favourable conditions, whilst locating the land areas with the potential to have the most ideal and stable conditions.

We employed an environmental envelope model to identify the areas that satisfy the central 50% of the range for the environmental variables affecting baboon distribution. The approach used follows that of Stone et al.[11] Briefly, we defined the environmental envelope using data derived from the Worldclim data set.[20] Then, based on the actual ranges inhabited by the animals, land satisfying certain requirements is identified. Stone et al.[11] identified mean maximum temperature of the warmest month, mean minimum temperature of the coldest month, mean annual rainfall and altitude as the most likely environment variables to affect the distribution. The components of the model described here differ from Stone et al.[11] in that the ranges of temperature and rainfall used here were defined only for the central 50% and altitude was excluded because of its minimal effect on the central range. We used the Worldclim data set (resolution 30 arc seconds),[20] as it provided surfaces of the climatic variables.

The environmental data sets were sampled using 459 locations (each with a 2.5-km buffer)[11] known to have been inhabited by baboons (Figure 1). Both current and historical (within the last 100 years) location data were used to establish the environmental variable ranges. Any location data with a positional error exceeding 1 km were excluded. To prevent bias from areas of spatially dense samples, all location data were declustered. The end result was 459 locations with a more uniform spatial density consistent with the majority of the data. Random selection was used to thin the clustered areas (more detail is provided in Stone et al.[11]). The environmental conditions were sampled for each of the 459 location points with a 2.5-km radius buffer around each point. This 2.5-km buffer represents the average day range (5 km) for the chacma baboon.[8,11,21-29]

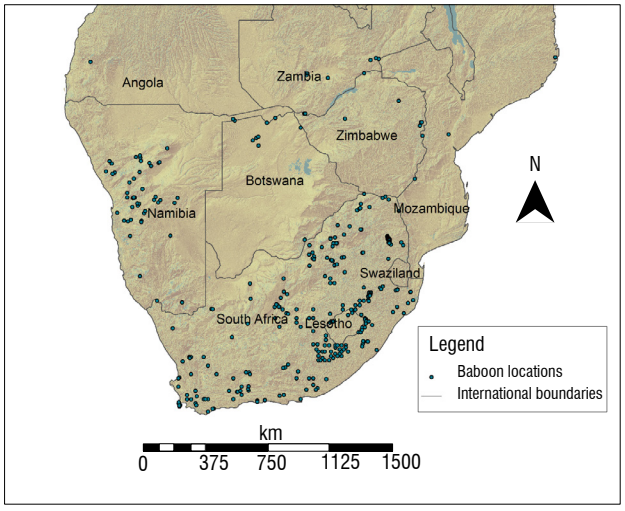


Figure 1: Location data used to sample the environmental variable data sets in order to model the environmental central niche of the chacma baboon in southern Africa.

For each environmental variable, the central 50% of the sampled range was identified using the 25th and 75th percentiles (Supplementary figure 1 online). All locations that fell within these bounds (Table 1) were given a value of 1 and those outside were given a value of 0 (Figure 2). The layers were then summed (using ArcGIS 10.0, ESRI, Redlands, CA, USA) to create a layer with a relative scale from 0 (outside the bounds for all variables) to 3 (within the bounds for all variables), with the latter being the predicted environmental central niche.

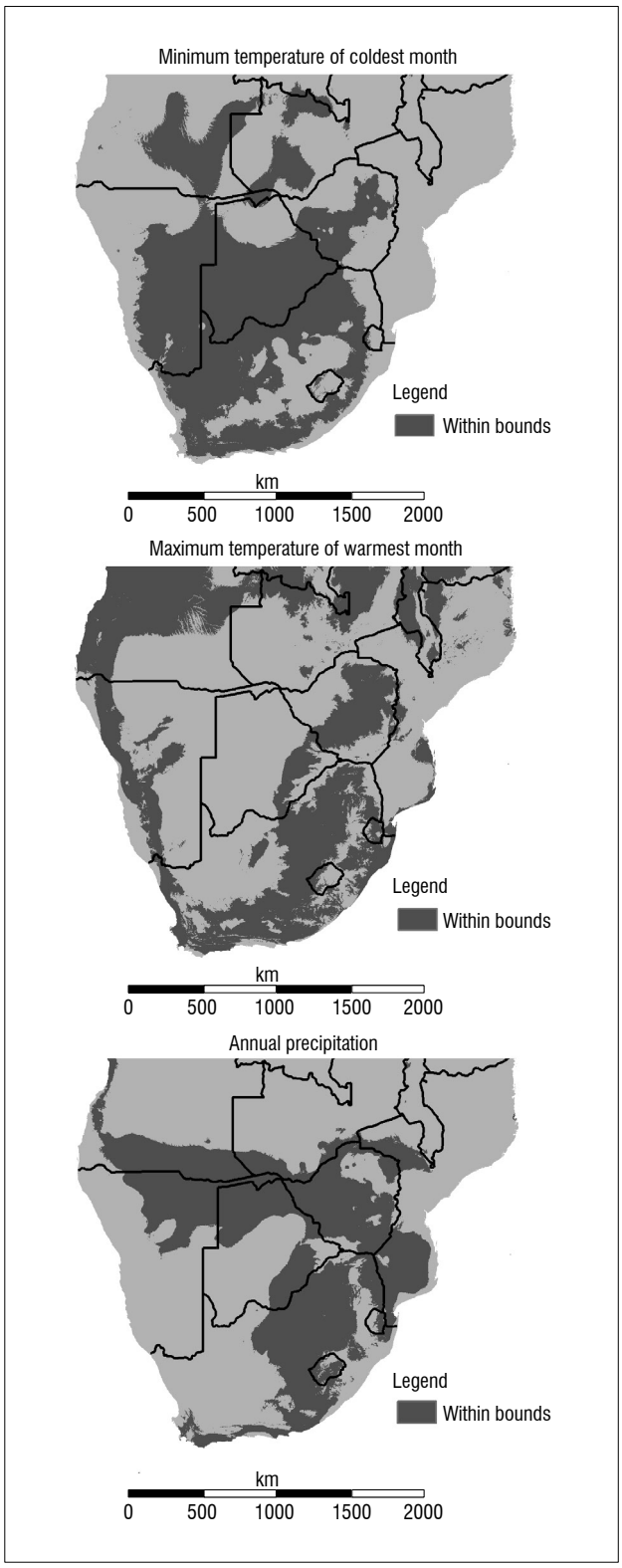


Figure 2: Layers depicting areas satisfying the criteria for each environmental variable. These areas are considered within bounds if the locations fall between the 25th and 75th percentiles of the distribution spanned by the observation locations and the 2.5-km buffer zones surrounding them.

Table 1: Parameters used to define the environmental central niche for the chacma baboon

Parameter	Minimum value	Maximum value
Minimum temperature (°C)	0.0	5.9
Maximum temperature (°C)	25.6	31.5
Rainfall (mm/year)	409	788

To assess the validity of the predicted central niche distribution, we needed to confirm if areas of estimated habitat, that are not impacted upon by cities and populated places, currently contained baboons or had contained baboons in the recent past. During fieldwork in August 2009, an area of central niche habitat was identified as satisfying these criteria with a resident population of baboons in rural KwaZulu-Natal, South Africa. Location data (visual sightings of animals and spoor) from throughout KwaZulu-Natal were collected using Garmin eTrex GPS units and additional location data were also recorded on 1:50 000 scale topographical maps as indicated by interviewees known to have reliable knowledge of the animal. Both areas known to be occupied by baboons and areas presumed to be uninhabited by baboons were surveyed. Further detail is provided in Stone et al.[7] These data were digitised and entered into the GIS to identify populations within the central niche area. Then, based on the non-habitat specific home range of 15.19 km^2,[7] we estimated how much of the central niche area in KwaZulu-Natal was inhabited by baboons.

We compared the predicted distribution of the chacma baboon central niche with the IUCN *Red List of Threatened Species* distribution for the chacma baboon[5] to identify areas that may require further investigation as a consequence of potential human impact. The IUCN map was used as it is easily accessible, widely known and generally accepted. The IUCN distribution map is at a continental extent, and depicts the animals as homogeneously distributed across the landscape.

The predicted central niche was then compared against a data set of human population estimates from the WorldPop Project[30] used under the Creative Commons Attribution 4.0 International Licence (http://creativecommons.org/licenses/by/4.0/). These data are provided at a resolution of 3 arc seconds, so were projected into Africa Albers Equal Area Conic at a resolution of 90 m to estimate the population density per km^2. We then calculated the fraction of the predicted central niche area that was occupied by human population densities exceeding a set of increasing thresholds (10 to 450 persons per km^2).

The central niche prediction was also compared with the distribution of large human settlements to assess how these areas have been affected by human occupation and land disturbance. Populated places were included in the assessment if they satisfied two criteria. Firstly, as a result of the southern African human population data being difficult and sometimes impossible to establish, we chose to use only settlements with population estimates that exceeded 20 000 people. Secondly, these large settlements needed to be contained by, or within 15 km of, predicted central niche areas.

For conservation purposes, the central niche areas in each selected country were overlaid and compared with protected areas[31] to gauge how much of the central niche is protected.

Finally, a tenfold cross validation was undertaken to test the reliability of the model. This validation used subsets that were formed by removing 10% of the original data points, randomly selected using R software (version 3.0.2). Each location point was given a buffer (2.5 km) and then converted back to points to sample the environmental data layers. These samples were analysed and the parameters identified and compared (Supplementary table 1 online). The parameters from the 10 samples were used to form layers that were then summed together. This summing resulted in one map (a confidence surface) for each environmental variable.

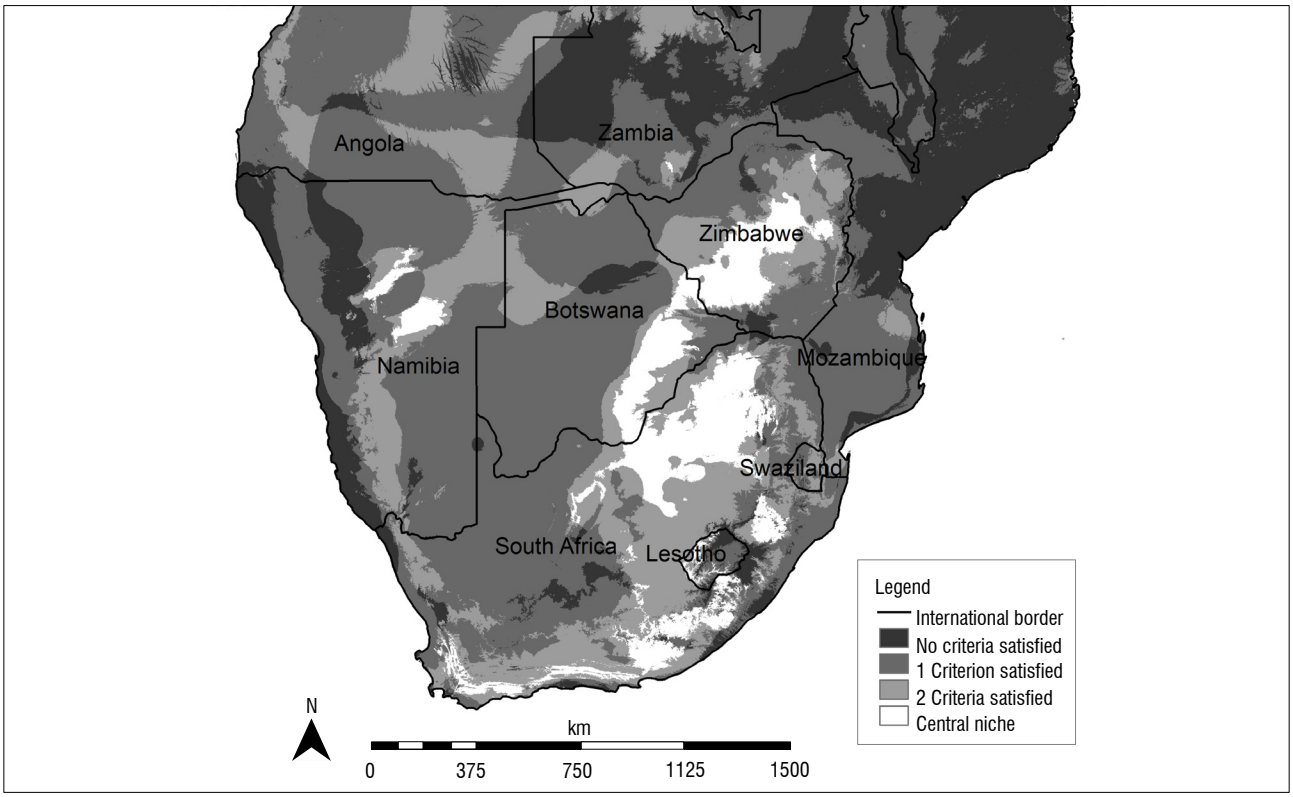


Figure 3: The predicted environmental central niche of the chacma baboon (*Papio ursinus*) in southern Africa.

Results

The central niche model identified approximately 389 000 km^2 of core habitat primarily located in six countries: Botswana, Lesotho, Namibia, South Africa, Zambia and Zimbabwe (Figure 3). South Africa and Zimbabwe contain approximately 78% of the predicted area. South Africa contains the largest area of central niche habitat, with approximately 219 570 km^2 and, at 462 km^2, the smallest area is within Zambia (Table 2). Lesotho also contains only a small area of approximately 3984 km^2. The habitat is a central region through northern South Africa, branching westward through southeastern Botswana, leading north to southwestern and central Zimbabwe. The distribution of the central niche is fragmented (which is expected), with large patches in central northern Namibia and more coastal patches leading from the Cape region of South Africa along the eastern coast to KwaZulu-Natal.

Table 2: The central niche habitat per country

Country	Central niche area (km^2)	Large settlements within the central niche	Area per settlement (km^2)	Central niche currently protected (km^2)
Botswana	59 053	11	5368	135
Lesotho	3984	7	569	1.1
Namibia	23 972	3	7991	10 956
South Africa	219 570	62	3541	17 944
Zambia	462	0	N/A	0
Zimbabwe	81 878	9	9098	2268
Total	**388 919**	**92**	**4227**	**31 304**

Rainfall determines the northern limits of the central niche habitat (Figure 2). With the higher rainfall to the north, coupled with low rainfall in the southwest and higher rainfall in the east, the predicted habitat is limited to a north–south corridor passing through eastern southern Africa and an east–west corridor sweeping from Mozambique to Namibia and marginally into Angola. Rainfall excluded 67% of southern Africa from the central niche.

Large areas, including >75% of Botswana and Mozambique and over 50% of Namibia, were above the bounds of maximum temperature in the hottest month (Figure 2). As an indicative value, approximately 58% of the land south of 10°S is outside the central niche because of this maximum temperature upper bound. In fact, the upper limits for both temperature variables were far more restricting than the lower limits. The upper bound for mean minimum temperature of the coldest month followed a similar pattern and excluded 57% of the land area south of 10°S.

None of the predicted central niche was located within Mozambique, as at least one variable was outside the envelope bounds in every location. Over half of Mozambique's total area was found to be outside the envelope bounds for all the variables, and both temperature variables were excluded as the upper bound was exceeded. Similarly, no central niche was predicted in Swaziland, as the land area was outside either the rainfall or minimum temperature bounds.

The majority (68%) of the central niche areas have a human population density of less than 10 people per km^2. A quarter of the central niche area has a density of between 10 and 100 people per km^2 and approximately 4% of the central niche area has a density of more than 200 people per km^2. A total of 5760 km^2 (1.5%) had a human population density of greater than 450 people per km^2. A total of 92 settlements (>20 000 people per settlement) were identified within the designated central niche area or within a distance of 15 km of its edge (Figure 4). Of these settlements, 30 were either provincial capitals (20), national capitals (5) or large cities (5) (Supplementary table 2 online). South Africa had the highest number of populated places, with 62 located within the predicted central niche areas, 4 of which are capital cities. Both Botswana and Lesotho had seven large cities within the central niche zones, including both national capitals, within or near to these areas.

A comparison of the predicted central niche areas with the IUCN distribution (Figure 4) suggests that a large proportion (>40%) of the central niche within Botswana is uninhabited by baboons. Botswana has the third largest area of predicted central niche and the third largest area of central niche land per large settlement (Table 2). Lesotho had the smallest amount of predicted area per settlement (Table 2), with all seven capitals in close proximity to the comparatively small central niche area.

South Africa had the largest area of central niche lands currently protected (approximately 8% of its core area) (Table 2). Namibia protects approximately 46% of its total central niche lands, which equates to the second largest holding of protected central niche. Interestingly, the majority of Namibia's protected central niche (92%) is non-government protected (private freehold and communal conservancies). Overall, approximately 8% of chacma baboon central niche is protected, including 0.03% of the central niche located in Lesotho, 3% in Zimbabwe and 0.2% in Botswana; none of the limited area in Zambia is protected.

Baboons in KwaZulu-Natal were recorded in unprotected central niche zones, despite the population's reliance on protected areas.[7] However, our findings suggest that less than 5% of KwaZulu-Natal's central niche is actually inhabited.

Lastly, the tenfold cross validation produced one confidence surface for each variable (Supplementary figure 2 online). A value of 10 identifies land that is always with the envelope and 0 identifies land that is never within the envelope bounds. The current central niche is predicted to cover approximately 389 000 km^2 of land (7% of available land south of 10°S); this tenfold cross validation results in central niche predictions ranging from approximately 315 000 km^2 to approximately 535 000 km^2 (or 5% to 9% of land south of 10°S). Areas of reduced confidence are not widespread and there are no substantial areas of conflict with the current central niche areas, thus supporting our prediction.

Discussion

Modelling and limitations

This model is preliminary and there are limitations where data is simply not available at such a large scale. However, we included the best available data and a search of the published literature aided the selection of environmental variables employed. There was a large interspecies variation in habitat mean annual rainfall across the genus,[32-35] thus suggesting some differentiation among *Papio* species. Rainfall indirectly affects plant productivity and species richness and thus food availability,[35-39] consequently foraging and ranging patterns, and indirectly distribution, are affected. However, rainfall seasonality was excluded as it was highly geographically restrictive, possibly as environmental extremes are more prevalent over an area as large as southern Africa. For example, rainfall in some inhabited arid areas is so low (<15 mm/year[11]) that the effect of seasonality becomes nullified. Both temperature variables are included because of their direct and indirect influence on distribution;[21,22,40-43] and altitude was excluded as it was found to have very little effect on the central range, possibly because of the already limiting effects of temperature.

This model predicts potential liveable environmental ranges and not an actual distribution. Therefore, the dates of the baboon location data are not required. More pertinent was the accuracy of the location data in order to sample the correct environmental conditions. Equally the human settlement data were chosen for accuracy. We found population estimates exceeding 20 000 to be more consistent and large enough to cause considerable land disturbance. The 15-km buffer allowed for urban, sub-urban and semi-rural spread from the datum and the baboons' ability to range towards the settlements (up to 19.3 km daily[22]).

There is no accurate up-to-date distribution map for the chacma baboon and we acknowledge the limitations arising from using the IUCN distribution, including undisclosed data accuracy. Whilst this map is useful for comparison, it does not consider regional or even international variations in distribution. Therefore, it can only be a guide for further research. Regional assessments have yet to be undertaken in most areas of southern Africa.

Modelling over a continental extent means variations in environmental conditions over time (from the last 100 years) are smoothed out and

there are likely to be areas included that are uninhabited by baboons, e.g. drought-affected areas. Furthermore, as Worldclim data are already temporally averaged, we expect the model to overestimate the land area. Although the model is mathematically derived, it is still based on the biological (environmental) requirements for the central section of the population (the central part of the range of the environmental conditions inhabited) and will encompass mean and median conditions encountered by actual populations. We expect the central niche areas to include only a small proportion of the entire species range. Nonetheless, these areas have the best potential to persist in adverse environmental conditions as a result of the large buffers (25%) afforded by the analysis. We cannot know whether the bounds include the entirety of the area most likely to promote species survival, nor can we determine if all less suitable areas are excluded. However, this central niche method does produce a quantitative, repeatable and testable hypothesis.

Core habitat and other baboon species

The area of core habitat is approximately 10% of the estimated total suitable habitat[11] as large areas of potential habitat are excluded from the central niche areas. This exclusion is because of a narrow range of rainfall (409–788 mm in the core habitat compared with a range of 15– 1555 mm for chacma baboon distribution[11]), and because temperatures are too hot over much of the potential range (e.g. in Mozambique).

The chacma baboon core habitat is extremely limited north of the Zambezi River, which is a putative transition or contact zone for Kinda baboons (*Papio kindae*), yellow baboons (*Papio cynocephalus*) and chacma baboons.[18,44,45] It is possible that species interaction could limit chacma baboon access to the central niche areas north of the Zambesi. However, assuming competitive exclusion, any contact should occur at the edges of the whole distribution, thus it is unlikely to occur at the edges of the central niche. Equivalent studies would need to be undertaken for both Kinda and yellow baboons to further investigate potential impact from species competition.[11]

Anthropogenic disturbance

The absence of baboons in Botswana's eastern central niche area, as depicted by the distribution shown in the IUCN map, may be related to the large number of major settlements within this region. Seven of the major settlements are capital cities including six provincial capitals and Gaborone, the national capital. It is possible that this land disturbance may have adversely affected baboon distribution. Botswana's central niche land area per settlement was third largest amongst the southern African countries (Table 2). Yet if the area per settlement is recalculated excluding the area that the IUCN reported as uninhabited by baboons, the land per settlement drops from 5368 km^2 to 3836 km^2, a reduction of 29% with nine cities remaining within the predicted central niche habitat.

The settlements in southeastern Botswana can be seen as a geographical cluster in the central niche area (Figure 4). This cluster includes or is adjacent to land that is uninhabited by baboons according to the IUCN distribution map. It is possible that the land in Botswana that is potentially the most valuable to the chacma baboon may be significantly compromised with regard to both ecological integrity and baboon access to the area. This land in southeastern Botswana is climatically compatible and has no major geographical barrier separating it from adjacent South African baboon populations, yet baboons are reportedly absent. Botswana's urban satellite villages are rapidly increasing in size, with urban villages growing faster than modern towns and cities.[46] In 1991, 50% of the national population of Botswana resided within 100 km of Gaborone.[47] Twenty years later the national population had reached over 2 million,[46] making Gaborone one of the fastest growing capital cities in Africa. The land disturbance within 100 km of Gaborone will likely be extensive as a result of this population growth. This 100-km zone encompasses a considerable proportion of the predicted central niche area in Botswana (Figure 5) and the central niche within this zone has a mean human population density of 21 people per km^2. This area of central niche has a higher mean density than 76% of the total central niche land and Gaborone itself has a population density of 1160 people per km^2.[30]

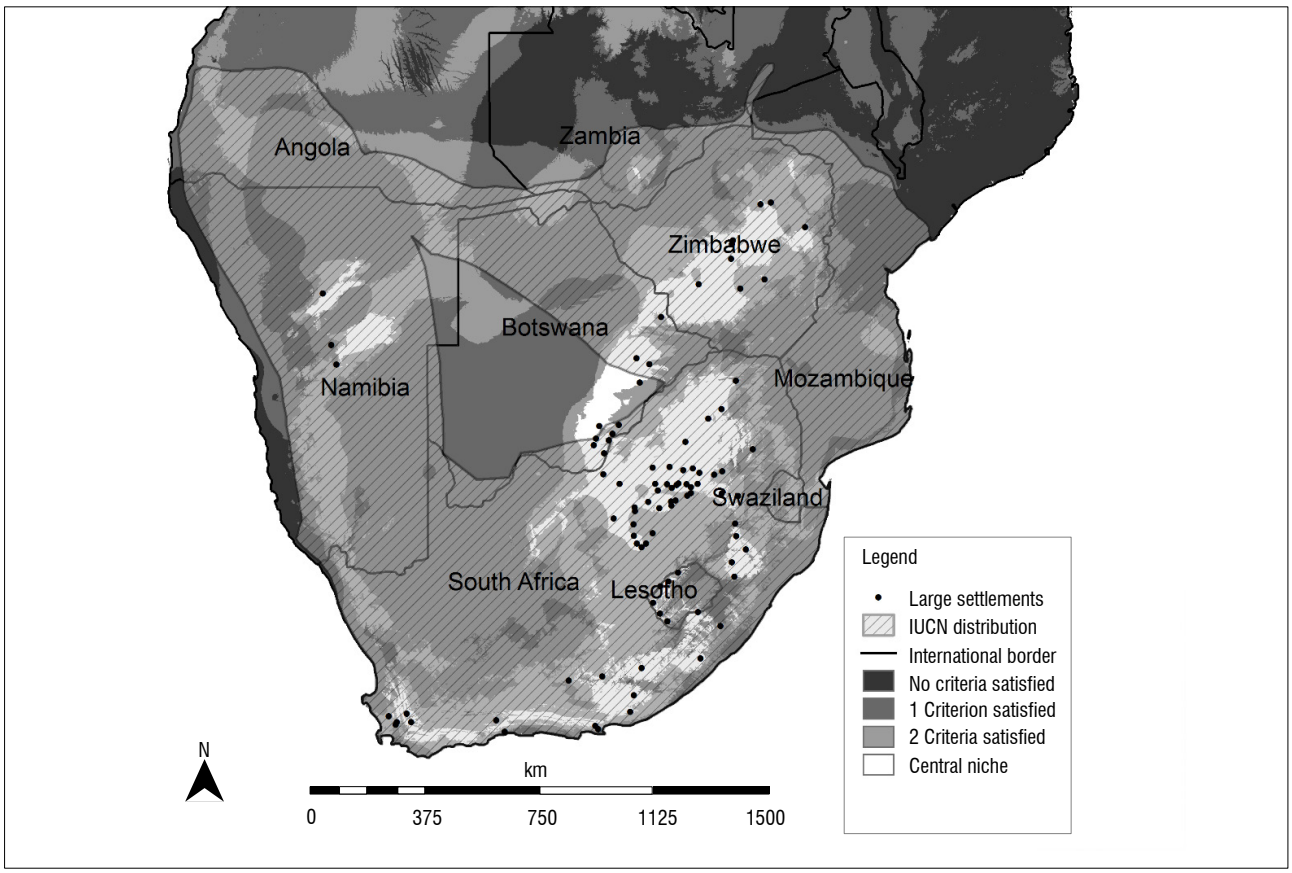


Figure 4: African settlements of >20 000 people within the predicted environmental central niche, overlaid with the IUCN distribution of the chacma baboon.[5]

Moreover there are many smaller settlements (with a human population <20 000), often in close proximity to the larger settlements, which are not depicted in Figure 5. These findings may help to explain why baboons are not found in some of these potentially viable areas.

Anthropogenic land disturbance in Lesotho may also be impacting baboons. The amount of central niche predicted within Lesotho is small. Our model predicts a total of 3984 km^2 of land in close proximity to seven major settlements (Supplementary figure 3 online). This figure provides an estimate of only 569 km^2 per city, all of which are provincial capitals. It is likely that these animals would be greatly affected by anthropogenic land alteration. The status of baboons in Lesotho deserves further investigation.

Conservation planning and protected areas

This model of predicted central niche areas has direct implications for conservation planning. The 92 large settlements identified within the central niche zone increases the likelihood of land disturbance in these areas. With conservation of the chacma baboon a growing concern, conservation authorities may have difficulty acquiring or even maintaining protected areas within central niche zones. Baboons are a highly adaptable, resilient species. Thus, if baboon populations are noted to decline over larger areas, then the more inflexible species will have already suffered. Human land use and human occupation can leave land transformed and deconditioned[48,49] and form physical barriers to wildlife. Although there are many protected areas within these zones, especially in South Africa, these areas are fragmented and separated by long distances. In addition, baboons are not always safe in these protected areas, as they often are culled when it is perceived that they have become problematic.

There are baboon populations occupying the core area to the east of Johannesburg and Pretoria. This area is under constant threat of development; however, the region is partly protected from major development by the Cradle of Humankind UNESCO World Heritage area and the Magaliesberg Protected Natural Environment. It is critical that we maintain these protected environments and protect similar and surrounding locations that are being encroached by new developments. As human population density and its resulting influence are increasing,[50-52] areas of core habitat with lower human population density and potentially less degradation become ideal conservation areas (Supplementary figure 4 online).

The review of this central niche prediction is a preliminary assessment which highlights the need for more detailed assessments in localised areas. Current and accurate location data are very limited and, with the exception of KwaZulu-Natal,[7] no regional assessments are published. There are spot assessments covering smaller areas and assessments of individual troops,[27,29,35,53,54] but nothing that would enable quantitative regional assessments. In addition, we have little detailed associated data regarding the surroundings of the very limited location data we can access. For the most part, we do not have enough information even for localised basic assessments. The lack of data includes but is not limited to some auxiliary data such as baboon abundance, permanent water sources (including small human-made farming adaptations or even streams or springs), flora, fauna and land usage at locations. As a preliminary assessment at a continental scale we did not presume to expect such data or detail; however the overall lack of such data severely limits further analysis. Further regional assessments need to be undertaken and it is unlikely that these will progress without prior

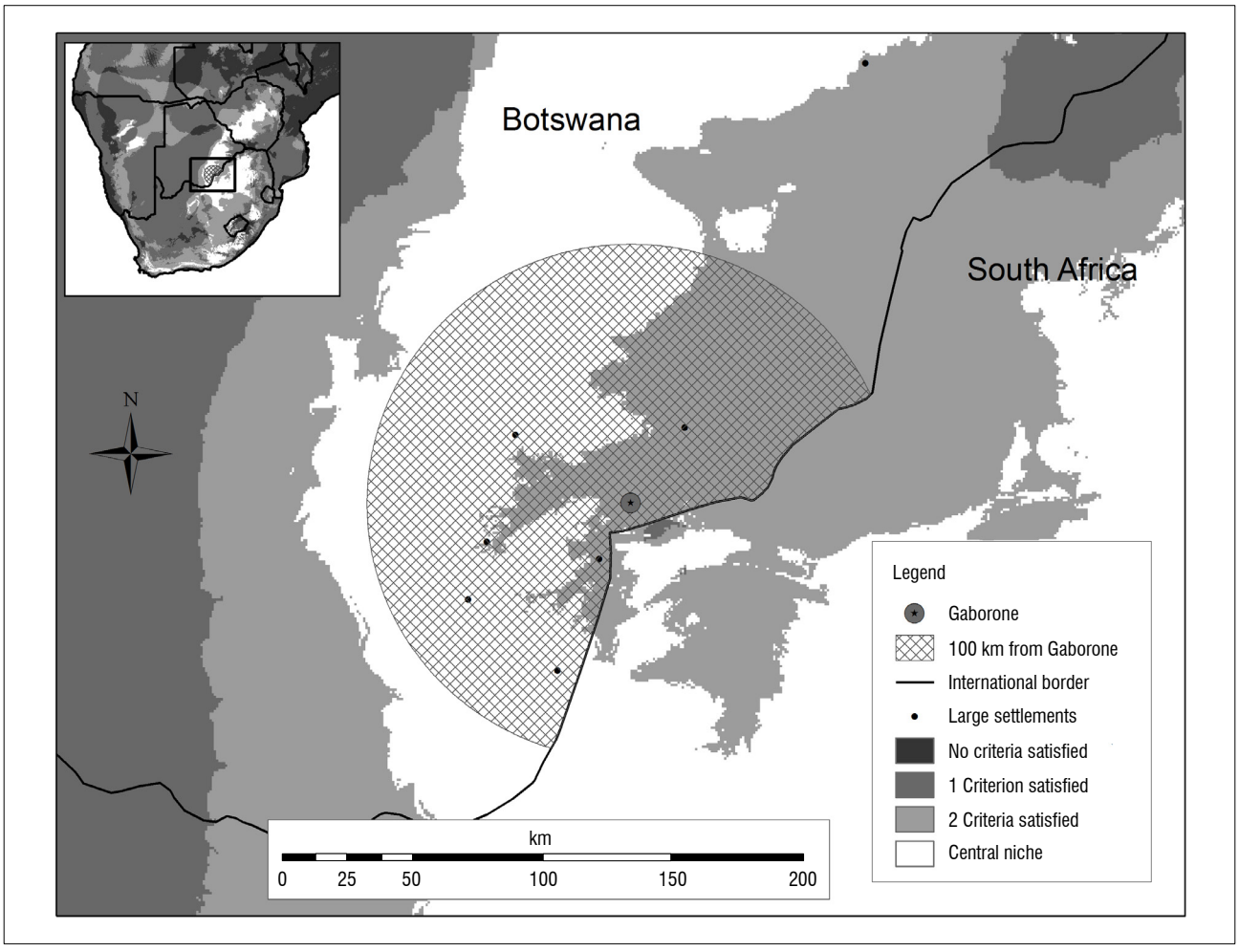


Figure 5: The 100-km zone surrounding Gaborone (Botswana), overlaid onto the predicted environmental central niche of the chacma baboon.

sub-regional assessments. These sub-regional assessments need to be specific to their local populations, and should assess viable population sizes and land requirements, including access to more valuable land areas.

This central niche prediction is a best estimate, formed using data sets that are themselves interpolated from raw data. This interpolation, combined with location data from the last 100 years, means that subtle changes in environmental conditions are smoothed out over the continental scale and the prediction is most likely to be an overestimate. We acknowledge the central niche is only the central habitat located within a much wider range occupied by baboons, but if future research finds these areas of key habitat are not able to support healthy populations then the concern must be that areas less conducive to survival will be undergoing a greater loss.

Acknowledgements

Location data used in the modelling were provided by Amathole Museum, Ezemvelo Wildlife, Iziko South African Museum, National Museum Bloemfontein and the Ditsong National Museum of Natural History (formerly the Transvaal Museum). We thank Ezemvelo KZN wildlife staff and residents of KwaZulu-Natal who assisted us in locating baboon troops and the anonymous reviewers whose comments much improved this manuscript. A.I.R.H. was supported by an Australian Research Fellowship as part of ARC Discovery Grant DP0877603 and ARC Future Fellowship FT 120100399.

Authors' contributions

All authors conceived the study; O.M.L.S., S.W.L. and A.I.R.H. collected the data; O.M.L.S. undertook the analysis; and O.M.L.S. led the writing with revisions undertaken by all authors.

References

1. IUCN. Primate fact sheet [document on the Internet]. c2008 [cited 2011 Oct 23]. Available from: http://cmsdata.iucn.org/downloads/primate_fact_sheet_2008___final.pdf;
2. Ewers RM, Didham RK. The effect of fragment shape and species' sensitivity to habitat edges on animal population size. Conserv Biol. 2007;21(4):926–936. http://dx.doi.org/10.1111/j.1523-1739.2007.00720.x
3. Lee PC, Priston NEC. Human attitudes to primates: perceptions of pests, conflict and consequences for primate conservation. In: Patterson JD, Wallace J, editors. Commensalism and conflict: The human-primate interface. Alberta: American Society of Primatologists; 2005. p. 1–23.
4. Jolly CJ. Species, subspecies and baboon systematics. In: Kimbel WH, Martin LB, editors. Species concepts and primate evolution. New York: Plenum Press; 1993. p. 67–107. http://dx.doi.org/10.1007/978-1-4899-3745-2_4
5. Hoffman M, Hilton-Taylor C. *Papio ursinus*: IUCN Red List of Threatened Species, Version 2011.1 [document on the Internet]. c2008 [cited 2011 July 06]. http://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T136856A4347282.en
6. Friedmann Y, Daly B. Red data book of the mammals of South Africa: A conservation assessment. Johannesburg: CBSG Southern Africa, Conservation Breeding Specialist Group (SSC/IUCN), Endangered Wildlife Trust; 2004.
7. Stone OML, Laffan SW, Curnoe D, Rushworth I, Herries AIR. Distribution and population estimate for the chacma baboon (*Papio ursinus*) in KwaZulu-Natal, South Africa. Primates. 2012;53(4):337–344. http://dx.doi.org/10.1007/s10329-012-0303-9
8. Anderson CM. Baboons below the tropic of capricorn. J Hum Evol. 1982;11(3):205–217. http://dx.doi.org/10.1016/S0047-2484(82)80037-7
9. Baldwin LA, Teleki G. Field research on baboons, drills, and geladas: An historical, geographical, and bibliographical listing. Primates. 1972;13(4):427–432. http://dx.doi.org/10.1007/BF01793662
10. Hall KRL. Variations in the ecology of the chacma baboon, *Papio ursinus*. Symp Zool Soc Lond. 1963;10:1–28.
11. Stone OML, Laffan S, Curnoe D, Herries AIR. The spatial distribution of chacma baboon (*Papio ursinus*) habitat based on an environmental envelope model. Int J Primatol. 2013;34(2):407–422. http://dx.doi.org/10.1007/s10764-013-9669-9
12. Goetz SJ, Jantz P, Jantz CA. Connectivity of core habitat in the northeastern United States: Parks and protected areas in a landscape context. Remote Sens Environ. 2009;113(7):1421–1429. http://dx.doi.org/10.1016/j.rse.2008.07.019
13. Linderman M, Bearer S, An L, Tan Y, Ouyang Z, Liu J. The effects of understory bamboo on broad-scale estimates of giant panda habitat. Biol Conserv. 2005;121(3):383–390. http://dx.doi.org/10.1016/j.biocon.2004.05.011
14. Williams-Guillén K, McCann C, Martínez Sánchez J, Koontz F. Resource availability and habitat use by mantled howling monkeys in a Nicaraguan coffee plantation: Can agroforests serve as core habitat for a forest mammal? Anim Conserv. 2006;9(3):331–338.
15. Blanco JC, Cortés Y, Virgós E. Wolf response to two kinds of barriers in an agricultural habitat in Spain. Can J Zool. 2005;83(2):312–323. http://dx.doi.org/10.1139/z05-016
16. Helferty N. Natural heritage planning for amphibians and their habitats: With reference to populations on the south slope of the Oak Ridges Moraine. Supplementary report for Oak Ridges Moraine Richmond Hill Ontario Municipal Board Hearing, Save the Rouge Valley System Inc. and the City of Toronto. Ontario: Natural Heritage Consulting; 2002.
17. Sithaldeen R, Bishop JM, Ackermann RR. Mitochondrial DNA analysis reveals Plio-Pleistocene diversification within the chacma baboon. Mol Phyl Evol. 2009;53(3):1042–1048. http://dx.doi.org/10.1016/j.ympev.2009.07.038
18. Zinner D, Wertheimer J, Liedigk R, Groeneveld LF, Roos C. Baboon phylogeny as inferred from complete mitochondrial genomes. Am J Phys Anthropol. 2013;150(1):133–140. http://dx.doi.org/10.1002/ajpa.22185
19. Pearson RG, Dawson TP. Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? Glob Ecol Biogeogr. 2003;12(5):361–371. http://dx.doi.org/10.1046/j.1466-822X.2003.00042.x
20. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965–1978. http://dx.doi.org/10.1002/joc.1276
21. Stoltz LP, Saayman GS. Ecology and behaviour of baboons in the Northern Transvaal. Ann Transvaal Mus. 1970;26(5):99–143.
22. Hall KRL. Numerical data, maintenance activities and locomotion in the wild chacma baboon, *Papio ursinus*. Proc Zool Soc Lond. 1962;139(2):181–220. http://dx.doi.org/10.1111/j.1469-7998.1962.tb01831.x
23. Brain C. Spatial usage of a desert environment by baboons (*Papio ursinus)*. J Arid Environ. 1990;18:67–73.
24. Davidge C. Ecology of baboons (*Papio ursinus*) at Cape Point. Zool Afr. 1978;13:329–350. http://dx.doi.org/10.1080/00445096.1978.11447633
25. Gaynor D. Foraging and feeding behaviour of chacma baboons in a woodland habitat [PhD thesis]. Durban: University of Natal; 1994.
26. Hamilton WJ. Namib Desert chacma baboon (*Papio ursinus*) use of food and water resources during a food shortage. Madoqua. 1986;14(4):397–407.
27. Marais AJ. Resource utilisation of the chacma baboon in different vegetation types in north-eastern mountain sour veld, Blyde Canyon Nature Reserve [dissertation]. Pretoria: University of South Africa; 2005.
28. Van der Weyde L. Ecology of chacma baboons, *Papio hamadryas ursinus*, in a semi-desert environment in South Africa. Sydney: University of New South Wales; 2004.
29. Whiten A, Byrne R, Henzi S. The behavioral ecology of mountain baboons. Int J Primatol. 1987;8(4):367–388. http://dx.doi.org/10.1007/BF02737389
30. Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One. 2012;7(2):e31743. http://dx.doi.org/10.1371/journal.pone.0031743
31. IUCN, UNEP. The World Database on Protected Areas (WDPA). Cambridge, UK: UNEP-WCMC; 2015.

32. Biquand S, Biquand-Guyot V, Boug A, Gautier J-P. The distribution of *Papio hamadryas* in Saudi Arabia: Ecological correlates and human influence. Int J Primatol. 1992;13(3):223–243. http://dx.doi.org/10.1007/BF02547815

33. Dunbar RIM, Dunbar EP. Ecological relations and niche separation between sympatric terrestrial primates in Ethiopia. Folia Primatol. 1974;21(1):36–60. http://dx.doi.org/10.1159/000155595

34. Barrett L, Henzi SP. An inter-population comparison of body weight in chacma baboons. S Afr J Sci. 1997;93(10):436–438.

35. Hamilton WJ, Buskirk RE, Buskirk WH. Defense of space and resources by chacma (*Papio ursinus*) baboon troops in an African desert and swamp. Ecology. 1976;57(6):1264–1272.

36. Kay RF, Madden RH, Van Schaik C, Higdon D. Primate species richness is determined by plant productivity: Implications for conservation. Proc Natl Acad Sci USA. 1997;94(24):13023–13027. http://dx.doi.org/10.1073/pnas.94.24.13023

37. O'Brien EM. Climatic gradients in woody plant species richness: Towards an explanation based on an analysis of southern Africa's woody flora. J Biogeogr. 1993;20(2):181–198. http://dx.doi.org/10.2307/2845670

38. Hamilton W. Demographic consequences of a food and water shortage to desert chacma baboons, *Papio ursinus*. Int J Primatol. 1985;6(5):451–462. http://dx.doi.org/10.1007/BF02735570

39. Barton RA, Whiten A, Strum SC, Byrne RW, Simpson AJ. Habitat use and resource availability in baboons. Anim Behav. 1992;43:831–844. http://dx.doi.org/10.1016/S0003-3472(05)80206-4

40. Henzi S, Byrne R, Whiten A. Patterns of movement by baboons in the Drakensberg mountains: Primary responses to the environment. Int J Primatol. 1992;13(6):601–629. http://dx.doi.org/10.1007/BF02551256

41. Korstjens AH, Lehmann J, Dunbar R. Resting time as an ecological constraint on primate biogeography. Anim Behav. 2010;79(2):361–374. http://dx.doi.org/10.1016/j.anbehav.2009.11.012

42. Hill RA. Thermal constraints on activity scheduling and habitat choice in baboons. Am J Phys Anthropol. 2006;129:242–249. http://dx.doi.org/10.1002/ajpa.20264

43. Hill RA. Day length seasonality and the thermal environment. In: Brockman DK, Van Schaik CP, editors. Seasonality in primates: Studies of living and extinct human and non-human primates. Cambridge, UK: Cambridge University Press; 2005. p. 197–214. http://dx.doi.org/10.1017/CBO9780511542343.008

44. Keller C, Roos C, Groeneveld LF, Fischer J, Zinner D. Introgressive hybridization in southern African baboons shapes patterns of mtDNA variation. Am J Phys Anthropol. 2010;142(1):125–136. http://dx.doi.org/10.1002/ajpa.21209

45. Jolly CJ, Burrell AS, Phillips-Conroy JE, Bergey C, Rogers J. Kinda baboons (*Papio kindae*) and grayfoot chacma baboons (*P. ursinus griseipes*) hybridize in the Kafue River Valley, Zambia. Am J Primatol. 2011;73(3):291–303. http://dx.doi.org/10.1002/ajp.20896

46. Majelantle A. Population & housing census: Preliminary results brief. Gaborone: Census Office, General Statistics Office Botswana; 2011.

47. Gwebu T. Patterns and trends of urbanization in Botswana and policy implications for sustainability. Gaborone: Department of Environmental Science, University of Botswana; 2004.

48. Mills A, Fey M. Declining soil quality in South Africa: Effects of land use on soil organic matter and surface crusting. S Afr J Sci. 2003;99(9):429–436.

49. Reynolds JF, Maestre FT, Kemp PR, Stafford-Smith DM, Lambin E. Natural and human dimensions of land degradation in drylands: Causes and consequences. In: Canadell J, Pataki D, Pitelka LF, editors. Terrestrial ecosystems in a changing world. Berlin: Springer-Verlag; 2007. p. 247–257. http://dx.doi.org/10.1007/978-3-540-32730-1_20

50. Lehohla P. Mid-year population estimates 2014. Pretoria: Statistics South Africa; 2014.

51. Steytler J. Namibia population projections 2011–2041. Windhoek: Namibia Statistics Agency; 2014.

52. Majelantle A. Population and housing census 2011: Analytical report. Gaborone: General Statistics Office Botswana; 2014.

53. Stoltz LP, Keith ME. A population survey of chacma baboon in the Northern Transvaal. J Hum Evol. 1973;2:195–212. http://dx.doi.org/10.1016/0047-2484(73)90076-6

54. Hoffman T, O'Riain MJ. The spatial ecology of chacma baboons (*Papio ursinus*) in a human-modified environment. Int J Primatol. 2011;32(2):308–328. http://dx.doi.org/10.1007/s10764-010-9467-6

Note: This article is supplemented with online only material.

Oxidative stability of blesbok, springbok and fallow deer *droëwors* with added rooibos extract

AUTHORS:
Maxine Jones[1,2]
Louwrens C. Hoffman[2]
Magdalena Muller[1]

AFFILIATIONS:
[1]Department of Food Science, Stellenbosch University, Stellenbosch, South Africa
[2]Department of Animal Sciences, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Louwrens Hoffman

EMAIL:
lch@sun.ac.za

POSTAL ADDRESS:
Department of Animal Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

DATES:
Received: 07 Oct. 2014
Revised: 19 Jan. 2015
Accepted: 20 Apr. 2015

KEYWORDS:
Aspalathus linearis; natural antioxidant; lipid oxidation; protein oxidation; dried sausage; venison

HOW TO CITE:
Jones M, Hoffman LC, Muller M. Oxidative stability of blesbok, springbok and fallow deer *droëwors* with added rooibos extract. S Afr J Sci. 2015;111(11/12), Art. #2014-0347, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140347

The addition of rooibos extract (RBE) (0%, 0.25%, 0.50%, 1.0% RBE) to improve the oxidative stability of blesbok, springbok and fallow deer *droëwors* (dried sausage) was studied. RBE treatments had no effects ($p>0.05$) on the lipid and protein oxidation of the dried product. With the addition of RBE 0.25%, lipid stability after drying showed the malonaldehydes decreased considerably. Haem-iron concentration increased after drying and differed ($p<0.05$) between RBE treatments within the dried stage within species. There were no differences ($p>0.05$) between the moisture, protein and fat contents between treatments within a specific processing stage. With the high polyunsaturated fatty acid content of the sausages, a high level of oxidation occurred. Even though RBE addition did not reduce oxidation significantly during the drying process, it could be a successful addition to the traditional South African meat product if it is shown to impart positive flavour attributes.

Introduction

Game that is hunted and harvested in South Africa is utilised both locally and internationally.[1] South Africa is known for its dried meat products such as biltong (a type of jerky) and *droëwors*. *Droëwors* is traditionally made from ground beef and animal fat (5–30% on a wet mass basis) and is a ready-to-eat dried seasoned meat sausage.[2] The *droëwors* market has expanded and is presently producing *droëwors* from different game species.[3]

Game meat is ideal for the production of *droëwors* as it is a lean meat with a low intramuscular fat content, has a favourable fatty acid profile and high iron content.[1,4] Blesbok (*Damaliscus pygargus phillipsi*) and springbok (*Antidorcas marsupialis*) are harvested throughout southern Africa. They are both used in the production of biltong and *droëwors*. It is becoming more popular to harvest feral fallow deer (*Dama dama*) and as this meat could also be used in the production of *droëwors*, it would be of interest to evaluate the suitability of this species for *droëwors*.

As *droëwors* is commonly stored at ambient temperatures for a long period and has added fat, oxidation is likely to occur. Lipid and protein oxidation in meat results in the development of off-flavours and aromas.[5,6] Natural antioxidants are commonly added to processed products to slow down oxidation and in turn, improve the shelf-life and flavour profile of the products.[7,8] Rooibos extract (RBE) is becoming more popular as a food ingredient, being used in products such as yoghurt, ready-to-drink iced teas and jams.[9] It is a powdered tea extract which is a waste product when producing commercial rooibos tea. The flavonoids found in rooibos have been shown to have the ability to inhibit lipid peroxidation in model systems.[10]

In commercial markets, synthetic and natural antioxidants are added to meat products. Butylated hydroxyanisole, butylated hydroxyl toluene, tertiary butylhydroquinone, sulphur dioxide and tocopherol are commonly added as a blend to enhance shelf-life via inhibition of oxidation.[11] Consumers are more aware of the addition of synthetic antioxidants and therefore by replacing these with RBE, this should overcome the negative perception of some of the meat products currently sold in today's market. According to Liu et al.[12] teas have been successful at low concentrations when used in combination with tocopherol to lower lipid oxidation. Research by Cullere et al.[13] and Hoffman et al.[14] suggests that the use of RBE at low concentrations could result in improved oxidative stability in meat products.

The aim of this study is therefore to investigate the effect of the addition of different levels of rooibos tea extract as a natural antioxidant to blesbok, springbok and fallow deer *droëwors* regarding its chemical properties and oxidation levels after drying.

Materials and methods

Droëwors *production*

Game meat of each species (blesbok, $n=6$; springbok, $n=6$; and fallow deer, $n=3$) was trimmed of all external fat and connective tissue and cut into approximately 100 mm x 100 mm cubes. No specific cuts were used. To increase the amount of fat in the product, sheep meat/fat (30 meat: 70 fat) was added. Three separate 36 kg replicates of *droëwors* from each species were made. Each 36 kg replicate was further divided into four treatment batches of 9 kg each. Each 9 kg batch consisted of 66.6% game meat and 33.3% sheep meat and fat. Each 9 kg batch was then combined with 1.5% (135 g) salt, 1% (90 g) pepper and varying amounts of powdered rooibos extract (RBE: 0%, 0.25%, 0.50% and 1.0% w/w concentrations) (Afriplex, Paarl, South Africa). The phenolic composition of the RBE used is depicted in Table 1. Each treatment batch was minced through a 10-mm grinder after which the spices (salt and pepper) and varying concentrations of RBE were added. The minced meat and spices were then mixed and minced through a 5-mm grinder. Natural sheep casings (22 mm diameter) were filled with the minced meat mixture. Half the sausage from each batch was dried in an environmentally controlled maturation chamber (Reich Unicontrol 2000S, Deli Spices, Cape Town, South Africa). Initially the sausage underwent a reddening stage at 30 °C with a 60% relative humidity for 15 min followed by a drying period of 48 h at a constant temperature of 30 °C and a relative

humidity of 30%. The drying process took 2 days during which the *droëwors* lost 45–50% of its mass. The other half of the sausage was left raw for analyses.

Table 1: Phenolic composition of fermented rooibos extract used in *droëwors* production

Flavonoid	Concentration (g compound/100 g extract)
Aspalathin	0.482
Nothofagin	0.056
Iso-orientin	0.956
Orientin	0.823
Quercetin-3-O-robinobioside	0.637
Isoquercitrin	0.138
Vitexin	0.179
Hyperoside	0.167
Rutin	0.179
Isovitexin	0.144
Luteolin-7-O-glucoside	0.196

Analysed according to the method of Beelders et al.[16]

Preparation and analysis of rooibos extract

Fermented rooibos extract was used and prepared as per the procedure reported by Joubert et al.[15] The phenolic composition of the rooibos extract (Table 1) was analysed according to the method of Beelders et al.[16] through an optimised reversed phase liquid chromatographic (RP-LC) separation technique.

Preparation of samples for analyses

To ensure a representative sample was tested, five replicate sub-samples (150–200 g) of both the raw and dried *droëwors* of each batch of *droëwors* were taken and individually homogenised in a blender for 3 min. The dried *droëwors* samples were taken within 30 min after the drying process was completed. The samples were stored in a -80 °C freezer until analysed. The homogenised samples were used for proximate analysis, fatty acid composition, haem-iron concentration, as well as for lipid and protein oxidation.

Proximate composition analysis

Samples were analysed according to the Association of Official Analytical Chemists (AOAC) International to determine the moisture[17] and ash[18] content. The protein content was determined by multiplying the amount of nitrogen (N) by a factor of 6.25; N levels were determined by the Dumas combustion method as described in procedure 992.15 of the AOAC.[19] Fat was determined using the chloroform/methanol (2:1) fat extraction method according to Lee et al.[20] All analyses were performed in duplicate.

Fatty acid composition

The fatty acids were methylated and then analysed using gas chromatography to determine the fatty acid composition and percentages. A modified method by Folch et al.[21] was used. An internal standard, heptadecanoic acid ($C_{17}H_{34}O_2$) (Cat. No. H3500, Sigma-Aldrich Inc, St.Louis, MO, USA.) was added to quantify the individual fatty acids. The fatty acid composition was expressed as percentage. Analysis was performed according to Hoffman et al.[14]

Haem-iron

The haem-iron concentration was determined according to the method of Hornsey[22] as referred to by Hoffman et al.[14] A 5 g homogenised sample was taken from each replicate and mixed with 1 mL distilled water, 20 mL acetone and 0.5 mL hydrochloric acid. Thereafter the flask was covered with parafilm and left in the dark overnight (±16 h). The sample was then filtered and the absorbance measured at 640 nm in a Cecil CE2021 2000 Series spectrophotometer (Lasec SA (Pty) Ltd, Durban, South Africa). The haem-iron concentration values were expressed in mg haem-iron per g sample.

Lipid oxidation

The oxidation was followed by measuring the thiobarbituric acid reactive substances (TBARS) using the spectrophotometric method described by Rosmini et al.[23] The samples, tested in duplicate, were prepared by homogenising a 1 g sub-sample taken from each replicate with 10 mL 0.15 M potassium chloride buffer for 20 s. The TBARS values were expressed as mg malonaldehyde (MDA) per kg product and mg MDA per g fat. For the full chemical analytical method description, refer to Hoffman et al.[14]

Protein oxidation

Carbonyls and proteins were measured on all samples. The samples, tested in duplicate, were prepared by homogenising a 1 g sub-sample taken from each replicate with 10 mL 0.15 M potassium chloride buffer for 20 s as with lipid oxidation. The carbonyls were determined according to the method outlined by Oliver et al.[24] The protein concentration values were expressed as μM carbonyl per mg protein. For the full chemical analytical method description, refer to Hoffman et al.[14]

Statistical analysis

To test the effects of species, treatment and stage (raw or dried) on the various measurements, mixed model repeated measures analysis of variance (ANOVA) was used. The stage effect was within the subject repeated effect. For the mixed model, species, treatment and stage were treated as fixed effects and the batches to which the treatments were applied as a random effect. Fisher LSD post hoc tests were used to further analyse significant differences when the main effects/interaction were significant. Statistical analyses were done using the VEPAC module of Statistica 11.[25] A 5% significance level was used as guideline for determining significant differences.

Results

Proximate composition analysis

Results of moisture, protein, fat and ash content for each species of *droëwors* (blesbok, springbok and fallow deer) are presented in Table 2. There were no differences ($p>0.05$) amongst treatments but significant differences ($p<0.05$) between raw and dried stages. Therefore, the results indicate the proximate composition as averaged means of each species at a stage.

Fatty acid composition

The total fatty acid profile of each species *droëwors* with the various levels of RBE added are presented in Table 3. The fatty acids composition should be the same for each treatment within a batch at the raw stage as no changes would have occurred; therefore the raw samples were tested using the control (RBE 0%) sample for each batch for each species. After drying, the replicates of each treatment's batches were tested so that differences could be analysed.

For each species, the same fatty acids were found, with the most prominent being oleic acid, stearic acid, linoleic acid and palmitic acid. As expected, the *droëwors* (irrespective of species) had a similar fatty acid profile to that of the sheep fat added.

There were differences ($p<0.05$) in the total polyunsaturated fatty acid (PUFA) profiles between the treatments of each of the species' *droëwors* after the drying process. When analysing the

blesbok species *droëwors* (Table 3), the PUFA values increased with increasing concentration of RBE. Springbok species *droëwors* (Table 3) did not follow a distinctive trend with the addition of the RBE; springbok with RBE 0.25% and RBE 0.50% had increased PUFA values whilst the other treatments both had similar percentages, which are also similar to that of the raw samples. With the fallow deer species *droëwors* (Table 3), the opposite trend to that of the blesbok species *droëwors* occurred; the PUFA values decreased with increasing RBE concentration.

Haem-iron

The mean values for the haem-iron concentration in the *droëwors* samples are depicted in Table 4. The treatments of blesbok and springbok species *droëwors* had similar initial haem-iron concentrations that increased after drying. Furthermore, after drying all treatments had similar haem-iron concentrations over all species types. Meat with RBE 0.25% had lower ($p<0.05$) haem-iron concentrations after drying for all species as noted in Table 4 and Figure 1. There were no significant differences between the species ($p=0.685$), therefore it would be acceptable to analyse the interaction ($p=0.00926$) between stage and treatment regardless of species, as seen in Figure 1. RBE 0.25% therefore had the lowest increase, which corresponds with the results of lipid oxidation given in Table 5. Although it was expected that the haem-iron would differ between raw and dry *droëwors*, when expressed on a dry mass basis, the *droëwors* still had higher haem-iron than the raw *droëwors* (Table 4).

Table 2: Average means*(%)±s.d. for proximate composition analyses of raw and dried *droëwors* with added rooibos extract (RBE) on a wet weight basis

	Moisture		Protein		Fat		Ash	
	Raw	Dried	Raw	Dried	Raw	Dried	Raw	Dried
Blesbok	74.6 ± 0.82	38.7 ± 0.36	12.8 ± 0.24	34.6 ± 0.32	10.7 ± 1.68	20.3 ± 0.18	2.6 ± 0.11	5.3 ± 0.50
Springbok	68.7 ± 0.71	36.1 ± 1.73	13.2 ± 0.22	34.3 ± 0.35	12.4 ± 0.64	20.1 ± 0.33	2.7 ± 0.18	9.8 ± 1.35
Fallow deer	72.1 ± 0.43	35.8 ± 1.70	14.5 ± 0.13	37.2 ± 0.22	10.7 ± 0.43	20.5 ± 0.47	2.2 ± 0.27	5.5 ± 0.62

**No differences ($p>0.05$) were noted amongst RBE treatments within the raw and dried stages within species. Results indicate the proximate composition as averaged means of each species at a stage.*

Table 3: Total fatty acid profile (%)±s.d. of game *droëwors* produced with increasing rooibos extract (RBE) levels

		SFA	MUFA	PUFA	PUFA:SFA	(n-6)/(n-3) PUFA ratio
Blesbok						
Raw	RBE 0%	57.5 ± 1.61	35.9 ± 2.33	6.6 ± 3.94	0.1 ± 0.071	4.5 ± 0.45
Dried	RBE 0%	40.4 ± 13.72	56.7 ± 12.90	2.9^{b} ± 0.84	0.08 ± 0.050	4.0 ± 0.31
	RBE 0.25%	45.5 ± 16.17	50.0 ± 17.80	4.5^{ab} ± 2.88	0.1 ± 0.051	3.6 ± 0.26
	RBE 0.50%	41.8 ± 15.76	53.1 ± 17.65	5.2^{a} ± 1.89	0.1 ± 0.0020	3.7 ± 0.72
	RBE 1.0%	39.4 ± 15.21	54.9 ± 17.55	5.8^{a} ± 2.35	0.1 ± 0.0047	3.7 ± 0.70
Springbok						
Raw	RBE 0%	48.7 ± 12.74	46.4 ± 12.83	4.9 ± 3.06	0.1 ± 0.063	3.6 ± 0.19
Dried	RBE 0%	29.1^{c} ± 5.18	66.0^{a} ± 5.91	4.9^{b} ± 0.77	0.2 ± 0.010	3.8 ± 0.12
	RBE 0.25%	56.8^{b} ± 1.36	32.7^{b} ± 1.22	10.5^{a} ± 0.23	0.2 ± 0.0078	4.6 ± 0.83
	RBE 0.50%	60.3^{a} ± 0.72	29.5^{c} ± 0.54	10.3^{a} ± 0.22	0.2 ± 0.0055	4.9 ± 0.70
	RBE 1.0%	32.6^{c} ± 9.67	61.4^{a} ± 11.27	6.0^{b} ± 1.62	0.2 ± 0.0070	3.8 ± 0.11
Fallow deer						
Raw	RBE 0%	54.7 ± 1.96	34.7 ± 3.23	10.7 ± 1.50	0.2 ± 0.023	5.3 ± 0.40
Dried	RBE 0%	61.9 ± 7.20	29.8 ± 5.51	8.4^{a} ± 2.16	0.1 ± 0.048	4.1 ± 0.11
	RBE 0.25%	54.8 ± 7.73	37.8 ± 10.00	7.4^{a} ± 2.27	0.1 ± 0.025	4.0 ± 0.19
	RBE 0.50%	50.3 ± 15.41	46.6 ± 13.89	3.2^{b} ± 1.53	0.08 ± 0.064	4.1 ± 0.13
	RBE 1.0%	62.9 ± 1.16	34.9 ± 1.24	2.2^{b} ± 0.11	0.03 ± 0.014	4.3 ± 0.30

SFA, saturated fatty acids; MUFA, monounsaturated fatty acids; PUFA, polyunsaturated fatty acids

[a,b,c]Means within a column per species at the dried stage with similar superscripts do not differ significantly ($p>0.05$)

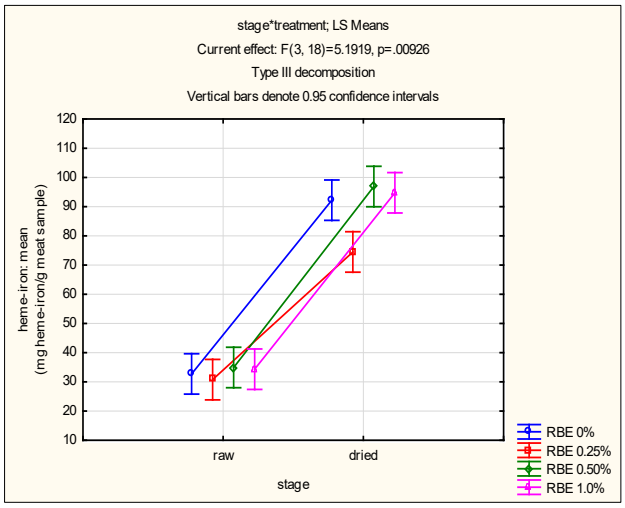


Figure 1: Haem-iron results indicating the interaction between stage and treatment apart from species ($p<0.05$) for *droëwors* prepared with varying concentrations of rooibos extract (RBE).

Lipid oxidation

Lipid oxidation results for blesbok, springbok and fallow deer *droëwors* are depicted in Table 5. The results for each species differed significantly from each other ($p=0.0231$). It is important to note that as *droëwors* is a dried meat product, it is expected to undergo a large level of oxidation and therefore the TBARS values are expected to be higher than those found in raw meat. Therefore, the results are compared across treatments rather than between raw and dried samples.

Blesbok species *droëwors* showed a general trend that after drying, each treatment had increased TBARS values, although the results also indicate that RBE 0.25% resulted in the lowest lipid oxidation and haem-iron concentration (Table 4) in blesbok species *droëwors*. The mean values for springbok species *droëwors* indicate that the TBARS values of the RBE-treated samples (0%, 0.25% and 0.50%) all increased with drying. RBE 1.0% decreases slightly (0.52 MDA per kg meat sample difference). As with the blesbok species *droëwors*, these results indicate that RBE 0.25% gave the best results in terms of minimum lipid oxidation. Fallow deer species *droëwors* mean values differed slightly from blesbok and springbok species *droëwors*, although the MDA levels fall within the same range as the blesbok and springbok *droëwors* of both the raw and dried stages, except that the best result is shown to be at RBE 1.0%.

Table 4: Means±s.d. for haem-iron (mg haem-iron/g meat sample) of *droëwors* with added rooibos extract (RBE).

	Raw				Dried			
	RBE 0 %	**RBE 0.25 %**	**RBE 0.50 %**	**RBE 1.00%**	**RBE 0 %**	**RBE 0.25 %**	**RBE 0.50 %**	**RBE 1.00 %**
As is† basis								
Blesbok	0.30 ± 0.42	0.28 ± 0.75	0.24 ± 0.42	0.36 ± 0.55	0.95^{a} ± 0.21	0.73^{b} ± 0.091	0.99^{a} ± 0.19	0.97^{a} ± 0.036
Springbok	0.29 ± 0.44	0.28 ± 0.11	0.23 ± 1.27	0.39 ± 0.86	0.94^{a} ± 0.098	0.72^{b} ± 0.12	0.98^{a} ± 0.36	0.95^{a} ± 0.24
Fallow deer	0.40 ± 0.51	0.36 ± 0.12	0.57 ± 0.33	0.31 ± 1.61	0.87^{a} ± 0.0026	0.76^{b} ± 0.081	0.93^{a} ± 1.15	0.90^{a} ± 0.20
Dry mass basis								
Blesbok	0.96 ± 0.11	0.88 ± 0.085	0.77 ± 0.097	1.16 ± 0.19	1.45^{a} ± 0.30	1.14^{b} ± 0.15	1.62^{a} ± 0.28	1.51^{a} ± 0.18
Springbok	0.94 ± 0.12	0.87 ± 0.090	0.79 ± 0.10	1.14 ± 0.19	1.48^{a} ± 0.23	1.18^{b} ± 0.12	1.55^{a} ± 0.30	1.52^{a} ± 0.098
Fallowdeer	1.44 ± 0.22	1.28 ± 0.12	2.01 ± 0.33	1.10 ± 0.089	1.33^{a} ± 0.13	1.20^{b} ± 0.13	1.45^{a} ± 0.17	1.41^{a} ± 0.26

†As is basis refers to the sample as would be purchased for consumption

[a,b]Means within a row per species at the dried stage with similar superscripts do not differ significantly ($p>0.05$)

Table 5: Means±s.d. for lipid oxidation (mg MDA/kg meat) of game *droëwors* with added rooibos extract (RBE).

	Raw				Dried			
	RBE 0 %	**RBE 0.25 %**	**RBE 0.50 %**	**RBE 1.00%**	**RBE 0%**	**RBE 0.25 %**	**RBE 0.50 %**	**RBE 1.00 %**
As is† basis								
Blesbok	0.57^{a} ± 0.0025	0.45^{b} ± 0.0036	0.37^{c} ± 0.0064	0.58^{a} ± 0.0015	0.92 ± 0.44	0.75 ± 0.023	0.76 ± 0.31	0.97 ± 0.21
Springbok	0.30^{c} ± 0.021	0.40^{b} ± 0.0056	0.33^{c} ± 0.028	0.50^{a} ± 0.023	0.69 ± 0.29	0.42 ± 0.13	0.40 ± 0.15	0.45 ± 0.12
Fallow deer	0.37^{b} ± 0.042	0.23^{c} ± 0.0055	0.23^{c} ± 0.0074	0.44^{a} ± 0.0034	0.73 ± 0.34	0.59 ± 0.24	0.70 ± 0.23	0.72 ± 0.19
Dry mass basis								
Blesbok	1.86^{a} ± 0.31	1.40^{b} ± 0.072	1.20^{c} ± 0.22	1.83^{a} ± 0.23	1.42 ± 0.67	1.15 ± 0.18	1.26 ± 0.47	1.51 ± 0.36
Springbok	0.99^{c} ± 0.13	1.24^{b} ± 0.13	1.06^{c} ± 0.21	1.59^{a} ± 0.32	1.05 ± 0.39	0.65 ± 0.22	0.66 ± 0.25	0.70 ± 0.22
Fallow deer	1.36^{b} ± 0.30	0.84^{c} ± 0.068	0.84^{c} ± 0.084	1.57^{a} ± 0.078	1.10 ± 0.41	0.94 ± 0.42	1.09 ± 0.43	1.12 ± 0.45

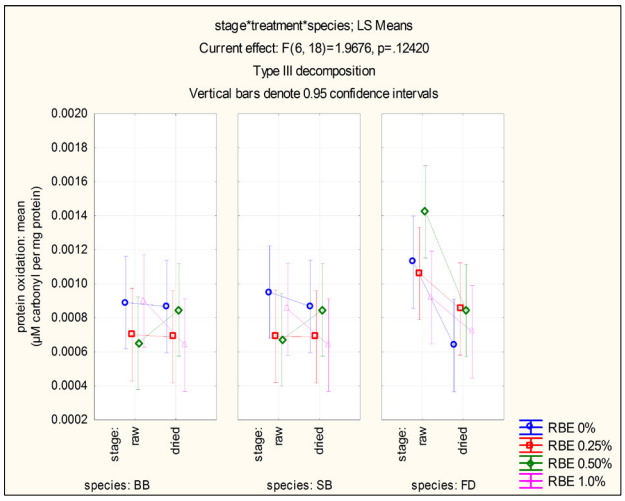
†As is basis refers to the sample as would be purchased for consumption

[a,b,c] Means within a row per species at the raw stage with similar superscripts do not differ significantly ($p>0.05$)

The addition of RBE to *droëwors* did not result in significant differences ($p>0.05$) between treatments after drying within a species.

Protein oxidation

There were significant differences ($p=0.0132$) in protein oxidation in terms of the two stages, raw and dried, which was expected as during drying, protein oxidation will occur (Figure 2). However, each treatment did not always result in an increase in carbonyls. When evaluating the species individually within a stage, the fallow deer species *droëwors* carbonyl levels decreased after drying (Figure 2). Blesbok and springbok species *droëwors* followed the same trends for each treatment, staying at a constant for RBE 0% and 0.25%, increasing for RBE 0.50% and decreasing for RBE 1.0%. As seen in Figure 2, there were visible differences between the species; mainly that of the fallow deer species *droëwors* in comparison with both blesbok and springbok species *droëwors*.



BB, blesbok; SB, springbok; FD, fallow deer.

Figure 2: Protein oxidation results indicating the interaction between stage, treatment and species interaction ($p>0.05$) for droëwors prepared with different concentrations of rooibos extract (RBE).

When evaluating within species, protein oxidation in blesbok species *droëwors* samples remained fairly constant with the addition of RBE 0% and 0.25% with mean values (expressed as μM carbonyl per mg protein) of 0.00089 (raw); 0.00087 (dried) and 0.00070 (raw), 0.00069 (dried) respectively (Figure 2). The protein oxidation in samples with the addition of RBE 0.50% increased over time from 0.00065–0.00085 μM carbonyl/mg protein. The samples with RBE 1.0% decreased in carbonyls by 0.00026 μM carbonyl/mg protein which is a very small (not significant) amount. Springbok species *droëwors* samples indicate similar results to blesbok species *droëwors* samples in that with the addition of RBE 0.25%, the protein oxidation products remained fairly equal over time with mean values (μM carbonyl/mg protein) of 0.00069 (raw) and 0.00069 (dried). Samples with RBE 0.50% again increased after drying by 0.00018 μM carbonyl/mg protein, whilst RBE 0% and RBE 1.0% decreased by 0.000085 μM carbonyl/mg protein and 0.00021 μM carbonyl/mg protein. All treatments of the fallow deer species *droëwors* samples decreased after drying with an initial average mean value of 0.0035 μM carbonyl/mg protein and 0.00076 μM carbonyl/mg protein average mean values after drying. The differences between these values were not significant. There were no significant differences amongst the treatments at each stage (raw and dried) for the individual species.

Discussion

The proximate composition results indicated that the moisture lowered after drying by approximately 50% which was expected as the *droëwors*, as per the procedure used, was dried to lose 45–50% of its original mass. Because of the moisture loss after drying, the protein and fat content was expected to become more concentrated.[26] With the moisture loss experienced during the drying process, as expected, the concentration of fat in the *droëwors* doubled (Table 2). The increase in concentration of the ash content with the drying can be explained by the drying process although it must be kept in mind that the higher than expected ash content would also be because of the added spices.

The fatty acid profile showed that the saturated fatty acids (SFA) initially made up approximately 50–60% of the sausages' total fatty acids. This was caused by the addition of sheep meat/fat to the lean game meat species. Sheep fat is high in stearic acid which contributes to the high total saturated fatty acid of the *droëwors*. The total unsaturated fatty acids made up the remaining 40–50% of the fatty profile. After drying, the fatty acid profile of the *droëwors* was expected to change because of oxidation; however, if the RBE functioned as an antioxidant, it was expected that the fatty acid profile would also differ in terms of percentage fatty acids between treatments within species. However, there were no significant differences between the PUFA values of the blesbok *droëwors* which would seem to indicate that the RBE did not function as an antioxidant. The fatty acid profile, especially that of the low PUFA, at treatments RBE 0% and RBE 1.0% of the dried springbok species *droëwors* (Table 3), would seem to indicate that more oxidation may have occurred at these concentrations. Fallow deer *droëwors*' PUFA decreased with increasing concentrations of RBE, which would indicate that there was increased oxidation with increasing RBE levels.

The haem-iron levels, lipid oxidation and protein oxidation need to be discussed in conjunction with each other for an accurate conclusion to be made as haem-iron and oxidation are closely related to each other.[27] Haem-iron is a catalyst for lipid oxidation, therefore the higher the haem-iron content, the more susceptible the product is to oxidation.[27] Protein oxidation products are also known to induce lipid oxidation as the protein oxidation products react with free lipid radicals resulting in increased lipid oxidation.[28] Lipid oxidation however may also initiate protein oxidation.[29] A standard Pearson's correlation indicated that the haem-iron was positively linked to the TBARS of the dried *droëwors* in all the species (blesbok 0.52, springbok 0.24, fallow deer 0.86).

During drying, protein oxidation occurs resulting in the formation of carbonyls.[29] Therefore it is expected that the oxidation products will increase after drying. Studies have shown that the 2,4-dinitrophenylhydrazine (DNPH) method is used to obtain a general overall measurement of protein oxidation in food systems.[30,31] This method is described as being accurate, especially for quantification of overall protein oxidation levels over time.[32] Decreases which are visible in Figure 2 may result from the protein carbonyls reacting with other cellular constituents. These cellular reactions result in the total protein carbonyls not being detected.[33] This corresponds with the fact that protein oxidation products are known to induce lipid oxidation[28] as the products will react with the free lipid radicals resulting in increased lipid peroxidation. Of importance are the changes in protein oxidation products over time per species. Because of the antioxidant effect of RBE, protein oxidation was expected to decrease with higher levels of RBE added, however, the level of μM carbonyl per mg protein, as seen with the blesbok and springbok species *droëwors*, remained fairly constant between treatments. Protein oxidation was thus not affected by the addition of RBE at any of the concentrations although the results (Figure 2) suggest that with the addition of RBE at concentrations higher than 1%, a large concentration of malonaldehydes will be detected (high oxidation level) with levels being similar to that of the control (RBE 0%). Natural polyphenols, such as those found in RBE, form strong complexes with metal ions such as iron. Polyphenols also have the ability to complex with proteins. This therefore will influence the antioxidant activity of the RBE as once the polyphenol binds to a molecule, it will no longer act primarily as an antioxidant. This is seen in Figure 2 as the carbonyls decrease with RBE 1.0% allowing the polyphenols in the RBE to bind to the remaining proteins available which in turn could inactivate the antioxidant activity. This supports the lipid oxidation results of RBE concentrations 0.50% and 1.0% which show an increase after drying. Some studies have proved that conventional antioxidant strategies do not necessarily apply to muscle proteins, as compounds that are able to prevent lipid oxidation are not always able to prevent protein oxidation.[28,29,34,35]

Oxidation in meat and meat products is also closely related to the break-down of haem-iron and the release of iron from the porphyrin ring.[36]

Haem-iron increases after drying which could be from moisture loss of the product. However, the possibility exists that the haem-iron is not released from the porphyrin ring because of the drying process. The principle that is followed involves the interaction between lipid and protein oxidation which subsequently results in lowered haem-iron values. Lipid oxidation occurs through a free radical chain reaction whereby oxygen is the most important factor in the development of lipid oxidation in meat and meat products.[37] *Droëwors* is a processed dried meat product and the process includes a mincing step. The mincing of the meat disrupts the cell membranes releasing pro-oxidants naturally found in meat which interact with the PUFA in the presence of oxygen and reactive oxygen species. This initiates lipid oxidation.[38] Lipid oxidation products include superoxide anions and hydrogen atoms which react to form hydrogen peroxide (H_2O_2). H_2O_2 acts as an oxidising agent by releasing the haem-iron molecule (via oxidative cleavage of the porphyrin ring) resulting in further lipid oxidation.[12,37,39,40] At this point however, if the antioxidants have bound to the binding sites of the myoglobin molecule in which the haem-iron is attached, this will be prevented. This process, called the Fenton reaction, is also involved in the reduction of PUFA.

At slaughter, reducing enzymes in meat are depleted, resulting in the accumulation of metmyoglobin. Metmyoglobin consists of haem-iron in its ferric state (Fe^{3+}), which has two binding sites.[37] RBE is made up of polyphenols,[41] therefore it can be speculated that they will attach to these binding sites of metmyoglobin, which primarily bind oxygen. With the RBE binding to these sites, a decrease in myoglobin oxidation occurs because oxygen cannot bind to the same sites. As the protein oxidation is a metal-ion catalysed reaction, there is a decrease in protein oxidation thereby decreasing lipid oxidation via protein oxidation products.

Aspalathin is unique to rooibos tea.[9,41] It acts as a potent antioxidant at low concentrations and a pro-oxidant at high concentrations. As an antioxidant, it has a radical scavenging ability which protects against oxidative damage. Plant antioxidants added to meat and meat products have beneficial effects for protection against excessive oxidative damage induced by reactive oxygen species.[42] High concentrations of total flavonoids, dihydrochalcones (aspalathin and nothofagin) and/or total polyphenols are a few reasons why RBE could also act as a pro-oxidant.[41] Assays have been demonstrated using RBE to inhibit lipid peroxidation which concludes that its efficacy is dependent on the total polyphenol content rather than the individual flavonoids.[10,41] However, the RBE used in this study had high concentrations of aspalathin and quercetin (Table 1) which can both have pro-oxidant properties when added in high concentrations, which could explain the increased lipid oxidation values with the addition of RBE 0.50% and 1.0%. Orientin and iso-orientin (analogues of aspalathin) could also play a role in the pro-oxidant effect of the RBE as these flavonoids are also found in high concentrations[41] (Table 1).

The low results for haem-iron and lipid oxidation at RBE 0.25% are because of the inhibition of haem-iron release via the binding of the antioxidant to its binding sites, which inhibits protein oxidation and in turn minimises the level of lipid oxidation. Lipid oxidation does occur to some extent because of the initiation of lipid oxidation during processing. Only with the addition of RBE 0.25% did this occur. The same trend was seen over all species. These results also relate to the protein oxidation results which show a minimal formation of carbonyls. With the addition of higher concentrations of RBE, this resulted in higher oxidation levels. This could be the result of the saturation of the binding sites of haem-iron. Once these sites have been saturated, the RBE begins to act as a pro-oxidant as it has a high total polyphenol content and plays a role in the generation of hydroxide (OH^-), which is a reactive oxygen species. Therefore, the more RBE present, the higher the level of oxidation that will occur. It can therefore be proposed that the addition of RBE 0.25% would be the threshold for usage as an antioxidant in *droëwors*. The control (RBE 0%) proved to have similar oxidation results as those of the RBE at high concentrations.

Droëwors is considered a high in fat product and has a high level of PUFA (Table 3); the higher the PUFA, the faster the oxidation process.[38] It must be noted that springbok species *droëwors* showed that RBE 1.0% decreases slightly (0.52 MDA per kg meat sample difference). This decrease could possibly be explained by experimental error as determination of MDA in a meat sample using the TBARS method is said to be inaccurate as it lacks specificity and sensitivity.[43,44] Therefore it has been suggested that the TBARS be used to assess the extent of lipid oxidation in general.[45] This method, which is specifically used for biological samples, muscle foods and fish oils,[46] should be regarded as a way to determine approximate lipid oxidation results. Variation in results has been seen in various studies which assess lipid oxidation in muscles food using the TBARS method.[47-49]

Both blesbok and springbok species *droëwors* showed that the addition of RBE 0.25% tended to slow down lipid oxidation. The addition of higher concentrations of RBE (0.50% and 1.0%) did not give expected results as oxidation activity was higher than the control (RBE 0%). Fallow deer species *droëwors* resulted in high oxidation activity with the addition of RBE at 0.25% but this does not correspond with the other species *droëwors* as the addition of RBE 1.0% gave the best results in terms of inhibition of oxidation activity after drying. This could be because of the chemical composition of the fallow deer differing from that of the other two species; an aspect that warrants further research.

Conclusions

Although the results obtained from this study did not prove the inhibition of lipid oxidation to be significant, there were some interesting trends observed. Lipid and protein oxidation showed that the addition of RBE at 0.25% gave the best results. Haem-iron concentrations after drying showed significant differences with RBE 0.25% resulting in a low haem-iron content in comparison with the higher RBE concentrations. Though the differences for lipid oxidation weren't significant, it is important to note that RBE 0.50% and 1.0% gave higher oxidation results but that these MDA values were never higher than the control MDA values. As some results were inconsistent across species, this indicates that species type (especially with the fallow deer *droëwors*) and formulation of the *droëwors* play an important role in the antioxidant activity of the added RBE. The fatty acid profile had high concentrations of oleic acid, stearic acid, linoleic acid and palmitic acid in both the raw and dry *droëwors* samples which is consistent with the sheep fat added. The high polyunsaturated fatty acids resulted in a higher oxidative activity as noted. The formulation used contains a high fat content of 33.3%. By reducing the fat content and re-formulating the *droëwors* recipe, this should allow for an improved fatty acid profile, as the *droëwors* tends to be more similar to the added fat. A sensory panel analyses should also be conducted to determine whether the oxidation is detectable and to give an indication on whether RBE will impart a specific flavour to the *droëwors*.

Acknowledgements

This work is based on research supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation (NRF) of South Africa. Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard. We thank the students and technical staff of the Department of Animal Sciences at Stellenbosch University for their assistance and support. Afriplex (Pty) Ltd, Paarl supplied the RBTE.

Authors' contributions

L.C.H. was the project leader and contributed to the manuscript; M.J performed the experimental work (both chemical and sensory work) and wrote the manuscript; and M.M. contributed to the sensory experimental design and manuscript.

References

1. Hoffman LC, Wiklund E. Game and venison – meat for the modern consumer. Meat Sci. 2006;74(1):197–208. http://dx.doi.org/10.1016/j.meatsci.2006.04.005

2. Burnham GM, Hanson DJ, Koshick CM, Ingham SC. Death of *Salmonella* serovars, *Escherichia coli* O157:H7, *Staphylococcus aureus* and *Listeria monocytogenes* during the drying of meat: A case study using biltong and *droëwors*. J Food Safety. 2008;28:198–209. http://dx.doi.org/10.1111/j.1745-4565.2008.00114.x

3. Carr MA, Miller MF, Daniel DR, Yarbrough CE, Petrosky JD, Thompson LD. Evaluation of the physical, chemical and sensory properties of jerky processed from emu, beef and turkey. J Food Quality. 1997;20:419–425. http://dx.doi.org/10.1111/j.1745-4557.1997.tb00484.x

4. Hoffman LC, Cawthorn D-M. What is the role and contribution of meat from wildlife in providing high quality protein for consumption? Anim Front. 2012;2(4):40–54. http://dx.doi.org/10.2527/af.2012-0061

5. Ladikos D, Lougovois V. Lipid oxidation in muscle foods: A review. Food Chem. 1989;35:295–314. http://dx.doi.org/10.1016/0308-8146(90)90019-Z

6. Shacter E. Quantification and significance of protein oxidation in biological samples. Drug Metab Rev. 2000;32(3/4):307–326. http://dx.doi.org/10.1081/DMR-100102336

7. Pokornỳ J. Preparation of natural antioxidants. In: Pokornỳ J, Yanishlieva N, Gordon M, editors. Antioxidants in food: Practical applications. Cambridge: Woodhead Publishing Ltd; 2001.

8. Sánchez-Escalante A, Djenane D, Torrescano G, Beltrán JA, Roncales P. Antioxidant action of borage, rosemary, oregano and ascorbic acid in beef patties packaged in modified atmosphere. J Food Sci. 2003;68(1):339–344. http://dx.doi.org/10.1111/j.1365-2621.2003.tb14162.x

9. Joubert E, De Beer D. Rooibos (*Aspalathus linearis)* beyond the farm gate: From herbal tea to potential phytopharmaceutical. S Afr J Bot. 2011;77:869–886. http://dx.doi.org/10.1016/j.sajb.2011.07.004

10. Joubert E, De Beer D. Phenolic content and antioxidant activity of rooibos food ingredient extracts. J Food Compos Anal. 2012;27:45–51. http://dx.doi.org/10.1016/j.jfca.2012.03.011

11. Shahidi F, Janitha PK, Wanasundra PD. Phenolic antioxidants. Crit Rev Food Sci Nutr. 1992;32(1):67–103. http://dx.doi.org/10.1080/10408399209527581

12. Liu F, Dai R, Zhu J, Li X. Optimizing colour and lipid stability of beef patties with a mixture design incorporating with tea catechins, carnosine, and a-tocopherol. J Food Eng. 2010;98:170–177. http://dx.doi.org/10.1016/j.jfoodeng.2009.12.023

13. Cullere M, Hoffman LC, Dalle Zotte A. First evaluation of unfermented and fermented rooibos (*Aspalathus linearis*) in preventing lipid oxidation in meat products. Meat Sci. 2013;95:72–77. http://dx.doi.org/10.1016/j.meatsci.2013.04.018

14. Hoffman LC, Jones M, Muller N, Joubert E, Sadie A. Lipid and protein stability and sensory evaluation of ostrich (*Struthio camelus*) *droëwors* with the addition of rooibos tea extract (*Aspalathus linearis)* as a natural antioxidant. Meat Sci. 2014;96:1289–1296. http://dx.doi.org/10.1016/j.meatsci.2013.10.036

15. Joubert E, Viljoen M, De Beer D, Manley M. Effect of heat on aspalathin, isoorientin, and orientin contents and color of fermented rooibos (*Aspalathus linearis*) iced tea. J Agric Food Chem. 2009;57:4204–4211. http://dx.doi.org/10.1021/jf9005033

16. Beelders T, Sigge GO, Joubert E, De Beer D, De Villiers A. Kinetic optimisation of the reversed phase liquid chromatographic separation of rooibos tea (*Aspalathus linearis*) phenolics on conventional high performance liquid chromatographic instrumentation. J Chromatogr A. 2012;1219:128–139. http://dx.doi.org/10.1016/j.chroma.2011.11.012

17. Association of Official Analytical Chemists (AOAC International). AOAC official method 934.01: Loss on drying (moisture) at 95–100°C for feeds. In: Official methods of analysis. 17th ed. Arlington, VA: Association of Official Analytical Chemists Inc; 2002.

18. Association of Official Analytical Chemists (AOAC International). AOAC official method 942.05: Ash of animal feed. In: Official methods of analysis. 17th ed. Arlington, VA: Association of Official Analytical Chemists Inc; 2002.

19. Association of Official Analytical Chemists (AOAC International). AOAC official method 992.15: Crude protein in meat and meat products including pet foods. Combustion method. J AOAC Int. 1992;76:787.

20. Lee CM, Trevino B, Chaiyawat M. A simple and rapid solvent extraction method for determining total lipids in fish tissue. J AOAC Int. 1996;79(2):487–492.

21. Folch J, Lees M, Sloane-Stanley GH. A simple method for the isolation and purification of total lipids from animal tissues. J Biol Chem. 1957;226:495–509.

22. Hornsey HC. The colour of cooked cured pork I – Estimation of the nitric oxide-haem pigments. J Sci Food Agric. 1956;7(8):534–540.

23. Rosmini MR, Perlo F, Pérez-Alvarez JA, Pagán-Moreno MJ, Gago-Gago A, López-Santoveña F, et al. TBA test by an extractive method applied to 'paté'. Meat Sci. 1996;42(1):103–110. http://dx.doi.org/10.1016/0309-1740(95)00010-0

24. Oliver CN, Ahn BW, Moerman EJ, Goldstein S, Stadtman ER. Aged-related changes in oxidised proteins. J Biol Chem. 1987;262:5488–5491.

25. StatSoft South Africa Analytics. STATISTICA Version 11 [homepage on the Internet]. c2012 [cited 2013 Jun 10]. Available from https://statisticasoftware.wordpress.com/

26. Nuñez de Gonzalez MT, Hafley BS, Boleman RM, Miller RM, Rhee KS, Keeton JT. Qualitative effects of fresh and dried plum ingredients on vacuum-packaged, sliced hams. Meat Sci. 2009;83:74–81. http://dx.doi.org/10.1016/j.meatsci.2009.04.002

27. Fernandez-Espla MD, O'Neill E. Lipid oxidation in rabbit meat under different storage conditions. J Food Sci. 1993;58:1262–1264. http://dx.doi.org/10.1111/j.1365-2621.1993.tb06161.x

28. Viljanen K, Kivikari R, Heinonen M. Protein–lipid interactions during liposome oxidation with added anthocyanin and other phenolic compounds. J Agric Food Chem. 2004;52:1104–1111. http://dx.doi.org/10.1021/jf034785e

29. Lund MN, Heinonen M, Baron CP, Estévez M. Protein oxidation in muscle foods: A review. Mol Nutr Food Res. 2011;55:83–95. http://dx.doi.org/10.1002/mnfr.201000453

30. Estévez M, Ollilainen V, Heinonen M. Analysis of protein oxidation markers α-aminoadipic and γ-glutamic semialdehydes in food proteins by using LC-ESI multi-stage tandem MS. J Agric Food Chem. 2009;57(9):3901–3910. http://dx.doi.org/10.1021/jf804017p

31. Estévez M. Protein carbonyls in meat systems: A review. Meat Sci. 2011;89(3):259–279. http://dx.doi.org/10.1016/j.meatsci.2011.04.025

32. Estévez M, Kylli P, Puolanne E, Kivikari R, Heinonen M. Fluorescence spectroscopy as a novel approach for the assessment of myofibrillar protein oxidation in oil-in-water emulsions. Meat Sci. 2008;80(4):1290–1296. http://dx.doi.org/10.1016/j.meatsci.2008.06.004

33. Baron CP, Klærsgård IVH, Jessen F, Jacobsen C. Protein and lipid oxidation during frozen storage of rainbow trout (*Oncorhynchus mykiss*). J Agric Food Chem. 2007;55(20):8118–8125. http://dx.doi.org/10.1021/jf070686f

34. Mercier Y, Gatellier P, Viau M, Remignon H, Renerre M. Effect of dietary fat and vitamin E on colour stability and on lipid and protein oxidation in turkey meat during storage. Meat Sci. 1998;48:301–318. http://dx.doi.org/10.1016/S0309-1740(97)00113-7

35. Estévez M, Cava R. Lipid and protein oxidation, release of iron from haem molecule and colour deterioration during refrigerated storage of liver paté. Meat Sci. 2004;68:551–558. http://dx.doi.org/10.1016/j.meatsci.2004.05.007

36. Estéves M, Ventanas S, Cava R. Protein oxidation in frankfurters with increasing levels of added rosemary essential oil: Effect on color and texture deterioration. J Food Sci. 2005;70(7):427–432. http://dx.doi.org/10.1111/j.1365-2621.2005.tb11464.x

37. Min B, Ahn DU. Mechanism of lipid peroxidation in meat and meat products – A review. Food Sci Biotechnol. 2005;14(1):152–163.

38. Vuorela S, Salminen H, Mäkelä M, Kivikari R, Karonen M, Heinonen M. Effect of plant phenolics on protein and lipid oxidation in cooked pork meat patties. J Agric Food Chem. 2005;53:8492–8497. http://dx.doi.org/10.1021/jf050995a

39. Renerre M, Labas R. Biochemical factors influencing metmyoglobin formation in beef muscles. Meat Sci. 1987;19:151–165. http://dx.doi.org/10.1016/0309-1740(87)90020-9

40. O'Grady MN, Monahan FJ, Brunton NP. Oxymyoglobin oxidation and lipid oxidation in bovine muscle-mechanistic studies. J Food Sci. 2001;66:386–392. http://dx.doi.org/10.1111/j.1365-2621.2001.tb16115.x

41. Joubert E, Winterton P, Britz TJ, Gelderblom WCA. Antioxidant and prooxidant activities of aqueous extracts and crude polyphenolic fractions of rooibos (*Aspalathus linearis*). J Agric Food Chem. 2005;53:10260–10267. http://dx.doi.org/10.1021/jf051355a

42. Middleton E, Kandaswami C, Theoharides T. The effects of plant flavonoids on mammalian cells: Implications for inflammation, heart disease, and cancer. Pharmacol Rev. 2000;52:673–751.

43. Raharjo S, Sofos JN. Methodology for measuring malonaldehyde as a product of lipid peroxidation in muscle tissues: A review. Meat Sci. 1993;35(2):145–169. http://dx.doi.org/10.1016/0309-1740(93)90046-K

44. Shahidi F, Zhong Y. Lipid oxidation: Measurement methods. In: Shahidi F, editor. Bailey's industrial oil and fat products. 6th ed. New York: Wiley-Interscience; 2005. p. 357–385. http://dx.doi.org/10.1002/047167849X.bio050

45. Gray JI, Monahan FJ. Measurement of lipid oxidation in meat and meat products. Trends Food Sci Technol. 1992;3:315–319. http://dx.doi.org/10.1016/S0924-2244(10)80019-6

46. Dobarganes MC, Velasco J. Analysis of lipid hydroperoxides. Eur J Lipid Sci Technol. 2002;104(7):420–428. http://dx.doi.org/10.1002/1438-9312(200207)104:7<420::AID-EJLT420>3.0.CO;2-N

47. Rhee KS, Anderson LM, Sams AR. Lipid oxidation potential in beef, chicken and pork. J Food Sci. 1996;61(1):8–12. http://dx.doi.org/10.1111/j.1365-2621.1996.tb14714.x

48. Soyer A, Özalp B, Dalmiş Ü, Bilgin V. Effects of freezing temperature and duration of frozen storage on lipid and protein oxidation in chicken meat. Food Chem. 2010;120:1025–1030. http://dx.doi.org/10.1016/j.foodchem.2009.11.042

49. Sampaio GR, Saldanha T, Soares RAM, Torres EAFS. Effect of natural antioxidant combinations on lipid oxidation in cooked chicken meat during refrigerated storage. Food Chem. 2012;135:1383–1390. http://dx.doi.org/10.1016/j.foodchem.2012.05.103

Putting fossils on the map: Applying a geographical information system to heritage resources

AUTHORS:
Merrill van der Walt[1]
Antony K. Cooper[2]
Inge Netterberg[3]
Bruce S. Rubidge[1]

AFFILIATIONS:
[1]Evolutionary Studies Institute, School of Geosciences, University of the Witwatersrand, Johannesburg, South Africa
[2]Built Environment Unit, Council for Scientific and Industrial Research, Pretoria, South Africa
[3]Netterberg Consulting, Leuven, Belgium

CORRESPONDENCE TO:
Merrill van der Walt

EMAIL:
Merrill.VanDerWalt@wits.ac.za

POSTAL ADDRESS:
[1]Evolutionary Studies Institute, Private Bag 3, University of Witwatersrand 2050, Johannesburg, South Africa

DATES:
Received: 23 Oct. 2014
Revised: 26 Mar. 2015
Accepted: 20 Apr. 2015

KEYWORDS:
Spatial data; Permo-triassic; database; Karoo Supergroup

HOW TO CITE:
Van der Walt M, Cooper AK, Netterberg I, Rubidge BS. Putting fossils on the map: Applying a geographical information system to heritage resources. S Afr J Sci. 2015;111(11/12), Art. #2014-0371, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140371

A geographical information system (GIS) database was compiled of Permo-Triassic tetrapod fossils from the Karoo Supergoup in South African museum collections. This database is the first of its kind and has great time applicability for understanding tetrapod biodiversity change though time more than 200 million years ago. Because the museum catalogues all differed in recorded information and were not compliant with field capture requirements, this information had to be standardised to a format that could be utilised for archival and research application. Our paper focuses on the processes involved in building the GIS project, capturing metadata on fossil collections and formulating future best practices. The result is a multi-layered GIS database of the tetrapod fossil record of the Beaufort Group of South Africa for use as an accurate research tool in palaeo- and geoscience research with applications for ecology, ecosystems, stratigraphy and basin development.

Introduction

The fossil record of the Karoo Supergroup, which comprises a largely unbroken temporal record of tetrapod evolution from the Middle Permian to the Middle Jurassic,[1,2] provides a unique opportunity to set up a GIS database of fossil occurrences which can be utilised to answer questions relating to ecological and biodiversity change through time. The Karoo fossil record is the best preserved ecological assemblage of pre-mammalian terrestrial tetrapods documenting the stem lineages of both mammals and dinosaurs.[3,4]

We geocoded palaeontological data for use in a geographical information system (GIS) for palaeoscience research to explore issues relating to the biodiversity of Permian and Triassic tetrapod faunas. This was the first time a GIS had been applied to the fossil records of the Karoo Supergroup. With the cooperation of seven South African museums and institutes (Council for Geoscience, Pretoria; Ditsong Museum, Pretoria; Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg; National Museum, Bloemfontein; Albany Museum, Grahamstown; Rubidge Collection, Wellwood, Graaff-Reinet; Iziko South African Museum, Cape Town) that curate collections of Karoo tetrapod fossils, a GIS incorporating the South African databases of fossil records collected from the Beaufort Group, Karoo Supergroup has been compiled.

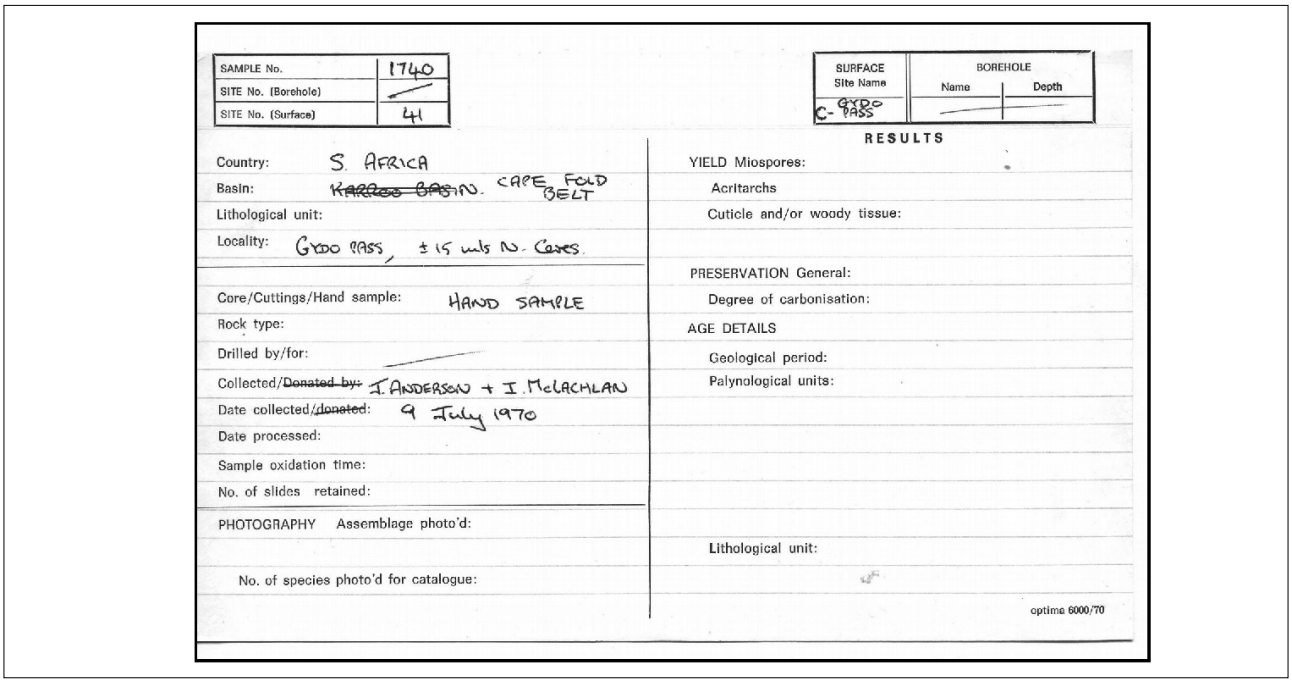
The hundreds of thousands of fossil artefacts stored and accessioned in museum collections are the foundation of our knowledge on past biodiversity. Great strides have been made in biodiversity informatics in providing digital access to extinct biodiversity data, for integration, interpretation, reconstruction and application objectives. Models for community data access are evident in abundant projects, such as:

- The Revealing Human Origins Initiative (RHOI)[5] Specimen Database, a collaboration of paleoanthropological and related projects studying Late Miocene (and Pliocene) hominins and other faunas in context, with the database including digital imagery and metadata that covers age, geology, collection elements and taxonomy;
- The digital@rchive of Fossil Hominoids[6], for which the primary mandate is to facilitate morphological investigations in the field of human evolution by providing digital data for the international scientific community;
- The Darwin Core metadata standard[7], a uniform standard designed to expedite the exchange of information about the geographic occurrence of species and specimen records in collections, with extensions for palaeontology.

These information systems, driven by distributed data retrieval, data capture and person-facilitated geospatial referencing, have enabled the investigation of novel research questions around ecological reconstruction, extinct biodiversity trends and predictive modelling.

Historically, details of fossils collected were recorded as hand-written descriptions on index cards or in catalogues (Figure 1). Such documentation included both data (e.g. species and location) and metadata (information about the record), such as who collected, prepared and/or identified the fossil, where the fossil is stored and who wrote up the index card.

There are a variety of standards available for metadata, such as the Dublin Core (ISO 15836:2009),[8] developed primarily for describing resources for discovery, and ISO 19115:2003,[9] for describing geographical data, of which the South African profile (subset) is SANS 1878-1:2005.[10] Dublin Core is primarily text-based, making it easy to enter information for its 15 metadata elements, while ISO 19115 makes extensive use of encoding, which facilitates automated processing and presenting the metadata in multiple languages. Metadata can be converted from one standard to another using an ontology or a cross-walk.[11] As ISO 19115 has encoded metadata and more detailed metadata elements, it is easy to convert its metadata to Dublin Core through a cross-walk (conversion table), but the reverse is difficult because of the need to subdivide metadata elements, text processing and, invariably, use human expertise. Hence, it would be better to use a metadata standard such as ISO 19115 for palaeontological records.



Source: *Evolutionary Studies Institute (ESI) Collections. Courtesy of Dr. Bernhard Zipfel.*

Figure 1: *Collected fossil recorded as a hand-written description on an index card.*

South African fossil-find field notes for the Beaufort Group (to be eventually reconfigured into museum index cards) were written up over the space of 150 years (since 1845[12,13]) and do not conform to any particular standard. The main disadvantage is that some records contain inadequate or ambiguous data, particularly relating to the precise location of the fossil provenance.

This paper focuses on the processes involved in establishing a GIS for tetrapod fossils from the Beaufort Group. It highlights the key challenges encountered during database establishment, as well as describing its main applications and future best practices for use as an accurate research tool in palaeontological research. This unique database is curated at the Evolutionary Studies Institute (ESI) at the University of the Witwatersrand and is available as a research tool to all bona-fide scientists

Creating a reliable product

Extensive fossil collections have been amassed from the rocks of the time-extensive Permo-Triassic Beaufort Group and curated in different museum collections in South Africa, providing a unique opportunity to incorporate these collections onto a GIS. Ultimately, this database will be expanded to include fossils from the Beaufort Group which are housed in overseas institutions such as the Natural History Museum, London; Smithsonian Institution, Washington DC and the Field Museum, Chicago. This GIS will enhance their utility in research relating to changing biodiversity patterns, both temporally and geographically, as well as stratigraphic and basin development modelling.

Problems that had to be overcome in setting up the GIS database related largely to a lack of consistency in the data, ambiguous locality data and outdated taxonomic records, requiring rigorous standardisation and updating.

While all the original data was provided in digital format, these were set up from manual records. This is the main drawback encountered when having to apply human interpretation verses the structured logic of the computer. The establishment of the GIS database highlighted the value of structuring data to suit GIS and other digital applications. The migration of paper records to useful electronic records could not simply be carried out verbatim as many of the data obtained from the contributing South African museums needed to be restructured to facilitate analysis through electronic means.

Table 1: Summary of processes for presenting data within a spatial system

Task order	Processes		
	Stage 1: Acquisition and processing of original data	**Stage 2: Establishing a GIS management system**	**Stage 3: Reconciliation**
1	Review and assimilation of original databases	Selection of map layers	Challenges encountered during mapping process
2	Creation of applicable data fields	Conversion of data	Future possible modes of action concerning standardising and fossil collection methods
3	Criteria established for the elimination of non-viable data with subsequent deletion	Establishing an alternative base data	Suggestions for improving the current success rate for digitising fossil localities
4	Division into automated and manual entry databases	Establishing a geodatabase	
5	Creation of map sheet shape file		

Table 2: Summary of specific processes for presenting digitised Beaufort Group data

Task order	Phases	
	Phase 1: Accessing and processing of data	**Phase 2: Development of spatial model**
1	Initial test case using the recorded fauna of the Evolutionary Studies Institute (ESI) database	Converting alphanumeric data to spatial data
2	Selection of datafields applicable for project objectives	Acquistion and application of selected map layers
3	Preliminary data cleaning	
4	Establishment of criteria list allowing for automatic validation	
5	Selection of data allowing for automatic insertion into GIS	
6	Establishment of manual insertion data	
7	Establishment of genera and species lists	
8	Acceptance or re-classification of taxonomic classification	
9	Establishment of final genera/species list for ESI database	
10	Assignment of fossil locality data to each specimen	
11	Repetition of tasks 2 – 10 for remaining six databases (ESI as template)	

Methodology

Mapping palaeontological specimens

As this database has been set up as a research tool to be used by palaeontologists, it is important to explain the methodology in detail so that users can fully understand why the GIS was created in this particular way.

The broad-spectrum processes were divided into three stages (Table 1): Stage 1: Acquisition and processing of original data; Stage 2: Establishing a GIS management system; Stage 3: Reconciliation.

More detailed processes involved in spatially mapping the fossils were subdivided into two phases (Table 2), Phase 1: accessing and processing of data and Phase 2: development of a spatial model.

Alphanumeric data was converted to spatial data because, for most of the records, the location was specified using geographical identifiers,[14] particularly farm names, rather than coordinates. Converting the data required rigorous 'cleaning', correction of spelling errors and standardisation of content to permit queries. Farm names with their corresponding farm numbers were aligned with the names registered with the Registrar of Deeds and the Surveyors General.

Once cleaning of data had been accomplished, selection of data fields applicable for spatial mapping was undertaken. Geospatial coordinates used for mapping species location and distribution are crucial for a reliable spatial system.[15] Access to geospatially referenced data from fossils provides a quantitative basis for biodiversity analyses over time and predictive niche modelling for determining sampling densities of various sites.

Providing locality coordinates proved a significant challenge. Most of the recorded specimens were associated with a georeference, but this reference was, in most instances, a worded description of the localities from where they were discovered with few records having geographic coordinates (Table 3).

To get to the point where data could be represented on a spatial map, two approaches were adopted. The first involved selecting records that qualified for automatic import into the system. The second approach involved records that could only be entered onto the system manually.

Table 3: Challenges posed by textual geo-referencing

Examples of textual geo-reference	Challenge posed
'Matjiesfontein' 'Bronkhorstspruit'	Same locality name for more than one record
'Ndanyana Hill on Hartebeespruit'	Topological nesting
'Tiparery (?Tipperary) Ndanayane? (Ndanyana) Hill'	Complex grammar
'0.25 miles NW of Mnweni River' 'Hantam Mts, Toren, 12 miles W of Calvinia'	Linear feature measurement
'On the road between Masite and Kolo'	Linear ambiguity
'Various sites in Bamboeshoek Valley' 'Possibly from Kiesbeen behind Platberg' 'Between the road and the Sundays River'	Vague localities
'Transvaal, Wiepe'	Change of political borders over time
''Fossil Bend' near Harrismith' 'Newcastle Admiralty Estate Group'	Historical and 'in-house' placename

Automated data entry procedure

Records with locality coordinates from a Global Positioning System (GPS) could be entered automatically. However, as the majority of records had only a farm name for the locality, a spatial database had to be created to allow records to be imported automatically to specific localities referenced as farm centroids. A farm centroid is the calculated gravitational centre of a polygon (farm boundaries are polygons). This centroid is calculated using the ArcMap field calculator which automatically sets a field value for a single record, or even all records.

Forcing such localities into a single point at the gravitational centre of the farm introduces error and inaccuracy into the data, but remains the best option to utilise locality data for the majority of fossils found prior to GPS usage. Current locality data were accurately captured by GPS.

Table 4: Methodology for automating spatial data

Phase		Task(s)			
1	**Obtain necessary data**	1. Landsat imagery	2. Geology (1:250 000)	3. Farms and farm boundaries	4. Fossil records
2	**Transform spatial data**	1. Data exploration	2. Check projections of spatial data for accuracy and relevance		
3	**Sort alphanumeric data**	1. Sort records per farm and per biozone			
4	**Digitise fossil record localities**	1. Digitise localities per farm using ArcGIS®	2. Align ID keys with alphanumeric database		
5	**Link alphanumeric database**	1. Link alphanumeric database to spatial database (compound key)			
6	**Output**	1. Spatial database of tetrapod fossil localities for the Beaufort Group of South Africa			

To allow data to be imported automatically, certain tasks had to be completed (Table 4). A geodatabase was created to house the spatial data of the farm, administrative, district and magisterial boundaries and local authorities' databases.[16] Various map layers (including Landsat 7 ETM+ Satellite imagery) were necessary as backdrop data to interpret the distribution patterns of fossil taxa.

Because most of the specimens in older collections lacked geographic coordinates for their place of discovery, the most accurate locality information in the majority of the databases was simply a farm and district name. To represent this locality information on the GIS, farm locality data was received in .FEA format from the Surveyor General and converted into shape file format. Alphanumeric data were exported as a point file and joined to the polygon data using a spatial join. The cadastre received from the Surveyor General contained farm boundaries and their farm numbers, but very few farm names. This lack of names posed a problem as localities for most of the specimens in the museum catalogues were given as locality names, which were assumed to correspond to the farm names. As such, the farm names were essential for the geocoding of the localities and thus the specimens.

To solve the number versus name problem, Environmental Potential Atlas (ENPAT 2004[17]) farm cadastre data was used as the new spatial layer to identify localities. For each farm, centroids were generated and used to geocode the specimens by linking the specimen locality names to farm names. Additional backdrop map layers included Surveyor General data for magisterial districts and provinces. These data were used to identify further localities, as farm names are not unique across the country. Digitised geological maps covering the extent of the Beaufort Group were included as additional backdrop data.

The Evolutionary Studies Institute (ESI) collection database was selected as the test case because of the high resolution of farm locality and map sheet data, to determine whether automated entry of palaeontological records was a feasible option. Unique localities were split into those with coordinates and those without, as the process for identifying the location of these two groups of localities was different[16].

Those localities with grid coordinates were extracted and all coordinate data converted to decimal degrees, imported into ArcGIS® as an event theme, converted to a shape file, and each specimen was located as a point in the spatial data file.

Localities that lacked coordinates were identified by districts, farm names and map sheet indices. As farm names are not unique and can be repeated for several districts, the map sheet index was used in addition to the farm names and districts as identifiers for the location of the farm localities. As an index shape file of the 1:50 000 map sheet series does not exist, a map sheet index shape file was created by digitising the sheets.

As a test run to determine how to automate the linkage of the locality name provided by the ESI data to farm records listed in the Surveyor General data, 13 distinct localities in the district of Beaufort West were selected (Table 5). According to the alphanumeric data, all these localities fall on the same map sheet except for the Winterberg (Gryskop) locality. Of the 13 localities, only seven were matched to the spatial data and of these only two localities fell on the correct map sheet,[16] indicating it would be difficult to automate the linking process.

Table 5: Unique localities in the Beaufort West District

Locality name	District
Beaufort West Commonage	Beaufort West
Biesjiespoort (on Nobelsfontein)	Beaufort West
Buffelsvlei	Beaufort West
Dunedin (formerly Quaggasfontein)	Beaufort West
Elandsfontein	Beaufort West
Heuninglaagte (Honinglaagte)	Beaufort West
Kuilspoort	Beaufort West
Leeu Kloof	Beaufort West
Leeufontein	Beaufort West
Leeukloof	Beaufort West
Rosary	Beaufort West
Uitspansfontein	Beaufort West
Winterberg (Gryskop)	Beaufort West

Another test was run to determine if 'selection by map sheet' could be used as a method to link alphanumeric data to spatial data. Map sheet 3123DD was randomly selected and alphanumeric records of the ESI collection located on this map sheet were selected, returning 41 records. These records were then queried such that only distinct localities would be returned, and resulted in 38 localities. As locality name should correspond to farm name, it follows that there should be 38 farms which intersect with this map sheet. A query was performed to select all the farms which lie wholly or partly on this map sheet and resulted in 16 farms - less than half the number of distinct localities. As the method of using locality name, farm name and farm centroid was not effective (because neither locality name nor farm name matched the government farm name), an alternative linkage solution needed to be created.

The database of Iziko South African Museum, which contains both locality names and formal government farm names, was used as the linkage mechanism. For each collection, a query for distinct records of specimens was run. Results from this query were input into a second query where the centroids with their government farm names were linked to the specimens with their locality names via the joining table derived from the Iziko South African Museum data. The results from these queries were imported into the GIS and merged to form one data file.

The automatic import of records yielded a poor success rate. Of the 19 718 specimens selected for automatic import, 8512 records were imported automatically. Against the total number of records (20 968) that required import into the GIS, this number indicates a success rate of 43.2%. The only solution was for the remaining data to be manually imported into the existing digital-palaeo system.

Manual data entry procedure and collation of Beaufort Group data

Data tables containing specimens that had failed to import automatically were established for each museum and imported into the spatial map under separate map layers. Each fossil entry was then systematically added in point format to the spatial map.

The final phase involved collating the point data of the seven collections into a single map layer to decrease the complexity of future queries. The original intention was to append all the data sets onto a single data set, but because of differing alphanumeric table structures and schemes of the contributing museums, a single data set was not directly possible. Accordingly, the alphanumeric table structures of the various data files were manipulated to conform to a designed standard structure (Table 6).[16]

Table 6: Alphanumeric table structure

Field name	Data type
Id	Short Integer (6)
Specimen_I	Text (50)
Current_Tax	Text (50)
Taxon_Assi	Text (50)
Biozone	Text (50)
Province	Text (50)
District	Text (50)
Locality_N	Text (50)
Mapsheet	Text (50)
Old_ID	Text (50)
Census_Dis	Text (50)
Farm_Name	Text (50)
Co_ord_X	Double (8.7)
Co_ord_Y	Double (8.7)

Shape files that had a multipoint geometry type, rather than a point geometry type, were converted to a point geometry type, and appended, resulting in a single point shape file. This file contains approximately 30 000 points and indicates where fossil vertebrates from the Beaufort Group were discovered in the field.[17]

Applying GIS to palaeontological research

Establishment of the Beaufort Group GIS has resulted in a research aid which can be utilised to answer questions relating to Permo-Triassic continental vertebrate biodiversity. The system displays taxonomic diversity of fossil tetrapods from the Beaufort Group, with specific reference to the housing of particular genera, including numbers of specimens of each genus and their respective locality and/or biozone data. This unique data set provides a record of fossil tetrapod biodiversity in the continental realm from the Middle Permian to the Middle Triassic, and shows accurate numbers of specimens of the various taxa which have been collected (e.g. Smith et al.[18]).

The spatial map allows for queries to be performed relating to geospatial distribution, which may assist in understanding the pattern of basin infill as well as significant biodiversity patterns. Because of the paucity of reliable lithological markers which extend over a large area as well as datable volcanic rocks, index fossils have been used as biozone markers to develop basin development models.[19-22] Now that the database of Karoo fossils is available on a GIS, biogeographic distribution patterns of index taxa can be determined and utilised in the development of basin development models. In addition, because the tropic level of each species is recorded in the database, changes in terrestrial vertebrate ecological relationships from the Middle Permian to the Mid Triassic have been explored.[23]

The database elucidates which genera occur in particular biozones, as well as numbers of individuals of each genus in successive biozones. Application of the GIS can highlight biodiversity changes across successive biozones, making it possible to calculate the extent of extinction of taxa within successive time slices, which allows for the determination of trends in biodiversity changes through time.

Through the application of the GIS database, the stratigraphic ranges of fossil specimens were calculated and a refined biostratigraphic subdivision of the *Eodicynodon*, *Tapinocephalus* and *Pristerognathus* assemblage zones (Middle Permian Beaufort Group) proposed.[24]

Previously, biostratigraphic maps of the Beaufort Group were compiled based on a rough estimate of the distribution of biozone-defining signature fossil genera. Through the utilisation of the GIS database, a far more precise biozone map of the Beaufort Group has been produced (Figure 2), with the capability of being continuously updated as new information is received. This method introduces an entirely new way of representing geographical fossil distribution data that can be used in basin development and tetrapod biogeographic studies.[25]

Future spatial map enhancement and maintenance

Given the scope of both the specimen locality data and the necessity for this data to be available in a readily usable form, efficiency and accuracy is of prime importance for the task of geospatial referencing. Individual institutions housing palaeo-collections typically lack the resources or informatics expertise to meet the challenges of georeferencing alone.[15] Designing a collaborative geospatial referencing methodology using the combined expertise of all palaeontologists in South Africa has yielded an accurate and reliable spatial map.

An important aspect of data management that will improve reliability of the data set is the accurate assignment of taxonomic data. This accuracy is crucial for trend analysis. The updating and verification of both biozone assessment and taxonomic assignment must be accomplished through ongoing collaboration of palaeontologists, each selected for their expertise in specific taxa of animals and/or geological expertise. Specimens in collections need to be checked and identifications must be updated using current taxonomic diagnoses.

The application of GIS technologies could have significant impact as it could open further avenues of GIS-based research in palaeontology. The data could be used for four-dimensional (4D) spatio-temporal modelling. These dimensions can be distorted by geological or other processes and hence there are different spatio-temporal dimensions that are relevant for the fossils, including the various 4D environments through which the fossil has been taken, changed, moved, etc., such as weathering, erosion, re-deposition, lithification, metamorphism, diagenesis, faulting, folding, etc.

Ongoing contribution to the GIS project involves the further development and refinement of the Spatial Map of Beaufort Group fossil specimens (e.g. Van der Walt et al.[25]). Continual refinement and upgrading will ensure a reliable product that may confidently be used as an analytical and research tool.

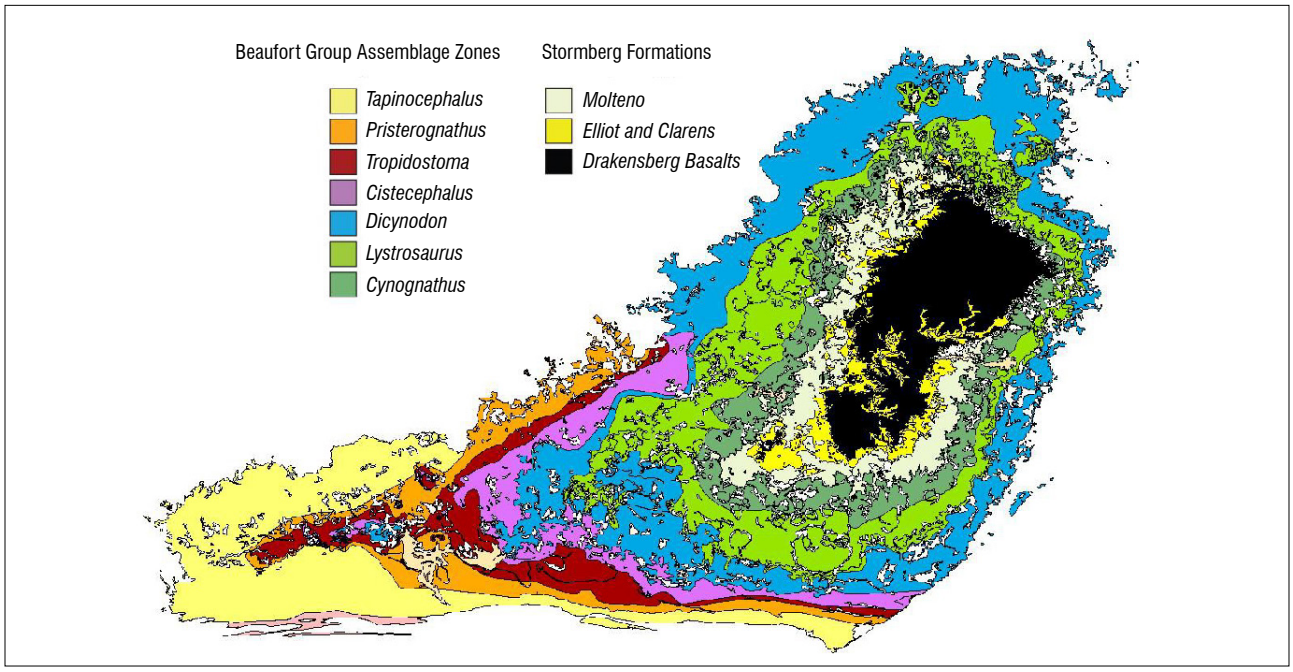


Figure 2: Defined biostratigraphic map of the Beaufort Group, Karoo Supergroup of South Africa – produced through the application of GIS technology.

Conclusion

Digital archiving and spatial cataloguing of artefact collections data are central to biodiversity research. The way forward for the digitised spatial map of the Beaufort Group of the Karoo Basin is to create easier access for researchers, where currently access to collections is limited by their distributed nature on museum shelves.

Combining all the South African databases of vertebrate fossils from the Beaufort Group onto a geographical information system is an important tool which could be utilised to address numerous different issues relating to Permo-Triassic continental tetrapod biodiversity in space and time. In the process of setting up this GIS, it became apparent that the course of action is not simply a matter of combining all the databases and coming up with answers.

Many metadata have been collected for recorded fossils from the seven large collections of vertebrate fossils in South Africa, but, particularly for specimens collected long ago, the quality of the metadata was not initially compliant with the requirements for a reliable GIS. Lack of complaince is largely because of a lack of consistency of data in the different collections, and also because precise GPS coordinates of localities of many of the fossils were not recorded, and only a farm locality was given. Certain specimens were listed more than once, despite the selection of distinct records. This is because of capture errors and lack of standardisation in the original data. In a number of the spreadsheets, the supposedly unique identifiers of some specimens were not actually unique, which made removing redundant data much more difficult.

Since overcoming these obstacles, the outcome is a GIS database that is reliable and applicable. The result is the establishment, for the first time, of a Geographical Information System incorporating all vertebrate fossils from the rocks of the Beaufort Group which are curated in South African Museums, have already been utilised to address questions relating to Permo-Triassic biodiversity and palaeoecological change.[18,23,26-28] As the Karoo geological succession preserves the most complete record of Middle Permian to Middle Triassic continental tetrapod biodiversity, a study of this nature is of great importance for an understanding of early continental tetrapod biodiversity changes as well as for Gondwanan basin modeling.

The objective of the established GIS database is for its utilisation as both a research tool and digital archive and as such, it requires continuous improvement, updating and refinement. Future work will include standardising the records for the fossils, for both data and the metadata.

Acknowledgements

The authors are indebted to the Council for Scientific and Industrial Research, the Department of Science and Technology, the National Research Foundation, the Centre of Excellence in Palaeoscience, the Palaeontological Scientific Trust and the University of the Witwatersrand for financial assistance. We thank and have benefitted from numerous discussions with many researchers: Fernando Abdala, Rose Prevec, Mike Raath, Roger Smith and Mike Day. Special thanks to Billy De Klerk (Albany Museum), Bernhard Zipfel (Evolutionary Studies Institute), Johann Neveling (Council for Geoscience), Elise Butler (National Museum), Richard, Robert and Marion Rubidge (Rubidge Collection), Sheena Kaal and Roger Smith (Iziko Museum) and Stephanie Potze (Ditsong) for allowing generous access to their museum records.

Authors' contributions

M.v.d.W. was the project leader and wrote the manuscript. A.C. made conceptual contributions and added to/edited the manuscript. I.N. performed the GIS queries; automated import and made experimental and conceptual contributions towards establishing the GIS database. B.R made conceptual contributions and added to/edited the manuscript.

References

1. Hancox PJ, Rubidge BS. Breakthroughs in the biodiversity, biogeography, biostratigraphy, and basin analysis of the Beaufort Group. J Afr Earth Sci. 2001;33:563–577. http://dx.doi.org/10.1016/S0899-5362(01)00081-1
2. Rubidge BS. 27th Du Toit Memorial Lecture Re-uniting lost continents–Fossil reptiles from the ancient Karoo and their wanderlust. S. Afr J Geol. 2005;108(1):135–172. http://dx.doi.org/10.2113/108.1.135
3. Broom R. The mammal-like reptiles of South Africa and the origin of mammals. London: HF & G. Witherby; 1932.
4. Nicolas MV, Rubidge BS. Assessing content and bias in South African Permo-Triassic Karoo tetrapod fossil collections. Pal Afr. 2009;44:13–20
5. The Revealing Human Origins Initiative (RHOI) [document on the Internet]. c2008 [cited: 2008 Oct 19]. Available from: http://rhoi.berkeley.edu

6. Virtual Anthropology. The digital@rchive of Fossil Hominoid [A-1909] [document on the Internet]. c2009 [2009 Jan 02]. Available from: http://www.virtual-anthropology.com/3d_data/3d-archive

7. Biodiversity Information Standards (TGWG). The Darwin Core [page on the Internet]. c2015 [2015 June 05]. Available from: http://wiki.tdwg.org/darwincore

8. International Organization for Standardization (ISO). ISO 15836:2009, Information and documentation — The Dublin Core metadata element set. Geneva:ISO; 2009.

9. International Organization for Standardization (ISO). ISO 19115:2003, Geographic information — Metadata. Geneva: ISO; 2013.

10. South African Bureau of Standards (SABS). SANS 1878-1:2005, South African spatial metadata standard, Part 1 — Core metadata profile. Pretoria: SABS; 2005.

11. Nogueras-Iso J, Zarazaga-Soria FJ, Lacasta J, Béjar R, Muro-Medrano PR. Metadata standard interoperability: Application in the geographic information domain. Comput Environ Urban. 2004;28(6):611–634. http://dx.doi.org/10.1016/j.compenvurbsys.2003.12.004

12. Bain AG. On the discovery of the fossil remains of bidental and other reptiles in South Africa. Transactions of the Geological Society of London. 1845:53–59. http://dx.doi.org/10.1144/transgslb.7.53

13. Cohen A. Mr. Bain and Dr. Atherstone: South Africa's Pioneer Fossil Hunters. Earth Sci Hist. 2000;19(2):175–191.

14. International Organization for Standardization (ISO). ISO 19112:2003, Geographic information — Spatial referencing by geographic identifiers. Geneva: ISO; 2003.

15. Beaman R, Wieczorek J, Blum S. Determining space from place for natural history collections. D-Lib Magazine. 2004;10(5):9. Available from: http://www.dlib.org/dlib/may04/beaman/05beaman.html

16. Cooper A, Netterberg I. Mapping palaeontological locations and specimens. icomtek document number: 0451-0001-701-A1. Pretoria: CSIR; 2004

17. Nicolas MV. Tetrapod biodiversity through the Permo-Triassic Beaufort Group (Karoo Supergroup) of South Africa. Johannesburg: University of the Witwatersrand; 2008.

18. Smith R, Rubige BS, Van der Walt, M. Therapsid biodiversity patterns and palaeoenvironments of the Karoo Basin, South Africa. In: Chinsamy-Turan A, editor. Forerunners of Mammals: Radiation Histology Biology. Indiana, IN: Indiana University Press; 2011. p. 223–246.

19. Catuneanu O, Hancox PJ, Rubidge BS. Reciprocal flexural behavior and contrasting stratigraphics: A new basin development model for the Karoo rhetoric foreland system, South Africa. Basin Res. 1998;10:417–439. http://dx.doi.org/10.1046/j.1365-2117.1998.00078.x

20. Rubidge BS. Biostratigraphy of the Beaufort Group. Pretoria: South African Committee for Stratigraphy, Biostratigraphic series No 1. Pretoria: Council for Geoscience; 1995; p. 1–2.

21. Hancox PJ. A stratigraphic, sedimentological and palaeonvironmental synthesis of the Beaufort-Molteno contact in the Karoo Basin. Johannesburg: University of the Witwatersrand; 1998.

22. Neveling J. Biostratigraphic and sedimentological investigation of the contact between the Lystrosaurus and Cynognathus Assemblage Zones (Beaufort Group: Karoo Supergroup). Johannesburg: University of the Witwatersrand; 2002.

23. Nicolas MV, Rubidge BS. Changes in Permo-Triassic terrestrial tetrapod ecological representation in the Beaufort Group (Karoo Supergroup) of South Africa. Lethaia. 2010;43(1):45–59. http://dx.doi.org/10.1111/j.1502-3931.2009.00171.x

24. Day MO. Middle Permian continental biodiversity changes as reflected in the Beaufort Group of South Africa: A bio- and lithostratigraphic review of the Eodicynodon, Tapinocephalus and Pristerognathus assemblage zones. Johannesburg: University of the Witwatersrand; 2013.

25. Van der Walt M, Day M, Rubidge B, Cooper AK, Netterberg I. A new GIS-based biozone map of the Beaufort Group (Karoo Supergroup), South Africa. Pal Afr. 2010;45:1–5.

26. Irmis RB, Whiteside JH, Kammerer CF. Non-biotic controls of observed diversity in the paleontologic record: An example from the Permo-Triassic Karoo Basin of South Africa. Palaeogeogr Palaeoclimatol Palaeoecol. 2013;372:62–77. http://dx.doi.org/10.1016/j.palaeo.2012.07.014

27. Fröbisch J. Vertebrate diversity across the end-Permian mass extinction – separating biological and geological signals. Palaeogeogr Palaeoclimatol Palaeoecol. 2013;372:50–61. http://dx.doi.org/10.1016/j.palaeo.2012.10.036

28. Abdala F, Rubidge BS, Van Den Heever JURI. The oldest therocephalians (Therapsida, Eutheriodontia) and the early diversification of Therapsida. Palaeontology. 2008;51(4):1011–1024. http://dx.doi.org/10.1111/j.1475-4983.2008.00784.x

Housing and population sprawl near tailings storage facilities in the Witwatersrand: 1952 to current

AUTHORS:
Melanie A. Kneen[1]
Matthew E. Ojelede[2]
Harold J. Annegarn[3]

AFFILIATIONS:
[1]Department of Geosciences, University of Texas at Dallas, Richardson, Texas, United States of America
[2]Department of Atmospheric Sciences, Digby Wells Environmental, Johannesburg, South Africa
[3]Energy Institute, Faculty of Engineering, Cape Peninsula University of Technology, Cape Town, South Africa

CORRESPONDENCE TO:
Melanie Kneen

EMAIL:
Melanie.kneen@gmail.com

POSTAL ADDRESS:
Department of Geology and Environmental Science, Spring Creek Campus 2800 E. Spring Creek Parkway Plano, TX 75074, United States of America

DATES:
Received: 04 June 2014
Revised: 19 Nov. 2014
Accepted: 25 Apr. 2015

KEYWORDS:
aeolian dust; residential population; quartz; silicosis risk; calculating exposed population;

HOW TO CITE:
Kneen MA, Ojelede ME, Annegarn HJ. Housing and population sprawl near tailings storage facilities in the Witwatersrand: 1952 to current. S Afr J Sci. 2015;111(11/12), Art. #2014-0186, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140186

Mining, tailings storage facilities (TSFs), dust pollution and growth in residential housing development are synonymous with the Witwatersrand, South Africa. Encroachment of housing onto land close to TSFs, i.e. areas rendered marginal because of the dust hazard and risk of structural failure, has continued unabated for decades, intensifying human exposure to windblown mineral dust. Recent research indicates that the finer milling used for modern gold extraction results in aeolian dust emanating from the TSFs which contributes to a higher proportion of inhalable particles in the source material. Air quality dispersion modelling, validated by ambient aerosol monitoring campaigns, indicates that episodic dust events generate particulate matter (PM_{10}) and, specifically, quartz dust concentrations that are unhealthy at distances of up to 2 km downwind from TSFs. This contribution documented residential development from 1952 to 2011 (using historical aerial photographs, census data from 2001 and 2011 and ancillary information) to determine the population exposed to dust emanations from the TSFs. Using the images, land use was classified into residential areas, TSF footprints and open areas, onto which a series of 500 m buffer zone contours were superimposed. The resulting statistics were used to assess the populations exposed to dust hazard within the defined buffer zones. Overall, housing development has experienced a growth of approximately 700% since 1952 at a rate of 14% per year. Analysis of recent monitoring campaign data has confirmed multiple occurrences of quartz-rich inhalable dust in residential settings at levels that exceed occupational health standards, extrapolated to values for population exposure.

Introduction

Evolution in social, economic and political landscapes often shapes population growth, migration and settlement patterns, which in turn often influence the net environmental risks to populations. This is particularly evident in Asian and African countries.[1-5] A growth projection of 2.4% over a twenty-year period (between 2000 and 2020) suggested for sub-Saharan Africa, is higher than for other developing regions (1.3%) and the developed world (0.1%). Cities in African countries will experience an additional 150 million residents because of 'high rates of rural-out migration'.[3] In South Africa, the transformation into an urban, socio-economic and political space has been ongoing since the discovery of gold in the mid-1890s, with gold mining initiating an unusually rapid population growth. Influxes of migrant labourers led to the proliferation of makeshift houses, the establishment of mining camps and eventually towns and cities – areas which continue to evolve. Events on the Witwatersrand have defied the norm whereby, on the cessation of mining, surrounding settlements usually become ghost towns. By contrast, prospecting, excavation, processing a range of ore minerals and the reclamation of legacy gold mine tailings continue to make mining a major regional industry. This process has made the Witwatersrand region a vibrant economic hub.[4]

Tailings storage facilities (TSFs) are intertwined with built-up areas and other mine features in the Witwatersrand.[6,7] With sparse vegetation cover and the reworking of older TSFs for their residual gold content, exposure to wind-generated mineral dust presents an intermittent but persistent environmental hazard for nearby residents.[6,8] As part of evolving technology for gold extraction, the reclaimed TSF material is milled to a much finer grade (down to 30 μm, considerably smaller than the previous 70 μm diameter). This greatly increases the fraction of inhalable material or PM_{10} (i.e. less than 10 μm aerodynamic diameter particles which can be inhaled) in the tailings material. Investigation into detailed size and chemical characterisation of material and ambient airborne dust exposure is ongoing, as is an epidemiological survey aimed at quantifying the associated human health risk of exposed residential populations developing silicosis.[8] Recent research indicates that dust management faces challenges[9], while the increasingly fine material size ranges[8,10] coupled with the results of this study (which identifies the extent of the population being exposed) makes mitigating the negative effects a priority.

While international standard practice and local mining regulations prohibit or avoid locating residential areas closer than 500 m from TSFs,[11] the earliest aerial photographic images of the areas within that distance around Johannesburg's TSFs, clearly show that housing development has been continuing despite the recent data on the emerging increase in health risks for residents.

This study assesses the growth of housing developments near the gold mine TSFs of the Witwatersrand over the past 50 years and evaluates changes in the population exposed to windblown dust emanations from the TSFs. This information forms a critical component of the process of evaluating net risk, and in quantitatively assessing benefit-to-cost ratio of proposed mitigation strategies that depend on scarce resources, such as water, for dust mitigation.

Background

Prior to the 1950s, the sites selected for deposition of gold mining residues were in sparsely populated areas mainly covered with savannah grassland, colloquially known as *veldt*, with grass or low shrubs and few trees. Nearest residential areas were 3-4 km distant from the margins of the TSFs. From the 1950s onwards, rapid

expansion of mining led to the socio-economic growth of Johannesburg and surroundings, with rapid urbanisation to the north and the simultaneous development of dormitory suburbs in the area south of the Witwatersrand ridge (through the forced resettlement policies of the apartheid-era government). In the post-apartheid era, after 1994, with newfound political freedom and removal of restrictions on movement of citizens of colour, there has been a surge in rural-urban migration by people in search of a better livelihood, leading to the increased development of formal and informal housing.

There is currently a housing backlog in South Africa and increased pressure for further housing developments, particularly on remaining open land within the greater city limits, along the Witwatersrand mining corridor. These historical forces have shaped the housing patterns and dynamics, resulting in extensive infill of vacant land that had constituted the safety buffer zone near both active and dormant TSFs, thereby aggravating the exposure to health hazards and environmental consequences of dust pollution from the TSFs.[8]

With 270 TSFs in the region,[12] several topics – such as the extensive development, continued surge in migrant labourers, great land pressure issues, public health concerns associated with exposure to airborne mineral quartz dust – have all been the subject of research initiatives.[13-20]

For the purpose of this research, three large TSFs were selected for investigation of historical changes in population density within the zones of maximum dust exposure: the Crown Gold Recoveries (CGR) complex in central Johannesburg; the Durban Roodepoort Deep (DRD) TSF on the west Witwatersrand; and the East Rand Proprietary Mine (ERPM) TSF complex on the east Witwatersrand. These TSFs, each with areal footprints exceeding 1 km^2, besides being the largest three TSFs in the Central Witwatersrand goldfield, represent a range of conditions. The CGR complex is the site of recent active deposition of reprocessed and finer grade material, reclaimed from numerous smaller TSFs which are a legacy of earlier mining efforts when gold extraction processes were less efficient. The DRD complex is dormant, and vegetation from an earlier rehabilitation campaign has been extensively eroded or buried by windblown sand. The ERPM complex is undergoing reclamation that commenced circa 2010, exposing previously vegetated surfaces to wind and water erosion.

In this study, historical and local 'at risk' population density and distribution trends are documented. Detailed algorithms – to expand the analysis to permit rapid application of population location-based risk analysis at regional (citywide) or specific (project or TSF based) scales – were explored.

Data and methods

Scanned aerial photographs, drawn from the records of the Chief Directorate of Survey and Mapping[21] in Cape Town, South Africa were employed to obtain historical information on residential patterns in relation to the selected TSFs. QuickBird Imagery for 2004 obtained from Southern Mapping Company (Pty) Ltd) and Google Earth historical images (Figure 1) were used to survey changes in residential land use. These were complemented with 1:50 000 topographic maps, which show the extent of formal housing developments, in shapefile format. Census data for 2001[22] and 2011[22], with an accurate count of the houses and people within demarcated census areas called 'enumerator areas', were obtained from Statistics South Africa (Stats SA).

Formal housing developments have the advantage that both the structure and housing density have remained fixed for reasonably long periods. With the dwelling density and the area of each dwelling space derived for 2001, the total population and the number of households for earlier periods were estimated. The assumption that the household density and dwelling space ratios, as calculated in the 2001 census data[22], remained the same over time was required, because census data coinciding with the image dates do not exist. Matching the historical residential areas in the newer images confirmed that for these formal residential developments, this assumption is reasonable. Because the relationship between land area and number of households has been constant within the study area, the household sizes, and thus the number of people associated with these homes, is also likely to have remained constant.

For several reasons the Census 2001 data[22] were the best estimate of the South African population, and therefore served as the basis for estimates of TSF-affected population in this study. This was the first census taken after the end of apartheid and consequently was less influenced by legal and political pressure. The 2001 census data[22] covered a wide range of parameters, including household size, income, individuals and their ages. Without the pressure of under-reporting (concealing) illegal occupation, the 2001 census results[22] could be treated with more confidence than any studies prior to the political changes in 1994.

Image pre-processing

The digital aerial photographic images and QuickBird imagery were geo-referenced using ground-control points derived from the 1:50 000 topographic maps. Images from the different historical periods were resampled to matching spatial resolutions. The pre-processing continued through making a mosaic of the different scenes, covering the area of interest using both aerial photo and remotely-sensed imagery in each period, typified in Figure 2. Across all images and for all time periods, the geo-referencing was accurate to 1 m or better. Separate images covering the area of interest for the selected years (1952, 1976, 1991 and 2002) for areas surrounding the three TSFs (CGR, DRD and ERPM) were extracted from the mosaic images. These have been illustrated by the series of aerial photographs for the CGR complex for the respective years, 1952, 1976, 1991 and 2002.

Delineation of buffers and polygon outlines and quantification of change

After the mosaic process, three different approaches were adopted to evaluate growth in residential housing developments over time. Firstly, using geographic information system (GIS) spatial analysis tools, buffer zones were defined at distances of 500 m, 1000 m, 1500 m and 2000 m from the edges of the TSFs under survey.

Secondly, on-screen digitising of the areal extent of residential housing developments, within each buffer zone, was performed for the different periods under survey. The areas covered by the digitised polygons were estimated using the statistical tools in TNTmips® (MicroImages 2009), with differences evaluated as the areas changed between successive periods.

Most of the housing developments within the buffer zones of the selected mega-TSFs are formal settlements with permanent structures. Informal settlements (comprising shack dwellings) occur extensively at other locations and around other TSFs in the city, but have not been rated in this study. The benefit of formal housing structures is that once a development has been completed, the number of dwellings remains remarkably stable. While the population might change over time, the number of households can be calibrated and compared with confidence.

The dramatic social and political changes in this region since the 1950s have made the comparison of census data problematic. In this case, population data are best extrapolated from housing changes, quantified as described above, and calibrated to population by the most reliable single census. The census data from 2001[22] provided the best available count of the *houses and people* within reporting areas – called *enumerator* areas – at that point in time. An observed problem is that the enumerator areas include both the open land and the residential areas. By contrast, the 1:50 000 topographic data show the *extent of the formal housing developments* (which excludes open land) but gives no indication of the actual *number of houses and people*. Combining these data sets in an ARC® GIS environment, the exact area occupied by residential areas, average dwelling space and the dwelling density in 2001 was computed. It was observed that in neighbourhoods reporting lower incomes (according to the 2001 census data[22]), properties occupied less space and were more densely populated; higher income neighbourhoods occupied larger expanses of land at a lower density.

Extrapolation to large scale or automated analysis

Once the detailed, specific and historical analysis of the three priority locations had been completed, the next stage was to extend the analysis,

Figure 1: (a) February 2000 and (b) December 2013 Google Earth images showing newest residential developments close to Crown Gold Recoveries tailings storage facilities. Red arrows indicate newest housing developments.

using the calibrated and hand checked data to include all the TSFs in the city and across the Witwatersrand. Population densities, derived from 2001 census data[22], and dwelling structure footprints from 2011[23] were used separately to estimate the regional impact of fine dust on residential areas adjacent to mine TSFs. These techniques to estimate automatically population numbers from census[22,23] and survey general topographic data, can be applied regionally or to any selected subset of locations.

Two approaches were used to estimate the population living close to the TSF. The 2001 census data[22], with the population densities (calculated as described previously), were intersected with the TSF buffers to compute population density polygons within each buffer zone. Within each buffer zone, the exposed population was calculated by multiplying the number of people (or households) per unit area by the area within that buffer zone (Figure 3a).

The first step was to designate buffers around the required land use areas. In this example, both types of TSF locations were used, i.e. older sand dumps, and more recent slimes dam deposits of hydraulically placed tailings material with a median particle grade of 70 μm or less. Buffer distances, of 0.5 km, 0.5–1.0 km, 1.0–1.5 km, 1.5–2.0 km, 2.0–2.5 km and 2.5–3.0 km, were used to assess the population distribution at a range of distances. This range corresponded to distances shown in dispersion modelling studies to be adversely influenced by transport of dust during high wind speed conditions.[24] Modelling results[24] have been validated against data from particulate matter monitoring campaigns

downwind of the studied TSFs. Monitoring results confirmed that the frequency and intensity of dust episodes are directly related to the strength and frequency of high wind speeds.[8]

The analysis undertaken with the 2001 census data[22] was then repeated using the 2011 census data[23]. Additionally, and of more interest, the 2011 census data[23] released for this project included access to a shapefile of the 'external dwelling frame' locations – i.e. the fixed structure, formal houses (small coloured squares, Figure 3b). To further investigate the stability of these population estimates, the number of dwellings identified within each buffer zone was computed. While this approach overlooks the significant homeless population of the city, it does provide a stable record of permanent structures, mostly brick or concrete block buildings, and therefore the long-term exposed population can be calculated. Buffer distances, in 50 m increments up to 250 m, were created to determine how closely residential dwellings had encroached onto marginal land (defined as land unsuitable for residential use because of exposure to heavy dustfall or risk of mechanical failure of the tailings impoundment).

Results and discussion

Changes in urban development around Crown Gold Recoveries TSFs

Changes in the number of households and the population within specific buffer distances (0–500 m, 500–1000 m, 1000–1500 m and 1500–2000 m) of each of the TSF complexes were derived from the changes in formal residential developments (fixed housing structures). Figure 4 illustrates the progression of developments around the Crown Gold Recoveries facilities. The most significant changes occurred between 1952 and 1976 – encompassing the rapid development of the Soweto dormitory city – and again between 1991 and 2002 – with urbanisation after the political transformation following the demise of the apartheid regime in 1994. Little undeveloped land remains around this location. The area of the TSF buffer zone occupied by residences has tripled, from 2.4 km^2 in 1952 to 9.6 km^2 in 2002 – detailed values of the change in area and number of households is shown in Table 1.

The latest available images from Google Earth indicate that, despite recent communication between official representatives, interested and affected parties and TSF site management's efforts to discourage or reduce the development of further housing projects on marginal land (land unsuitable for human settlement for environmental or geo-structural reasons), development in such areas has continued.

Changes in urban development around Durban Roodepoort Deep TSFs

In 1952, the area to be covered by the Durban Roodepoort Deep (DRD) tailings facility and surroundings was still undisturbed veldt. By 1976, the encroachment of residential development towards the boundaries of the TSF had begun. Development areas within TSF buffer zones remained minor through 1991 (Figure 5). Development accelerated after 1991, because of the scarcity of open land close to the city centre and the changing political climate. The Reconstruction and Development Plan of the new government post-1994 incorporated a policy of providing free houses (25–35 m^2) to indigent citizens. These government-financed housing developments often encroached on the TSF buffer zones (brown area, Figure 5). Overall, residential development in this region has increased from zero in 1952 to 0.63 km^2 in 1976 and 8.46 km^2 in 2002.

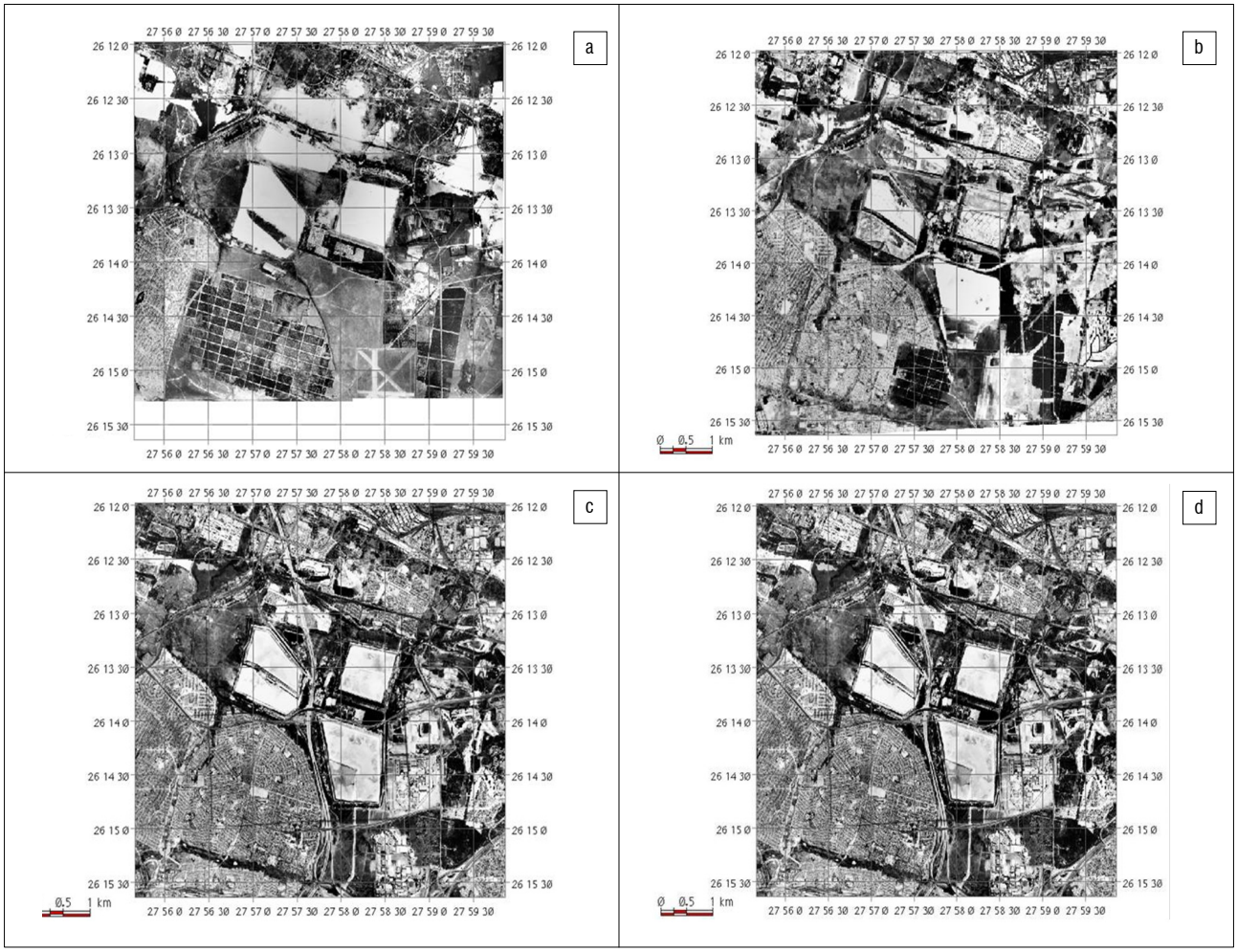


Figure 2: Aerial photograph of the Crown Gold Recoveries tailings complex for (a) 1952, (b) 1976, (c) 1991 and (d) 2002.

In contrast with the CGR TSF, there are indications that the 500 m buffer zone around the DRD TSF has been maintained (as can be seen along the eastern and southern margins of this TSF (Figure 5)). Detailed values of the change in area and number of households is shown in Table 1.

Changes in urban development around East Rand Properties TSFs

Around the East Rand Properties Mining (ERPM) TSF, even from the earliest images, there has been no buffer zone between the tailings facilities and the residential developments (Figure 6). In 1952 there was already 1.67 km^2 of residential developments close to the ERPM TSF. The coloured forced resettlement area of Reiger Park, 0.43 km^2 in extent, was situated closer than 500 m from the northern margin of the TSF. Suburbs to the south east, intended for habitation by whites, were located 0.5–1 km or more, distant from the TSF margins. The total residential area increased gradually and by 2002, had grown to 10.2 km^2 of developments, with 1.1 km^2 lying within 500 m of the TSF footprint. Later images from Google Earth (2012 Nov 29) do not indicate further residential developments occurring in the buffer zones defined around this TSF. Detailed values of the change in area and number of households are shown in Table 1.

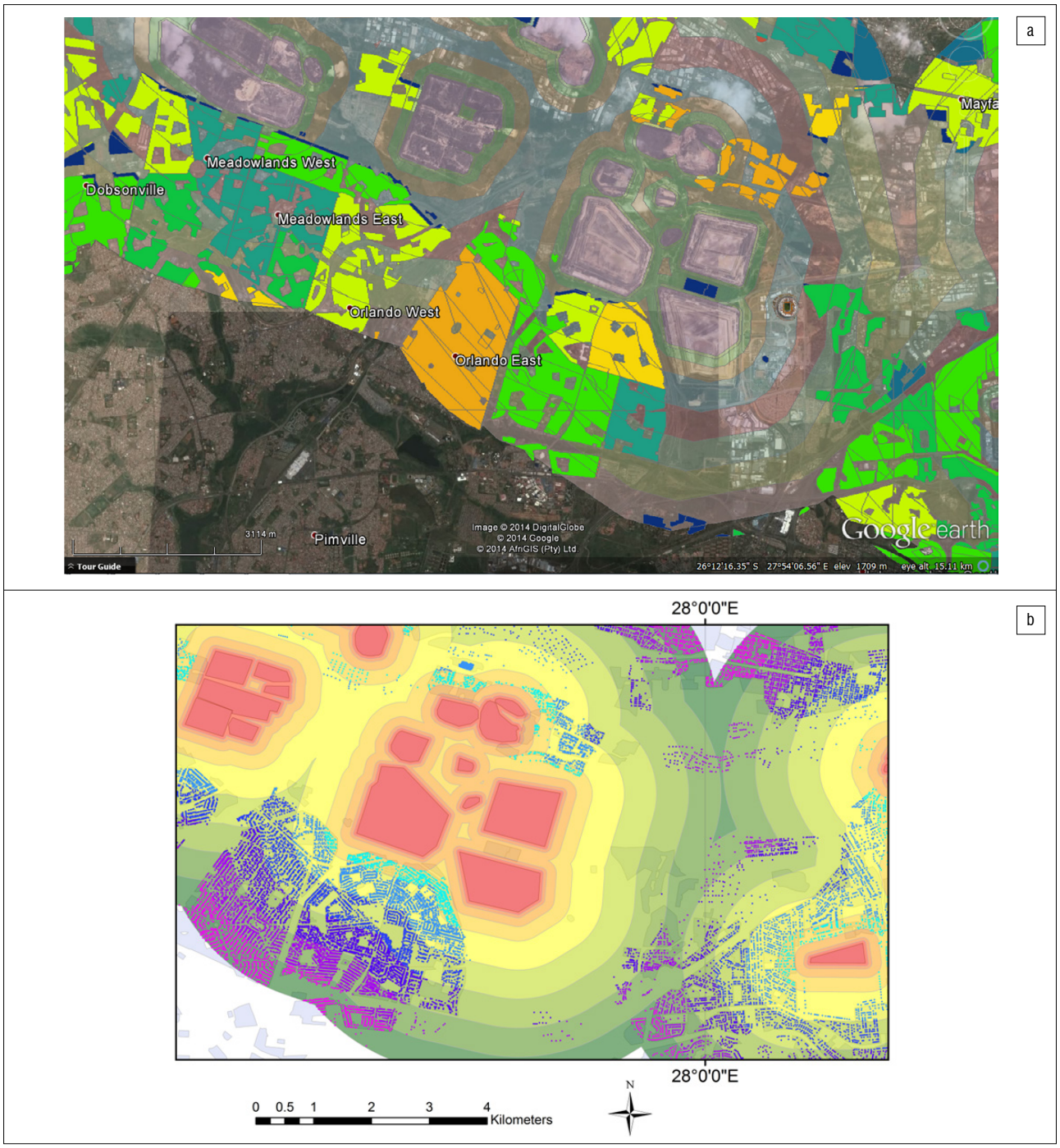


Figure 3: (a) Map of Crown Gold Recoveries region showing overlay of distance buffers (transparent and changing colour every 500 m) and residential areas coloured (blue = low, to orange = high) to reflect the average population density in the relevant census 2001[22] enumerator area. (b) Map showing individual dwellings recorded for the 2011 census,[23] coloured based on distance to the tailings storage facilities. Buffers coloured red to green in 500 m steps up to 3 km, dwellings as small squares coloured cyan to purple.

Table 1: Summary of residential development around Crown Gold Recoveries (CGR), Durban Roodepoort Deep (DRD) and East Rand Proprietary Mine (ERPM) tailings storage facilities, showing changes in the area and number of households or dwellings between 1952 and 2008

Location	Area (km²)					Number of households				
Distance (m)	1952	1976	1991	2002	2008	1952	1976	1991	2002	2008
CGR Total	2.42	6.44	8.48	9.65	10.65	38 847	68 151	85 602	91 850	95 109
500	0.11	1.39	1.75	1.89	1.86	87	10 328	9808	10 104	8201
500–1000	0.54	1.26	1.96	2.27	2.60	811	12 075	15 553	18 163	20 167
1000–1500	0.75	1.96	2.50	2.82	3.15	12 574	20 344	28 087	31 630	32 893
1500–2000	1.02	1.83	2.27	2.67	3.03	25 375	25 404	32 154	31 953	33 848
DRD Total		0.63	0.88	8.46	10.25		1960	2889	46 731	58 669
500				0.66	0.85				4310	5599
500–1000				1.46	2.58				9171	16 654
1000–1500		0.07	0.25	2.73	3.21		230	929	15 411	18 575
1500–2000		0.56	0.63	3.61	3.61		1729	1960	17 839	17 841
ERPM Total	1.67	8.52	9.28	10.23	10.47	4534	16 298	19 557	24 669	25 078
500	0.43	0.68	0.77	1.00	1.29	2147	3806	4139	5260	5732
500–1000	0.30	1.87	2.02	2.48	2.43	1169	4153	4990	7312	7246
1000–1500	0.30	2.69	3.00	3.23	3.23	327	3336	3903	5396	5397
1500–2000	0.64	3.28	3.50	3.52	3.52	891	5003	6524	6701	6702

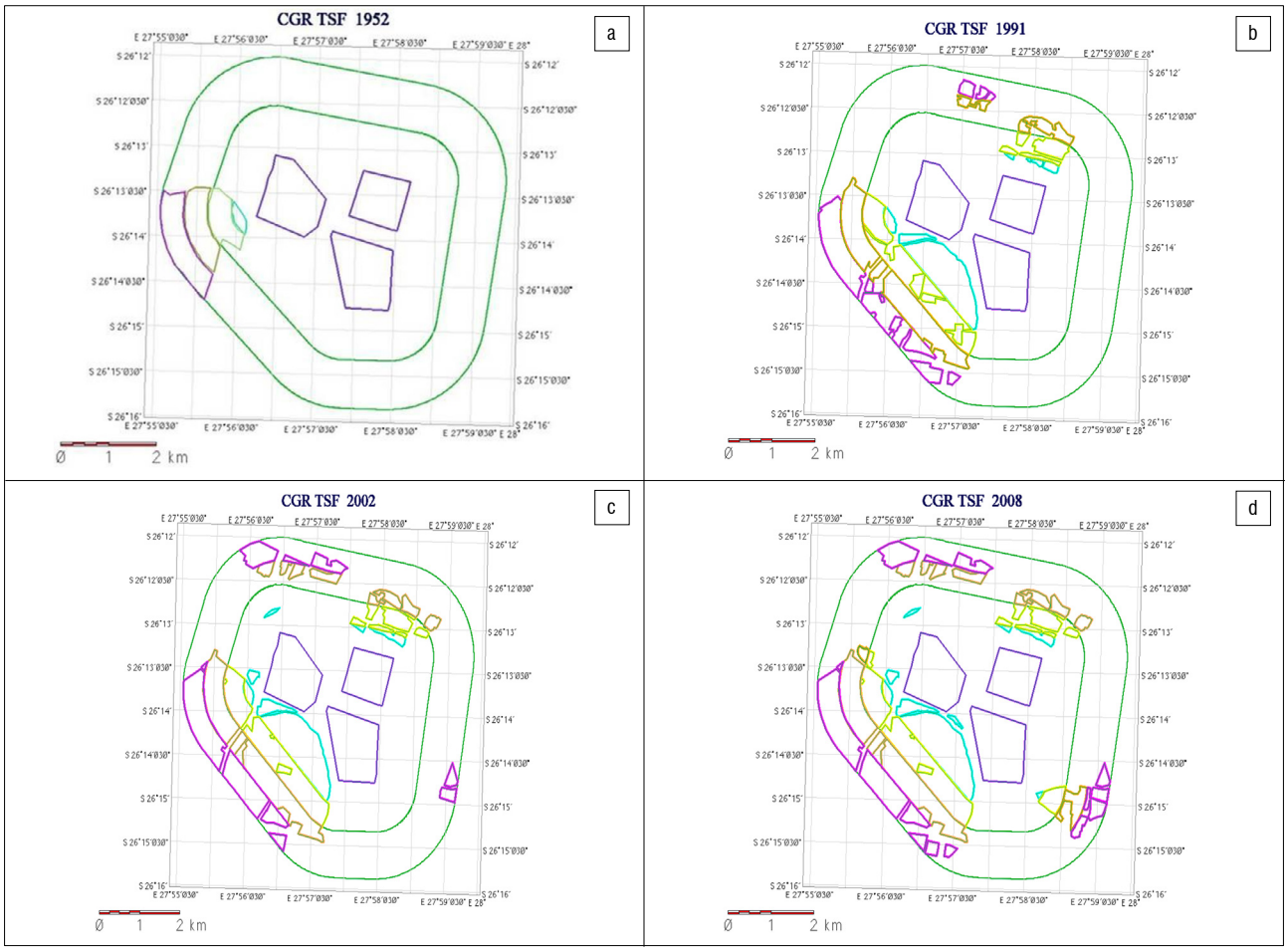


Figure 4: Time series progression of residential developments around the Crown Gold Recoveries tailings storage facilities (inner purple): (a) 1952; (b) 1991; (c) 2002 and (d) 2008. The 1 km and 2 km buffer zones are shown in dark green and the development polygons are coloured by distance from tailings storage facilities (TSF): cyan ≤ 500 m; yellow between 500 m and 1 km; orange between 1 km and 1.5 km; cerise between 1.5 and 2 km; purple = TSF footprints.

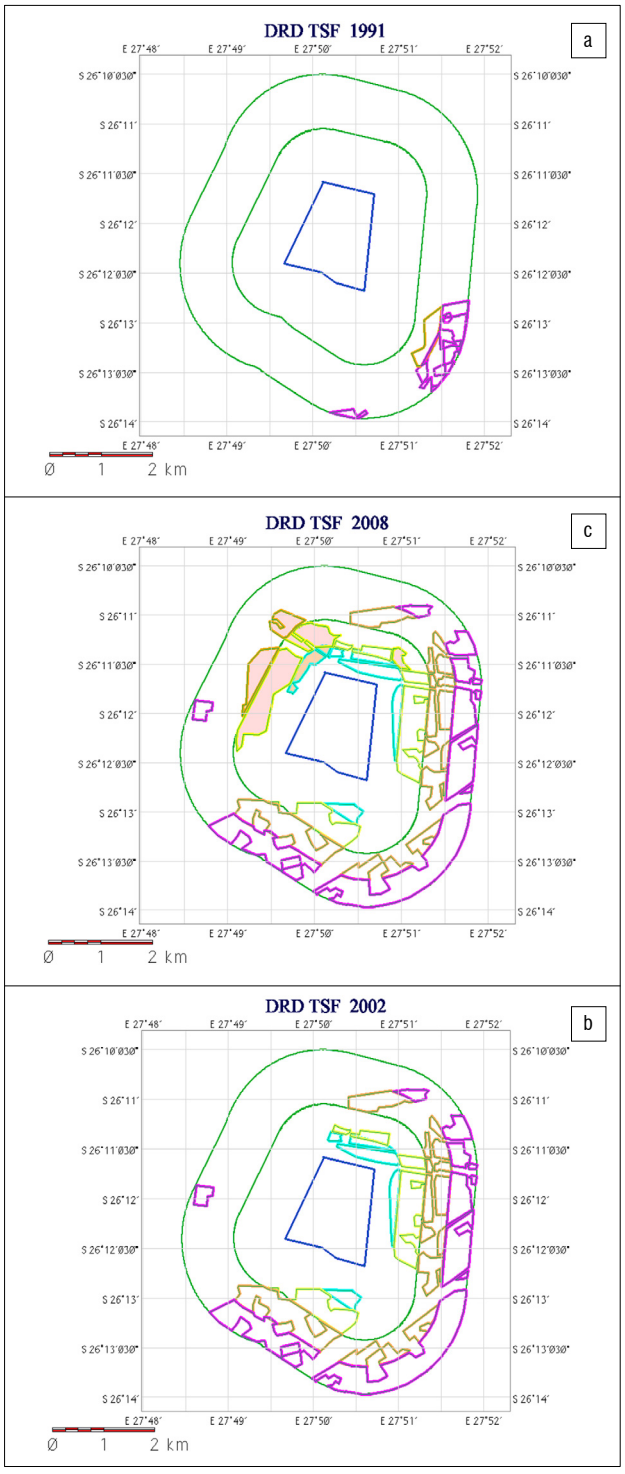


Figure 5: Time series progression of residential developments around the Durban Roodepoort Deep tailings storage facilities (inner blue polygon), with 1 km and 2 km buffer zones (dark green): (a) 1991; (b) 2002 and (c) 2008, cyan ≤ 500 m; yellow between 500 m and 1 km; orange between 1 km and 1.5 km; cerise between 1.5 km and 2 km, purple = tailings storage facilities (TSF) footprints. Brown coloured polygons indicate where houses were being built in 2008.

Changes in population around TSFs since 2001

Table 1 summarises the change in residential development around the TSFs from 1952 to 2008. By 1952 there was already a small population living around the newly established CGR and ERPM tailings storage facilities: the principal increase in near-TSF residential population occurred after 1952.

Steady growth has been evident after 1976, with the DRD location lagging somewhat in time. Urbanisation and development of low-cost housing caused a boom in residential developments after 1994 and this is reflected in the increased population for 2002. While there is only a short time interval between 2002 and 2008, the images for 2008 are evidence that the housing boom has continued.

Based on the images, it is obvious that there was no (enforced) policy in place in Johannesburg to prevent human settlements close to these tailings facilities. Since these facilities were created, buffer zones around the TSFs, as prescribed under mining regulations, have not been enforced. After more than half a century of mining and mine tailings in the city, Johannesburg's residents were confident they could live in the shadow of the TSFs. Since 1952, for over 50 years, this has been true. While dust was a constant presence in these areas during spring (August, September and October), that dust was mainly coarse, settleable particles which were not inhaled and remained primarily a nuisance. Since 2002, finer milling and generation of inhalable dust have transformed this nuisance into a health hazard. This contributes to the current need to invest in dust suppression and the importance of educating the local population about not damaging or disturbing the dust mitigation measures (water sprays and vegetation on TSF surfaces).

Table 2 shows the population estimate calculations based on household size information from the Census 2001[22] data set. This Table also contains an additional column reporting the number of households comprised of more than 10 individuals, according to the Census 2001[22] reports. Earlier survey data collected within Johannesburg, on a small scale and with return times more frequent than the national census, indicate that household size has increased since the 1950s. As explained above, the earliest reliable household size data are Census 2001[22], which is the basis of the population calculations presented here. The large household concentrations for 2008 have been provided as a reference metric for the potential effect of smaller households for earlier years.

Table 2: Population estimates for residential development around Crown Gold Recoveries (CGR), Durban Roodepoort Deep (DRD) and East Rand Proprietary Mine (ERPM) tailings storage facilities (TSFs) showing changes in the area and number of households or dwellings between 1952 and 2008

Location & distance (m)	Area (km²)		Number of households		Number of people		Large families (>10)
	1952	2008	1952	2008	1952	2008	2008
CGR Total	**2.42**	**10.65**	**38 847**	**95 109**	**120 611**	**319 055**	**2 516**
500	0.11	1.86	87	8201	388	30 639	264
500–1000	0.54	2.60	811	20 167	3119	72 805	468
1000–1500	0.75	3.15	12 574	32 893	38 967	110 735	817
1500–2000	1.02	3.03	25 375	33 848	78 136	104 876	968
DRD Total		**10.25**		**58 669**		**182 033**	**605**
500		0.85		5599		15 534	1
500–1000		2.58		16 654		47 898	92
1000–1500		3.21		18 575		57 983	191
1500–2000		3.61		17 841		60 617	301
ERPM Total	**1.67**	**10.47**	**4 534**	**25 078**	**18 348**	**88 798**	**418**
500	0.43	1.29	2147	5732	9518	23 839	189
500–1000	0.30	2.43	1169	7246	5034	27 252	168
1000–1500	0.30	3.23	327	5397	1092	17 419	43
1500–2000	0.64	3.52	891	6702	2703	20 289	18

Table 3: Households and residential areas around tailings storage facilities during 2001 and 2011 with estimates of population for the City of Johannesburg

Buffer distance (m)	Number of households 2001	Population estimate 2001	Number of households 2011	Population estimate 2011
500	26 232	80 700	22 370	71 600
500–1000	24 809	82 600	39 524	126 500
1000–1500	21 597	69 300	36 192	115 800
1500–2000	16 575	52 000	31 510	100 800
2000–2500	16 144	51 700	28 445	91 000
2500–3000	16 489	53 900	24 952	79 800
Grand total	121 846	390 200	182 993	585 500

Comparing the 2001 and 2011 populations on a regional basis, focusing on the permanent structures, also shows continuous growth of the residential developments. With the underlying housing deficit in Johannesburg and the ongoing urbanisation of South Africa's population, this was not unexpected. Residential housing needs, especially those within easy travel distances from employment opportunities and infrastructure, will continue to motivate infill development of marginal lands within the urban boundary. Based on the population estimates in Table 3, there are already more residents in these marginal zones than could possibly be relocated and it is unlikely that any policy or health requirement will slow this growth for as long as open space continues to exist around any of the TSFs within the city limits. Pressure to find work opportunities (and housing within reasonable distance) will override health concerns related to quartz inhalation.

One potentially significant weakness of using the population density to calculate the exposed households is the sliver effect that small polygons add to the population calculation. This can occur when a narrow strip of land along the buffer edge is classified as residential but does not contain any actual houses. To quantify this effect, smaller buffer zones within 500 m of the TSF boundary were evaluated (Table 4). Comparisons, made between the dwelling frame based data (2011) and population density data (2001) show that this effect is important at the 50 m and 100 m distances but by 500 m, the data sets are in closer agreement.

Further, it was found that household densities also vary, along with the property sizes, with implications for the population per unit area, thus indicating a strong link between socio-economic status and household sizes. Over time, there has been a substantial growth in residential areas around the tailings storage facilities. This ongoing growth has created tension between the need for affordable housing within existing urban zones, and the protection of public health. While there are currently a large number of homes within 2–3 km of these tailings storage facilities, there is still a considerable proportion of open land under continuing development. Quantifying the population currently (and potentially) living in these regions highlights the vital nature of research into the health hazard potential.

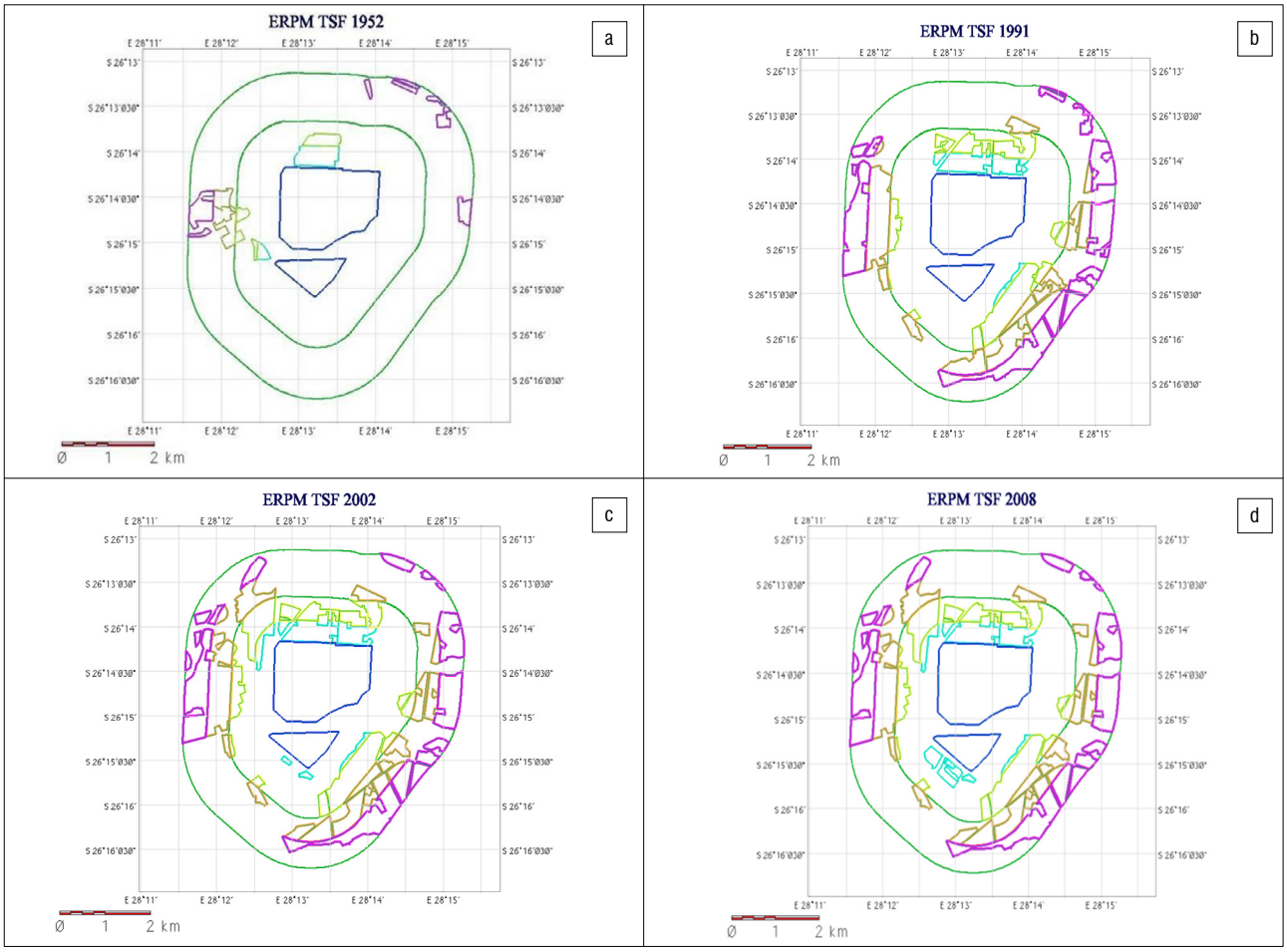


Figure 6: Time series progression of residential developments around the East Rand Proprietary Mine tailings storage facilities: (a) 1952; (b) 1991; (c) 2002 and (d) 2008, cyan ≤ 500 m; yellow between 500 m and 1 km; orange between l km and1.5 km; cerise between 1.5 km and 2 km; purple = tailings storage facilities (TSF) footprints.

Table 4: Households and residential areas very close to tailings storage facilities during 2001 and 2011, with estimates of population for the City of Johannesburg

Buffer distance (m)	Number of households 2001	Population estimate 2001	Number of households 2011	Population estimate 2011
50	1924	5585	960	3072
50–100	1889	5702	902	2886
100–250	8055	23 671	5427	17 366
250–500	14 365	45 729	15 081	48 259

Conclusion

The research detailed above indicates that continued human encroachment onto marginal lands poses a growing risk of human health consequences (because of the surge in residential dwellings near TSFs). Increasing hazard from inhalable dust around TSFs sites[8] requires continued monitoring, quantification and mitigation efforts.

The growth of residential housing developments on marginal land near tailings storage facilities has been appraised over a 50-year period, using historical aerial photographic images and contemporary satellite retrievals. The spatial evolution associated with the residential housing and the number of households, coupled with the population dynamics over time, has been analysed.

Residential units within a 2 km radius of the three researched TSFs: Crown Gold Recoveries complex, Durban Roodepoort Deep and East Rand Proprietary Mine now number 163 250, 46 731 and 24 669 households respectively (Table 3). The combined human population within a radius of 2 km, exposed to airborne dust from the TSFs, grew from 138 959 in 1952 to 585 600 in 2011. The implication of so many residents downwind of what is now potentially an important health risk pollutant source, justifies ongoing monitoring efforts, resource allocation for dust control measures (including water resources even in times of drought) and education programmes about the causes, prevention and treatment of silicosis.

A steady growth in housing developments edging closer towards the various tailings storage facilities has been confirmed by analysis of aerial and satellite images in conjunction with census data,[22,23] indicating the combined number of households within TSF buffer zones has increased from 43 381 in 1952 to 163 250 in 2004 (Table 3). The area covered by residential housing development has increased approximately 700% since 1952 at a rate of 14% per year. These changes have led directly to increased human exposure to increasingly hazardous inhalable dust from the TSFs. Ideally, no further development of residential areas should be permitted within a certain perimeter – i.e. within a distance of less than 2 km from the edge of the TSFs. Similarly, relocation of the current residential population would be desirable, but is unlikely to be feasible within the economic and political landscape. Socio-economic pressure, fostered by the need for low-income homes close to work opportunities, will continually cause pressure for further residential development. Hence, increased emphasis on preventing wind erosion of TSF is likely to be the most cost-effective solution to this increasing hazard – which currently affects half a million residents of South Africa.

Acknowledgements

Funding for this work was provided by Focus Area Grant No. FA2005040600018 from the South African National Research Foundation to H.J. Annegarn, and by the University of Johannesburg. We thank the Chief Directorate of Survey and Mapping in Cape Town, South Africa for providing the aerial photographic images and Statistics South Africa for the census data.

Authors' contributions

H.J.A. was the project leader, M.E.O. performed the image processing and M.A.K. performed the calculations and spatial analysis.

References

1. Parnell S. Building developmental local government to fight poverty: Institutional change in the city of Johannesburg. Int Dev Plann Rev. 2004;26(4):377–399. http://dx.doi.org/10.3828/idpr.26.4.3
2. Parnell S, Robinson J. Development and urban policy: Johannesburg's city development strategy. Urban Stud. 2006;43(2):337–355. http://dx.doi.org/10.1080/00420980500406710
3. Parnell S, Walawege R. Sub-Saharan African urbanisation and global environmental change. Global Environ Change. 2011;21(Suppl.1):S12–S20. http://dx.doi.org/10.1016/j.gloenvcha.2011.09.014
4. Schäffler A, Swilling M. Valuing green infrastructure in an urban environment under pressure – The Johannesburg case. Ecol Econ. 2013;86:246–257. http://dx.doi.org/10.1016/j.ecolecon.2012.05.008
5. Population Reference Bureau. World population data sheet. Washington, DC: Population Reference Bureau; 2007.
6. Annegarn HJ, Surridge AD, Hlapolosa HSP, Swanepoel DJDV, Horne AR. Review of 10 years of environmental dust monitoring at Crown Mines. J Mine Vent Soc S Afr. 1991;44(3):46.
7. Blight GE, Caldwell JA. Abatement of pollution from abandoned gold residue dams. J S Afr Inst Min Metall. 1984;84(1):1–9.
8. Ojelede ME, Annegarn HJ, Kneen MA. Evaluation of aeolian emissions from gold mine tailings on the Witwatersrand. Aeolian Res. 2012;3(4):477–486. http://dx.doi.org/10.1016/j.aeolia.2011.03.010
9. JJ Martins. A critical evaluation of the challenges facing dust management within gold mining regions of South Africa [MSc thesis]. Potchefstroom: North-West University; 2014.
10. Oguntoke O, Ojelede ME, Annegarn HJ. Frequency of mine dust episodes and the influence of meteorological parameters on the Witwatersrand area, South Africa. Int J Atmos Sci. 2013;3(4):602–611. http://dx.doi.org/10.1155/2013/128463
11. Williams DA, Jones H. Tailings storage facilities. Developments in Mineral Processing 2005;15:729–751. http://dx.doi.org/10.1016/S0167-4528(05)15030-3
12. Anglogold Ashanti. Anglogold Ashanti environmental report, Woodlands Project – Good Progress being made with Phytoremediation Project. Johannesburg: Anglogold Ashanti; 2004
13. Annegarn HJ, Zucchiatti A, Sellschop JPF, Booth-Jones P. PIXE characterization of airborne dust in the mining environment. Nucl Instrum Meth B. 1987;22(1–3):325–330. http://dx.doi.org/10.1016/0168-583X(87)90351-X
14. Hnizdo E, Murray J, Sluis-Cremer GK, Thomas RG. Correlation between radiological and pathological diagnosis of silicosis: An autopsy population based study. Am J Ind Med. 1993;24(4):427–445. http://dx.doi.org/10.1002/ajim.4700240408
15. Hnizdo E. Risk of silicosis in relation to fraction of respirable quartz. Am J Ind Med. 1994;25(5):771–772. http://dx.doi.org/10.1002/ajim.4700250517
16. Donaldson K, Borm PJA. The quartz hazard: A variable entity. Ann Occup Hyg. 1998;42(5):287–294. http://dx.doi.org/10.1093/annhyg/42.5.287
17. Belle B, Phillips HR. Influence of section return dust levels as an exposure indicator and assessment parameter of engineering control in coal mines. J Mine Vent Soc S Afr. 2009;62(4):6–9.
18. Belle BK, Du Plessis J. Recent advances in dust control technology on South African underground coal mines. J Mine Vent Soc S Afr. 2002;55(4):138–144.
19. Belle BK. Comparison of three side-by-side real-time dust monitors in a duct using average and peak display dust levels as parameters of performance evaluation. J Mine Vent Soc S Afr. 2010;63(1):10–17.
20. Gibbs GW, Du Toit RSJ. Estimating the quartz exposure of South African gold miners. Ann Occup Hyg. 2002;46(7):597–607. http://dx.doi.org/10.1093/annhyg/mef084
21. Chief Directorate of Survey and Mapping/National Geo-spatial Information. Aerial Photography (1926–2008). Cape Town: Department of Rural Development and Land Reform; no date.
22. Statistics South Africa. Census 2001. Pretoria: Stats SA; 2003
23. Statistics South Africa. Census 2011. Pretoria: Stats SA; 2011
24. Liebenberg-Enslin H. A functional dependence analysis of wind erosion modelling system parameters to determine a practical approach for wind erosion assessments [PhD thesis]. Johannesburg: University of Johannesburg; 2014.

Exploring the connections between green economy and informal economy in South Africa

AUTHORS:
Suzanne Smit[1]
Josephine K. Musango[1,2]

AFFILIATIONS:
[1]School of Public Leadership, Stellenbosch University, Stellenbosch, South Africa
[2]Centre for Renewable Energy Studies, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Suzanne Smit

EMAIL:
suzann3sm1t@gmail.com

POSTAL ADDRESS:
School of Public Leadership, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

DATES:
Received: 04 Dec. 2014
Revised: 04 Feb. 2015
Accepted: 07 May 2015

KEYWORDS:
informal green jobs; sustainable livelihoods; second economy; sustainable development

HOW TO CITE:
Smit S, Musango JK. Exploring the connections between green economy and informal economy in South Africa. S Afr J Sci. 2015;111(11/12), Art. #2014-0435, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20140435

The notion of an inclusive green economy in the context of sustainable development and poverty eradication requires an approach that engages with the informal economy. However, the informal economy is generally ignored or undervalued in discussions on the green economy. This paper set out to bolster this argument by identifying the ways in which the green economy and the informal economy may be connected by establishing the extent to which policies and plans relating to green economy connect with the informal economy, and recognising several informal green activities. The barriers and opportunities for connecting the two spheres were also explored as well as possible ways in which such activities may be supported at different levels of organisation. In the case of South Africa, many informal green activities that contribute to sustainable livelihoods are recognised. However, issues pertaining to procedure, process and participation hinder the transition to a truly inclusive green economy.

Introduction

The South African government recognises a green economy as providing potential to transition to a low carbon economy, resource efficiency and the creation of pro-poor jobs.[1,2] Generally, the concept of green economy is not new but a re-emerging one, and it is gaining prominence in international and national policy debates.[3-5] The United Nations Environment Programme[4] defines the green economy 'as one that results in improved human well-being and social equity, while significantly reducing environmental risks and ecological scarcities'. According to Death[6], there are four green economy discourses that are theoretically distinct, but which in practice tend to overlap. These discourses are green revolution, green transformation, green resilience and green growth. Some studies in the South African context have focused on only one of these discourses. For instance, Resnick et al.[5] focused on green growth and utilised four case studies in Southern Africa; Swilling et al.[7] specified some studies that could be classified in each of the four different discourses.

Death[6] argues that these four discourses are characteristic of South Africa and that 'large-scale structural transformation' will be required to orient the country towards green economy. While Death's[6] view is agreeable to many, Swilling et al.[7] differ with some arguments and contend that: (1) green transformation in a developing country would require some element of growth; (2) green economy should be part of wider sustainability transition that is not about decarbonisation and resource efficiency, but also about redistributive measures that reduce inequality and poverty.

At Rio+20, the green economy concept was further expanded to represent the social dimension of sustainable development leading to the notion of an inclusive green economy[8] as taking place 'in the context of sustainable development and poverty eradication'[9]. The implication for an inclusive green economy is that both the formal and the informal economy need to be prioritised in green economy discussions.

It is thus within the context of finding redistributive measures as argued in Swilling et al.[7] and as representing the social dimension as per the Rio+20 outcome document, that this paper relates to, and argues the need for connecting the green economy with the informal economy.

There are a number of definitions pertaining to the informal economy that impact on data collection and measurement of its size and monetary value.[10-13] The following definition by the International Labour Organisation[14], is preferred as it recognises the heterogeneity of activities occurring in the informal economy that cut across different sectors. Accordingly, the informal economy is recognised as:

> *…economic activities by workers and economic units that are – in law or in practice – not covered or insufficiently covered by formal arrangements. Their activities are not included in the law, which means that they are operating outside the formal reach of the law; or they are not covered in practice, which means that – although they are operating within the formal reach of the law, the law is not applied or not enforced; or the law discourages compliance because it is inappropriate, burdensome, or imposes excessive costs.*[14]

Literature pertaining to the green economy and the informal economy indicate that the latter is generally undervalued or ignored in discussions on the former.[15,16] It is further argued that the informal economy 'offers both opportunities and lessons on resilience and innovation' and that connecting the two is imperative in the context of sustainable development and poverty eradication.[15]

To the best of our knowledge, there are limited theoretical and empirical studies that connect the green economy with the informal economy. Smit and Musango's[15] study seems to be among the initial studies providing such an understanding in the South African context. In their paper, Smit and Musango[15] critically review the macro perspective and understanding of the green economy landscape, and how these relate to the reality of a green economy in South Africa. They deepen this understanding by further tracing various perspectives and approaches to 'managing' the informal economy and identify the role and value of the informal economy for sustainable development and a green economy.

Based on their critical review, Smit and Musango[15] argue that an inclusive green economy requires the conceptualisation and implementation of a green economy that engages with the informal economy. By overlooking the informal sector, policy and planning initiatives are not fully engaged with the real everyday experiences and activities of those living in the survivalist economy, leading to a narrow conception of what a green and inclusive economy may be. The informal economy, however, is a very real phenomenon that warrants greater inclusion and consideration in planning for sustainable development.

We set out to explore the range of possible connections between the green economy and the informal economy within the context of a developing country, taking South Africa as a case. Although the informal economy is markedly smaller in South Africa than in the rest of Africa,[17] it is no less valuable to an expanded and contextually relevant green economy policy environment. The need to engage the informal economy is especially pressing given that South Africa has to consider transitioning towards a low carbon path whilst curbing the high levels of inequality and unemployment.[18,19] Thus, in this paper, the key questions explored, in connecting the green economy to the informal economy are: (1) To what extent does green economy policies and plans connect with the informal economy? (2) What 'green' activities are taking place within the informal economy that contribute to sustainable livelihoods? (3) How can these activities be supported and developed to expand the current planning and policy environment?

Methods and tools

This paper was informed by grounded theory methodology; we used multiple methods, thereby allowing for a more holistic understanding of the problem. The initial literature analysis by Smit and Musango[15] indicated a gap in terms of research that connected the informal economy with the green economy, leading to the proposition that a new theory would have to emerge from the study itself. In this way, the study lent itself to the grounded theory methodology. Data collection and data analysis tools utilised are discussed in the sections that follow. A detailed step-by-step process can be found in Smit[20].

Content analysis

The content analysis involved identifying and analysing a set of fourteen national policies and plans in South Africa that relate to the green economy in order to establish the extent of informal economy coverage and to identify the informal green activities therein. According to Babbie[21(p.309)], content analysis 'is particularly well suited to the study of communications' in order to establish what is said, to whom, why, how and with what effect. In the case of policies and plans relating to the green economy in South Africa, the aim was to ascertain the level and quality of engagement with the informal economy. The units of analysis were appropriately identified as those documents that guide national, regional and local interventions for a green economy as recognised by Montmasson-Claire[22], Musango et al.[2] and Sustainlabour, COSATU and TIPS[23]. The sampling method involved reducing the scope as per the limitations of the study to appropriate national policies and plans created in the period 2008–2013. These included the following:

- Ten Year Innovation Plan 2008–2018[24]
- Framework for South Africa's response to the international economic crisis[25]
- Medium Term Strategic Framework 2009–2014[26]
- Integrated Resource Plan (IRP) 2010–2030[27]
- National Skills Development Strategy III 2011–2016[28]
- National Climate Change Response White Paper[29]
- National Waste Management Strategy[30]
- New Growth Path, Green Economy Accord[31]
- National Strategy for Sustainable Development and Action Plan (NSSD) 2011–2014[32]
- National Development Plan (NDP) 2030[33]
- Draft 2012 Integrated Energy Plan[34]
- Bioeconomy Strategy[35]
- National Water Resource Strategy[36]
- Industrial Policy Action Plan 2014[37]

In order to extract the relevant data, each key document was evaluated on the number of times that particular terms occurred throughout each document, referred to as evaluating the manifest content.[21] These terms were initially coded as relating to direct references to the informal economy (Figure 1); however, through the coding process, it became apparent that it was necessary to consider a wider formulation of the concept of informal economy based on its defining characteristics. Such occurrences were coded as indirect references to the informal economy (Figure 2). The resulting data were then coded and analysed using ATLAS.ti.

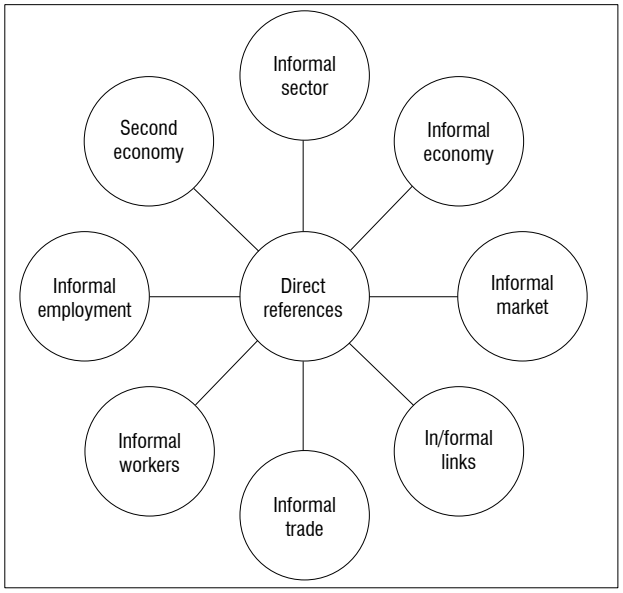


Figure 1: Network view of terms considered direct references to the informal economy

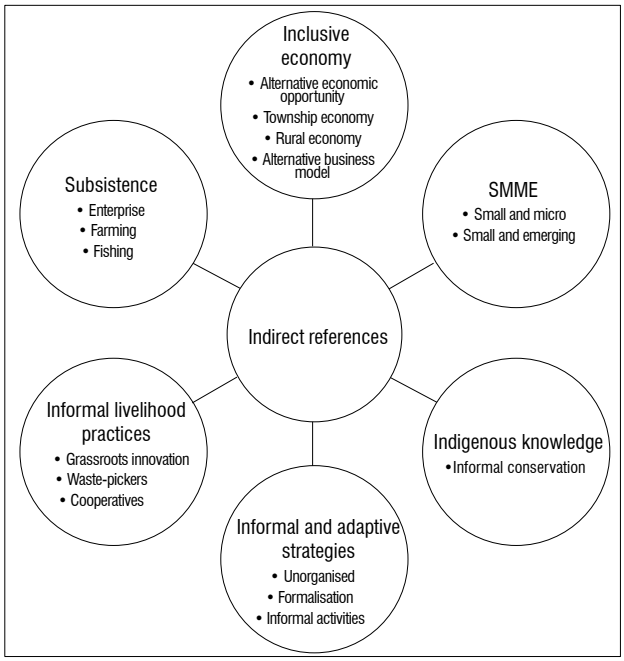


Figure 2: Network view of terms considered indirect references to the informal economy.

Table 1: Ranking and frequency of direct and indirect references to the informal economy

Policy document or plan	Direct references	Indirect references	Totals	Overall reference %	Placement
National Development Plan[33]	33	76	109	44%	1
Industrial Policy Action Plan[37]	3	21	24	10%	2
Bioeconomy Strategy[35]	2	16	18	7%	3
National Water Resource Strategy[36]	0	16	16	7%	4
Medium-Term Strategic Framework[26]	2	12	14	6%	5
National Growth Path, Green Economy Accord[31]	1	12	13	5%	6
National Waste Management Strategy[30]	0	11	11	4%	7
National Strategy for Sustainable Development and Action Plan[32]	0	11	11	4%	8
National Climate Change Response[29]	1	8	9	4%	9
National Skills Development Strategy[28]	0	9	9	4%	10
Ten-Year Innovation Plan[24]	0	5	5	2%	11
Framework for South Africa's Response to the Economic Crisis[25]	1	2	3	1%	12
Draft Integrated Energy Plan[34]	1	2	3	1%	13
Integrated Resource Plan[27]	0	0	0	0%	14
TOTALS:	**44**	**201**	**245**	**100%**	

Semi-structured interviews

Semi-structured interviews were conducted with experts in the field of the informal economy as well as those involved in facilitating, planning or implementing green economy policies, plans, projects or activities.

Initially, interviewees were identified through networking opportunities at green economy meetings or workshops and subsequent candidates were identified from mailing lists obtained from the organisers of some of the events. The sampling method therefore involved nonprobability sampling, utilising purposive or judgemental sampling in combination with snowball sampling.

A list of pertinent open-ended and closed questions was formulated and posed, yet the semi-structured nature of the questions allowed for deeper probing and exploration (these questions are available in Appendix 1 as online supplementary material). The responses to these questions were recorded and coded. The coding process was both inductive and deductive. Inductive coding related to codes and concepts as emerging directly from the interview data, whereas deductive coding related to codes and concepts that emerged from the survey results, content analysis and observations and which were then searched for reiteratively within the interview data.

Surveys

An online questionnaire was circulated to a number of potential informants ranging from academics, to NGO practitioners and national, provincial and municipal level government officials (the questionnaire is available in Appendix 2 as online supplementary material). The questionnaire was designed to extract data relevant to each of the three research questions and therefore consisted of multiple question types. A number of questions lent themselves to quantitative data analysis through descriptive statistics, such as frequency tables and bar graphs, whilst the open-ended questions were considered to constitute qualitative data requiring computer-assisted qualitative data analysis using ATLAS.ti and were presented either as matrices or, where appropriate, as direct quotations.

Respondents were also asked to rate the level of engagement between a list of policies and plans relating to the green economy and the informal economy on a scale including Very High, High, Moderate, Low and Non-existent. Participants were asked to only rate those policies that they were familiar with and were given the option of choosing 'Not Applicable' in order to indicate the policies that they were unfamiliar with. The valid responses per policy were presented as a frequency table. The valid response count was calculated by deducting the number of 'Not Applicable' responses from the response count per policy.

Direct observation

Within this study, observations were obtained with different levels of participation, including attendance of four green economy related workshops or meetings ranging from national to local level. Direct observations were also made of green activities taking place in two informal settlements, thereby triangulating data from the interviews and questionnaires with observations from the field.

Results and discussion

Extent of engagement with the informal economy

Policies and plans relating to the green economy 'set the tone', so to speak, for understanding the extent of engagement between the green economy and the informal economy. Table 1 represents the results from the content analysis. The categories 'Direct references' and 'Indirect references' were considered to be indicative of the level and quality of engagement between a particular policy or plan and the informal economy.

Based on the combined direct and indirect references to the informal economy, the policies and plans were ranked in terms of their comparative level of engagement. The results indicate the National Development Plan[33] as the plan with the most overall references and the Integrated Resource Plan[27] as the plan with the fewest references to the informal economy.

Table 1 further illustrates that there are significantly fewer direct than indirect references to the informal economy, which may be indicative of a lack of conceptual or definitional understanding of what the informal economy is. Alternatively, it may signal that the informal economy is not recognised as having value within the green economy. For example, the Green Economy Accord[31], ranked sixth, made only one

direct reference to the informal economy, even though the Economic Development Department has a mandate in terms of developing the 'second economy'.

Assuming that direct references are preferable to indirect references in terms of recognising and engaging with the informal economy, several policies were found to be problematic, in particular the National Strategy for Sustainable Development[32], which ranked eighth yet aims to build sustainable communities and work towards a green economy. The Ten-Year Innovation Plan[24] was ranked 11th and made several indirect references, yet all of the references pertained to indigenous knowledge systems only, and therefore this plan does not explore other possibilities within the informal economy.

There also seems to be an overall disconnect between policies and plans that drive the green economy and the informal economy. The low level of engagement is evidenced by the lack of overall reference to the informal economy (that is either direct or indirect). In contrast to the majority of the policies and plans, however, the National Development Plan[33] ranked first and stands out as recognising and engaging with the informal economy. The quality of its engagement is also evidenced by its identification of challenges and opportunities relating to the informal economy as well as its recognition of the need for a more inclusive economy.

In order to triangulate the findings from the content analysis process, participants in the online questionnaire were asked to rate a set of policies and plans in terms of each policy's engagement with the informal economy (Table 2). Respondents were asked to rate only those policies that they were familiar with. The valid response count therefore took into account the number of people who responded to each question less those who were unfamiliar with the particular policy. The results therefore pertain to the rating by respondents who were familiar with each policy.

From Table 2, the following observations are pertinent. The National Development Plan[33], which most participants (12 out of 14) were familiar with, was rated by the majority of respondents (8 out of 12, or 67%) as having a 'High'' (25%) to 'Very High' (42%) level of engagement with the informal economy. This finding supports the results from the content analysis, which found the National Development Plan[33] to have the highest number of direct and indirect references to the informal economy as compared to the other policies in the sample. Data from the semi-structured interviews, however, rate the National Development Plan[33] at a moderate level of engagement. Interviewee 1 explained that on first reading, the National Development Plan[33] seemed to have little engagement with the informal economy but on more detailed reading and considering more 'holistic comments', the level of engagement was more moderate. Interviewee 1 also indicated that the National Development Plan[33] seemed to designate responsibility for the informal economy at local government level.

A conflicting result, however, relates to the National Strategy for Sustainable Development[32], which was found in the content analysis to have a surprisingly low number of references to the informal economy, whereas the survey indicated that more than half of the respondents (7 out of 11 or 63%) rated it as having 'High' (36%) to 'Very High' (27%) engagement.

In terms of the Integrated Resource Plan[27], the content analysis indicated that there were no references to the informal economy, and although more respondents (5 out of 11 or 45%) rated it as having 'Low' (36%) to 'Non-existent'' (9%) engagement, there were some respondents (4 out of 11 or 36%) who considered it to have 'High' (27%) to 'Very High' (9%) engagement.

The Green Economy Accord[31] was rated by more than half of the survey respondents (7 out of 12 or 58%) as having 'High' (33%) to 'Very High' (25%) engagement. This accord was rated 'Moderate' by 25% of the respondents and only 16% rated it as 'Low' (8%) to 'Non-existent' (8%). In comparison, the content analysis indicated that the Accord made only 13 overall references to the informal economy, which placed it sixth in the overall ranking of policies. This discrepancy in the comparison between the respondents' perception and the results from the content analysis may be indicative of the divide between the expectations and the reality of an inclusive green economy.

Table 2: Results for question 16 on rating green economy policies' engagement with the informal economy

Question 16: The following is a sample of national policies and plans related to the green economy in South Africa. How would you rate each policy or plan in terms of its level of engagement with the informal economy? Please only rate the policies and plans that you are familiar with.

Answer options	Very high	High	Moderate	Low	Non-existent	Valid responses
Framework for Environmental Fiscal Reform	0%	22%	22%	44%	11%	9
Ten-Year Innovation and Global Research Plan	11%	11%	22%	44%	11%	9
Framework for South Africa's response to the International Economic Crisis	10%	30%	30%	20%	10%	10
Medium-Term Strategic Framework 2009–2014 and 12 Outcomes	11%	44%	22%	11%	11%	9
Industrial Policy Action Plan	27%	18%	27%	18%	9%	11
New Growth Path, Green Economy Accord and Green Jobs Report	25%	33%	25%	8%	8%	12
Integrated Resource Plan 2010–2030	9%	27%	18%	36%	9%	11
Draft Integrated Energy Plan	9%	27%	9%	45%	9%	11
National Water Resource Strategy	10%	30%	20%	30%	10%	10
National Climate Change Response	30%	10%	40%	10%	10%	10
National Strategy for Sustainable Development	27%	36%	18%	9%	9%	11
National Skills Development Strategy III: 2011–2016	18%	27%	36%	9%	9%	11
National Development Plan	42%	25%	17%	8%	8%	12
					Answered question	**14**
					Skipped question	**7**

Green activities in the informal economy

From the second question, we aimed to identify 'green' activities taking place in the informal economy and to gauge the range of possible connections of the green economy to the informal economy.

An overwhelming majority of survey participants (15 of 17) and all interviewees (9 of 9) agreed that there are green activities taking place in the informal economy. These activities are described as mainly being driven by NGOs, and researchers or community organisations (Interviewee 3, 4 and 6). Although a number of projects have been initiated by different levels of government, Interviewees 3, 4 and 6 indicated that the level of commitment and dedication to such government-run projects is less than desirable. Regardless of outside intervention, Interviewee 5 suggested that some informal economy activities may even be 'greener' than their formal counterparts.

However, the notion that green activities occur in the informal economy is not uncontested. From the survey, 88% (15 of 17) of respondents agreed that there are green activities taking place in the informal economy, whereas 12% (2 of 17) disagreed. One such contestation revolved around the lack of regulation whereas another related to the motivation or intent of informal workers to be green. In contrast to this view, Interviewee 3 suggested that there is willingness and an understanding by some informal actors to participate in green activities. Interviewee 9, an informal settlement resident, demonstrated this willingness and understanding by describing the problems with 'earth and air pollution' and suggesting 'we need to do something about it ourselves'.

When asked to describe green activities taking place in the informal economy, online survey respondents identified the following: (1) subsistence farming; (2) bio-processing, management and trade; (3) recycling; (4) waste picking; (5) waste to compost activities; (6) use of renewable energy and other fuel saving innovations; and (7) rainwater harvesting.

Green economy activities in national policies and plans

A number of the green activities taking place in the informal economy were also recognised by national policies and plans related to the green economy as well as by green economy and informal economy practitioners who were interviewed.

Bioprospecting and biotrade

The Bioeconomy Strategy[35] makes several references to bioprospecting as connected to indigenous knowledge systems, whilst the Ten Year Innovation Plan[24(p.11)] illustrates the connection as an 'interface between biodiversity heritage, indigenous knowledge systems and new development in genomics'. The advantages of bio-composites are described in the Industrial Policy Action Plan[37(p.128)] as including 'a reduced environmental footprint, reduced energy consumption in production, and health and safety benefits' as well as offering 'potential for new enterprise development and job creation, including in rural communities'.

Recycling and waste picking

Within the waste sector, recycling and waste picking activities are recognised as contributing to the green economy and as involving the informal economy. The National Waste Management Act, for example, aims to 'stimulate job creation and broaden participation by SMEs [small and medium enterprises] and marginalised communities in the waste sector'.[30(p.27)]

Waste collection and recycling was recognised by the majority of survey respondents (7 of 11) and all interviewees (9 of 9) as providing 'informal green jobs' making it the most recognisable of the activities. There is also much potential for further exploration, for example, through waste to compost activities and waste to fuel projects (Respondent 3). Interviewee 2 also suggested that within the value-chain for the recycling of waste in South Africa, certain levels of activity are already taking place within the informal economy with much scope for expansion.

Field observations indicated that some recycling activities are taking place within informal settlements. The recycling project at Site B is driven by a single member of the community. The recyclable material is collected from the individual households and then stored until it can be collected by a formal recycling business. The level of recycling however varies with community cooperation and is therefore both sporadic and small-scale in nature. From the issues pertaining to the level and quality of recycling in informal settlements, there may be a need for greater incentives and support for those individuals and communities that want to expand their efforts.

A further connection could also be made between recycling activities and the craft sector. The Industrial Policy Action Plan[37] makes no explicit connection between the craft industry and the green economy, however, it does place emphasis on the benefits of localisation. The craft industry is described as being predominantly small-scale and informal, thus linking it to the informal economy. The potential here lies in connecting it with recycling activities, by using recycled materials to create marketable artefacts. For example projects such as Trashback and the Waste to Art Market (Respondent 6), which connect to the informal economy, localisation and the green economy.

Small-scale and subsistence farming

Along with several survey respondents (Respondent 1, 6, 7), Interviewees 1 and 2 described subsistence farming as green activities taking place in the informal economy. The connection between subsistence or small-scale farming and climate change is also made explicit by the National Climate Change Response Paper[29(p.23)], which proposes the following response:

> *Educate subsistence and small-scale farmers on the potential risks of climate change, and support them to develop adaptation strategies with on-farm demonstration and experimentation. Adaptation strategies will include conservation agriculture practices including water harvesting and crop rotation, and will prioritise indigenous knowledge and local adaptive responses.*

In the context of climate change, the National Development Plan[33(p.89)] emphasises the need for innovative responses whilst also referring to the role of subsistence farmers:

> *The Alliance for a Green Revolution in Africa, for example, is working to achieve food security for Africa by promoting sustainable agricultural growth through smallholder farmers. Recognising that smallholders – mostly women – produce most of Africa's food today with minimal resources and little government support...*

As noted by the final point above, there is a disconnect between agricultural policies and the informal economy. Interviewee 1 describes the food system in South Africa as being 'managed and driven in a very entrepreneurial way, [and] the State is very absent from any of those structures'. The implication is that there are 'very few policies structurally that seek to enable the informal economy'.

Interviewee 2 suggested that a lack of support for small-scale farmers may have an impact on food security, and therefore have wider repercussions. In order to address the disconnect from the informal economy and to promote wider food security, Interviewee 1 proposed that the city formally adopts a food strategy that engages with the food system across all of its complexity; and that it adapts and adopts a city-level mandate from the bottom up. This strategy may involve alternative processes where citizens come together to assert a different kind of agency in terms of city governance. A third component revolved around the criminalisation of certain activities, but which are not necessarily acted upon or enforced, because it is recognised that action could have negative effects on the wider and often impoverished communities they serve. As indicated by Interviewee 1, there are, however, numerous repercussions and complications associated with this final point that should also be acknowledged.

Small-scale and subsistence farming activities were observed in both the informal settlements (Site A and Site B) (Figures 3 to 5). The type, quantity and quality of produce varied according to plot size and aspect, season, soil quality, access to tools, and individual and financial constraints. In general, all

food was grown for personal consumption by the individual households and the surplus then either sold or donated to the community members.

Photo: Suzanne Smit

Figure 3: Household level food production (Site A).

Photo:Shanelle de Villiers

Figure 4: Marginally larger scale food production (Site B).

Photo: Shanelle de Villiers

Figure 5: Community level food production (Site B).

Green infrastructure, technology and construction

One way of connecting the green economy with the informal economy is through the introduction of green infrastructure and technologies. The use of renewable energy, rainwater harvesters, bio digesters, construction through eco-design, and bio-mimicry for solid and waste water treatment were recognised as projects taking place in a number of informal settlements (Interviewees 4, 6, 7 and 9). Small-scale solar installations were observed as being promoted and trialled in both informal settlements in order to help address the lack of access to energy.

Another energy related project observed at both sites incorporated the use of a bio-digester. Principally, the gas emissions from human waste are captured and transformed into gas for cooking, thereby creating a more sustainable form of energy. This project may provide several social, economic and environmental benefits, but community buy-in at Site B was problematic. Observations before and after installation indicated that there were certain benefits to upgrading the community toilets and although residents were initially excited by and involved in the upgrading process; one year later, only one toilet and bio-digester unit remains operative, partly because of community apathy and interpersonal conflict.

The difficulties associated with green projects in informal settlements reiterate the need for greater and better quality engagement with informal actors and entities. Nonetheless, if handled well, such projects may produce several social and economic gains in addition to environmental benefits. A number of interviewees indicated that green projects have the potential to create informal green jobs and also to assist municipalities with their infrastructure and resource constraints (Interviewees 4, 6 and 7). 'Decentralised infrastructure is about a completely different model of management and ownership' (Interviewee 7), and contains a variety of opportunities in relation to 'upcycling'; creating economic opportunities and assets that the poor can capitalise on; and can positively impact people's living conditions (Interviewees 4, 6 and 7).

It is felt that there is much scope for the informal economy in the green economy 'due to the number of sectors of work that could be included beyond just the construction industry' (Interviewee 2) and in terms of 'unlocking new value streams' (Interviewee 4). The job creating potential of green infrastructure may therefore be a topic for future study as well as a descriptive study of the number of informal worker groups in the green economy, their make-up, and the social, environmental and economic impact that these technologies have for those working and living in informal conditions.

Mechanisms to improve current planning and policy environment

The third research question aimed at recognising the ways in which the current planning and policy environment could improve. Therefore, it was deemed necessary to first explore both the challenges and the opportunities associated with connecting the green economy to the informal economy. Combining results from the online survey and the semi-structured interviews, the following SWOT analysis (Table 3) identifies the range of strengths, weaknesses, opportunities and threats pertaining to connecting the green economy to the informal economy in South Africa.

The SWOT analysis considered a range of factors or conditions that may either help or hinder such connections (Table 3). The strengths and weaknesses relate to conditions within government or organisational control, whereas the opportunities and threats relate to factors outside or external to direct control.

From Table 3 it may be surmised that there are two types of conditions that negatively affect connections between the green economy and the informal economy. These connections may be considered as either related to organisational processes, structures and procedures, which fall within the purview of an organisation; or to individual and cultural differences, which fall outside the scope of organisational control. Table 3 also indicates that the benefits of connecting the green economy to the informal economy may be considered as intentional objectives, whereas the opportunities relate to indirect or unintended consequences.

Many of these benefits and opportunities are interconnected and cut across the three pillars of sustainable development, that is, the social, economic and environmental spheres. These benefits and opportunities align with international[4] and national[28] definitions of green economy and, it may be argued, echo the ideals of an inclusive green economy. Connecting the green economy to the informal economy in South Africa is recognised by participants as contributing to an inclusive green economy and therefore warrants the call for greater engagement between the informal economy and the green economy in the context of sustainable development and poverty eradication.

Support for connecting green economy and informal economy

Beyond identifying positive and negative factors associated with connecting the green economy to the informal economy, survey respondents and interviewees were asked to identify how informal green activities taking place in the informal economy could be supported. The following section describes these support mechanisms at different levels of organisation.

Table 3: SWOT analysis of connecting green economy to informal economy

Strengths (benefits)	Weaknesses (barriers)
• Environmental benefits – reduction in waste – improved ecosystem services – resource-use efficiency – reduced carbon emissions • Socio-economic benefits – Decent work – Poverty eradication • Job creation • Localisation • Food security • Development goals – Improved health and wellbeing – Access to basic services – Reduced inequality – Increased skills and level of education • Economic benefits – new value streams – inclusive growth – reduced welfare payments	• Horizontal and vertical co-ordination in terms of both policies and practices • Shortage of funding and difficulties in accessing funds • Capacity or ability (manpower and time) to engage • Power relations – representational issues – democratic processes – structural issues – marginalisation – crony-ism – corruption • Bureaucracy – lack of flexibility or adaptability – over/under emphasis on regulation • Skills shortages in terms intellectual ability to engage • Competition for resources • Lack of mandate, or vision to engage • Conceptual and definitional uncertainty – regarding the informal economy and how to 'manage' it – regarding the green economy and how to 'achieve' it • Lack of models for successful engagement or modes for bridging the informal and formal
Opportunities	**Threats (challenges)**
• Alternative economic opportunities and/ or systems • Social economy • Barter system • Circular economy / closed loops • Inclusivity • Human dignity and empowerment • Sustainable livelihoods • Bridging the human-nature divide	• Interpersonal collaboration and willingness to engage • Politics – inter-personal – inter-group – intra-group • Personality, ego and attitude – autocratic – self-congratulatory – individualistic • Cultural differences

Local government or city level

A number of challenges were recognised at municipal level that impede a supportive or enabling environment for engaging the informal economy. These challenges included procedural and bureaucratic processes; lack of vision or leadership; lack of long-term commitment; lack of transparency; various forms of internal and external politics; competition for resources; lack of capacity and/or capability; pressure to conform and comply; and individual and interpersonal issues (Interviewees 1 to 9).

For a number of participants, recognition of and engagement with the informal economy were key support mechanisms, particularly in terms of inclusivity. Other support mechanisms related to infrastructure provisioning, improved coordination, funding, and facilitating processes for formalisation and organisation (see Table 4).

The notion of organisation was a recurring theme underlying the process of engagement. For some, the 'lack of organisation' in the informal economy was a major challenge. Several interviewees (1, 2 and 5) however indicated that the informal economy in general is not 'unorganised', as there may be a number of organisational structures at any given time, for example, those based on gender, seniority, and tribal or political affiliation. At the same time, however, it was recognised that not all forms of organisation were conducive to wider or meaningful engagement. It would seem that both the level and type of organisation were important factors in determining the level and quality of engagement between green economy practitioners and the informal economy.

Organisational and procedural challenges aside, personal perceptions and attitudes may also play an important role in the quality and level of engagement between municipalities and informal actors. Interviewee 6 pointed out that 'it requires a specific mind-set or attitude to make something work in an informal settlement', a mind-set which several interviewees (2, 5, 6 and 7) felt was very much absent at local level. Support mechanisms at local level were also identified as relating to the capacity and capability of municipalities to deal with complexity; to reconcile formality with informality; and to develop new models of engagement (Interviewees 1, 4, 6 and 7).

Provincial government

As previously described, some of the main barriers for connecting the green economy to the informal economy are related to lack of capacity, mandate and awareness at local level. Many respondents therefore indicated that provincial level support mechanisms need to address such challenges (Table 5).

Table 4: Local level support mechanisms that could contribute towards a more inclusive economy

Respondent	Response
1,6	Provide infrastructure to facilitate informal green activity
1	Build relationship with civil society groups
2,3,7,9	Engagement and co-ordination – building connections or partnerships among the private sector, industry, municipal departments and informal actors
4,12	Recognise role and value of informal sector workers in green services and promote their involvement
10	Facilitate, recognise and support co-operatives
8,9,10,12	Facilitate formalisation and organisation
7	Funding of projects

Table 5: Provincial level support mechanisms that could contribute towards a more inclusive economy

Respondent	Response
1,2	Skills development
1,6	Capacity building – knowledge sharing, providing expertise
1,6,10	Network forming, facilitating partnerships at higher level
1,8	Funding
1	Piloting of new innovations
3	Create mandate
4,6	Exercise influence over industries and promote supply chain contracts to procure from informal economy
6	Monitor and evaluation
7,8	Support facilitation of formalisation, engagement and dialogue with informal workers

It was also suggested that provincial governments have a responsibility to assist in bridging the gap between national policy and local practice and to recognise and support the informal economy, particularly in relation to facilitating partnerships and networks at a national level.

National government

Some of the barriers and challenges for connecting the green economy to the informal economy at national level related to lack of interdepartmental cooperation, lack of coordination and alignment of policies and in some cases, lack of capacity or willingness to engage with the informal economy (Interviewee 1, 2 and 5). Survey responses indicated the need for improving the regulatory environment, facilitation of partnerships at various levels (including at international level) and facilitation of cooperatives, as key support mechanisms (Table 6).

As previously indicated, there is a need for greater consistency in the interface between polices and/or practices by national government. According to Interviewee 2, in contrast to current 'top-down tendencies' support mechanisms may include a 'type of multi or inter ministerial structure' which is mandated to engage with the informal economy. A second supportive measure included recognising the different informal worker groups operating in the green economy (Interviewee 5). It is worth noting that the former support mechanism may assist in horizontal coordination and the latter with vertical coordination. For example, Interviewee 5 explained that identifying the different worker groups 'helps to navigate [the complexity]' and that 'by understanding it by worker group and then thinking it through per worker group' will improve sectoral or vertical coordination.

Table 6: National level support mechanisms that could contribute towards a more inclusive economy

Respondent	Response
1,6	Financial subsidies and incentives
2,6,9	Facilitate partnerships (at various levels)
4,6,9	Improve regulatory environment
6	Supply chain contracts to procure from informal economy
6	Build capacity
6	Improve reporting and evaluation
7,8,9	Facilitate co-operatives
10	Develop strategies that will support the informal industry/economy

NGOs or civil society groups

Several interviewees (1, 2, 3, 4 and 5) recognised the role of civil society groups or NGOs in connecting the informal economy to the green economy in South Africa. A number of support mechanisms pertaining to NGOs were recognised by survey respondents (Table 7).

Table 7: NGO level support mechanisms that could contribute towards a more inclusive economy

Respondent	Response
1	Skills development
1	Network forming
1	Knowledge sharing
2	Support local initiatives
3,7	Society to be proactive and take responsibility for its environment
4,10,11	Mobilise around value-adding activities, circular economy and localisation and enter into environmental forums
6	Advocate for principles of sustainability
7,8,11	Formalise, organise or create co-operatives
9	Engage in dialogue, communication and reaching groups not reachable by public authorities

Key support mechanisms reported by survey respondents related to the facilitation of organisations within the informal economy and the mobilisation of NGOs around green activities. In previous discussions on connecting the green economy with the informal economy, the focus fell on green economy practitioners and their role in engaging informal actors. As indicated above, and as reiterated by Interviewee 5, there is also a need for NGOs that represent and advocate for the informal economy to engage with the actors and entities promoting the green economy.

Such engagements would require practitioners to reconsider the green economy space in terms of the opportunities it may hold for informal economy actors (Interviewee 5). NGOs can further contribute to a more

enabling policy and planning environment by assisting in 'mapping the size of the opportunity to enable further interventions' (Respondent 6); identifying different informal worker groups that operate in the green economy; and by participating at different levels, local to international, in discussions on the green economy (Interviewee 5).

Although NGOs may offer much assistance in connecting the green economy to the informal economy, they may also unwittingly impede such processes through poor representation; ideological differences; conflicting agendas and the tendency to 'roll-out a blueprint formula that works in some places' but that does not take context into account (Interviewees 1, 3, 4 and 5). These challenges therefore also warrant further attention if a more enabling policy and planning environment is to ensue.

Private sector organisations

Finally, survey respondents were asked to consider the role of private sector organisations in contributing to a more inclusive green economy and several support mechanisms were identified (Table 8).

Table 8: Private sector level support mechanisms that could contribute towards a more inclusive economy

Respondent	Response
1,3,6,7,11	Funding, investment
1,6,10	Skills transfer, mentoring
2,3	Infrastructure support
2,7	Start-up or incubation support
6,7,8,12	Partnership, value chain development
8,9	Supporting formalisation
9,10	Employ best practices and ensure informal suppliers adhere to decent work criteria

The most pertinent mechanisms related to formalising relationships through partnerships and investment in small and micro enterprises in terms of both financial support and skills transfer, as well as the development of broader value-chains that incorporate informal actors or entities. The notion of an expanded value-chain was also supported by a number of interviewees. For example, Interviewee 4 pointed to opportunities in the green infrastructure and technology sector, whereas Interviewee 2 referred to opportunities within the waste sector.

Overall, support mechanisms for connecting the green economy to the informal economy would require a paradigm shift in terms of what an inclusive green economy is; who the stakeholders and role-players are; as well as in addressing how and to what extent they are engaged.

Conclusions

We explored the connections between the green economy and the informal economy in the South African context using multiple methods. Results from observations, an online survey and interviews indicated that the informal economy is mostly invisible, in part because of a lack of conceptual consensus and interpretation but also as a result of insufficient recognition or awareness of the informal economy's role within the green economy space or how to engage with it successfully.

The paper revealed a number of possible connections between the green economy and the informal economy in the form of green activities, indicating that the informal economy may add much value to an inclusive green economy in the context of sustainable development and poverty eradication. The results also indicated a number of opportunities and benefits related to connecting the green economy to the informal economy, which warrant further investigation.

In terms of improving, enabling or supporting connections between the green economy and the informal economy, the survey and interview of several green economy and informal economy practitioners and policy influencers revealed a number of support mechanisms. These mechanisms were identified as addressing several barriers and challenges at various levels of organisation, and if incorporated, may assist in bringing about more empowering and inclusive practices or enabling policy and planning environments.

Although every effort was made to enhance the validity and reliability of the study through triangulation, the size and quality of the sample may be improved in terms of greater representation of informal economy actors or entities. Greater representation would involve cutting across a greater geographical region with varying conditions as related to climate, local bylaws, and social and economic practices and procedures, to allow for broader assertions and comparisons between provinces. It is further recommended that future studies include a focus on (1) profiling informal worker groups within the green economy; (2) exploring 'organisation' in the informal economy; (3) the role of NGOs and collectives in the green economy; (4) government process and change management for an inclusive green economy; and (5) a value chain analysis of green economy opportunities at the bottom of the pyramid.

Although the list of possible future research may be infinite, the implications of the study are relatively clear: for South Africa to achieve an inclusive green economy in the context of sustainable development and poverty eradication, it needs to engage with the actors and entities operating in the informal economy. To do so, both informal economy and green economy practitioners and policy advisors would need to apply a new 'lens' for considering the informal economy and its contribution to the green economy. This new lens involves a paradigm shift in terms of how the informal economy is perceived and approached, as well as pertaining to what an inclusive green economy entails and proposes.

Furthermore, a more open and enabling policy environment may greatly assist in not only recognising current opportunities, but also in leading to an alternative economic model. That is to say, an economic model in contrast to the dominant neo-liberal approach and which embraces the plurality of South African society; is open to alternative forms of development; promotes inclusivity and diversity; and which may be more conducive to socially equitable and environmentally sustainable development.

Acknowledgement

This work was supported by funds received from the National Research Foundation's Community Engagement Programme awarded to Stellenbosch University.

Authors' contributions

S.S. conducted the study and wrote the manuscript; J.K.M. made conceptual contributions and assisted with editing.

References

1. United Nations Environment Programme and Department of Environmental Affairs. Green economy modelling report for South Africa (SAGEM): Focus on the sectors of natural resource management, agriculture, transport and energy [document on the Internet]. c2013 [cited 2013 Sept 09]. Available from: https://www.environment.gov.za/sites/default/files/docs/greeneconomy_modelingreport.pdf.
2. Musango JK, Brent AC, Bassi AM. Modelling the transition towards green economy in South Africa. Technol Forecast Soc. 2014;87:257–273. http://dx.doi.org/10.1016/j.techfore.2013.12.022
3. Oelofse C, Scott D, Oelofse G, Houghton J. Shifts within ecological modernization in South Africa: Deliberation, innovation and institutional opportunities. Loc Environ: Int J Justice Sustain. 2006;11:61–78. http://dx.doi.org/10.1080/13549830500396214
4. United Nations Environment Programme. Towards a green economy: Pathways to sustainable development and poverty eradication [document on the Internet]. c2011 [cited 2014 Jan 07]. Available from: http://www.unep.org/greeneconomy/GreenEconomyReport/tabid/29846/language/en-US/Default.aspx

5. Resnick D, Tarp F, Thurlow J. The political economy of green growth: Cases from Southern Africa. Public Admin Develop. 2012;32:215–228. http://dx.doi.org/10.1002/pad.1619

6. Death C. The green economy in South Africa: Global discourses and local politics. Politikon. 2014;41:1–22. http://dx.doi.org/10.1080/02589346.2014.885668

7. Swilling M, Musango JK, Wakeford J. Introduction: Deepening the green economy discourse in South Africa. Cape Town: UCT Press; Forthcoming 2015.

8. Allen C. A guidebook to the green economy. Issue 2: Exploring green economy principles. Division for Sustainable Development, UNDESA, [document on the Internet]. c2012 [cited 2014 Jan 07]. Available from: http://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=743&menu=35.

9. United Nations. The future we want. Outcome document for the United Nations Conference on Sustainable Development (Rio+20); 2012 June 13–22; Rio de Janeiro, Brazil [document on the Internet]. c2012 [cited 2014 Jan 05]. Available from: https://rio20.un.org/sites/rio20.un.org/files/a-conf.216l-1_english.pdf.pdf.

10. Devey R, Skinner C, Valodia I. Informal economy employment data in South Africa: A critical analysis. Cape Town: Development Policy Research Unit; 2003.

11. Statistics South Africa. Quarterly labour force survey: Quarter 4. Pretoria: Statistics South Africa, 2013

12. Economic Development Department. Economic policy development: Second economy. [Document on the Internet]. c2015 [cited 2015 April 20]. Available from: http://www.economic.gov.za/about-us/programmes/economic-policy-development/second-economy

13. Rogerson CM. 'Second economy' versus informal economy: A South African Affair. Geoforum. 2007;38:1053–57. http://dx.doi.org/10.1016/j.geoforum.2007.01.005

14. International Labour Organisation (ILO). Report of the Seventeenth International Conference of Labour Statisticians, 2003 Nov 24–Dec 03; Geneva, Switzerland. Geneva: ILO.

15. Smit S., Musango, JK. Towards connecting green economy with informal economy in South Africa: A review and way forward. Ecological Economics. 2015;116:15–159. http://dx.doi.org/10.1016/j.ecolecon.2015.04.022].

16. Dawa FO, Kinyanjui MN. Green economy and sustainable development: Which way for the informal economy? Think piece produced as part of a series for the United Nations Research Institute for Social Development (UNRISD) [document on the Internet]. c2012 [cited 2013 Nov 15]. Available from: http://www.unrisd.org/unrisd/website/newsview.nsf/%28httpNews%29/43B2464B82F227A1C1257A41002D14B2?OpenDocument

17. Davies R, Thurlow J. Formal-informal economy linkages and unemployment in South Africa. S Afr J Econ. 2010;78(4):437–458. http://dx.doi.org/10.1111/j.1813-6982.2010.01250.x

18. Trade and Industrial Policy Strategies. Second economy strategy: Addressing inequality and economic marginalisation [document on the Internet]. c2009 [cited 2014 Feb 06]. Available from: http://www.tips.org.za/files/Second_Economy_Strategy_Framework_Jan_09_0.pdf.

19. Borel-Saladin JM, Turrok IN. The impact of the green economy on jobs in South Africa. S Afr J Sci. 2013;109(9/10), Art. #a0033, 4 pages. http://dx.doi.org/10.1590/sajs.2013/a0033

20. Smit S. Towards an 'inclusive green economy' for South Africa: Engaging with the informal economy. Stellenbosch: Faculty of Economics and Management Sciences, Stellenbosch University; 2015.

21. Babbie E. The practice of social research. 8th ed. Belmont, CA: Wadsworth; 1998.

22. Montmasson-Clair G. Green economy policy framework and employment opportunity: A South African case study. TIPS Working Paper Series. Pretoria: TIPS; 2012.

23. Sustainlabour. Congress of South African Trade Unions and Trade and Industrial Policy Strategies. Green jobs and related policy frameworks: An overview of South Africa [document on the Internet]. c2013. [cited 2014 Feb 08]. Available from: http://www.sustainlabour.org/documentos/Green%20and%20decent%20jobs-%20An%20Overview%20from%20Europe%20FINAL.pdf

24. Department of Science and Technology. The ten-year plan for science and technology [document on the Internet]. c2008 [cited 2014 June 22]. Available from: http://www.gov.za/documents/index.php?term=&dfrom=&dto=&yr=0&subjs%5B0%5D=70&p=3.

25. Republic of South Africa. Framework for South Africa's response to the international economic crisis [document on the Internet]. c2009 [cited 2014 January 27]. Available from: http://www.polity.org.za/article/framework-for-south-africas-response-to-the-international-economic-crisis-february-2009-2009-02-23.

26. National Planning Commission. Together doing more and better. Medium term strategic framework. A framework to guide government's programme in the electoral mandate period (2009–2014). Pretoria: The Presidency of the Republic of South Africa; 2009.

27. Department of Energy. Integrated resource plan: 2010–2013. [document on the Internet]. c2011 [cited 2014 June 22]. Available from: http://www.doe-irp.co.za/content/IRP2010_promulgated.pdf.

28. Department of Higher Education and Training. National skills development strategy III: 2011–2016 [document on the Internet]. c2012. [cited 2014 March 10]. Available from: http://www.dhet.gov.za/LinkClick.aspx?fileticket=AOv8W%2Bm3xHI%3D&tabid=1131.

29. Republic of South Africa. National climate change response white paper [document on the Internet]. c2011 [cited 2014 June 22]. Available from: http://www.gov.za/documents/detail.php?cid=315009.

30. Department of Environmental Affairs. National waste management strategy [document on the Internet]. c2011 [cited 2014 June 22]. Available from: http://www.gov.za/documents/index.php?term=&dfrom=&dto=&yr=0&subjs%5B0%5D=49&p=6.

31. Economic Development Department. The new growth path: Accord 4 – Green economy accord [document on the Internet]. 2011 [cited 2014 June 22]. Available from: http://www.gov.za/documents/index.php?term=&dfrom=&dto=&yr=0&subjs%5B%5D=92.

32. Department of Environmental Affairs. National strategy for sustainable development and action plan (NSSD1) 2011–2014. Pretoria: Department of Environmental Affairs; 2011.

33. National Planning Commission. National development plan, Vision 2030: Our future - make it work [document on the Internet]. c2012 [cited 2014 April 22]. Available from: http://www.npconline.co.za/pebble.asp?relid=25.

34. Department of Energy. Draft 2012 integrated energy planning report [document on the Internet]. c2013 [cited 2014 March 10]. Available from: http://www.energy.gov.za/files/IEP/IEP_Publications/Draft-2012-Integrated-Energy-Plan.pdf.

35. Department of Science and Technology. The Bio-economy strategy [document on the Internet]. c2013 [cited 2014 June 22]. Available from: http://www.dst.gov.za/images/ska/Bioeconomy%20Strategy.pdf.

36. Department of Water Affairs. National water resource strategy: Water for an equitable and sustainable future. 2nd ed [document on the Internet]. c2013 [cited 2014 March 10]. Available from: http://www.dwaf.gov.za/nwrs/LinkClick.aspx?fileticket=3AVrHanrkfw%3d&tabid=91&mid=496

37. Department of Trade and Industry. Industrial policy action plan 2014/15–-2016/17 [document on the Internet]. c2014 [cited 2014 June 22]. Available from: http://www.gov.za/documents/index.php?term=&dfrom=&dto=&yr=0&subjs%5B%5D=112

Note: This article is supplemented with online only material.

Evaluating procedural modelling for 3D models of informal settlements in urban design activities

AUTHORS:
Victoria Rautenbach[1]
Yvette Bevis[1]
Serena Coetzee[1]
Carin Combrinck[2]

AFFILIATIONS:
[1]Centre for Geoinformation Science, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa
[2]Department of Architecture, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Victoria Rautenbach

EMAIL:
victoria.rautenbach@up.ac.za

POSTAL ADDRESS:
Centre for Geoinformation Science, Department of Geography, Geoinformatics and Meteorology, Geography Building 311, University of Pretoria, Private bag X20, Hatfield 0028, South Africa

DATES:
Received: 11 Mar. 2015
Revised: 29 Apr. 2015
Accepted: 20 May 2015

KEYWORDS:
shanty town; cost-effective modelling; visualisation characteristics; spatial data infrastructure; geovisualisation

HOW TO CITE:
Rautenbach V, Bevis Y, Coetzee S, Combrinck C. Evaluating procedural modelling for 3D models of informal settlements in urban design activities. S Afr J Sci. 2015;111(11/12), Art. #2015-0100, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20150100

Three-dimensional (3D) modelling and visualisation is one of the fastest growing application fields in geographic information science. 3D city models are being researched extensively for a variety of purposes and in various domains, including urban design, disaster management, education and computer gaming. These models typically depict urban business districts (downtown) or suburban residential areas. Despite informal settlements being a prevailing feature of many cities in developing countries, 3D models of informal settlements are virtually non-existent. 3D models of informal settlements could be useful in various ways, e.g. to gather information about the current environment in the informal settlements, to design upgrades, to communicate these and to educate inhabitants about environmental challenges. In this article, we described the development of a 3D model of the Slovo Park informal settlement in the City of Johannesburg Metropolitan Municipality, South Africa. Instead of using time-consuming traditional manual methods, we followed the procedural modelling technique. Visualisation characteristics of 3D models of informal settlements were described and the importance of each characteristic in urban design activities for informal settlement upgrades was assessed. Next, the visualisation characteristics of the Slovo Park model were evaluated. The results of the evaluation showed that the 3D model produced by the procedural modelling technique is suitable for urban design activities in informal settlements. The visualisation characteristics and their assessment are also useful as guidelines for developing 3D models of informal settlements. In future, we plan to empirically test the use of such 3D models in urban design projects in informal settlements.

Introduction

The promise of work and a better life attracts thousands of South Africans to the country's major cities. However, various factors such as the housing backlog, shortage or inaccessibility of housing subsidies and affordability of housing, force hopefuls to live in informal housing or shack settlements.[1–3] Informal settlements in South Africa continue to regularly make news headlines because of lack of adequate housing and service delivery. The National Planning Commission (NPC) produces annual reports providing some insight into the magnitude of the problem: in 2013, 7.6% of the South African population (3–4 million people) lived in informal settlements.[4]

The South African government is party to the United Nations Millennium Development Goals (MDG) that aim to significantly improve the lives of at least 100 million inhabitants by the year 2020 (MDG 7, target 4).[5] As a result, the government is obligated to assist the inhabitants of informal settlements or slums to improve their living conditions. In response to the MDGs, the National Housing Code by the Department of Human Settlements proposes a process for the upgrading of informal settlements.[6] The South African Housing Programme is described in Part 3 of this Code and aimed to improve the lives of 2.2 million households living in informal settlements by 2014. Additionally, the South African National Development Plan (NDP) identified informal settlement upgrading or relocation as one of its main objectives[7(p. 489)]: by 2030 all informal settlements located on suitable land should be upgraded, or else relocated to suitable land.

To achieve the objectives set out by the MDG and NDP, urban design of informal settlement upgrades, such as the construction of a community hall, taxi rank, health clinic or playground, needs to consider both humans and the natural environment. Urban design is concerned with shaping the physical environment of cities, towns and villages, while considering human behaviour, aesthetics and functionality.[8,9] In highly dense urban regions (i.e. cities, such as New York or Hong Kong), urban design aims to create public spaces from the areas between buildings where activities, for example street markets, can take place.[10,11] Similarly in urban planning, the human is the central component in the process. Urban planners tend to focus on designing and planning large areas for future use, e.g. new infrastructure,[12] and set out a strategy for crisis management.[13,14] While there are urban design visionaries, such as Lynch[9] and Montgomery[15], who describe how urban design can shape the development of a city so that it becomes a functional and aesthetically pleasing environment, urban design tends to focus on enhancing the current localised environment. An important component of urban design in a democratic society is involving and communicating planned changes to various stakeholders. Geovisualisation of spatial data is a powerful tool in this communication. Over the years, spatial data have been represented in various formats ranging from two dimensional (2D) maps, three dimensional (3D) physical scale models, to 3D digital representations. Recently, 3D geovisualisations have grown in popularity, partly through the availability of applications such as Google Earth (https://www.google.com/earth/) and NASA World Wind (http://worldwind.arc.nasa.gov/java/).

Three dimensional (3D) models are increasingly being used in urban design applications.[16-19] 3D models can present information that could not previously be visualised in 2D maps and designs. In response to these benefits, most large German municipalities provide level of detail (LoD)2 models of their municipal areas and in some cases, LoD3 models of their city centres. Many of the German states aim to provide LoD2 models of all their urban areas by 2016. Other European cities, such as Monaco, Geneva, Zurich and Leeuwarden, use 3D city models as a means of representing and exchanging data; similarly in Asia.[18]

Numerous studies have been conducted on various aspects of 3D city models, but these are limited to urban environments (i.e. city centres). 3D models of cities are used in various application fields, such as urban planning, disaster management, computer games, entertainment and education.[18,20] 3D models of informal settlements could also be used in these application fields, however research on their use is limited to exploratory studies in a few countries[21] and to our knowledge has not been done in South Africa.

Modelling large 3D cities manually requires an enormous amount of labour,[22] making this approach expensive. Owing to financial constraints and the dynamic nature of informal settlements, this approach is not feasible for South African informal settlements. Procedural modelling is a form of automatic generation that has been suggested as an alternative to traditional manual methods for creating 3D models. Algorithms are used sequentially to generate 3D objects in order to generate mass 3D models.[23,24] Through procedural modelling, an intricate object is created by iteratively applying a rule to each resulting component of the object.[25] An additional benefit of procedural modelling is that it can be used to automatically prepare 3D models from spatial data (e.g. topography and terrain models) made available through the spatial data infrastructure of a city or country.[26,27]

The goal of this article is to describe a procedural modelling process to generate a 3D model of an informal settlement in South Africa, and to evaluate the applicability of the resulting model in urban design activities. We adapted visualisation characteristics of 3D models of buildings in city planning for the specific requirements of urban design activities during informal settlement upgrades in South Africa. Then we assessed the importance of each characteristic for such activities. We developed a prototype 3D model of a part of Slovo Park, an informal settlement in Johannesburg, South Africa, and evaluated its visualisation characteristics.

Study area: Slovo Park settlement

Slovo Park is an informal settlement situated roughly 10 km south of Soweto, in the City of Johannesburg Metropolitan Municipality, South Africa (See Figure 1). It is located adjacent to a large freeway and is surrounded by industrial factories. There are approximately 1100 stands in Slovo Park, laid out by the community (i.e. not officially surveyed). The local municipality provides only rudimentary services in the form of waste and sewerage collection, and communal taps. The residents do not have access to formal electricity (See Figure 2).

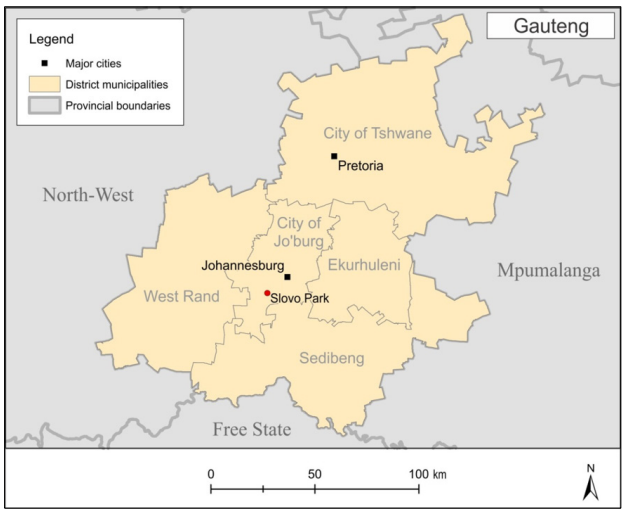


Figure 1: Location of Slovo Park within Gauteng, South Africa.

The 3D model described in this article is merely a proof of concept, and thus only a sub-section of the settlement was selected as the study area. The study area contains various types of human-made constructions, such as a community hall (central meeting place for the community), a *shebeen* (local term for a bar), and surrounding dwellings (typical shelters found in an informal settlement).

Figure 2: Photos depicting life in Slovo Park. (a) Informal vendor. (b) Outside toilet and recycling for monetary gain.

A procedural modelling process for informal settlements

Overview of the procedural modelling process

A procedural modelling process was developed as an alternative to manual modelling of informal settlements for planning upgrades. Procedural modelling iteratively applies algorithms and rules to an initial 2D shape that is replaced by a 3D object afterwards.[28,29] Natural real world systems can be realistically replicated with procedural modelling as it utilises parameterisation that allows infinite variation in the generated model.[29,30] However, identifying the correct parameters and values to generate a photo-realistic model is not feasible.

Procedural modelling has proven to be successful for generating various models, such as urban environments, road networks, infrastructure, vegetation, and terrains.[30-33] Urban modelling is a prominent reason for the popularity of procedural modelling, as cities are not only big, but also incredibly detailed.[34] With manual modelling, it would be too tedious and time-consuming to generate realistic 3D models.[33]

The procedural modelling process described in this article was derived from the visualisation pipeline and adapted for procedural modelling. The visualisation pipeline defines the process required to create visualisation from raw data.[35-37] The basic visualisation pipeline consists of filtering, mapping and rendering.[36,37] During the filtering phase, the data are converted and a subset of the data is selected. Thereafter, the data are transformed into geographic primitives during the mapping phase. Lastly, the geometric primitives are converted into realistic images, i.e. they are rendered. Our procedural modelling process is depicted in Figure 3.

Level of detail

3D objects can be characterised according to the level of detail (LoD) with which they represent objects in the real world. This LoD ultimately determines the model's applicability.[38] One of the simplest reasons for using LoD in 3D models is to reduce the complexity of the geometry

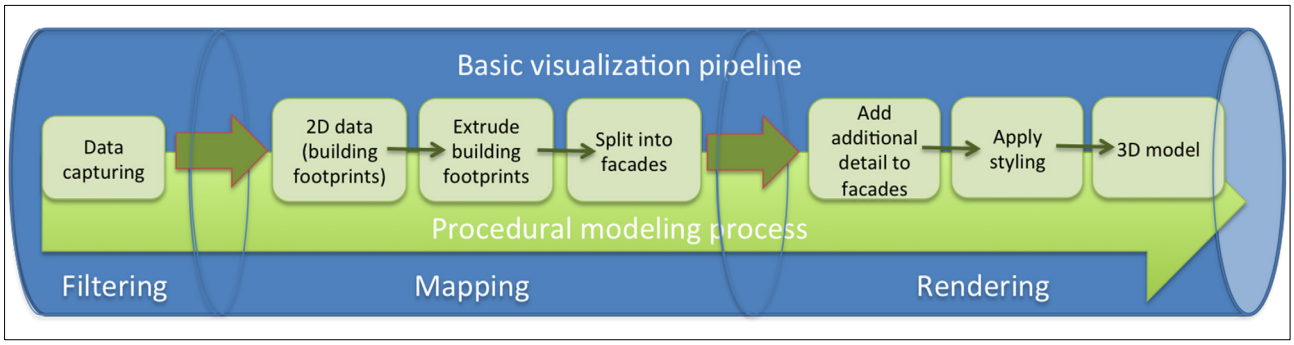


Figure 3: Overview of the procedural modelling process.

before rendering. Biljecki et al.[39] suggested additional advantages for using LoD, e.g. to specify the level at which 3D data are captured, and for generalisation during rendering.

Various approaches exist for specifying LoD for urban models. However, none of the approaches are without criticism. For example, some LoDs are considered too coarse, thereby hindering smooth transition between the levels.[39] In this paper, the Open Geospatial Consortium (OGC) CityGML specification of LoD will be used (see Figure 4)[40]:

- *LoD0* is the 2.5D representation of the basic outline of objects, typically a flat polygon.
- *LoD1* is achieved when the volume of objects is modelled in a generalised fashion with the inclusion of vertical walls and flat roofs.
- *LoD2* builds on LoD1 with the addition of a roof structure, as well as the inclusion of texture or photography on surfaces and terrain.
- *LoD3* is an extension of LoD2 with the inclusion of openings, such as windows and doors, and by adding more detailed roof structures.
- In *LoD4,* the interiors of buildings are added to the model.

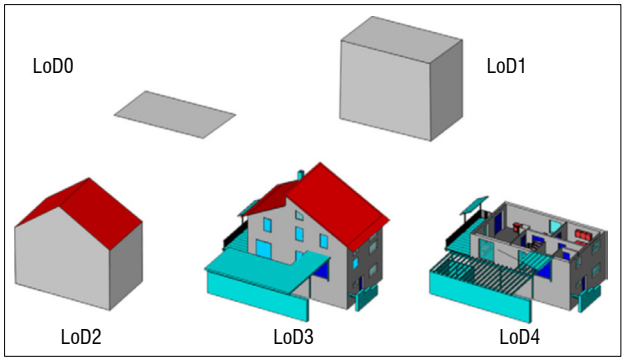


Figure 4: Levels of Detail (LoD) as specified by CityGML.[18]

Implementation of the procedural modelling process

During implementation, the steps identified in the section above on the procedural modelling process were refined and a detailed procedural modelling process was developed, depicted in Figure 5. The process was implemented in CityEngine (http://www.esri.com/software/cityengine) and CGA (Computer Generated Architecture). After completion of a phase in the process, the resulting model can be related to a specific LoD. The prototype model of Slovo Park was modelled up to LoD3. The interiors of the informal structures (LoD4) were not considered, as these are deemed unnecessary for urban design.

LoD0 is the most simplistic level; it consists of building footprints and terrain. Building footprints are commonly represented by 2D simple geometry. LoD1 was attained through the extrusion of the 2D building footprints into 3D cuboids. Each 2D geometry (or building footprint) in LoD0 was assigned a height value. The CGA *extrude* command was then executed to produce 3D geometric objects with the generalised volume and depth of the real world features. The dwellings were extruded to a height of 1.9 m, as this was observed to be the average height in the field. Figure 6a is an example of such a dwelling. In-situ measurements were taken of the community hall and the height of the structure was found to be approximately 2.3 m (See Figure 7a). The community hall was extruded accordingly.

Procedural modelling allows users to extrude all of the building footprints simultaneously, compared to manual modelling which requires individual extrusion. With procedural modelling, structures can either be extruded to a predefined height, or a random height within a specified range. The result of the LoD0 to LoD1 transformation (i.e. the LoD1 model) is shown in Figures 6b and 7b.

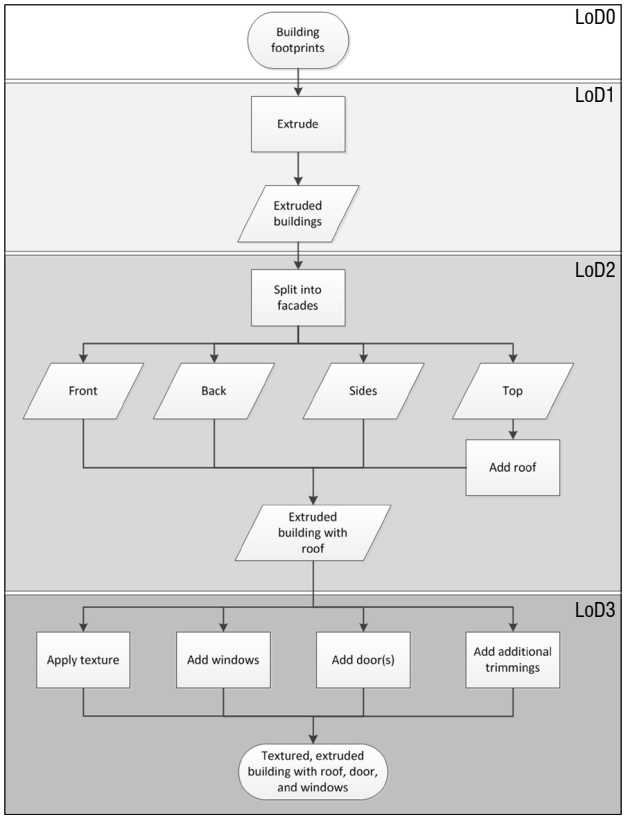


Figure 5: The detailed procedural modelling process.

The 3D object at LoD2 starts to resemble its true form. A LoD2 model is attained when a roof is added to the LoD1 shape or object.[18] Typically, roofs have a triangular shape, however, most structures in informal settlements have flat roofs. Gröger & Plumer[18] argue that a horizontal roof is an attribute of a LoD1 model (refer to Figure 7b). To enable the modelling of more complex objects (at higher LoDs) using procedural techniques, the objects have to be separated into different components or objects, called facades. In our model, we used the CityEngine *split*

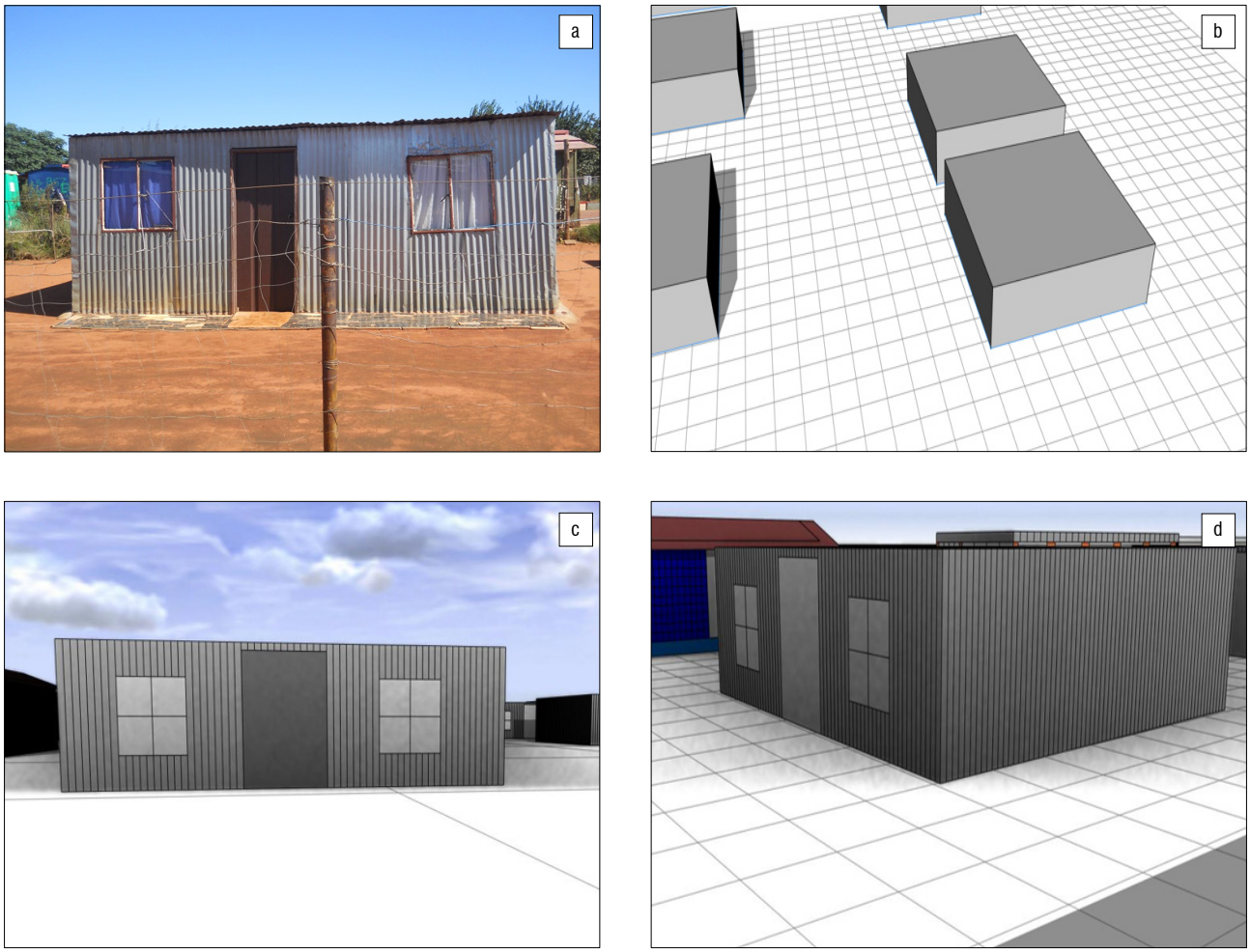


Figure 6: Results of modelling an informal dwelling. (a) The generic informal dwelling. (b) Informal dwelling modelled at LoD1. (c) The front of an informal dwelling modelled at LoD3. (d) The side of the informal dwelling modelled at LoD3.

command to divide an object into facades. The following facades (or objects) were generated: front, back, sides, and top. An advantage of procedural modelling is that these facades are now independent objects that can be modelled separately (e.g. different combinations of facades). Because there is an explicit top facade (the roof) after the execution of the split command, structures with flat roofs are now LoD2 objects. Semantic information (e.g. descriptive attributes) can be added to each of the facades so that complex analysis can be performed on the 3D model.

LoD3 was achieved by adding windows and doors to the objects and by adding texture, resulting in a more complex geometric object that resembles the physical environment more closely. For the informal dwellings in our model, the front facade was split into seven sections, as depicted in Figure 8. Sections *S2* and *S6* in Figure 8 were divided into three sections, each along the y-axis so that two windows could be created in the wall. Each window was then split in half along the x- and y-axes to form burglar bars, a common feature in South Africa, even in informal settlements.

The informal structures are constructed mainly from corrugated iron. The corrugated iron was modelled recursively by splitting the front facade (excluding the windows and door) along the x-axis at an interval of 0.10 m. The remaining facades did not contain any windows and therefore only the corrugated iron was modelled for them.

The resulting 3D objects at LoD3 can be seen in Figure 6c and 6d. The same recursive splitting technique was used to generate the texture of the community hall (Figures 7c and 7d). The other objects in the 3D model, such as outside toilets and post boxes, were generated using similar techniques. The final results of the model can be seen in Figure 9.

Visualisation characteristics of 3D informal settlement models

Visualisation can support the urban design process at various stages, including the initial gathering of information about the environment; when designing or planning new developments; and for communicating the planned changes to stakeholders. In this section, we describe the visualisation characteristics of 3D models of informal settlements. The characteristics are adapted from those described by Dickmann and Dunker[41] for 3D building models in urban planning. They defined three categories of characteristics: visualisation content; visualisation application environment and functionality; and visualisation performance. Drawing on discussions with experts, informal settlement inhabitants, as well our experience in urban design projects and informal settlement upgrades, the characteristics by Dickmann and Dunker[41] were adapted for use in 3D models of informal settlements in urban design (as opposed to urban planning) activities in South Africa.

First, we expanded the visualisation content category by including human-made structures, landmarks, the movement network and service delivery infrastructure. We split vegetation and street furniture into two distinct characteristics. While vegetation in urban areas tends to be cultivated and arranged in an orderly fashion (similar to street furniture), vegetation in informal settlements usually comprises remainders of natural vegetation and haphazardly distributed trees, shrubs and patches of grass. Therefore, it needed to be considered separately.

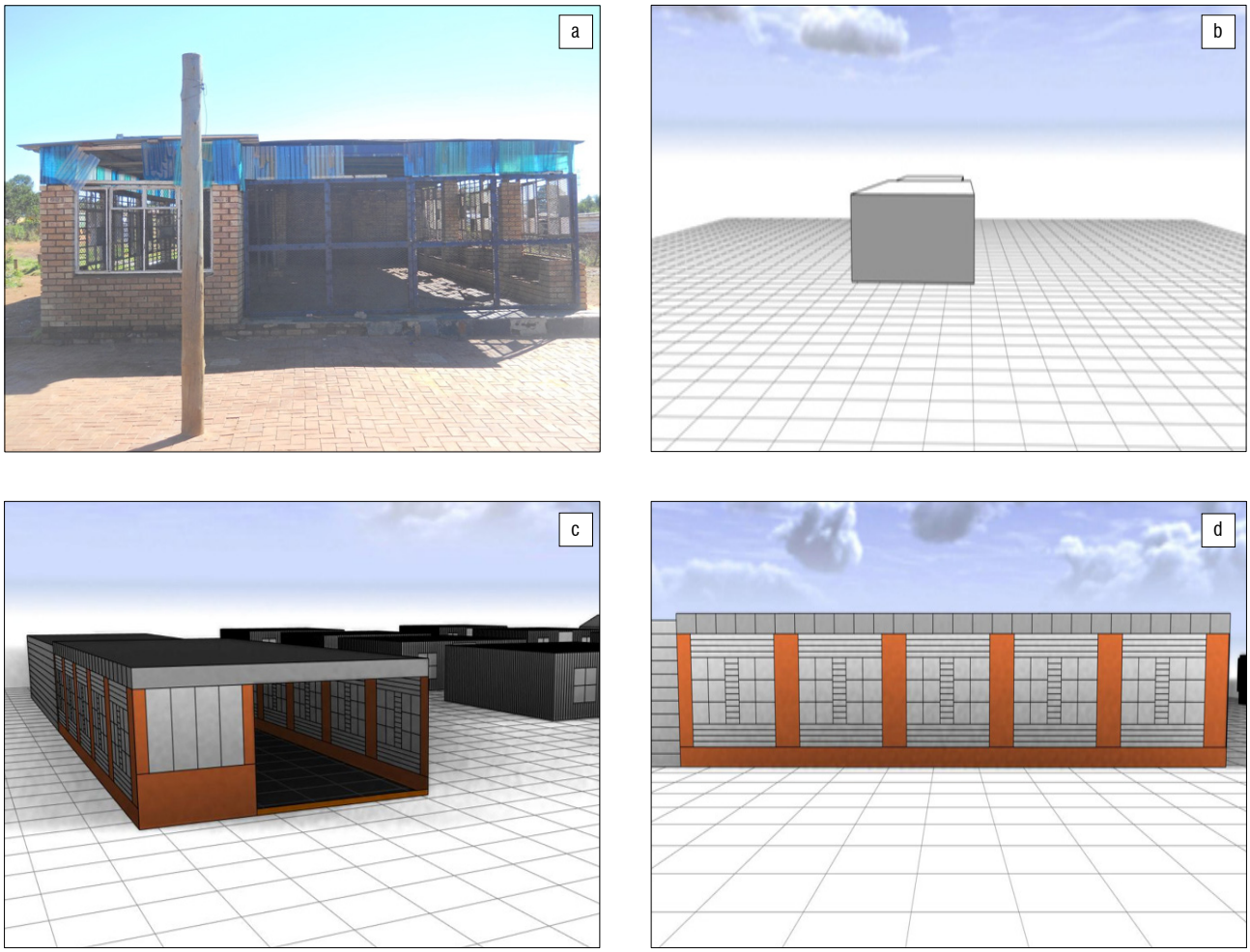


Figure 7: Results of modelling the community hall. (a) Front of the community hall. (b) Community hall modelled at LoD1. (c) Community hall modelled at LoD3. (d) The side of the community hall modelled at LoD3.

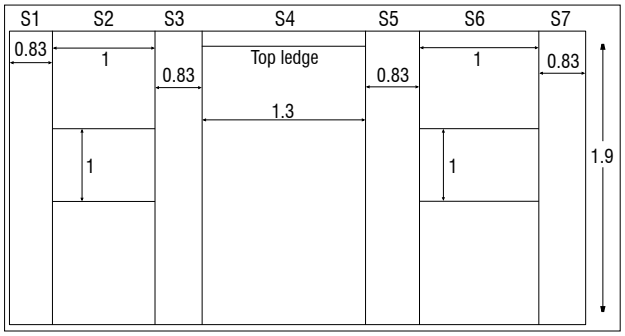


Figure 8: A preliminary plan of the informal structure (not drawn to scale).

In the category of visualisation application environment and functionality, we combined the pedestrian perspective and bird's eye view into a single characteristic. The remaining characteristics in this category and the visualisation performance category were adapted.

Dickmann and Dunker[41] concluded that the importance of visualisation characteristics varies among different users of 3D models of buildings in city planning. For example, LoD1 or LoD2 are sufficient for planning experts, but community members prefer more detail and textures. Based on their findings and our experience in urban design and informal settlement upgrade projects, each criteria identified in Table 1 was assigned an importance ranking, ranging from *no contribution* to *major contribution* (refer to Table 2). Contribution refers to our assessment of how important the specific visualisation characteristic is regarded by various stakeholders (e.g. settlement dwellers, planners and officials) when using the 3D model in urban design activities.

Evaluation of the visualisation characteristics

For communicating planned upgrades, it is important that 3D models, specifically for the area of interest, be generated at a high *LoD*, such as LoD3. The higher LoD creates a more realistic and relatable 3D model. If stakeholders can recognise aspects of the environment, they can more easily relate to the model. Consequently, they will have a better understanding of the effect that proposed upgrades may have on their daily lives and environment. The surrounding area is acceptable at a lower LoD. Similarly, a lower LoD is typically sufficient for planning experts when gathering information about the current environment before planning and designing upgrades.

Our research demonstrated that procedural modelling is a viable alternative for generating a 3D model of a typical South African informal settlement at LoD3 (Refer to Figures 6 and 7). Roofs, windows and doors were modelled procedurally through either adding geometry or splitting up existing geometry. Substantial initial preparation is required for procedural modelling, however, once the correct dimensions of the objects have been calculated, the procedural technique to model the objects is straightforward and can be repeatedly applied.

The *size or extent of the model* is determined by the planner, and influenced by upgrades where the integration of the area of interest with the surrounding community needs to be considered. However, we consider the size or extent less important for urban design activities,

which tend to be localised, in comparison to regional or city planning which needs to consider a wider area. Procedural modelling provides an effortless method of populating the background (space beyond the area of interest) of a 3D model through the repetitive execution of the same procedures with different parameters. There is no limit on the size of the model that can be created with procedural modelling.

Facades and/or texture are less important in informal settlements (similar facades and textures on human-made structures) than in urban areas (more variety in building facades and textures) when collecting information about existing environment or designing upgrades. LoD3 provides sufficient detail about 3D objects to make a model suitable for communicating information about the existing informal settlement. Additional details could prove useful when it is important to communicate the specific facade or texture of a planned building. Detailed facades comprising irregular shapes are time-consuming and tedious to create with procedural modelling.

Table 1: Visualisation characteristics for 3D models of informal settlements in South Africa

Category	Characteristic	Origin	Description
Visualisation content	1. Level of detail (LoD)	Adopted	Distinct levels according to which 3D models can be generated. The level specifies object complexity and granularity regarding geometric representation and thematic differentiation.
	2. Size or extent of the model	Adopted	The area or extent of the model is the physical land covered by the model. Smaller models focus on the area of interest, while larger models 'see the bigger picture'. For example, in informal settlements a larger area is required to see how a community is or can be integrated into the surrounding area.
	3. Facade and/or texture	Adopted	The outward appearance of the object. The desired appearance can be achieved using textures (imported from photographs or custom textures using graphic editors) or it is modelled using a combination of geometry and colour.
	4. Terrain model	Adapted	The representation of elevation information (the bare ground) without any details about objects on the surface. A terrain model represents the physical characteristics of the natural features in an area, i.e. landform.43 The terrain provides import information for landscape design and flood drainage in an informal settlement.
	5. Topography	Adapted	Commonly known as the physical lay of the land. The earth's surface features, such as relief, terrain, vegetation and any features created in the landscape by human endeavours.43 Topographical information thus provides import information about the surface in an informal settlement, e.g. natural or human-made barriers.
	6. Vegetation	Adapted	The total plant cover of an area, comprising one or more plant species.43 Vegetation may be considered as landmarks in some cases, e.g. three abnormally tall trees in a specific area could be considered a landmark.
	7. Human-made structure	Additional	A human-made structure is any building or dwelling constructed by humans. In informal settlements a wide variety of materials are used, including bricks, wood, corrugated iron and even road signs. Some of these could be classified as landmarks, street furniture or service delivery infrastructure, or could form part of the movement network (e.g. taxi rank).
	8. Landmark	Additional	A landmark is a notable object (stands out from its surroundings) used as a point of reference when orientating oneself within an environment, e.g. taxi ranks, cemeteries, or playgrounds.
	9. Street furniture	Adapted	Street furniture generally refers to pieces of equipment located along a thoroughfare for various purposes, e.g. light poles or road signs. In this case, we use the term to refer to any objects along a thoroughfare that are deemed necessary in the 3D model.
	10. Movement network	Additional	A movement network is defined broadly as the public right-of-way network, accommodating land-based movement by a range of movement modes.42 It is the combination of the traditional road layout and footpaths within a settlement. The movement network provides an indication of the accessibility of the informal settlement, as well as the accessibility of locations within the settlement.
	11. Service delivery infrastructure	Additional	In informal settlements where households do not have running water, water distribution points at various points in the settlement are a common feature. These water distribution points are often just a single tap. Informal electricity connections are also a common feature of informal settlements.
Visualisation application environment and functionality	1. Model navigation	Adopted	The ability to traverse or move around in the virtual environment. Movement through the model is either user-directed or predefined by the designer of the model.
	2. Perspective	Adapted	A perspective, such as the pedestrian perspective or bird's eye view, refers to the specific viewpoint or camera position from which the model is observed.
	3. Shading and illumination	Adopted	Shading and illumination add to the realism of the virtual environment, and can also be used to do certain analysis within the model.
	4. Measurement tools	Adopted	Allows the user or viewer to determine the length, area or height of an object.
Visualisation performance	1. Time cost	Adopted	The time and effort required for developing the 3D model; to determine this a cost-benefit assessment may be required.
	2. Application performance	Adopted	Refers to the real-time visualisation performance of the application rendering the 3D model.

However, once the facade details have been modelled, they can be flexibly combined into a variety of different facades. Various texture libraries are available online and these can be included in the models, as necessary.

Table 2: Assessment of the importance of each visualisation characteristic to stakeholders in urban design activities

Category	Characteristic	Minor contribution	Moderate contribution	Major contribution
Visualisation content	1. Level of detail			Informal Urban
	2. Size or extent of the model		Informal	Urban
	3. Facade and/or texture	Informal	Urban	
	4. Terrain model			Informal Urban
	5. Topography		Urban	Informal
	6. Vegetation	Informal	Urban	
	7. Human-made structure			Informal
	8.Landmarks		Informal	
	9. Street furniture	Informal	Urban	
	10. Movement network			Informal
	11. Service delivery infrastructure	Informal		
Application environment and functionality	1. Model navigation			Informal Urban
	2. Perspectives			Informal Urban†
	3. Shading and illumination	Urban	Informal	
	4. Measurement tools		Informal Urban	
Performance	1. Time cost			Informal Urban
	2. Application performance			Informal Urban

†Moderate contribution by the pedestrian perspective; major contribution by the bird's eye view

A *model of the terrain* also provides useful information when considering integration. For example, it could allow the identification of possible natural barriers (e.g. a *donga,* hill or mountain) to integration into the surrounding area. However, a terrain model also aids in identifying physical risks to the settlement, such as the risk of landslides or flooding. Informal settlements often arise on land not suitable for development, so physical risks may be significant. Therefore, the terrain should always form part of the 3D model of an informal settlement. Aerial imagery draped over height maps can be used to depict various landforms. Such 'draping' is possible with procedural modelling. Terrain modelling, however, was not relevant in this particular study area (Slovo Park) as it is relatively flat.

Topographical information adds the required detail resulting in a more realistic and relatable model. Topography may also provide reference points to the audience in the public participation process. When considering the integration of the settlement with the surrounding community, topographical information may point out how certain features separate the settlement from the surrounding area. For example, South African 1:50 000 topographic maps include prominent rock outcrops, mine dumps or excavations, graves and powerlines. Such information would already have been considered during planning in a developed urban area, but may significantly impact the location and distribution of planned upgrades in an informal settlement that arises spontaneously. Therefore, we consider this characteristic more important in informal settlements than in urban areas. Procedural modelling provides tools to automate the conversion of topographical information from vector data (e.g. in SHP files) into 3D objects.

Vegetation adds realism to the model, and in some cases, could act as a landmark. Vegetation is, however, a minor consideration when modelling an informal settlement because much of the vegetation is usually destroyed during the construction of the settlement. If a certain element of the vegetation is also a landmark (e.g. an unusually tall tree), then it should be included. In procedural modelling, vegetation is added to the model as an object and then manipulated procedurally. For example, the size of the tree can be increased or additional trees can be added recursively.

Informal settlements comprise *human-made structures*, mainly in the form of informal dwellings. Such dwellings are typically conglomerations of cuboids. Procedural modelling allows developers to generate large quantities of informal dwellings with a single command call in a simple script. More complex structures can also be generated, however, they may require initial preparation (similar to the process for higher LoDs). The human-made structures provide context to understand how planned upgrades relate to and impact on the current environment. For example, it could help inhabitants to understand how accessible a planned community hall, clinic or playground is to their own dwelling. Human-made structures are also considered by planning experts to ensure that the planned upgrades 'fit' with existing dwellings (e.g. avoid destruction of, or are in too close proximity to, dwellings).

Informal settlement dwellers place great value on certain *landmarks* within their community. These landmarks could be of social (e.g. community hall, *shebeen* or water distribution point) and/or cultural (e.g. cemetery or church) importance to them. It is important to include these landmarks in the model, as they provide context for the stakeholders.

As there typically is no *street furniture* in an informal settlement, its contribution to 3D informal settlement models is limited. Street furniture should only be included when referred to specifically and when included in planned upgrades.

The *movement and access network* needs to be considered when planning upgrades. For example, access to economic centres, education and health services should be taken into account. Informal settlement dwellers commonly do not own cars and often walk or take public transportation to their destinations. All aspects of the movement network, such as roads, taxi ranks and footpaths, need to be included. The movement network is important for determining how best to integrate the settlement into the surrounding community, and also to determine accessibility of a planned upgrade within the informal settlement itself. Existing streets can either be imported into the model (e.g. from a SHP file or spatial database) or grown algorithmically to align with existing buildings and terrain.

Service delivery infrastructure is of concern when the planned upgrades involve upgrades to the infrastructure, the construction of new infrastructure or if planned upgrades are close to existing infrastructure. Infrastructure objects, such as water meters, street lighting poles and electricity distribution boxes, are typically scarce or non-existent in informal settlements. However, it makes sense to at least consider

Figure 9: A section of the Slovo Park informal settlement modelled at LoD3

'other' types of service delivery objects, such as water distribution points and illegal connections to the electricity network (also known as *izinyoka*), as they often pose significant risks to inhabitants. In procedural modelling, 3D objects can be added from libraries (collections of objects that may be shared and reused). A South Africa-specific object library does not exist. Thus 3D objects resembling South African service delivery (e.g. *izinyoka*), vegetation (e.g. indigenous trees) and street furniture (e.g. road signs) had to be developed from scratch for this study.

The ability to explore or *navigate* through the 3D model is a significant consideration when selecting a visualisation application to plan the upgrades (planners) and to demonstrate the planned upgrades to stakeholders (officials and inhabitants). Although interoperability issues still exist, most applications can currently export 3D models into diverse formats for use in a variety of applications. An intuitive mechanism for moving within the model allows unskilled individuals to explore the model and thus better understand the planned upgrades. Alternatively, videos of different view paths through a 3D model allow users to experience specific parts of the model, for example, the area of interest. Such videos can be made available on the Internet for easy access by stakeholders.

Different *perspectives*, such as bird's eye view or the pedestrian perspective, allow the 3D model to be viewed from several angles or viewpoints. Different perspectives are important for officials and experts during the planning stage (e.g. to evaluate the impact of a large building) and for communicating upgrades (e.g. to understand the impact of a large building) to stakeholders.

Shading and illustration are common functionalities in most 3D modelling applications, allowing analyses to be performed, for example, to evaluate solar radiation availability. Shading and illustration are more important to planners than to the public and are used, for example, to analyse the availability of solar radiation on the roofs of informal dwellings after construction of a double storey community hall. Owing to the increased use of solar energy in informal settlements, we consider this characteristic more important in 3D models of informal settlements than in 3D models of urban areas, and also more important in urban design than in urban planning.

The importance of *measuring tools* depends on the purpose of the 3D model (e.g. to communicate an initial proposal vs the final plan). Measuring tools are generally more important during planning than during communication with stakeholders. Procedural modelling produces proportionally accurate models, as calculated dimensions are used to iteratively generate the objects in the model. Navigation, perspective, shading and illumination, and measuring tools are tool-specific functionalities and are not relevant during the procedural modelling process.

Using procedural techniques for generating intricate 3D models, especially if they contain complex geometric details, is initially time-consuming *(time cost)*, as it requires extensive planning and preparation. But once the initial model is generated, modifications can be made by adjusting parameters, commands and scripts. Large quantities of 3D objects can be generated with a single script. Procedural modelling makes it possible to prepare models with semi-realistic structures with minimal effort, and yet, the models contain enough detail for planning, analysis and communication during urban design activities. For financially constrained municipalities in South Africa, procedural modelling could be a cost-effective alternative to expensive manual modelling techniques. Initially, procedural modelling has a steep learning curve for users who are not familiar with scripting and the concepts of spatial data and 3D modelling. This needs to be considered when evaluating return on investment from the use of procedural modelling.

Procedural modelling is performed using scripts that extrude and generate 3D objects from 2D geometries. The real-time *performance* of the visualisation application relies on the computer specification. The application must be able to handle large amounts of data and render high quality 3D models in real-time. For the Slovo Park model, a standard desktop PC was sufficient.

Conclusion

In this article, we discussed the procedural modelling process followed to develop a 3D model of the Slovo Park informal settlement. The visualisation characteristics of the resulting model were evaluated for use in urban design activities in informal settlement upgrade projects in South Africa.

The description and assessment of visualisation characteristics indicated a number of unique challenges for the use of 3D informal settlement models in urban design activities. For example, visualisation content, such as human-made structures and the movement network, are considered important for informal settlement upgrades, but not for modelling typical city centres. On the other hand, facades and/or textures, vegetation and street furniture are less important in informal settlement models than in models of city centres. Increased use of solar energy in informal settlements raises the importance of shading in 3D models of informal settlements. The visualisation characteristics and their assessment in this article are also useful as guidelines for developing 3D models of informal settlements.

Procedural modelling was found to be a viable alternative to the traditional time-consuming manual modelling process for typical South African informal settlements. With procedural modelling, planners can produce cost-effective 3D models for planning and communication during urban design. Although procedural modelling

requires initial preparation, once the groundwork is complete, high quantity and quality 3D objects can be generated in near real-time. The scripts can also be reused and customised to prepare models for more than one settlement. However, some challenges exist, such as the initial preparatory work, the steep learning curve and the lack of a 3D object library for South Africa. The resulting model will always be an abstraction and never truly photorealistic as models resulting from LiDAR data or pictometry images are. Nevertheless, research has suggested that non-photorealistic models can still be a valuable resource for urban design.

This article described the frameworks for the development and evaluation of 3D informal settlement models in urban design activities during informal settlement upgrades. In future, we plan to design and perform user studies to empirically evaluate the use of 3D models during preparation, planning, design and communication of informal settlement upgrades in urban design projects. In addition, the development of a South Africa-specific 3D object library for procedural modelling of informal settlements needs to be explored.

Authors' contributions

Y.B. developed the 3D model and procedural modelling process as part of her postgraduate studies under the supervision of V.R. and S.C.; V.R. refined the visualisation characteristics as part of her PhD research under the supervision of S.C.; C.C. provided urban design input; V.R. and S.C. wrote the majority of the article; and the evaluation of the visualisation characteristics was done jointly by the four authors.

References

1. Richards R, O'Leary B, Mutsonziwa K. Measuring quality of life in informal settlements in South Africa. Soc Indic Res. 2006;81(2):375–88. http://dx.doi.org/10.1007/s11205-006-9014-1
2. Huchzermeyer M, Karam A. Informal settlements: A perpetual challenge? Cape Town: Juta; 2006.
3. Jürgens U, Donaldson R, Rule S, Bähr J. Townships in South African cities – literature review and research perspectives. Habitat Int. 2013;39:256–60. http://dx.doi.org/10.1016/j.habitatint.2012.10.011
4. National Planning Commission. Millennium development goals: Country report 2013. Pretoria: Statistics South Africa; 2013.
5. United Nations. United Nations Millennium Development Goals [document on the Internet]. c2011 [cited 2015 March 10]. Available from: http://www.un.org/millenniumgoals/
6. Department of Human Settlements. The National Housing Code. Part 3: Upgrading informal settlement. Pretoria: Department of Human Settlements; 2009. Available from: http://www.dhs.gov.za/sites/default/files/documents/national_housing_2009/4_Incremental_Interventions/5%20Volume%204%20Upgrading%20Infromal%20Settlement.pdf
7. National Planning Commission. National Development Plan 2030: Our future – make it work [document on the Internet]. c2012 [cited 2015 October 30]. Available from: http://www.gov.za/sites/www.gov.za/files/Executive%20Summary-NDP%202030%20-%20Our%20future%20-%20make%20it%20work.pdf
8. Kim D, Bejleri I. Using 3D GIS simulation for urban design. ESRI Users Conference; 2005 Nov 04; San Diego, CA, USA; 2005. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.5804&rep=rep1&type=pdf
9. Lynch K. The Image of the City. Cambridge, MA: MIT Press; 1960. p. 202.
10. Gehl J. Life between buildings: Using public space In: LeGates RT, Stout F, editors. The City Reader. 5th ed. London: Routledge.; 2011. p. 202.
11. Badland H, Whitzman C, Lowe M, Davern M, Aye L, Butterworth I, et al. Urban liveability: Emerging lessons from Australia for exploring the potential for indicators to measure the social determinants of health. Soc Sci Med. 2014;111:64–73. http://dx.doi.org/10.1016/j.socscimed.2014.04.003
12. Rautenbach V, Coetzee S, Coltekin A. Towards evaluating the map literacy of planners in 2D maps and 3D models in South Africa. AfricaGEO 2014 Conference; 2014 July 1–3; Cape Town, South Africa. Cape Town: CONSAS; 2014.
13. Chen R. The development of 3D city models and its applications in urban planning. 2011 19th International Conference on Geoinformatics; 2011 June 24–26; Shanghai, China. Shanghai: IEEE; 2011. p. 1–5 . http://dx.doi.org/10.1109/GeoInformatics.2011.5981007
14. Engel J, Döllner J. Immersive visualisation of virtual 3D city models and its applications in E-planning. IJEPR. 2012;1(4):17–34. http://dx.doi.org/10.4018/ijepr.2012100102
15. Montgomery C. Happy City. New York: Farrar, Straus and Giroux; 2013.
16. Cartwright W, Pettit C, Nelson A, Berry M. Towards an understanding of how the "geographical dirtiness"(complexity) of a virtual environment changes user perceptions of a space. In: Zerger A, Argent R, editors. Proceedings of the International Congress of Modelling and Simulation: Advances and Applications for Management and Decision Making. Canberra, Australia: Modelling and Simulation Society of Australia and New Zealand; 2005. p. 3037–3044.
17. Cartwright W, Pettit C, Nelson A, Berry M. Community collaborative decision-making tools: Determining the extent of "Geographical Dirtiness" for effective displays. 21st International Cartographic Conference; 2005 July 9–16; Corona, Spain. Corona: International Cartographic Association, 2005.
18. Gröger G, Plümer L. CityGML – Interoperable semantic 3D city models. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci. 2012;71:12–33. http://dx.doi.org/10.1016/j.isprsjprs.2012.04.004
19. Wergles N, Muhar A. The role of computer visualisation in the communication of urban design: A comparison of viewer responses to visualisations versus on-site visits. Landscape Urban Plan. 2009;91(4):171–182. http://dx.doi.org/10.1016/j.landurbplan.2008.12.010
20. Parish YIH, Müller P. Procedural modelling of cities. Proceedings of the 28th annual conference on Computer Graphics and Interactive Techniques – SIGGRAPH '01. New York, USA. New York: ACM Press; 2001. p. 301–308. http://dx.doi.org/10.1145/383259.383292
21. Buehler M. Making of Favela [blog on the Internet]. c2014 [cited 2015 March 10]. Available from: http://www.ronenbekerman.com/making-favela/
22. Müller P, Wonka P, Haegler S, Ulmer A, Van Gool L. Procedural modelling of buildings. ACM T Graphic. 2006;25(3):614. http://dx.doi.org/10.1145/1141911.1141931
23. Ganster B, Klein R. An integrated framework for procedural modelling. Proceedings of the 23rd Spring Conference on Computer Graphics – SCCG '07. New York, USA. New York: ACM Press; 2007. p. 123–130. http://dx.doi.org/10.1145/2614348.2614366
24. Ilčík M, Fiedler S, Purgathofer W, Wimmer M. Procedural skeletons. Proceedings of the 26th Spring Conference on Computer Graphics – SCCG '10. New York, USA. New York: ACM Press; 2010. p. 157. http://dx.doi.org/10.1145/1925059.1925087
25. Talton JO, Lou Y, Lesser S, Duke J, Měch R, Koltun V. Metropolis procedural modelling. ACM T Graphic. 2011;30(2):1–14. http://dx.doi.org/10.1145/1944846.1944851
26. Hjelmager J, Moellering H, Cooper A, Delgado T, Rajabifard A, Rapant P, et al. An initial formal model for spatial data infrastructures. Int Geog Inf Sci. 2008(11/12):1295–1309. http://dx.doi.org/10.1080/13658810801909623
27. Cooper A, Coetzee S, Rapant P, Laurent D, Danko DM, Iwaniak A, et al. Exploring the Impact of a spatial data infrastructure on value-added resellers and vice versa. In: Andrienko G, Andrienko N, Schumann H, Tominski C, editors. Cartography from pole to pole. Berlin: Springer; 2014. p. 395–404. http://dx.doi.org/10.1007/978-3-642-32618-9_28
28. He S, Besuievsky G, Tourre V, Patow G, Moreau G. All range and heterogeneous multi-scale 3D city models. In: Leduc T, Moreau G, Billen R, editors. Usage, usability, and utility of 3D city models – European COST action TU0801. Les Ulis: EDP Sciences; 2012:02006. http://dx.doi.org/10.1051/3u3d/201202006
29. Glass K, Morkel C, Bangay S. Duplicating road patterns in South African informal settlements using procedural techniques. Cape Town: AFRIGRAPH; 2006. http://dx.doi.org/10.1145/1108590.1108616
30. Smelik RM, Tutenel T, Bidarra R, Benes B. A survey on procedural modelling for virtual worlds. Comput Graph Forum. 2014;33(6):31–50.
31. Fletcher D, Yue Y, Al Kader M. Challenges and perspectives of procedural modelling and effects. 2010 14th International Conference on Information Visualisation; 2010 Jul 10; London, UK. London: IEEE; 2010. p. 543–550. http://dx.doi.org/10.1109/IV.2010.80

32. Krecklau L, Kobbelt L. Procedural modelling of interconnected structures. Comput Graph Forum. 2011;30(2):335–344. http://dx.doi.org/10.1111/j.1467-8659.2011.01864.x

33. Beneš B, Šťava O, Měch R, Miller G. Guided procedural modelling. Comput Graph Forum. 2011;30(2):325–34. http://dx.doi.org/10.1111/j.1467-8659.2011.01886.x

34. Watson B, Müller P, Veryovka O. Procedural urban modelling in practice. IEEE Comp Graph. 2008;28(3):18–26. http://dx.doi.org/10.1109/MCG.2008.58

35. Wood J, Kirschenbauer S, Döllner J, Lopes A, Bodum L. Using 3D in visualisation. In: Dykes J, MacEachren AM, Kraak M-J, editors. Exploring geovisualisation. Oxford: Elsevier Ltd; 2005. p. 293–312. http://dx.doi.org/10.1016/B978-008044531-1/50432-2

36. Upson C, Faulhaber TA, Kamins D, Laidlaw D, Schlegel D, Vroom J, et al. The application visualisation system: A computational environment for scientific visualisation. IEEE Comp Graph. 1989;9(4):30–42. http://dx.doi.org/10.1109/38.31462

37. Haber R, McNabb D. Visualisation idioms: A conceptual model for scientific visualisation systems. In: Nielson G, Shriver B, Rosenblum L, editors. Visualisation in scientific computing. Los Alamitos, CA: IEEE Computer Society Press; 1990. p. 74–93.

38. Morton PJ, Horne M, Dalton RC, Thompson EM. Virtual city models : Avoidance of obsolescence. Digital physicality – Proceedings of the 30th eCAADe Conference. Czech Technical University, Prague. Prague: Czech Technical University; 2012. p. 213–24.

39. Biljecki F, Zhao J, Stoter J, Ledoux H. Revisiting the concept of level of detail in 3D city modelling. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci. 2013;II-2/W1:63–74. http://dx.doi.org/10.5194/isprsannals-II-2-W1-63-2013

40. Open Geospatial Consortium (OGC). City Geography Markup Language (CityGML) Encoding Standard 1.0.0. city:OGC; 2008.

41. Dickmann F, Dunker S. Visualisierung von 3D-Gebäudemodellen – Welche Ansprüche stellt die Planung an dreidimensionale Stadtansichten? [Digital 3D building models – which demands have planners?]. Kartographische Nachrichten; 2014;1:10–16. German.

42. Whittow J. The Penguin dictionary of physical geography. 2nd ed. London: Penguin Books; 2000. p. 590.

43. Council for Scientific and Industrial Research (CSIR): Building and construction technology. Guidelines for human settlement planning and design: Volume 1. Pretoria: CSIR: Building and Construction Technology; 2003. p. 283.

Assessing the effects of climate change on distributions of Cape Floristic Region amphibians

AUTHORS:
Mohlamatsane M. Mokhatla[1,2]
Dennis Rödder[3]
G. John Measey[2]

AFFILIATIONS:
[1]Department of Zoology, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
[2]Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Stellenbosch, South Africa
[3]Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany

CORRESPONDENCE TO:
Mohlamatsane Mokhatla

EMAIL:
m.mokhatla@mail.com

POSTAL ADDRESS:
Centre of Excellence for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Matieland 7602, South Africa

DATES:
Received: 05 Nov. 2014
Revised: 29 Apr. 2015
Accepted: 20 May 2015

KEYWORDS:
amphibian community; species range shifts; Cape Floristic Region; species distribution models; biotic velocity

HOW TO CITE:
Mokhatla MM, Rödder D, Measey GJ. Assessing the effects of climate change on distributions of Cape Floristic Region amphibians. S Afr J Sci. 2015;111(11/12), Art. #2014-0389, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140389

Climatic changes have had profound impacts on species distributions throughout time. In response, species have shifted ranges, adapted genetically and behaviourally or become extinct. Using species distribution models, we examined how changes in suitable climatic space could affect the distributions of 37 endemic frog species in the Cape Floristic Region (CFR) – an area proposed to have evolved its megadiversity under a stable climate, which is expected to change substantially in future. Species distributions were projected onto mean climate for a current period (1950 to 2000), hindcasted to palaeoclimate (Last Glacial Maximum; LGM ≈ 21 kya and Holocene Glacial Minimum; HGM ≈ 6 kya) and forecasted for two emissions scenarios (A2a and B2a) for the year 2080. We then determined the changes in area sizes, direction (longitude and latitude), fragmentation index and biotic velocity, and assessed if these were affected by life-history traits and altitude. We found that the biotic velocity at which the CFR amphibian community is expected to shift north (A2a ≈ 540.5 km/kya) and east (B2a ≈ 198 km/kya) far exceeds historical background rates (≈1.05 km/kya, north and west ≈ 2.36 km/kya since the LGM). Our models further suggest that the CFR amphibian community has already lost about 56% of suitable climate space since the LGM and this loss is expected to accelerate under future emission scenarios (A2a ≈ 70%; B2a ≈ 60%). Lastly, we found that highland species were more fragmented than lowland species between the LGM and current period, but that the fragmentation of lowland species between current and future climates is expected to increase.

Introduction

Climatic changes have had profound impacts on the survival and extinction rates of species through time.[1] One of the ways in which species have dealt with these climatic changes has been to adjust their distribution ranges[2], although changes in behaviour and genetic adaptation can also play a role[3,4]. Evidence from ice cores and other climate reconstruction studies[5,6] has provided better insight into how past climatic changes may have structured distributions as climates were changing. Recent evidence from fossil records suggests that past climatic fluctuations may have brought about changes at broadscale regional or biome level. Examples of such changes include the occurrence of lakes and 'forest-elements' within the current Sahara region (between 11 kya and 4 kya, under warmer and wetter conditions) and the replacement of forest by savannah vegetation in equatorial west Africa when conditions are predicted to have been warmer and drier than present.[7]

Studies undertaken in southern Africa, particularly in the Cape Floristic Region (CFR) suggest that climate has remained relatively stable throughout the Quaternary.[5] Indeed, areas such as the CFR that are proposed to have experienced relatively stable climate coincide with contemporary centres of richness and narrowly distributed endemic species.[2] These areas may have served as climate refugia during times of climatic perturbations.[8,9] Moreover, such areas are also proposed to have experienced relatively slow climate change velocities.[10,11] Conversely, areas that have experienced rapid changes in climate are suggested to harbour biota that have coupled biotic velocities: the velocity of shifts in species' climate envelopes.[12] Understanding how communities respond to these changes is very important given the biological importance of megadiverse regions under threat.[13] Amphibians of the CFR in particular provide an excellent group to study the role of climate in structuring species community distribution because: (1) they are biogeographically distinct at a global scale[14] and (2) amphibians are generally considered to be poor dispersers[15], and are thus more directly affected by changes in climate compared to other vertebrates.

While attempting to elucidate the importance of ecological factors in describing amphibian species distribution patterns in southern Africa, Poynton[16] hypothesised that the climatic oscillations resulting from Milankovitch cycles would have driven the expansion of the CFR amphibian assemblage inland as a response to an increase in climatic suitability during glacial periods. Dynesius and Jansson[2] suggest that large-scale climatic oscillations may have favoured species with large distribution ranges, because these species were able to migrate to climatically suitable areas or persist in climatically unsuitable areas[17,18]. Furthermore, other past climate models for the CFR argue that during glacial periods, the extent of the CFR's winter rainfall area increased as a result of westerlies migrating north towards the equator, providing moisture to areas of the current Karoo region,[19] thus supporting Poynton's hypothesis. Until recently, it had been very difficult to test such hypotheses. However, using species distribution modelling techniques, a recent test of Poynton's hypothesis concluded that he was correct in predicting range expansion of the CFR assemblage during glacial periods.[20] Nevertheless, instead of expanding inland, the assemblage was predicted to have made use of the exposed Agulhas bank during periods when sea level was much lower.[5]

Future climate models for southern Africa predict increased drying in the winter rainfall region[21,22], caused by the increase in both Indian and Atlantic sea surface temperatures and the El Niño Southern Oscillation[23]. Understanding how changes in climate may have structured distribution is of particular relevance as it will help in: (1) predicting how current changes in climate might affect distributions as well as interactions among species[24] in the CFR and (2) adequately planning for these changes considering the climatic history of the region[2] and other biodiversity threats such as habitat modification[25]. The aim of this study is to determine how climate (palaeo-, current and future) affects the distribution of the CFR endemic amphibian community. To do this we adopt an ensemble

modelling approach using general circulation models (GCMs) of two palaeoclimates and four future climate GCMs, projected under two emission scenarios. In particular, we determine how the distributions of the CFR amphibian community have changed in terms of: (1) direction (latitude and longitude), (2) area of occupancy (changes in areas occupied at different climate periods), (3) biotic velocities (the rate at which communities shift with changes in climate)[12] and (4) fragmentation. Furthermore, we estimate how life-history traits and altitudinal specialisation affect these outcomes. We assume that shifts (direction of movement and size) in climatically suitable space across climatic periods will lead to coupled species range shifts in the CFR amphibian community.[20]

Methods

Data

The most recent taxonomic treatment of species was used to ensure that all species of interest were included.[26-29] Approximately 40 species are endemic to the CFR (here taken as areas representing both the Fynbos and the Succulent Karoo Biomes[30]; Figure 1). We excluded three recently described species (*Breviceps branchi*[31], *Cacosternum aggestum* and *C. australis*[32]) because of the relative lack of distribution records, leaving 37 CFR endemic amphibian species from six families for further analysis. Some amphibians do not display a biphasic typical amphibian lifestyle, for example members of the direct developing genus *Arthroleptella* endemic to the CFR[33], which leads to a variation in the ways each life-history trait will be affected by threats[34]. To assess the effect of life-history traits on species distributions resulting from changes in climate, species were grouped according to their modes of egg deposition and development[35] (Supplementary table 1 online).

There was a high variation in the number of available species distribution records, from *Cacosternum platys*, Rose 1950 (4 records) to *Tomopterna delalandii*, Tschudi 1838 (541 records). Occurrence records were obtained from the South African Frog Atlas Project (SAFAP)[33], CapeNature and iSpot, as well as records published in recent literature[26,29,36] and our own observations. We included only records with accuracy to the nearest second or higher. Records were projected into ArcMap GIS 10.2[37] to identify any doubtful records that may have been recorded outside known ranges, using Global Amphibian Assessment (IUCN 2008; http://www.iucnredlist.org/; accessed September 2013) polygons as 'known distribution ranges'. Based on these results, we judged whether the outlying records represented credible sightings or should be excluded from further analyses.

Climate data

Environmental predictors were compiled from a set of 19 bioclimatic variables at a spatial resolution of 2.5 arc minutes available from WorldClim.[38] Inter-correlation structure among predictors throughout the study area was assessed by computing pairwise squared Spearman's rank correlation coefficients, and in cases in which R^2 exceeded 0.75, only the putatively biologically most important variables were retained. After computing the Spearman's rank correlations, 10 variables were retained: isothermality (mean diurnal range/temperature annual range; BIO3), temperature annual range (BIO7), temperature of wettest quarter (BIO8), mean temperature of driest quarter (BIO9), mean temperature of warmest quarter (BIO10), mean temperature of coldest quarter (BIO11), precipitation of wettest month (BIO13), precipitation seasonality (BIO15), precipitation of driest quarter (BIO17) and precipitation of coldest quarter (BIO19); see Supplementary table 2 online for percentage variable contributions.

Species distribution modelling

Species distribution model projections were created at a 2.5 arc minute resolution, using current distribution data to reconstruct hypothetical distributions for palaeo- and future climate. To reconstruct past distributions, we projected species distribution models onto palaeo-climate for the Last Glacial Maximum (LGM $\approx$ 21 kya) and the Mid-Holocene/Holocene Glacial Minimum (HGM $\approx$ 6 kya) climate periods. We used two GCMs: the Community Climate System Model (CCSM)[39] and the Model for Interdisciplinary Research on Climate (MIROC)[40], available through http://pmip2.lsce.ipsl.fr/[41]. Palaeoclimate data were downscaled to a spatial resolution of 2.5 arc minutes following the delta method.[42] Current distributions were based on species distribution modelling projections onto mean climate from 1950 to 2000. To forecast future distributions for the year 2080, species distribution models were projected using four GCMs, to capture the variability associated with predicting climate change: (1) General Circulation Global Climate Model (CGCM2)[43], (2) Hadley Centre Coupled Model (HadCM3)[44], (3) Commonwealth Scientific and Industrial Research Organisation Mark-2b (CSIRO-Mk2)[45] and (4) National Institute for Environmental Studies (NIES 99)[46]. Each GCM was projected onto two Intergovernmental Panel on Climate Change emission scenarios, both representing a more heterogeneous world: (1) A2a representing a business as usual climate outlook with average temperature increases between 2.0 °C and 5.4 °C and (2) B2a representing a more optimistic outlook with a moderate increase in future temperature of between 1.4 °C and 3.8 °C.[47]

The selected environmental variables were used to model both past and future potential distributions of amphibian species using an ensemble species distribution modelling framework, BIOMOD2[48] – a multi-model platform implemented in R.[49] The BIOMOD2 procedure uses an ensemble of modelling techniques including generalised linear models, general additive models, generalised boosted models, classification tree analysis, artificial neural networks, surface range envelope, flexible discriminant analysis, multi-adaptive regression splines, random forests, including maximum entropy to produce consensus models, which are weighted averages between all models.[50] To ensure that our models were comparable across scale, they were rescaled using a binomial generalised linear model.

We randomly selected 1000 pseudo absence records, three times within a circular buffer of 100 km enclosing the species records as environmental background. For model evaluation, we divided the species records into 80% used to train the models and 20% for model evaluation using three different performance criteria (area under the receiver operating characteristic curve, AUC[51]; true skills statistics; and Cohen's kappa[52]). As a quality threshold for the final ensemble, we selected only those models with an AUC score greater than 0.7 (Supplementary table 3 online). The final ensemble was built using a proportional weighting of the models according to their predictive performance. As a presence-absence threshold, we selected the minimum training presence as the non-fixed threshold.[53] When projecting the ensemble across space and time, non-analogous climatic conditions that exceeded the training range of the models were quantified, wherein the numbers of these predictors were highlighted using multivariate environmental similarity surfaces (MESS)[54] (see online Supplementary figures 1 and 2: palaeo-, current and future projections for *Arthroleptella bicolor* as an example). Projections into non-analogous climates may increase the potential for extrapolation and interpolation errors.[54] Potential areas of occurrence were computed based on presence-absence maps applying the minimum training presence threshold using the respective functions in the raster package for R (i.e. 'area') but excluding extrapolation areas as identified by MESS. To determine the final consensus area projection for each climate period across all models (different GCMs), we projected all models on to ArcMap 10.2[37] and the consensus was taken as a combination of all grid cells that were identified as suitable by 50% or more of the models at each climate period[55].

Statistical analyses

To assess changes in species distributions across climatic periods, we calculated the total area, perimeter and the centroid coordinates for each species across climatic periods from suitable climate space projections. Area and perimeter were then used to determine the fragmentation index (i.e. the landscape dissection index, DI´)[56]. This was achieved by determining the edge of each species distribution range, in relation to the core of the distribution using the formula:

$$DI' = \sum_{i=1}^{n} P / \left(2\sqrt{\pi \sum_{i=1}^{n} A}\right), \qquad \text{Equation 1}$$

where P is the perimeter and A is the area. DI´ compares the amount by which any 'island' or a distribution range is dissected relative to the total

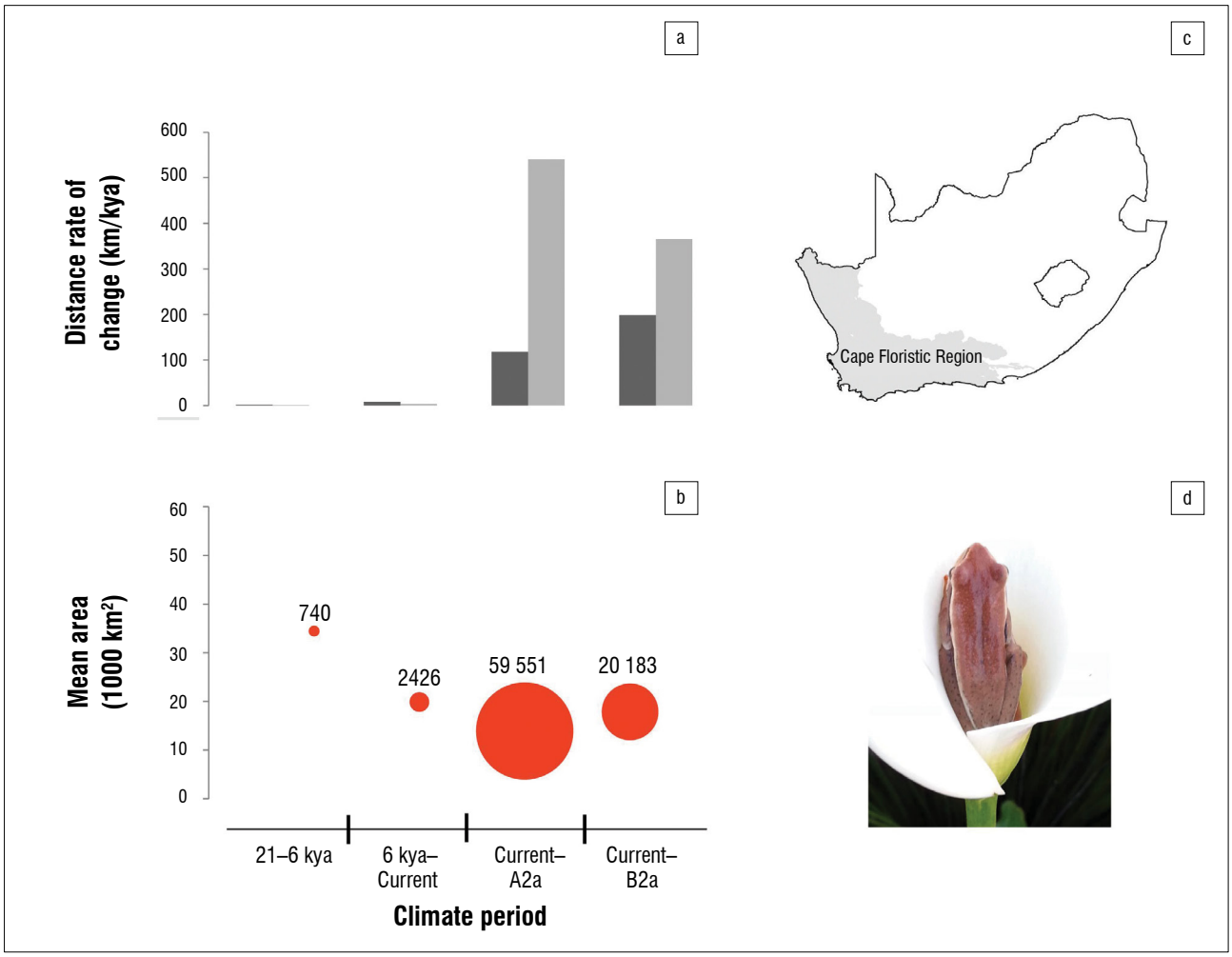


Figure 1: Changes in distribution predicted by species distribution models for the endemic amphibian community of the Cape Floristic Region (CFR). (a) Average longitude (grey bars) and latitude (black bars) rate of change between climate periods. (b) Mean area distributions range size across climatic periods. The size of the circle is proportional to the rate of area change within the climate period and the numbers above the circles are the mean predicted rates of area change between climatic periods in km²/kya. (c) Map of South Africa showing the extent of the CFR (Fynbos and Succulent Karoo Biomes). (d) *Hyperolius horstockii*, the Arum lily frog, is endemic to the CFR.

area, with a DI´ of 1 representing a circle, suggesting a less fragmented shape with a low perimeter to area ratio. We determined changes in area occupancy, rate of area change, the direction of range shifts (across both longitudes and latitudes) and fragmentation index (DI´) between climate periods.[57,58] We ran our results in two ways: first with a full suite of 37 species and second with 34 species (excluding those species with less than 10 species records, to avoid model over-fitting[59]). We found that the results in both attempts were almost identical and that patterns remained largely similar and thus report results of the analysis considering a full suite of 37 species. Lastly, we checked whether these changes were affected by altitude and life-history trait. All the tests were performed in R.[49]

Results

Range shifts

Our results suggest that the amphibian community of the CFR would have responded to Quaternary climatic changes by shifting distributions. Longitudinally, we found that between the LGM and current periods, amphibian communities have shifted their distribution northwards, and are expected to shift further northwards, for both A2a and B2a emission scenarios. Biotic velocity has been relatively slow since the LGM to current, but it is expected to increase dramatically if amphibian communities are assumed to track climate suitable space under anthropogenically induced climate change towards the end of this century (Figure 1a; e.g. LGM–HGM = 1.058 km/kya compared with current–A2a = 540.554 km/kya). Across latitudes, our models suggest that the CFR amphibian community has shifted its range westward between the LGM and HGM, as well as between the HGM and current. However between current and 2080, our models suggest that the CFR amphibian community will shift eastward under both A2a and B2a emission scenarios, although we found this shift between climate periods to be significant for the B2a emission scenario only (paired *t*-test; $t=2.119$; $p=0.039$). Biotic velocity across latitudes between current and B2a climate periods is expected to occur at rapid rates compared with historical rates (Figure 1a; LGM–HGM = 2.360 km/kya compared with current–B2a = 198.319 km/kya).

Changes in area of occupancy

We found that the CFR amphibian community has progressively lost ≈56% of its suitable climate space since the LGM. Furthermore it is expected to lose ≈70% and 60% of climatically suitable space under A2a and B2a emission scenarios, respectively (Figure 2a). However, we found that the rate at which the CFR endemic amphibian community has lost climatically suitable space has been relatively slow between the LGM and HGM and HGM and current, but this rate is expected to increase significantly under future climate change (in particularly, under the A2a emission scenario; Figure 1b).

Landscape dissection index

Fragmentation of potential distributions increased slightly from LGM to HGM, but decreased between HGM and current climates (Figure 2b). Our models predict that fragmentation will increase under future scenarios. We did not find any significant difference between life-history traits and DI´. In contrast, we found that the fragmentation of highland species was significantly higher than that of lowland species between HGM and the current climate (Mann–Whitney–Wilcox test; $W=37$, $p=0.002$). However, between current and future climate periods under the A2a emission scenario, fragmentation of lowland species is expected to be higher than that of highland species ($W=18$, $p=0.033$). When looking at each climate period separately, we found that highland species were significantly more fragmented than lowland species during the LGM ($W=113$, $p=0.035$), HGM ($W=162$, $p=0.012$) and current scenarios ($W=233$, $p=0.023$) (Figure 3). Lastly, we found that the rate at which DI´ changed was only significant between HGM and the current (paired *t*-test; $t=2.065$, $p=0.047$).

Discussion

Quaternary climate changes have had significant impacts on the distributions of amphibian communities in the CFR as predicted[16] and modelled[20]. Our results suggest that climatically suitable areas for the CFR amphibian community have shifted northwards since the LGM, and are expected to shift further northwards under both A2a and B2a future emissions scenarios. Latitudinally, we found that the amphibian community shifted westward from the LGM through to current, but this trend is expected to change to an eastwards directional community shift under both A2a and B2a emission scenarios. Furthermore, our models indicate that fragmentation is expected to increase under future climatic scenarios (Figure 2b) as amphibian communities are predicted to continuously lose suitable climatic space (Figure 2a). These results largely concur with previous studies undertaken in the region.[60]

One multi-taxon study which included birds, mammals, reptiles and insects for the whole of South Africa suggested that about 78% of the species included in the analysis are expected to experience range contractions, while 41% are expected to shift their ranges eastward, in response to future climatic changes by the year 2050.[60] Other studies assessing the impacts of future climate change in the CFR also predicted an eastward directional shift and concluded that this shift will lead to large-scale spatial mismatches with current protected areas portfolio.[61] The predicted range shifts and associated impacts are expected to result in as much as 50% loss in current climate space[62] and as much as 10% of the endemic plant species.[63] In addition, we found that the CFR amphibian community has lost ≈56% of its suitable climate. Indeed, several studies looking at anurans of southern Africa suggest that they have already lost ≈60% of suitable climate space[64,65], such that current distributions could very well be representative of refugial distributions for most of the endemic Cape species[20].

We found that fragmentation (as measured by the index DI´) between highland and lowland species was more pronounced during the LGM, HGM and current distributions (Figure 3), with highland species being more fragmented than lowland species. When looking at fragmentation between climate periods, we found that the fragmentation index of lowland species is expected to be higher than that of highlands species, suggesting that projected changes in climate will exert pressure on the CFR lowland species in addition to substantial changes in land use.[66] This result is of significance considering that the CFR lowlands have already been transformed,[66] mainly for agriculture and human settlements, and will become more fragmented in the future as suitable climate space is lost to climate change. These results are supported by several studies suggesting that the coupled threat posed by climate change and land use will be very important, particularly for sub-Saharan African amphibians in the future.[25,67,68] Furthermore, the suggested increase in fragmentation of lowland species points to the likely increase in movement along altitudinal gradients[69] as species migrate to highlands in response to projected climate change[65]. This finding further emphasises the importance of improving habitat linkages between the CFR highlands and lowlands,[70] thus retaining the opportunity for altitudinal migrations, which may be critical for poorly dispersing taxa such as amphibians.

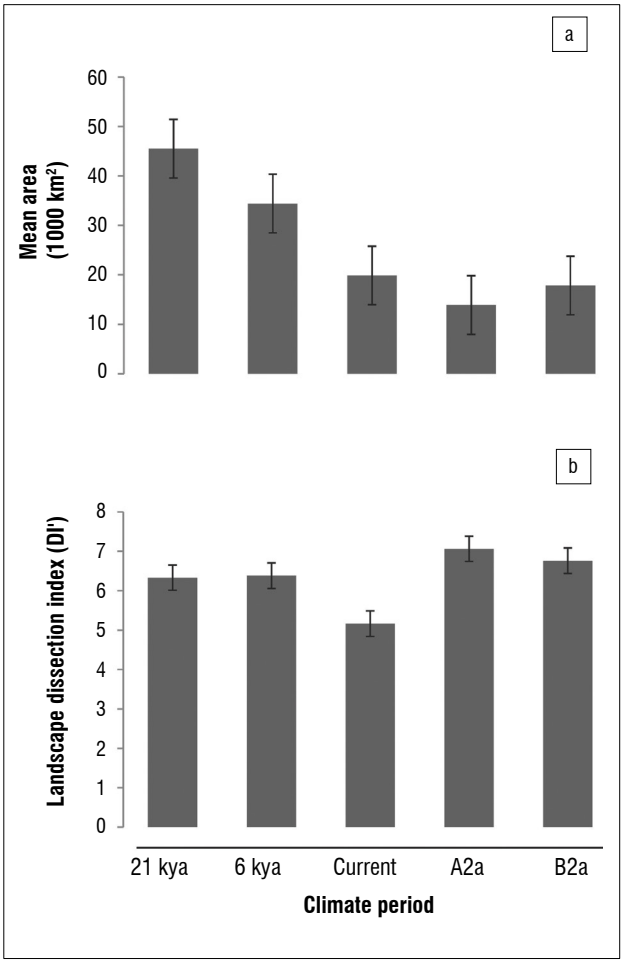


Figure 2: (a) Mean (± standard error) of predicted area occupied by the Cape Floristic Region (CFR) endemic amphibian community suggests that most are already in refugia. (b) Mean (± standard error) of predicted fragmentation (DI´) in the CFR endemic amphibian community under different climate scenarios.

Our results, based on future climate change models for the CFR, suggest significant range reductions for the CFR amphibian community as well as rapid rates of suitable area loss. However, certain issues must be taken into consideration when interpreting these results. Firstly, despite their importance in conservation biology, correlative models such as species distribution modelling techniques are largely based on the assumption that species distributions are in tune with current climatic factors, and this assumption could lead to an under-estimation of what the real future impact might be[15], specifically when considered in the light of other biodiversity threats such as habitat destruction[71]. Secondly, our landscape dissection index DI´ does not standardise for area size.[71] However, controlling for area size in predictive modelling studies may be problematic as these areas are largely speculative. Nonetheless, these methods are still informative and our predictions are in line with those of previous studies.[64] Thirdly, the rate of change (as determined in this study) for past and future models, are simple averages taken across climatic periods. We are, however, of the view that past climatic changes, resulting from Milankovitch cycles, provided species with time to track climatically suitable areas because landscapes were untransformed. We are also aware that species range shifts are but one of several strategies amphibian communities can employ in response to climatic change: other strategies include behavioural, physiological and genetic adaptations.[3,4]

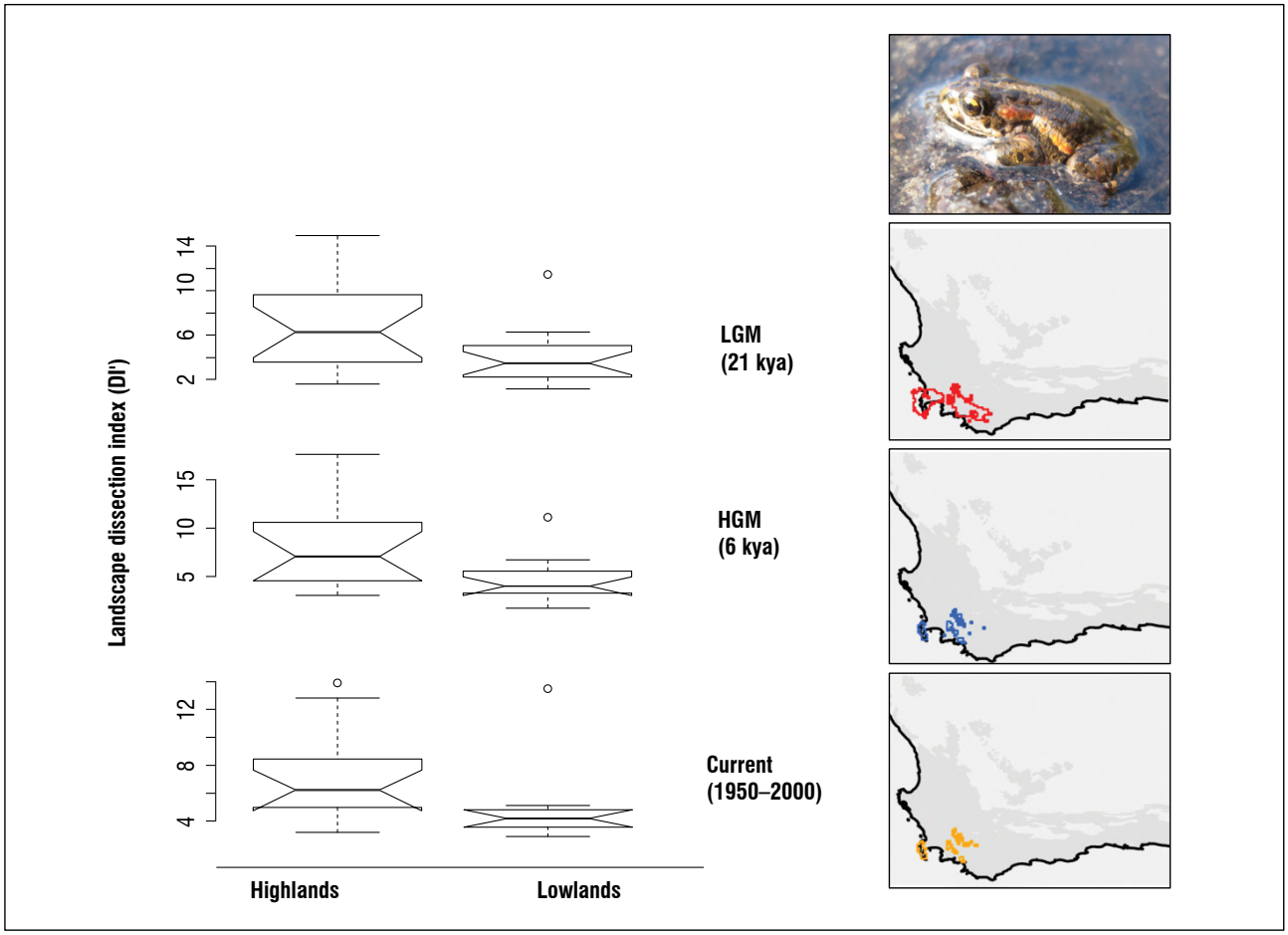


Figure 3: Landscape dissection index (DI´) compared between lowland and highland species of the Cape Floristic Region (CFR) endemic amphibian community during the Last Glacial Maximum (LGM), Holocene Glacial Minimum (HGM) and current period. (Right) Hypothetical distribution of a highland species, *Capensibufo rosei* (Rose's mountain toadlet), across the respective climate periods. The figure shows a continuous distribution during the LGM to a more fragmented distribution during the HGM and currently.

In conclusion, we suggest that the results of this study provide important baseline information for a region-wide assessment of the predicted impacts of climate change on a highly diverse vertebrate community of reduced vagility. If we are to reduce the threat of biodiversity loss as a result of climate change within the CFR, several factors need to be addressed. Firstly, climate change should be included in future protected areas planning efforts.[72,73] Secondly, efforts aimed at improving the connectivity between highland and lowlands as well as the breeding (e.g. aquatic) and the non-breeding (e.g. terrestrial) habitats of amphibians within the CFR need to be intensified. Thirdly, solutions such as the land stewardship programmes entered into with private landowners should be accelerated[74], considering that most biodiversity is found in privately owned land in the CFR[75]. These solutions are particularly important if we consider that amphibians have strong habitat specificity, which is largely driven by their generally poor dispersal abilities.[76] This specificity results in deep phylogenetic structures at relatively small spatial scales[15] which could be an outcome of both climate stability and topography[77]. We advocate that future models should try to incorporate the use of different GCMs[55] because such consensus models outperform individual GCMs in predicting consensus climatic space. We also suggest that, if possible, these GCMs should be augmented with mechanistic models[78] as species are more likely to use different responses available (behaviour, physiology and range shifts) when dealing with predicted changes in climate.

Acknowledgements

We thank the numerous volunteers who contributed their amphibian species distribution data to SAFAP, CapeNature, and iSpot, and Werner Conradie, Lemmy Mashinini, Sarah Davies, Andrew Turner and Tony Rebelo. We also thank Peter Bradshaw for advice on GIS and Giovanni Vimercati for reading through previous drafts of this manuscript. We acknowledge the editor and two anonymous reviewers, whose comments helped to improve the quality of the manuscript. M.M.M. acknowledges support from the NRF Innovation bursary and the Nelson Mandela Metropolitan University. G.J.M. acknowledges the DST-NRF Centre of Excellence for Invasion Biology for support.

Authors' contributions

G.J.M. and M.M.M. conceived the ideas. D.R. and M.M.M. constructed species distribution models. All authors analysed the data and wrote the manuscript.

References

1. Wake DB, Vredenburg VT. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. Proc Natl Acad Sci USA. 2008;105:11466–11473. http://dx.doi.org/10.1073/pnas.0801921105
2. Dynesius M, Jansson R. Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. Proc Natl Acad Sci USA. 2000;97:9115–9120. http://dx.doi.org/10.1073/pnas.97.16.9115
3. Davis MB, Shaw RG. Range shifts and adaptive responses to Quaternary climate change. Science. 2001;292:673–679. http://dx.doi.org/10.1126/science.292.5517.673
4. Jump AS, Penuelas J. Running to stand still: Adaptation and the response of plants to rapid climate change. Ecol Lett. 2005;8:1010–1020. http://dx.doi.org/10.1111/j.1461-0248.2005.00796.x

5. Compton JS. Pleistocene sea-level fluctuations and human evolution on the southern coastal plain of South Africa. Quaternary Sci Rev. 2011;30:506–527. http://dx.doi.org/10.1016/j.quascirev.2010.12.012

6. Zachos J, Pagani M, Sloan L, Thomas E, Billups K. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science. 2001;292:686–693. http://dx.doi.org/10.1126/science.1059412

7. Willis KJ, Bennett KD, Burrough SL, Macias-Fauria M, Tovar C. Determining the response of African biota to climate change: Using the past to model the future. Philos T Roy Soc B. 2013;368, Art. #20120491, 9 pages. http://dx.doi.org/10.1098/rstb.2012.0491

8. Araújo MB, Nogués-Bravo D, Diniz-Filho JAF, Haywood AM, Valdes PJ, Rahbek C. Quaternary climate changes explain diversity among reptiles and amphibians. Ecography. 2008;31:8–15. http://dx.doi.org/10.1111/j.2007.0906-7590.05318.x

9. Jansson R. Global patterns in endemism explained by past climatic change. P Roy Soc Lond B Bio. 2003;270:583–590. http://dx.doi.org/10.1098/rspb.2002.2283

10. Loarie SR, Duffy PB, Hamilton H, Anser GP, Field CB, Ackerly DD. The velocity of climate change. Nature. 2009;462:1052–1055. http://dx.doi.org/10.1038/nature08649

11. Sandel B, Arge L, Dalsgaard B, Davies RG, Gaston KJ, Sutherland WJ, et al. The influence of Late Quaternary climate-change velocity on species endemism. Science. 2011;334:660–664. http://dx.doi.org/10.1126/science.1210173

12. Ordonez A, Williams JW. Climatic and biotic velocities for woody taxa distributions over the last 16 000 years in eastern North America. Ecol Lett. 2013;16:773–781. http://dx.doi.org/10.1111/ele.12110

13. Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J. Biodiversity hotspots for conservation priorities. Nature. 2000;403:853–858. http://dx.doi.org/10.1038/35002501

14. Holt BG, Lessard J-P, Borregaard MK, Fritz SA, Araújo MB, Dimitrov D, et al. An update of Wallace's zoogeographic regions of the world. Science. 2013;339:74–78. http://dx.doi.org/10.1126/science.1228282

15. Tolley KA, De Villiers AL, Cherry MI, Measey GJ. Isolation and high genetic diversity in dwarf mountain toads *(Capensibufo)* from South Africa. Biol J Linn Soc. 2010;100:822–834. http://dx.doi.org/10.1111/j.1095-8312.2010.01469.x

16. Poynton J. The biotic divisions of southern Africa, as shown by the Amphibia. In: Davis DHS, editor. Ecological studies in southern Africa. The Hague: Junk; 1964. p. 206–218.

17. Jansson R, Dynesius M. The fate of clades in a world of recurrent climatic change: Milankovitch oscillations and evolution. Annu Rev Ecol Syst. 2002;33:741–777. http://dx.doi.org/10.1146/annurev.ecolsys.33.010802.150520

18. Van Bocxlaer I, Loader SP, Roelants K, Biju SD, Menegon M, Bossuyt F. Gradual adaptation toward a range-expansion phenotype initiated the global radiation of toads. Science. 2010;327:679–682. http://dx.doi.org/10.1126/science.1181707

19. Chase BM, Meadows ME. Late Quaternary dynamics of southern Africa's winter rainfall zone. Earth-Sci Rev. 2007;84:103–138. http://dx.doi.org/10.1016/j.earscirev.2007.06.002

20. Schreiner C, Rödder D, Measey GJ. Using modern models to test Poynton's predictions. Afr J Herpetol. 2013;62:49–62. http://dx.doi.org/10.1080/21564574.2013.794865

21. Engelbrecht C, Engelbrecht F, Dyson L. High-resolution model-projected changes in mid-tropospheric closed-lows and extreme rainfall events over southern Africa. Int J Climatol. 2013;33:173–187. http://dx.doi.org/10.1002/joc.3420

22. Giannini A, Biasutti M, Held IM, Sobel AH. A global perspective on African climate. Climatic Change. 2008;90:359–383. http://dx.doi.org/10.1007/s10584-008-9396-y

23. Hoerling M, Hurrell J, Eischeid J, Phillips A. Detection and attribution of twentieth-century northern and southern African rainfall change. J Clim. 2006;19:3989–4008. http://dx.doi.org/10.1175/JCLI3842.1

24. Blois JL, Zarnetske PL, Fitzpatrick MC, Finnegan S. Climate change and the past, present, and future of biotic interactions. Science. 2013;341:499–504. http://dx.doi.org/10.1126/science.1237184

25. Newbold T, Hudson LN, Hill SLL, Contu S, Lysenko I, Senior RA, et al. Global effects of land use on local terrestrial biodiversity. Nature. 2015;520:45–50. http://dx.doi.org/10.1038/nature14324

26. Channing A, Wahlberg K. Distribution and conservation status of the desert rain frog *Breviceps macrops*. Afr J Herpetol. 2011;60:101–112. http://dx.doi.org/10.1080/21564574.2011.608383

27. Frost DR, Grant T, Faivovich J, Bain RH, Haas A, Haddad CFB, et al. The amphibian tree of life. B Am Mus Nat Hist. 2006;297:1–291. http://dx.doi.org/10.1206/0003-0090(2006)297[0001:TATOL]2.0.CO;2

28. Turner AA, Channing A. A new species of *Arthroleptella* Hewitt, 1926 (Anura: Pyxicephalidae) from the Klein Swartberg Mountain, Caledon, South Africa. Afr J Herpetol. 2008;57:1–12. http://dx.doi.org/10.1080/21564574.2008.9635564

29. Turner AA, De Villiers AL, Dawood A, Channing A. A new species of *Arthroleptella* Hewitt, 1926 (Anura: Ranidae) from the Groot Winterhoek Mountains of the Western Cape Province, South Africa. Afr J Herpetol. 2004;53:1–12. http://dx.doi.org/10.1080/21564574.2004.9635493

30. Colville JF, Potts AJ, Bradshaw PL, Measey GJ, Snijman D, Picker MD, et al. Floristic and faunal Cape biochoria: Do they exist? In: Allsopp N, Colville JF, Verboom GA, editors. Fynbos: Ecology, evolution, and conservation of a megadiverse region. Oxford: Oxford University Press; 2014. p. 73–92. http://dx.doi.org/10.1093/acprof:oso/9780199679584.001.0001

31. Channing A. A new species of rain frog from Namaqualand, South Africa (Anura: Brevicipitidae: *Breviceps*). Zootaxa. 2012;3381:62–68.

32. Channing A, Schmitz A, Burger M, Kielgast J. A molecular phylogeny of African Dainty Frogs, with the description of four new species (Anura: Pyxicephalidae: *Cacosternum*). Zootaxa. 2013;3701:518–550. http://dx.doi.org/10.11646/zootaxa.3701.5.2

33. Minter LR, Burger M, Harrison JA, Braak HH, Bishop PJ, Kloepfer D. Atlas and red data book of the frogs of South Africa, Lesotho, and Swaziland. Washington DC/Cape Town: Smithsonian Institute/Avian Demography Unit; 2004.

34. Mokhatla MM, Measey GJ, Chimimba CT, Van Rensburg BJ. A biogeographical assessment of anthropogenic threats to areas where different frog breeding groups occur in South Africa: Implications for anuran conservation. Divers Distrib. 2012;18:470–480. http://dx.doi.org/10.1111/j.1472-4642.2011.00870.x

35. Wells KD. The ecology and behavior of amphibians. Chicago, IL: University of Chicago Press; 2007. http://dx.doi.org/10.7208/chicago/9780226893334.001.0001

36. Measey GJ, Tolley KA. Investigating the cause of the disjunct distribution of *Amietophrynus pantherinus*, the endangered South African western leopard toad. Conserv Genet. 2011;12:61–70. http://dx.doi.org/10.1007/s10592-009-9989-7

37. ArcGIS Desktop version 10.2. Redlands, CA: Environmental Systems Research Institute; 2011.

38. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965–1978. http://dx.doi.org/10.1002/joc.1276

39. Otto-Bliesner BL, Brady EC, Clauzet G, Tomas R, Levis S, Kothavala Z. Last Glacial Maximum and Holocene climate in CCSM3. J Climate. 2006;19:2526–2544. http://dx.doi.org/10.1175/JCLI3748.1

40. Hasumi H, Emori S. K-1 coupled gcm (miroc) description. Tokyo: Center for Climate System Research, University of Tokyo; 2004.

41. Braconnot P, Otto-Bliesner B, Harrison S, Joussaume S, Peterchmitt J-Y, Abe-Ouchi A, et al. Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum Part 1: Experiments and large-scale features. Clim Past. 2007;3:261–277. http://dx.doi.org/10.5194/cp-3-261-2007

42. Peterson AT, Nyari AS. Ecological niche conservatism and Pleistocene refugia in the thrush-like mourner, *Schiffornis* sp., in the neotropics. Evolution. 2008;62:173–183.

43. Flato G, Boer G. Warming asymmetry in climate change simulations. Geophys Res Lett. 2001;28:195–198. http://dx.doi.org/10.1029/2000GL012121

44. Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, et al. The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dynam. 2000;16:147–168. http://dx.doi.org/10.1007/s003820050010

45. Gordon HB, O'Farrell SP. Transient climate change in the CSIRO coupled model with dynamic sea ice. Mon Weather Rev. 1997;125:875–908. http://dx.doi.org/10.1175/1520-0493(1997)125<0875:TCCITC>2.0.CO;2

46. Emori S, Nozawa T, Abe-Ouchi, A, Numaguti A, Kimoto M, Nakajima T. Coupled ocean–atmosphere model experiments of future climate change with an explicit representation of sulfate aerosol scattering. J Meteor Soc Japan. 1999;77:1299–1307.

47. Nakicenovic N, Swart R, editors. Special report on emissions scenarios. Cambridge, UK: Cambridge University Press; 2000.

48. Thuiller W, Georges D, Engler R. biomod2: Ensemble platform for species distribution modeling. R package version 2; 2013.

49. R Development Core Team. R: A language and environment for statistical computing; 2014.

50. Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W. Evaluation of consensus methods in predictive species distribution modelling. Divers Distrib. 2009;15:59–69. http://dx.doi.org/10.1111/j.1472-4642.2008.00491.x

51. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988;240:1285–1293. http://dx.doi.org/10.1126/science.3287615

52. Allouche O, Tsoar A, Kadmon R. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). J Appl Ecol. 2006;43:1223–1232. http://dx.doi.org/10.1111/j.1365-2664.2006.01214.x

53. Liu C, Berry PM, Dawson TP, Pearson RG. Selecting thresholds of occurrence in the prediction of species distributions. Ecography. 2005;28:385–393. http://dx.doi.org/10.1111/j.0906-7590.2005.03957.x

54. Elith J, Phillips SJ, Hastie T, Dudik M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Divers Distrib. 2011;17:43–57. http://dx.doi.org/10.1111/j.1472-4642.2010.00725.x

55. Forester BR, DeChaine EG, Bunn AG. Integrating ensemble species distribution modelling and statistical phylogeography to inform projections of climate change impacts on species distributions. Divers Distrib. 2013;19:1480–1495. http://dx.doi.org/10.1111/ddi.12098

56. Bowen GW, Burgess R. Quantitative analysis of forest island pattern in selected Ohio landscapes. Oak Ridge, TN: Oak Ridge National Laboratory; 1981. http://dx.doi.org/10.2172/6367140

57. Van Der Wal J, Murphy HT, Kutt AS, Perkins GC, Bateman BL, Perry JJ, et al. Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change. Nat Clim Change. 2013;3:239–243. http://dx.doi.org/10.1038/nclimate1688

58. Wollenberg KC, Vieites DR, Glaw F, Vences M. Speciation in little: The role of range and body size in the diversification of Malagasy mantellid frogs. BMC Evol Biol. 2011;11:217. http://dx.doi.org/10.1186/1471-2148-11-217

59. Platts PJ, Garcia RA, Hof C, Foden W, Hansen LA, Rahbek C, et al. Conservation implications of omitting narrow-ranging taxa from species distribution models, now and in the future. Divers Distrib. 2014;20:1307–1320. http://dx.doi.org/10.1111/ddi.12244

60. Erasmus BF, Van Jaarsveld AS, Chown SL, Kshatriya M, Wessels KJ. Vulnerability of South African animal taxa to climate change. Glob Change Biol. 2002;8:679–693. http://dx.doi.org/10.1046/j.1365-2486.2002.00502.x

61. Hannah L, Midgley G, Hughes G, Bomhard B. The view from the Cape: Extinction risk, protected areas, and climate change. BioScience. 2005;55:231–242. http://dx.doi.org/10.1641/0006-3568(2005)055[0231:TVFTCE]2.0.CO;2

62. Altwegg R, West A, Gillson L, Midgley GF. Impacts of climate change in the Greater Cape Floristic Region. In: Allsopp N, Colville JF, Verboom GA, editors. Fynbos: Ecology, evolution, and conservation of a megadiverse region. Oxford: Oxford University Press; 2014. p. 299–320. http://dx.doi.org/10.1093/acprof:oso/9780199679584.003.0013

63. Midgley G, Hannah L, Millar D, Rutherford M, Powrie L. Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. Global Ecol Biogeogr. 2002;11:445–451. http://dx.doi.org/10.1046/j.1466-822X.2002.00307.x

64. Botts EA, Erasmus BF, Alexander GJ. Methods to detect species range size change from biological atlas data: A comparison using the South African Frog Atlas Project. Biol Conserv. 2012;146:72–80. http://dx.doi.org/10.1016/j.biocon.2011.10.035

65. Botts EA, Erasmus BF, Alexander GJ. Observed range dynamics of South African amphibians under conditions of global change. Austral Ecol. 2015;40(3):309–317. http://dx.doi.org/10.1111/aec.12215

66. Bomhard B, Richardson DM, Donaldson JS, Hughes GO, Midgley GF, Raimondo DC, et al. Potential impacts of future land use and climate change on the Red List status of the Proteaceae in the Cape Floristic Region, South Africa. Glob Change Biol. 2005;11:1452–1468. http://dx.doi.org/10.1111/j.1365-2486.2005.00997.x

67. Foden WB, Butchart SHM, Stuart SN, Vié J-C, Akçakaya HR, Angulo A, et al. Identifying the world's most climate change vulnerable species: A systematic trait-based assessment of all birds, amphibians and corals. PLoS One. 2013;8, e65427. http://dx.doi.org/10.1371/journal.pone.0065427

68. Hof C, Araújo MB, Jetz W, Rahbek C. Additive threats from pathogens, climate and land-use change for global amphibian diversity. Nature. 2011;480:516–519. http://dx.doi.org/10.1038/nature10650

69. Parmesan C. Ecological and evolutionary responses to recent climate change. Annu Rev Ecol Evol Syst. 2006;37:637–669. http://dx.doi.org/10.1146/annurev.ecolsys.37.091305.110100

70. Williams P, Hannah L, Andelman S, Midgley GF, Araújo MB, Hughes G, et al. Planning for climate change: Identifying minimum-dispersal corridors for the Cape Proteaceae. Conserv Biol. 2005;19:1063–1074. http://dx.doi.org/10.1111/j.1523-1739.2005.00080.x

71. Thomas CD, Cameron A, Green RE, Bakkenes M, Beaumont LJ, Collingham YC, et al. Extinction risk from climate change. Nature. 2004;427:145–148. http://dx.doi.org/10.1038/nature02121

72. D'Amen M, Bombi P. Global warming and biodiversity: Evidence of climate-linked amphibian declines in Italy. Biol Conserv. 2009;142:3060–3067. http://dx.doi.org/10.1016/j.biocon.2009.08.004

73. Holness SD, Biggs HC. Systematic conservation planning and adaptive management. Koedoe. 2011;53:34–42. http://dx.doi.org/10.4102/koedoe.v53i2.1029

74. Egoh BN, Reyers B, Carwardine J, Bode M, O'Farrell PJ, Wilson KA, et al. Safeguarding biodiversity and ecosystem services in the Little Karoo, South Africa. Conserv Biol. 2010;24:1021–1030. http://dx.doi.org/10.1111/j.1523-1739.2009.01442.x

75. Gallo JA, Pasquini L, Reyers B, Cowling RM. The role of private conservation areas in biodiversity representation and target achievement within the Little Karoo region, South Africa. Biol Conserv. 2009;142:446–454. http://dx.doi.org/10.1016/j.biocon.2008.10.025

76. Tolley KA, Bowie RCK, Measey GJ, Price BW, Forest F. The shifting landscape of genes since the Pliocene: Terrestrial phylogeography in the Greater Cape Floristic Region. Oxford: Oxford University Press; 2014. p. 142–163.

77. Cowling RM, Procheş Ş, Partridge TC. Explaining the uniqueness of the Cape flora: Incorporating geomorphic evolution as a factor for explaining its diversification. Mol Phylogenet Evol. 2009;51:64–74. http://dx.doi.org/10.1016/j.ympev.2008.05.034

78. Kearney M, Porter W. Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. Ecol Lett. 2009;12:334–350. http://dx.doi.org/10.1111/j.1461-0248.2008.01277.x

Note: This article is supplemented with online only material.

Shark–Cetacean trophic interaction, Duinefontein, Koeberg, (5 Ma), South Africa

AUTHOR:
Romala Govender[1]

AFFILIATION:
[1]Natural History Department, Iziko Museums of South Africa, Cape Town, South Africa

CORRESPONDENCE TO:
Romala Govender

EMAIL:
rgovender@iziko.org.za

POSTAL ADDRESS:
Natural History Department, Iziko Museums of South Africa, PO Box 61, Cape Town 8000, South Africa

DATES:
Received: 15 Dec. 2014
Revised: 14 Apr. 2015
Accepted: 27 May 2015

KEYWORDS:
baleen whale, odontocete, white shark, tooth marks, Zanclean

HOW TO CITE:
Govender R. Shark–Cetacean trophic interaction, Duinefontein, Koeberg, (5 Ma), South Africa. S Afr J Sci. 2015;111(11/12), Art. #2014-0453, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140453

This study forms part of a larger project to reconstruct the Mio-Pliocene marine palaeoenvironment along South Africa's west coast. It documents the shark–cetacean trophic interaction during the Zanclean (5 Ma) at Duinefontein (Koeberg). The damage described on the fragmentary cetacean bones was compared with similar damage observed on fossils from Langebaanweg, a Mio-Pliocene site on the west coast of South Africa, and data present in the literature. This comparison showed that the damage was the result of shark bites. The state of preservation makes it difficult to determine if the shark bite marks were the cause of death or as a result of scavenging. The presence of the bite marks on the bone would, however, indicate some degree of skeletonisation. Bite marks on some cranial fragments would suggest that the cetacean's body was in an inverted position typical of a floating carcass. The preservation of the material suggests that the bones were exposed to wave action resulting in their fragmentation as well as abrasion, polishing and rolling. It also suggests that the cetacean skeletons were exposed for a long time prior to burial. The morphology of the bites suggests that the damage was inflicted by sharks with serrated and unserrated teeth. Shark teeth collected from the deposit include megalodon (*Carcharodon megalodon)*, white (*Carcharodon carcharias*) as well as mako (*Isurus* sp. and *Cosmopolitodus hastalis*) sharks, making these sharks the most likely predators/scavengers.

Introduction

Along the nearly 2000 km of southern African west coast there are few onshore deposits, but where they do occur they are rich in palaeontological and archaeological material.[1,2] During the 1970s when the foundations for the Koeberg Power Station at Duinefontein, located on the farm Duynefontyne 34[3,4] just inland of the west coast of South Africa,[5] were being dug, a subsurface locality was uncovered[3] (Figure 1a–b). This locality also contained a rich marine vertebrate faunal assemblage, for example sharks, cetaceans, seals and fish. Currently there is only one study published focusing on the fossil seal from Duinefontein (Koeberg).[6]

Previous studies show that there was a rich and diverse cetacean and seal fauna along the coast during the Mio-Pliocene.[7-9] During the current study to identify the cetacean fauna preserved at Koeberg, damage to various skeletal fragments was analysed and identified as bites. These bites were different from those identified as terrestrial carnivore damage seen at Langebaanweg (e.g. Hendey[10,11] and Govender unpublished data). There was no direct association between shark teeth collected at Koeberg and the cetacean fossils although they co-occur. When compared with the material from Langebaanweg[8] and other studies,[12-18] the most parsimonious conclusion was that the damage was produced by shark bites. This is the second study documenting the interactions between sharks and cetaceans along the south-western Cape coast during the Zanclean (early Pliocene, 5 Ma).[8] These studies will help build our knowledge of the marine mammals on the South African west coast as well as improve our understanding of the palaeoenvironment along the west coast during the Mio-Pliocene.[19,20]

Materials and methods

Geological and palaeontological setting

The fossils preserved in Duynefontyn member of the Varswater Formation are about 8.5 m below sea level and the member is divided into five beds.[3-5] The fossils that are part of the present study are preserved in the shark tooth bed (Figure 1b) which contains fossil sharks, teleost fish, marine mammals and birds; Rogers[21] interpreted this as a tsunamite deposit. The fossils provide a Zanclean age (5 Ma) for the deposit[4] which is equivalent to the Muishond Fontein Pelletal Phosphorite Member of the Varswater Formation at Langebaanweg.[5]

A barrier spit developed parallel to the coast with each successive regression and absorbed the energy from the wave action characteristic of the west coast's open ocean.[3,4] The shark tooth bed was concentrated into a placer deposit after the barrier spit was overtopped by storms or spring tides; and the retreating water scoured the intertidal flats.[3,4] The intertidal mixed flats were drained by a subtidal channel.[3] The presence of sub-Antarctic seabirds suggests that the marine temperature was colder than present,[22] while the presence of entirely pelagic, migrating and non-breeding birds indicates that the area was open to the ocean at times.[23]

Palaeontological material

The material described here was recovered during excavations of the foundations for the Koeberg Reactor site, 10–12 m below the surface.[3,5] The rich fossil material recovered from the shark tooth bed includes terrestrial mammals and reptiles[3,4] while the terrestrial pollen is too sparse to identify.[3,4] The cetacean fossil material, like that described from Langebaanweg,[8] is fragmentary; however, unlike Langebaanweg, identification of the Koeberg material is difficult; particularly those with bites.

The cetacean fossils consist of vertebral centra, tympanic bullae, periotics, isolated teeth and cranial fragments. A preliminary analysis of the cetacean fossils suggests that there are mysticetes (balaenopterids), odontocetes (e.g. sperm whale, porpoises and delphinids) present at Koeberg. These specimens form part of a separate

taxonomic study (Govender unpublished data). Fossils from Duinefontein (Koeberg) show evidence of having been rolled, some have a polished surface and some are abraded (stage 2)[24] (Figure 1c–e). Most of the damage seen on the bones suggests that the breaks occurred prior to burial. The fragmentary nature of the material would also suggest that the fossils were transported prior to deposition and after the skeletons

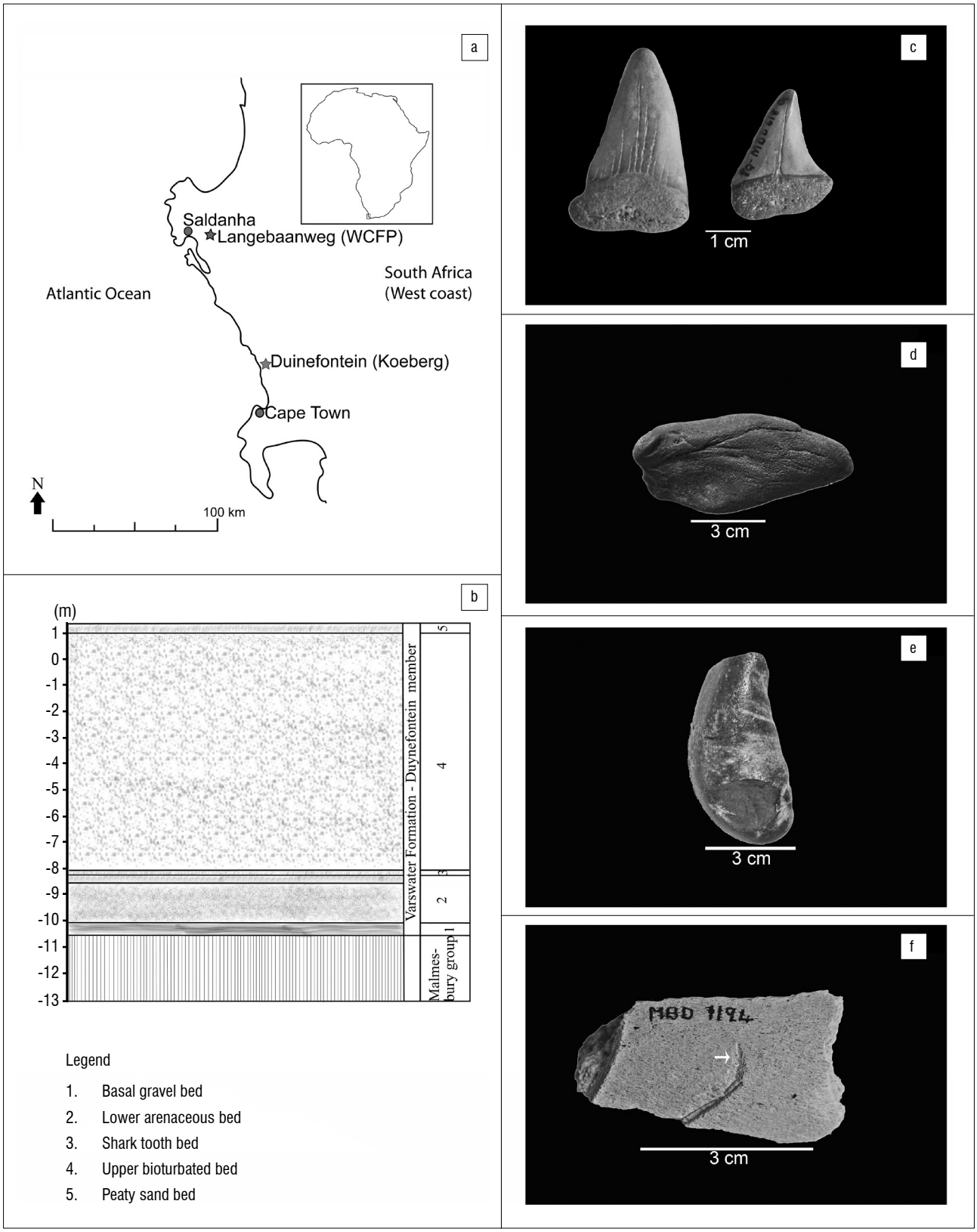


Figure 1: (a) Location of Duinefontein (Koeberg). (b) Schematic showing the stratigraphy for Koeberg – Duynefontein Member, Varswater Formation (based on Rogers[3]). (c) Shark teeth (SAMPQMB-D-618G, SAMPQMB-D-618Q) showing evidence of abrasion. (d) Rolled periotic posterior process (SAMPQMB-D-65). (e) Tympanic bulla showing a polished surface (SAMPQMB-D-1197). (f) CF1a bite mark on cranial fragment (arrow) (SAMPQMB-D-1194).

had become dissociated.[25] The rolled fossils initially came to rest on the beach where they were exposed to wave action before deposition.[25].

The terminology used follows Cigala-Fulgosi[13] and Bianucci et al.[18] Their experimental analyses of the bite action of extant sharks demonstrated four types of damage left by serrated and unserrated teeth. The designation of the type of bite follows the modification of Govender and Chinsamy[8], CF for Cigala-Fulgosi[13] and B for Bianucci et al.[18] Four types of damage were recognised as being caused by the serrated teeth of White Sharks (*Carcharodon carcharias*)[13]:

- CF1[8], damage was sub-divided into two types (designated CF1a and CF1b[8])

  CF1a is a simple, superficial groove with 'dotted' markings left by serrations

  CF1b is a deeper groove with ridges and grooves caused by tooth serrations;

- CF2[8], results in a simple groove with tapered end and no trace of serrations;
- CF3[8], damage has numerous sub-parallel ridges and grooves corresponding with the tooth's serrated edge. There is no cut groove.
- CF4[8], damage suggests cutting and/or scraping action with rotating movement. This action leaves curvilinear markings caused by the rotation of the tooth.

Unserrated shark teeth also produced four types of damage (using *Isurus oxyrinchus* and *Cosmopolitodus hastalis*)[18]:

- B1[8], produced a straight side (labial margin) and a curved side (lingual margin).
- B2[8], produced a more or less elongated incision with wide terminal extremities.

The depth of these two types depends on the tooth's position in the tooth row and the part of the crown used for the bite action that passes into the bone.[8]

- B3[8], the tooth edge had been dragged perpendicular to the dental axis, resulting in no grooves with ridges; however pseudo ridges and grooves can be created by damaged or worn teeth.
- B4[8], the tooth had also dragged perpendicular to the dental surface in a rippled or waved movement visible as parallel incisions resulting from repeated movement across the bone.

Although the tooth morphology of *Isurus oxyrinchus* and *Cosmopolitodus hastalis* is different, the experimental impressions in 'plasticine' showed that it was difficult to distinguish between the two tooth forms because of only slight variation and the number of variables involved, making it difficult to use for identifying particular sharks.[18]

Results

Description of bites on cetacean fossils

The specimens described are fragmentary and are not associated with any of the identifiable mysticete and odontocete specimens. The orientation and the depth of the bites vary amongst the specimens. The bites on the cetacean fossils were produced by serrated and unserrated teeth.

Cranial fragments

The cranial fragments, which include mandibular fragments, have superficial damage to the bone surface. This has resulted in the surface chipping off without penetrating the bone (Figure 1f, 2a-c) (and see Type 1 in Cigala-Fulgosi[13] and CF1a in Govender and Chinsamy[8] fig. 2A–B). There are instances where the teeth have shallowly penetrated the bone surface on the cranial fragment (South African Museum Quaternary Palaeontology, Melkbosstrand, Duinefontein (SAMPQMB-D)-1194; Figure 1f, arrowed) and the cranial fragment SAMPQMB-D-1341 (see Figure 2b).

There are elongated grooves with no serrations and tapered ends which represent CF2 bites caused by serrated teeth (Figure 2d–f, 3a–b) (and see Type 2[13] and fig. 3A–D in Govender and Chinsamy[8]). On SAMPQMB-D-1182, the CF2 bite mark follows the ventral surface of the mandibular fragment (Figure 2d, arrowed). On SAMPQMB-D-1194, the bite intersects the bottom of the CF1a bite (Figure 2e, arrow 1). It is possible that the fragment broke off along another CF2 bite (SAMPQMB-D-1194, Figure 2e, arrow 2). There is evidence of two different shark bites on SAMPQMB-D-1339, one having serrated teeth (Figure 2f, arrowed; CF2 bite). On the surface behind the CF1a and ventral surface of SAMPQMB-D-1342 above the CF3 bite, there is evidence of a very deep CF2 bite with bone having been removed (Figure 3a, arrow 1, 3b) and another CF2 bite above the B2 bite mark (Figure 3a, arrow 2) (and see Type 2[13] and fig. 3A-D[8]).

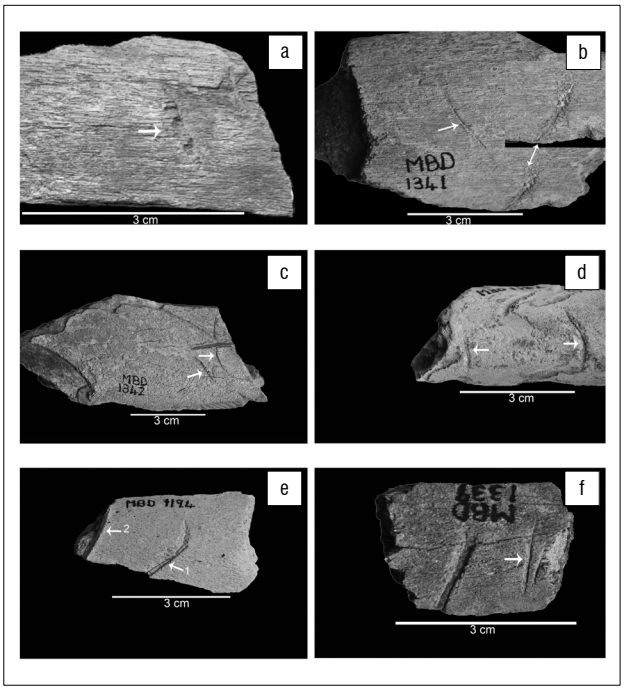


Figure 2: (a) CF1a bite mark on cranial fragment (SAMPQMB-D-1340). (b) Two CF1a bite marks on cranial fragment (SAMPQMB-D-1341) (inset close-up of second). (c) CF1a bite mark on a cranial fragment (SAMPQMB-D-1342) (arrowed). (d) CF2 bite mark on mandible fragment SAMPQMB-D-1182 (arrow). (e) CF2 bite marks on cranial fragment (SAMPQMB-D-1194) (arrow 1 and 2). (f) Two CF2 bite marks on the lateral surface of the cranial fragment (SAMPQMB-D-1339) (arrow).

On a mandibular fragment (SAMPQMB-D-1182) a bite mark tapers at one end and remains wide on the other (Figure 3c) which could be a result of the movement of the shark or the cetacean body (see Bianucci et al. fig. 3J, RR6i[18]). This bite resembles a B1 bite mark (see Type 2[18] and B1[8]). There are a number of wide grooves without ridges and grooves with wide ends (Figure 3d–f) which resemble B2 bite marks left by a shark with unserrated teeth (see Type 2[18] and fig. 3E–F[8]). On the surface of SAMPQMB-D-1342; a CF1a bite is bisected by a B2 bite mark that has damage along groove edges (see Figure 3f).

There are parallel ridges and grooves on the surface of SAMPQMB-D-1182 that are not very deep and are faint (see Figure 4a). These types of damage are similar to that caused by unserrated shark teeth scraping across the bone surface (fig. 3Hb, LR10[18] and fig. 4D[8]). The ridges and grooves could be the result of damage to the teeth and consequently it is not clear if this is a B3 or B4 bite mark (see Type 3, Type 4[18] and fig. 4D[8]). On the ventral surface of SAMPQMB-D-1342, there is a group of sub-parallel ridges

and grooves (Figure 4b) which resemble a CF3 bite caused by teeth with serrated edges (Cigala-Fulgosi[13] and Govender and Chinsamy[8] fig. 4A-B).

Vertebra

SAMPQMB-D-71 is an isolated caudal vertebra with taphonomic damage to the ventral, lateral and dorsal surfaces on the left side. Ventrally on SAMPQMB-D-71, the bone surface has been damaged without penetrating the bone. This resembles a CF1a bite mark (Figure 4c) (and see Type 1 in Cigala-Fulgosi[13] and CF1a in Govender and Chinsamy[8] fig. 2A–B). Along the left lateral side of the anterior articulation, the bone surface has been removed with tapered ends which resemble CF2 caused by serrations on the shark teeth (see Figure 4d) (and see Type 2 in Cigala-Fulgosi[13] and Govender and Chinsamy[8]). Along the left lateral surface close to the posterior articulation is a deep groove that has no ridges and grooves and tapers at the ends which resembles CF2 bite marks caused by teeth with serrations (see Figure 4d) (and see Type 1 in Cigala-Fulgosi[13] and CF1a in Govender and Chinsamy[8] fig. 2A, B).

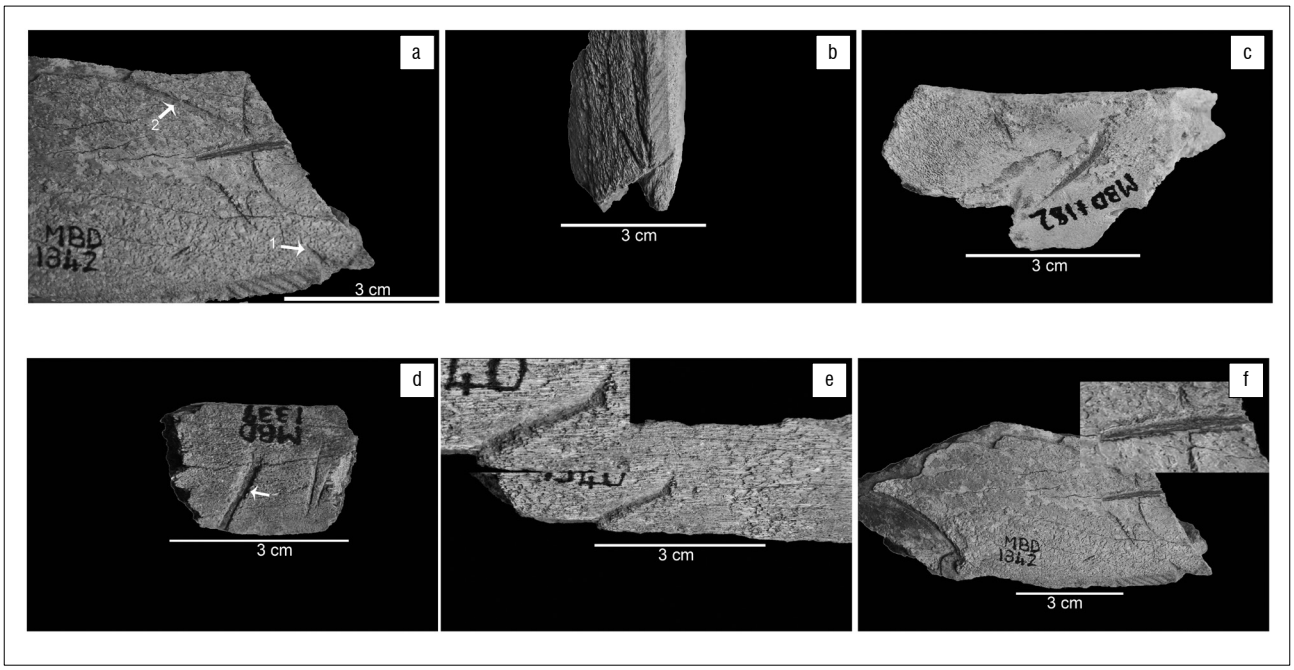


Figure 3: (a) CF2 bite marks on the lateral surface of cranial fragment (SAMPQMB-D-1342) (arrow 1, arrow 2). (b) CF2 bite mark on ventral surface of SAMPQMB-D-1342. (c) B1 bite mark on the lateral surface of a mandibular fragment (SAMPQMB-D-1182). (d) B2 bite mark on the small cranial fragment (SAMPQMB-D-1339) (arrow). (e) B2 bite mark on cranial fragment (SAMPQMB-D-1340) (inset close-up of damage). (f) B2 bite mark on the lateral surface of cranial fragment (SAMPQMB-D-1342) (inset close-up of damage).

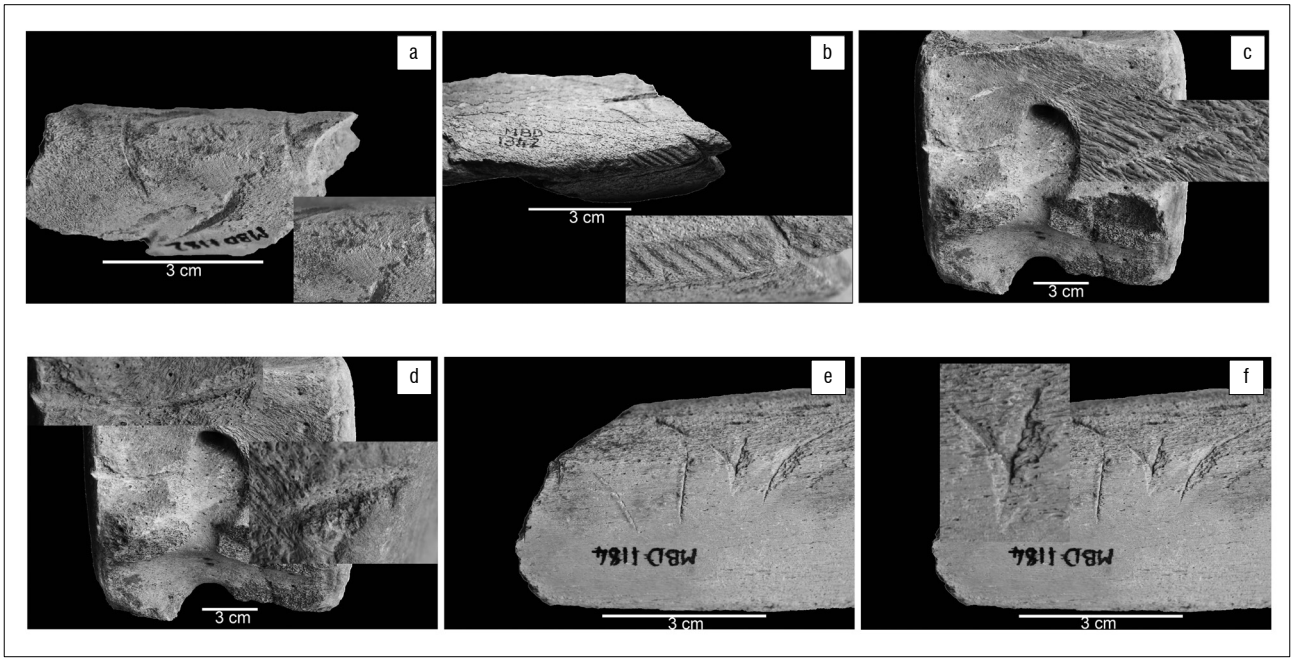


Figure 4: (a) B3/B4 bite mark caused by unserrated teeth scraping across the mandible fragment (SAMPQMB-D-1182) (inset close-up of damage). (b) CF3 bite mark on a cranial fragment (SAMPQMB-D-1342. (c) CF1a bite mark on a caudal vertebra centrum (SAMPQMB-D-71) (inset close-up). (d) CF2 bite marks on caudal vertebra centrum (SAMPQMB-D-71) (inset shows close-up of damage). (e) CF1a bite marks on a rib fragment. (f) Unusual bite mark caused by serrated teeth forming an upside down 'V' (SAMPQMB-D-1184) (inset close-up of damage).

Rib

On SAMPQMB-D-1184, a rib fragment (Figure 4e), there is a shallow groove that only slightly damaged the surface of the bone causing some of the surface to chip off. This type of bite mark resembles CF1a in fig 2A–B[8] (and see Type 1[13]). Other CF1a bites have shallowly penetrated the bone surface (Figure 4e). There is an unusual trace where two bite marks intersected forming an upside down 'V' (see Figure 4f).[8,13] This unusual trace may have been caused either by the shark or the cetacean being in motion which resulted in the shark losing its grip on the prey and having to bite down a second time.

Discussion

Great white (*Carcharodon carcharias*), tiger (*Galeocerdo cuvier*), Zambezi (bull) (*Carcharhinus leucas*), whale (*Rhincodon typus*), ragged tooth (sandtiger, *Carcharias tarsus*) and Shortfin Mako (*Isurus oxyrinchus*) sharks, to name a few, currently occur along the South Africa coasts and occupy various habitats. Numerous shark teeth have been collected from Duinefontein (Koeberg); however, there has been no detailed study of this material. Based on informal identifications[3-5,21] and a comparison with the Langebaanweg collection,[26] white shark (*Carcharodon carcharias*), mako shark (*Isurus* sp. and *Cosmopolitodus hastalis*), ragged tooth (sandtiger, *Carcharias tarsus*) and megalodon (*Carcharodon megalodon*) have been identified in the sample. There are also teeth that resemble lemon shark (*Negaprion brevirostris*) in the collection (author observations, 2014).

White sharks most likely had a worldwide distribution in the geological past[12,13,15-18] as they are capable of exploiting a wide range of habitats and temperatures.[27,28] At present the adults live along the cold west coast while the pups and juveniles inhabit the warmer east coast of South Africa.[27,29,30] Modern white sharks are considered to exclude other sharks while feeding,[31] however, there are documented cases where white sharks and tiger sharks have fed concurrently on a whale carcass.[32,33] These sharks were all the same size, not exceeding 3.5 m.[32] Mako are highly active, mobile sharks[27] that inhabit warm coastal and oceanic waters that range in depth from shallow coastal to 500 m.[28]

Shark diets are highly varied and include teleost fish, marine mammals and other sharks.[27,28,33-35] Some large sharks obtain most of their food by actively hunting small odontocetes and by scavenging large mysticetes.[27,28,33-35] Off South Africa's coast there are documented attacks on odontocetes.[36] Body fluids leaching from large whale carcasses are thought to attract sharks from as far as 10 km away.[30] Sharks also feed on cetacean carcasses at sea and are not always observed as carcasses may remain afloat for a number of weeks.[25,37] Large whale carcasses that have a high fat content remain afloat immediately after death and for an extended period,[25,37] while some sink and refloat from the build-up of gases resulting from decomposition.[22] Others may sink into an anaerobic environment and remain there.[22]

The nature of the preservation of the Duinefontein (Koeberg) material only allows one to extrapolate a discussion of the shark–cetacean interaction from the damage observed. Duinefontein (Koeberg) was open to the ocean during the transgression; however, during the regression a barrier spit developed along the coast.[3,4] The beach at Koeberg would have been open to wave action to varying degrees. As a result, whale carcasses could have become beached and remained long enough for gas to build up and allow the carcass to be refloated or moved during high tides or surf backwash.[25] Eventually, tensile stress would cause the stretched skin 'bag' to rupture and scatter skeletal elements already separated by decay.[25] It is also the most parsimonious reason for there being rolled and abraded fossil remains at the site; however, the flooding of the intertidal region could also be responsible for this as material was moved around by retreating water.

The cetacean fragments show evidence of superficial scrapes to penetrating bites. This variation in the depth of the bite marks is potentially as a result of the shark and cetacean both being in motion because of the currents and wave activity. There are a few CF2[8] (see Type 2[13]) bite marks. These were probably as a result of a shark propelling itself forward to bite and then reversing straight back, a behaviour seen particularly in white sharks.[38] Only the points of the teeth contact the bone surface leaving no evidence of serrations.[38] SAMPQMB-D-1182 is a mandible fragment that has bite marks on the ventral surface. The mandible probably became separated from the skull early in the decaying process.[25] In most instances, the head of the whale is the focus of feeding as the carnivores target the tongue.[30] The nature of the preservation suggests that two possible feeding scenarios could be extrapolated from the damage. The first is that the sharks were scavenging on a floating inverted whale carcass[8] prior to the mandible and skull becoming detached from the rest of the skeleton. The second would indicate a possible predatory attack on a whale as sharks also approach carcasses from below the water.[38] Other cranial fragments show bite marks on the lateral or medial surfaces. This would lend support to them being part of whale carcasses being scavenged rather than actively hunted. None of the bites show signs of healing which would suggest that the cetacean was most likely scavenged; however, the nature of the preservation does not eliminate active hunting as cause of death.

The rib fragment has bite marks on the lateral surface and an isolated caudal vertebra also shows bite marks on the ventral surface giving no context to the damage to the vertebra. The damage suggests that there are two possible scenarios; that the whale carcass was being scavenged or that the whale had been attacked by a shark. The bite marks would suggest an attack from the side as sharks do attack from the side and below.[35] The skeleton of cetaceans is protected by blubber and muscle therefore the presence of bite marks on the bones suggests that the cetacean body was in an advanced state of decomposition and becoming skeletonised, allowing the shark to penetrate the tissue and reach the bone.

A comparison with the cetacean fossils from Langebaanweg as well as other studies indicates the bite marks identified on the Koeberg cetacean fossils typically resemble bites caused by sharks with serrated teeth CF1a[8] (and see Type 1[13]), CF2[8] (and see Type 2[13]), CF3[8] (and see Type 3[13]), sharks with unserrated teeth B2[8] (and see Type 2[18]), B3/B4[8] (and see Type 3, Type 4[18]) and some unusual damage.[8,13] The damage caused by serrated teeth most closely resembles the damage caused by white sharks described by Cigala-Fulgosi[13] and Govender and Chinsamy.[8] White sharks are known to roll onto their ventral surfaces when feeding,[31,36] which can cause the shark's teeth to slip and scrape over the bone only damaging the surface (see CF3 in Figure 4b).

On SAMPQMB-D-1182 there is damage that suggests the shark with unserrated teeth first bit into the bone (B2 bite in Figure 4a) and possibly as a result of the movement of the shark and/or prey, the shark lost its hold on the prey item causing the tooth/teeth to slip across the surface of the bone leaving a scrape with very faint ridges and grooves (Figure 4a). The damage from both serrated and unserrated shark teeth documented on the fragmentary cetacean remains suggests that more than one shark taxon fed on the cetaceans.

Most of the damage inflicted on the bones from Duinefontein (Koeberg) closely resembles that described for white sharks[8,13,39] as there are no secondary serrations within the grooves like those described by Cigala-Fulgosi[13] for tiger sharks. Currently, no tiger shark teeth have been described from the Koeberg collection although they have been described from the contemporaneous site of Saldanha Steel,[40] so the most parsimonious explanation is that the whales were fed on by white sharks indicated by the numerous white shark teeth in the collections, although megalodon teeth have also been found.

There are also some B2 and B3/B4[8] bite marks on the remains. It is difficult to assign these to a specific shark as the damage on the Langebaanweg cetaceans caused by sharks with unserrated teeth retains no diagnostic information that allows exact species identification.[18] Examples of mako shark *(Isurus sp.* and *Cosmopolitodus hastalis)* are present in the collections makes them the most likely sharks to have left the traces of unserrated teeth. The unusual bites may be as a result of the shark losing traction while biting and having to grab at bone or flesh of the prey item a second or third time. The sharks may have also left evidence of their feed on partially skeletonised carcasses that eventually came to rest on the beach in the area and were later refloated and restranded on the beach resulting in the complete disarticulation of the skeleton and taphonomic damage.

Conclusion

Whale carcasses would have beached along the coast at Duinefontein (Koeberg) and refloated a number of times resulting in the breaking up of the carcasses. The wave action also resulted in some material being fragmented into small pieces, rolled, abraded and polished. Bites on a mandibular, cranial and rib fragments as well as an isolated vertebra preserve evidence of shark–cetacean interaction. The presence of the bite marks on the ventral surface of the mandibular fragment and lateral surfaces of the cranial fragments would strongly suggest that the cetaceans were in an inverted position when these bite marks were made, supporting scavenging action on floating carcasses. The fragmentary nature of the fossils, however, does not preclude the possibility of the bites being the cause of death. A comparison of the damage on the Koeberg fossil cetacean remains with other studies[8,12-18,39,41] suggested that the damage caused by serrated teeth was produced by white sharks. The grooves that do not have ridges and grooves have tapered ends and the superficial damage is similar to that described for white sharks.[8,13] There are no secondary serrations in the grooves as produced by tiger sharks.[13] Other bite marks were caused by sharks with unserrated teeth; however, definitive identification is difficult. The shark most likely to have caused the bite marks is the mako as a number of teeth are found in the collection. This second study[8] shows more evidence of shark–cetacean trophic interaction in the geological past.

Acknowledgements

The author would like to thank the Cenozoic collections staff, particularly Angus Rayners and Mark de Benedictis, at Iziko South African Museum, Cape Town for their assistance. The author is grateful to Dr Margaret Avery for advice. I would like to thank John Rogers for his comments and help with the schematic of the stratigraphic section. I am also grateful to the two reviewers for their constructive comments.

References

1. Pether J, Roberts DL, Ward JD. Deposits of the West Coast. In: Partridge TC, Maud RR, editors. The Cenozoic of southern Africa. Oxford: Oxford University Press; 2000. p. 33–54.
2. Roberts DL, Botha GA, Maud RR, Pether J. Coastal Cenozoic deposits. In: Johnson MR, Anhaeusser CR, Thomas RJ, editors. The geology of South Africa. Johannesburg/Pretoria: Geological Society of South Africa/Council for Geoscience; 2006. p. 605–628.
3. Rogers J. The sedimentary succession at the Koeberg Nuclear Power Station, Melkbosstrand. Geokongres 79: 18th Congress of the Geological Society of SA. Johannesburg: Geological Society of South Africa; 1979. p. 310–322.
4. Rogers J. First report on the Cenozoic sediments between Cape Town and Eland's Bay. Open file. Geological Survey of South Africa. 1980;165:64.
5. Rogers J. Lithostratigraphy of Cenozoic sediments between Cape Town and Eland's Bay. Palaeoecol Afr. 1982;15:121–137.
6. Avery G, Klein RG. Review of fossil phocid and otariid seals from the southern and western coasts of South Africa. T Roy Soc S Afr. 2011;66(1):14–24. http://dx.doi.org/10.1080/0035919X.2011.564490
7. Govender R, Chinsamy A, Ackermann RR. Anatomical and landmark morphometric analysis of fossil phocid seal remains from Langebaanweg, West Coast of South Africa. T Roy Soc S Afr. 2012;67(3):135–149. http://dx.doi.org/10.1080/0035919X.2012.724471
8. Govender R, Chinsamy A. Early Pliocene (5 Ma) shark–cetacean trophic interaction from Langebaanweg, western coast of South Africa. Palaios. 2013;28(5):270–277. http://dx.doi.org/10.2110/palo.2012.p12-058r
9. Govender R. Preliminary phylogenetics and biogeographic history of the Pliocene seal, *Homiphoca capensis* from Langebaanweg, South Africa. T Roy Soc S Afr. http://dx.doi.org/10.1080/0035919X.2014.984258
10. Hendey QB. The Late Cenozoic Carnivora of the south-western Cape Province. Ann S Afr Mus.1974;63:1–369.
11. Hendey QB. Palaeoecology of the Late Tertiary fossil occurrences in 'E' Quarry, Langebaanweg, South Africa, and a reinterpretation of their geological context. Ann S Afr Mus. 1981;84(1):1–104.
12. Deméré TA, Cerutti R. A Pliocene shark attack on a cethotheriid whale. J Paleontol. 1982;56:1480–1482.
13. Cigala-Fulgosi F. Predation (or possible scavenging) by a great white shark on an extinct species of bottlenosed dolphin from the Pliocene of Italy. Tert Res. 1990;12:17–36.
14. Noriega JI, Cione AL, Aceñolaza F. Shark tooth marks on Miocene balaenopterid cetacean bones from Argentina. Neues Jahrb Geol Palaeontol Abh. 2007;245:185–192.
15. Aguilera OA, García L, Cozzuol MA. Giant-toothed white sharks and cetacean trophic interaction from the Pliocene Caribbean Paraguaná Formation. Paläontol Z. 2008;82:204–208. http://dx.doi.org/10.1007/BF02988410
16. Ehret DJ, Macfadden BJ, Salas-Gismondi R. Caught in the act: Trophic interactions between a 4-million-year-old white shark (*Carcharodon*) and mysticete whale from Peru. Palaios. 2009;24:329–333. http://dx.doi.org/10.2110/palo.2008.p08-077r
17. Esperante R, Muñiz Guinea F, Nick KE. Taphonomy of a Mysticeti whale in the Lower Pliocene Huelva Sands Formation (Southern Spain). Geol Acta. 2009;7:489–505. http://dx.doi.org/10.1344/105.000001451
18. Bianucci G, Sorce B, Storai T, Landini W. Killing in the Pliocene: Shark attack on a dolphin from Italy. Palaeontology. 2010;53:457–470. http://dx.doi.org/10.1111/j.1475-4983.2010.00945.x
19. Bianucci G, Lambert O, Post K. A high diversity in fossil beaked whales (Mammalia, Odontoceti, Ziphiidae) recovered by trawling from the sea floor off South Africa. Geodiversitas. 2007;29:561–618.
20. Bianucci G, Post K, Lambert O. Beaked whale mysteries revealed by seafloor fossils trawled off South Africa. S Afr J Sci. 2008;104:140–142.
21. Rogers J. Sedimentology of Late Cenozoic sediments, including the Pliocene Dynefontyn Member of the Varswater Formation, Koeberg Nuclear Power Station, Melkbosstrand, Cape Town. Afr Nat Hist. 2006;2:194–196.
22. Olson SL. Fossil seabirds and changing marine environments in the Late Tertiary of South Africa. S Afr J Sci. 1983;79(10):399–402.
23. Olson SL. An early Pliocene marine avifauna from Duinefontein, Cape Province, South Africa. Ann S Afr Mus. 1985;95(4):147–164.
24. Boessenecker RW, Perry FA, Schmitt JG. Comparative taphonomy, taphofacies, and bonebeds of the Mio-Pliocene Purisima Formation, Central California: Strong physical control on marine vertebrate preservation in shallow marine settings. PLoS One. 2014;9(3), e91419. http://dx.doi.org/10.1371/journal.pone.0091419
25. Schäfer W. Ecology and palaeoecology of marine environments. Chicago, IL: University of Chicago Press; 1972.
26. Tulu Y, Chinsamy-Turan A. Langebaanweg Quarry, Western Cape, South Africa: The Elasmobranch fauna and comparisons to faunas of PCS (Lee Creek) phosphate mine and Sharktooth Hill. In: 71st Annual Meeting of the Society of Vertebrate Paleontology program and abstracts book; 2011 Nov 2–5; Las Vegas, NV, USA. Bethesda, MD: Society of Vertebrate Paleontology; 2011. p. 207.
27. Compagno LJV. Sharks of the world: An annotated and illustrated catalogue of shark species known to date: Bullhead, mackerel and carpet sharks (Heterodontiformes, Lamniformes and Orectolobiformes). FAO species catalogue for fishery purposes v. 2. Rome: FAO; 2001.
28. Compagno LJV, Dando M, Fowler S. Field guide: Sharks of the world. London: HarperCollins; 2005.
29. Dewar H, Domeier M, Nasby-Lucas N. Insights into young of the year white shark, *Carcharodon carcharias*, behavior in the Southern California Bight. Environ Biol Fish. 2004;70:133–143. http://dx.doi.org/10.1023/B:EBFI.0000029343.54027.6a
30. Dicken ML. First observations of young of the year and juvenile great white sharks (*Carcharodon carcharias*) scavenging from a whale carcass. Mar Freshwater Res. 2008;59:596–602. http://dx.doi.org/10.1071/MF07223
31. Pratt HL, Casey JG, Conklin RE. Observations on large white sharks, *Carcharodon carcharias*, off Long Island, New York. Fish Bull. 1982;80:153–156.
32. Dudley SFJ, Anderson-Reade MD, Thompson GS, McMullen PB. Concurrent scavenging off a whale carcass by great white sharks, *Carcharodon carcharias*, and tiger sharks, *Galeocerdo cuvier*. Fish Bull. 2000;98:646–649.

33. Bornatowski H, Wedekin LL, Heithaus MR, Marcondes MCC, Rossi-Santos MR. Shark scavenging and predation on cetaceans at Abrolhos Bank, eastern Brazil. J Mar Biol Assoc UK. 2012;92(08):1767–1772. http://dx.doi.org/10.1017/S0025315412001154

34. Cliff G, Dudley FJ. Sharks caught in the protective gill nets off Natal, South Africa 4: The bull shark *Carcharhinus leucas* Valenciennes. S Afr J Mar Sci. 1991;10:253–270. http://dx.doi.org/10.2989/02577619109504636

35. Lowe CG, Wetherbee BM, Crow GL, Tester AL. Ontogenetic dietary shift and feeding behavior of the tiger shark, Galeocerdo cuvier, in Hawaiian waters. Environ Biol Fish. 2002;47:203–211. http://dx.doi.org/10.1007/BF00005044

36. Wcisel M, Chivell W, Gottfried MD. A potential predation attempt by a great white shark on an Indo-Pacific humpback whale. S Afr J Wildl Res. 2011;40(2):84–187.

37. Long DJ, Jones RE. White shark predation and scavenging on cetaceans in eastern North Pacific Ocean. In: Klimley A, Ainsley D, editors. Great white sharks: Biology of *Carcharodon carcharias*. San Diego, CA: Academic Press; 1996. p. 293–307. http://dx.doi.org/10.1016/B978-012415031-7/50028-8

38. Tricas TC. Feeding ethology of the white shark, *Carcharodon carcharias*. Mem Calif Acad Sci. 1985;9:81–91.

39. Bianucci G, Bisconti M, Landini W, Storai T, Zuffa M, Giuliani S, et al. Trophic interactions between white sharks (*Carcharodon carcharias*) and cetaceans: A comparison between Pliocene and recent data. In: Vacchi M, La Mesa G, Serena F, Séret B; editors. Proceedings of the 4th Meeting of the European Elasmobranc Association; 2000 Sep 27–30; Livorno, Italy. Abbeville, France: Imprimerie F. Paillart; 2002. p. 33–48.

40. Kock A. Preliminary identification of Neogene fossil shark teeth from the Varswater Formation of the west coast of South Africa [unpublished Honours report]. Cape Town: University of Cape Town; 2002.

41. Bianucci G, Gingerich PD. *Aegyptocetus tarfa*, n. gen. et sp. (Mammalia, Cetacea), from the middle Eocene of Egypt: Clinorhynchy, olfaction, and hearing in a protocetid whale. J Vertebr Paleontol. 2011;31(6):1173–1188. http://dx.doi.org/10.1080/02724634.2011.607985

Food habits of the aoudad *Ammotragus lervia* in the Bou Hedma mountains, Tunisia

AUTHORS:
Jamel Ben Mimoun[1]
Saïd Nouira[1]

AFFILIATION:
[1]Laboratory of Animal Ecology, Department of Biology, Faculty of Science, Tunis El Manar University, Tunis, Tunisia

CORRESPONDENCE TO:
Jamel Ben Mimoun

EMAIL:
jamelbmimoun@yahoo.fr

POSTAL ADDRESS:
Laboratory of Animal Ecology, Department of Biology, Faculty of Science, Tunis El Manar University, Farhat Hached, Tunis University Campus BP 94, Rommana 1068, Tunisia

DATES:
Received: 12 Dec. 2014
Revised: 13 Mar. 2015
Accepted: 16 June 2015

KEYWORDS:
Barbary sheep; food habits; north Africa; seasonal variation; ungulate

HOW TO CITE:
Ben Mimoun J, Nouira S. Food habits of the aoudad *Ammotragus lervia* in the Bou Hedma mountains, Tunisia. S Afr J Sci. 2015;111(11/12), Art. #2014-0448, 5 pages. http://dx.doi.org/10.17159/sajs.2015/20140448

A micro-histological analysis of composite faecal samples was used to determine the food plants consumed by the aoudad *Ammotragus lervia* (Barbary sheep) in the mountains of Bou Hedma National Park, as well as to study food preferences and seasonal variation of the diet of this ungulate. A total of 19 plant species was identified in the faecal samples: 8 grasses, 6 browse types and 5 forbs. The annual diet of this ungulate is composed of approximately 67% grasses, 17% browse and 16% forbs. Two grass species, *Stipa parviflora* and *Stipa tenacissima*, comprised 97% of grasses consumed and 63% of the annual diet, and were eaten in the year with considerably greater frequency than any other plant species. The diet of the aoudad at Bou Hedma National Park showed a seasonal variation. For example, grasses were consumed the entire year with a preference in spring (72.57%) and summer (78.31%), but their occurrence in the diet dropped in winter (38.7%). Browse was eaten preferentially in autumn (21.86%) and forbs were utilised most during winter (43.22%) and least during summer (7.53%). Our results show the plasticity of the Tunisian aoudad's diet and identify it predominantly as a grazer.

Introduction

The Barbary sheep or aoudad *(Ammotragus lervia)* is a bovid within the Tribe Caprini (sheep and goats) and is endemic to North Africa.[1] It may be found in any rugged terrain or mountain chain throughout the northern part of the continent.[2-6] Its distribution within the Sahara region extends 'south about 14° near the Niger and through N. Tchad and Sudan to the Red Sea'.[6,7] The aoudad has become rare and has been locally extirpated from its original range as a result of human encroachment, overgrazing and constant hunting.[8,9]

In Tunisia, the aoudad is the only representative of wild caprinids and it is considered the flagship species of Tunisian wildlife. According to the Red List of Threatened Species of the International Union for Conservation of Nature (IUCN), the aoudad is classified as a globally 'vulnerable' species.[6] It has disappeared from most of its original environment and was considered by Schomber and Kock to be on the verge of extinction in Tunisia as a result of overhunting and a reduction in its habitat. This species is considered in Tunisia by the National Register of Wild Species as 'critically endangered'. Indeed, the Barbary sheep is exposed to threats that may cause its extinction in the wilderness such as overhunting for its meat and overgrazing of its habitat by livestock.[11] The National Park of Bou Hedma is one of the country's biotopes that has succeeded in rehabilitating a very small number of aoudad. In a report of the General Directorate of Forests,[12] the population density of the aoudad in the park was estimated at about 50 individuals.

In its native habitat in northern Africa, the aoudad is largely unknown apart from the little research done regarding its distribution.[3,4,9,11] The social organisation of the aoudad population of Chambi in Tunisia was recently studied by Ben Mimoun et al.[13] The study of diet is a necessary step in understanding the eco-biology of herbivorous animals. Indeed, besides its applicability in quantifying the food ingested by animals, it can also contribute to effective management and conservation of endangered herbivores such as the aoudad. No sophisticated food habit studies have been made on this species in North Africa, but general references note that both browse and herbaceous vegetation are eaten.[14,15] In Tunisia, the diet of this endangered North African wild caprid has never been studied.

Our objective was to study the feeding behaviour of the aoudad occupying the mountains of the Bou Hedma National Park in central Tunisia. The knowledge of the trophic regime of this ungulate is an opportunity to increase our understanding of the feeding behaviour and the requirements of this species.

Material and methods

Study area

The feeding behaviour of the aoudad in Tunisia was determined for a wild population of this species occupying the mountain steppe of Bou Hedma National Park (16 448 ha) of Sidi Bouzid in the centre of the country (Figure 1). The habitat of the National Park is represented by the Djebel of Bou Hedma, culminating at 790 m, and is characterised by vegetation including *Lygeum spartum, Stipa tenacissima, Stipa parviflora, Artemisia herba alba, Pistacia lentiscus, Retama raetam and Acacia tortillis*. This habitat is located in a bio-climate of lower semi-arid to lower arid, with an annual average rainfall of about 150 mm.[3]

Sample collection and analysis

Monthly collections of fresh faecal pellets of the aoudad were made between March 2009 and February 2010 at Bou Hedma National Park. Samples were collected in an opportunistic manner throughout the mountain of the park. A total of 180 piles of droppings was collected, with an average of 15 piles of droppings per month. Within each pile, an average of five fresh droppings was randomly collected. Droppings were harvested and preserved in 95% ethanol in screw-top jars labelled with the month and locality of collection. For each month, a composite sample was prepared by combining the five fresh faeces from the 15 piles.[16,17]

Composite sample analysis was performed in the Animal Ecology Laboratory at The Arid Land Institute, Medenine Tunisia as described in Simpson et al.[18] This method was also applied by Wagner et al.[17] for bighorn sheep (*Ovis canadensis*) in North America. A reference atlas of the epidermis of the harvested plants in the habitat of the aoudad was prepared by applying the experimental protocol followed by Sparks et al.[19] Reference material and faecal samples were prepared for analysis following the micro-histological slide preparation technique described by Free et al.[20] Examination and identification procedures followed Sparks et al.[19] As indicated by Williams[21] and Storr[22], no staining was performed for microscopic observation. Four slides were prepared for each composite sample. About 275 epidermal fragments were recognised for each composite sample by systematic scanning of the blades.[23]

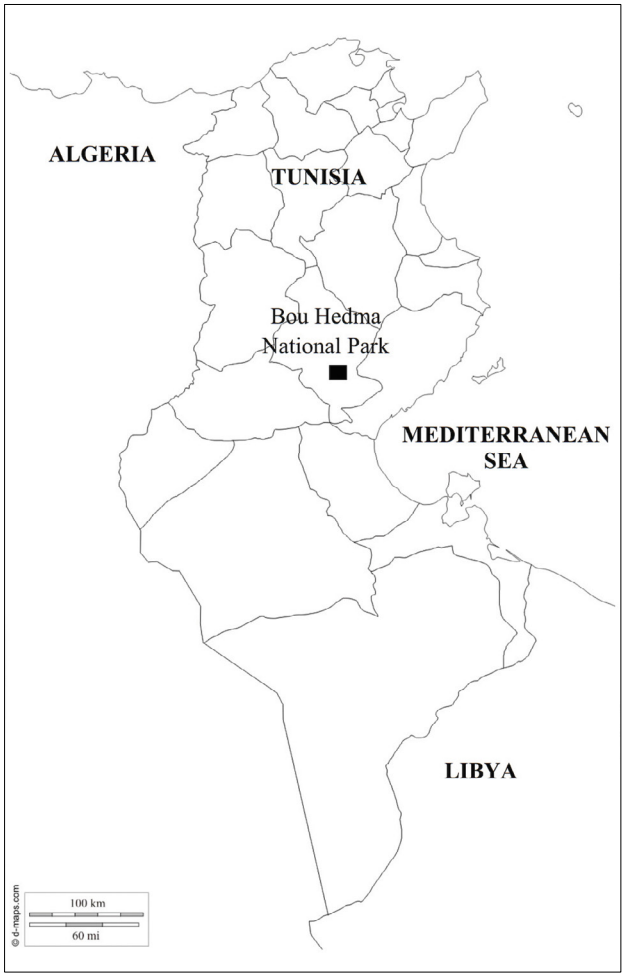


Figure 1: Location of Bou Hedma National Park within Tunisia.

The results of the diet analysis are expressed in average frequencies (AF%) of plant fragments in each composite sample of droppings in accordance with the work of Simpson et al.[18] Seasonal and annual average frequency of each plant were calculated using the formula:

$$AF\% = (ni / Ni) \times 100 \qquad \text{Equation 1}$$

where ni is the total number of fragments of the species (i) identified and Ni is the total number of fragments counted.

The average frequencies (AF%) provide a good estimate of the diet by indicating the range of plant species collected by the animal from its habitat and they also show food use within the year. Species with an AF<1.0 were not a major part of the annual diet and were classified as trace foods.[18] Non-parametric Kruskal–Wallis tests[24] were applied to compare average frequencies of the different forage classes and plant species consumed by the aoudad; $p<0.05$ was considered significant.

Results

Average diet composition

The food spectrum of wild aoudad in Bou Hedma National Park comprised 19 plant species. The average annual diet of the aoudad sampled in this study was found to consist of 67.12% grasses, 16.70% browse and 15.81% forbs (Table 1). Unidentified species represented only 2.36% of the diet.

The principal forage plants identified in the diet of aoudad in Bou Hedma National Park throughout the year were two grass species *(Stipa parviflora and Stipa tenacissima)* four browse species *(Periploca leavigata, Juniperus phoenicea, Pistacia lentiscus and Retama raetam)* and two forbs *(Helianthemum kahiricum and Globularia alypum)* (Table 1).

Seasonal diet variation

The diet of the aoudad at Bou Hedma National Park shows a seasonal variation. The spring food spectrum is the largest with 14 plant species in comparison to 13 in autumn, 12 in winter and only 9 species in summer (Table 1).

There was no significant difference in the consumption of each forage class among the four seasons for grasses ($K=1.44$; $df=3$; $p>0.05$), for browse ($K=2.43$; $df=3$; $p>0.05$) and for forbs ($K=3.16$; $df=3$; $p>0.05$). However, interaction between the effects of season and forage class was identified ($K=6.57$; $df=2$; $p=0.037$). Indeed, the aoudad shows a higher use of grasses during the spring (72.57%), summer (78.31%) and autumn (58%), than in winter when the diet is concentrated on forbs (43.22%). Furthermore, a net increase in the consumption of browse was observed in autumn (21.86%) (Table 1).

Considering each forage class separately, the Kruskal–Wallis test indicated a significant difference in the consumption of grass species by the aoudad in spring ($K=23.59$; $df=7$; $p=0.002$), summer ($K=22.86$; $df=7$; $p=0.002$), autumn ($K=22.69$; $df=7$; $p=0.002$) and winter ($K=21.26$; $df=7$; $p=0.003$). In fact, eight species of grasses were consumed in the National Park, but only three species – *Stipa tenacissima*, *Stipa parviflora* and *Ampelodesma mauritanica* – had an AF of >1.0 in the annual diet. The genus of *Stipa* alone represents 97% of grasses consumed and 63.22% of the annual diet of animals, and was eaten throughout the year with considerably greater frequency than that of any other species (Table 1).

No significant seasonal variation was observed in the consumption of browse species ($K=10.82$; $df=5$; $p>0.05$), but during each season the aoudad showed a preference for only a few species: in spring ($K=14.07$; $df=5$; $p=0.015$), summer ($K=15.54$; $df=5$; $p=0.008$), autumn ($K=12.73$; $df=5$; $p=0.026$) and winter ($K=13.87$; $df=5$; $p=0.016$). Six browse species were eaten by the aoudad in Bou Hedma National Park: a single tree species *(Olea europea)* and five shrubs *(Periploca laevigata, Juniperus phoenicea, Pistacia lentiscus, Retama raetam and Salsola longifolia)*. Among this forage class, only two species – *Periploca laevigata and Juniperus phoenicea* – were eaten by the aoudad during the year with significant relative frequency. The main browse species used by the aoudad was *Periploca leavigata* which occurred the most in the samples of droppings analysed in spring (7.37%) (Table 1).

No significant seasonal variation was observed in the consumption of forbs ($K=9.36$; $df=4$; $p>0.05$) (Table 1). A clear seasonal preference for certain species of forbs was shown: spring ($K=11.22$; $df=4$; $p=0.024$); summer ($K=11.12$; $df=4$; $p=0.025$); autumn ($K=11.88$; $df=4$; $p=0.018$); and winter ($K=11.44$; $df=4$; $p=0.022$). Indeed, five forb species were identified in the diet of the aoudad in the Bou Hedma National Park but only two species, *Helianthemum kahiricum and Globularia alypum*, appeared during all seasons with significant relative frequencies compared to the rest of the plants in this forage class. *Helianthemum kahiricum* of the Cistaceae family was consumed the most throughout the year, except in summer. The consumption of the second species, *Globularia alypum*, increased markedly in winter (14.57%) (Table 1).

Table 1: Average frequencies (AF%) of different forage species consumed by aoudad in the mountains of Bou Hedma National Park (2009–2010)

Forage species	Annual AF%	Spring AF%	Summer AF%	Autumn AF%	Winter AF%
Stipa parviflora	47.5 (1568)	45.6 (507)	63.26 (353)	48.88 (700)	4 (8)
Stipa tenacissima	15.72 (519)	23.65 (263)	15.05 (84)	7.75 (111)	30.65 (61)
Ampelodesma mauritanica	1.12 (37)	2.7 (30)	0	0.48 (7)	0
Allium rosum	0.39 (13)	0	0	0.77 (11)	1(2)
Panicum turgidum	0.18 (6)	0	0	0	3 (6)
Hyparhenia hirta	0.12 (4)	0.36 (4)	0	0	0
Urginea maritima	0.06 (2)	0.18 (2)	0	0	0
Asphodelus teniufolius	0.03 (1)	0.08 (1)	0	0	0
% Grasses	**65.13**	**72.57**	**78.31**	**58**	**38.70**
Periploca laevigata	5.93 (196)	7.37 (82)	5.2 (29)	5.16 (74)	5.53 (11)
Juniperus phoenicea	4.27 (141)	2.52 (28)	4.84 (27)	5.86 (84)	1 (2)
Pistacia lentiscus	3.36 (111)	0.45 (5)	0	7.40 (106)	0
Retama raetam	2.36 (78)	3.15 (35)	1.43 (8)	2.44 (35)	0
Salsola longifolia	0.48 (16)	0	0	0.63 (9)	3.5 (7)
Olea europea	0.27 (9)	0.18 (2)	0	0.35 (5)	1 (2)
% Browse	**16.70**	**13.67**	**11.47**	**21.86**	**11.05**
Helianthemum kahiricum	8.93 (295)	6.3 (70)	0.9 (5)	13.55 (194)	13.06 (26)
Globularia alypum	4.27 (141)	4.68 (52)	2.15 (12)	3.35 (48)	14.57 (29)
Helianthemum semiglabrum	1.39 (46)	0	3.94 (22)	0	12.06 (24)
Teucrium pollium	1.12 (37)	1.26 (14)	0	1.11 (16)	3.52 (37)
Anabasis europodium	0.09 (3)	0	0.54 (3)	0	0
% Forbs	**15.81**	**12.23**	**7.53**	**18.02**	**43.22**

The total numbers (ni) of fragments of the species (i) identified are shown in parentheses.

Discussion

We have shown that the aoudad in Tunisia consumes different types of plants including grasses, forbs, browse, trees and shrubs. These results reveal that the aoudad is a generalist herbivore with a very flexible diet.[25-27] The number of plants used by the aoudad in Bou Hedma is very small (19 plant species) compared to that of Barbary sheep that have been introduced elsewhere. In the Canary Islands, 41 taxa were identified[28] and in the Guadalupe Mountains of New Mexico, the annual diet of this ungulate includes 74 plant species.[18] However, the limited number of plants constituting the diet of the aoudad at Bou Hedma is more comparable with results obtained in Morocco[29] and in Palo Duro Canyon, Texas, in which a total of 25 food plants was identified.[30]

This variation in the number of plants used by the aoudad between these regions can be related to differences in the specific richness of vegetation available. The scarcity of plant species constituting the diet of the Tunisian aoudad could potentially be associated with plant identification problems from faecal samples and the apparent ease of digestibility of tender grass and succulent forbs.[31] In Bou Hedma National Park, the aoudad shows a clear preference for grasses (67% of the annual diet), followed by browse (17%) and forbs (16%), which allows this ungulate to be identified as a grazer.[31] This preference for grasses can be explained by the abundance of these plants in the mountain steppe of Bou Hedma National Park.[32] Our results concur with those obtained for this ungulate in the Edwards Plateau region of Texas where the annual diet consisted of 61% grasses, 21% browse and 18% forbs[31], and in California where grass classes represented 97% of the annual diet.[33] Furthermore, a relatively high level of grasses (42%) in the diet of aoudads was observed by Ogren[34] in the Canadian River Gorge of northeastern New Mexico. The low frequency of forbs in the diet of the aoudad at Bou Hedma National Park may be the result of the high pressure of overgrazing by domestic animals as demonstrated by Ogren[34] in the Canadian River Gorge and by Simpson et al.[27] in Palo Duro Canyon. Indeed, Ramsey et al.[35] found that in pastures with low forbs availability, the aoudad preferred grasses to browse, which led the authors to surmise that this forage class was the alternative food source in overgrazed areas.

Nevertheless, our results differ from those obtained in Palo Duro Canyon, where the annual diet of the aoudad is composed of 70% browse, 20% forbs and only 10% grasses,[30] and in the Guadalupe Mountains where the aoudad consumes 42% browse, 38% forbs and only 20% grasses.[18] According to another study conducted in Texas,[27] woody browse is the most important forage class in spring (53%), summer (68%) and autumn (48%), while browse (42%) and grasses (43%) are equally important during winter. Also, in the Eastern High Atlas National Park in Morocco, browse was found to be the major dietary component during winter (52.7%) and autumn (56.6%). Grasses and forbs represented 27.4% and 19.9% of the diet, respectively, in autumn and 21.0% and 23.4% of the diet, respectively, in winter.[29] As has been demonstrated in

the Guadalupe Mountains by Simpson et al.[18], the class of grass in our study has a fairly high rate of consumption in summer and spring, and is at its lowest in winter. Evans[36] also recorded low winter use of grasses in Palo Duro Canyon.

The high consumption of grass species during the breeding season, which tends to be focused in spring,[25] can be explained by the fact that these plants reach a peak level of their protein content and digestibility during this period of the year.[17] Moreover, the Tunisian aoudad tend to prefer open land[13] with abundant grass species in the breeding season. The importance of the consumption of grasses with high nutritional value[32] in summer can be justified by the increased requirements of the animals during this season, which is characterised by peak lamb drop and lactation.[17] In our study, *Stipa parviflora and Stipa tenacissima* were the only major grass species consumed in all seasons with the highest consumption in spring and summer and the lowest in winter.

The genus *Stipa* was also among the grasses preferred by the aoudad in Largo Canyon, New Mexico[37] and was observed in the diet of this ungulate in Morocco in northern Africa.[29] The great affinity to this grass plant is explained by its high pastoral value.[32]The increased consumption of browse in the autumn can be explained by the reduced availability of herbaceous plants during this season and highlights the plasticity of the diet of this ungulate. Our study showed that the highest use of forbs was in winter and the lowest use was during summer, mirroring the results obtained by Simpson et al.[18] in the Guadalupe Mountains, New Mexico.

This observation can be explained by the decrease in the availability of herbaceous plants in summer because of the effects of heat, drought and overgrazing by domestic animals in the mountains of Bou Hedma.[32] The variation in the availability of plant resources in Bou Hedma National Park influenced the aoudad's choice of forage class, as was also demonstrated for this ungulate in the Edwards Plateau of Texas.[33]

Acknowledgements

We thank the forest directors for their support and for their role in facilitating access to Bou Hedma National Park, the park employees for their voluntary collaboration, and the members of the Animal Ecology Laboratory of Arid Land Institute, Medenine (Tunisia) who gave permission for the micro-histological faecal analyses.

Authors' contributions

J.B.M. was responsible for the experimental design, for preparing the samples and for writing the manuscript. S.N. performed some of the experiments and made conceptual contributions.

References

1. Brentejs B. The Barbary sheep in ancient North Africa. In: Simpson CD, editor. Symposium on ecology and management of Barbary sheep. Lubbock, TX: Texas Tech University Press; 1980. p. 25–26.
2. Gray GG. Status and distribution of *Ammotragus lervia*: A worldwide review. In: Hoefs M, editor. Wild sheep. Distribution, abundance, management and conservation of the sheep of the world and closely related mountain ungulates. Whitehorse, Canada: Northern Wild Sheep and Goat Council; 1985. p. 95–126.
3. Le Houérou HN. Outline of the biological history of the Sahara. J Arid Environ. 1992;22:3–30.
4. Loggers CO, Thevenot M, Aulagnier S. Status and distribution of Moroccan wild ungulates. Biol Conserv. 1992;59:9–18. http://dx.doi.org/10.1016/0006-3207(92)90708-U
5. Shackleton DM, editor. Wild sheep and goats and their relatives: Status survey and conservation action plan for Caprinae. Gland: International Union for Conservation of Nature; 1997.
6. Cassinello J, Cuzin F, Jedeidi T, Masseti M, Nader I, De Smet K. *Ammotragus lervia*. In: International Union for Conservation of Nature (IUCN). IUCN Red List of Threatened Species. Version 2013.2 [document on the Internet]. c2008 [cited 2015 March 13]. http://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T1151A3288917.en
7. Corbet GB. The mammals of the palaearctic region: A taxonomic review. Ithaca, NY: Cornell University Press; 1978. p. 314.
8. Ansell WFH. Order Artiodactyla. In: Meester J, Setzer HW, editors. The mammals of Africa: An identification manual. Washington DC: Smithsonian Institution Press; 1971. p. 1–84.
9. De Smet K. Tunisia. In: Shackleton DM, editor. Wild sheep and goats and their relatives: Status survey and conservation action plan for Caprinae. Gland: International Union for Conservation of Nature; 1997. p. 45–47.
10. Schomber HW, Kock D. The wild life of Tunisia: Part 2. Some larger animals. Afr Wildlife. 1960;14:277–282.
11. Clarck JL. The great arc of the wild sheep. Norman, OK: University of Oklahoma Press; 1964. p. 247.
12. General Directorate of Forests (DGF). Stratégie Nationale Tunisienne pour la conservation et la restauration des antilopes sahélo-sahariennes et leurs habitats (2001–2020) [Tunisian National Strategy for the Conservation and Restoration of Sahelo-Saharan antelopes and their habitats (2001–2020)]. Tunisia: Direction Générale des Forêts, Ministère de l'Agriculture Tunisie; 2001. p. 29. French.
13. Ben Mimoun J, Nouira S. Social organization of Barbary sheep *(Ammotragus lervia)* population in the Chambi National Park, Tunisia. Int J Biodivers Conserv. 2013;5(1):15–19. http://dx.doi.org/10.5897/IJBC11.104
14. Brouin G. Contribution à l'étude de l'Aïr: Notes sur les ongulés du cercle d'Agadez et leur chasse. Institut Français d'Afrique noire [Contribution to the study of the Aïr: Notes on ungulates of the circle of Agadez and their hunting. French Institute of Black Africa]. Mémoires. 1950;10:425–455. French.
15. Malbrant R. Faune du centre Africain Français (Mammifères et oiseaux) [Fauna of the French African center (mammals and birds)]. In: Lechevalier P, editor. Encyclopédie Biologique. 2nd ed. Paris; 1952. p. 59–61. French.
16. Hansen RM, Lucich GC. A field procedure and study design for fecal collections to be used to determine wildlife and livestock food habits. Fort Collins, CO: Composition Analysis Lab; 1978. p. 15.
17. Wagner GD, Peek JM. Bighorn sheep diet selection and forage quality in central Idaho. Northwest Sci. 2006;80(4):246–258.
18. Simpson CD, Krysl LJ, Dickinson TG. Food habits of Barbary sheep in the Guadalupe Mountains, New Mexico. In: Simpson CD, editor. Symposium on ecology and management of Barbary sheep. Lubbock, TX: Texas Tech University Press; 1980. p. 87–91.
19. Sparks DR, Malechek JC. Estimating percentage dry weights in diets using a microscopic technique. J Range Manage. 1968;21:264–265. http://dx.doi.org/10.2307/3895829
20. Free CJ, Hansen RM, Sims PL. Estimating dry weights of food plants in feces of herbivores. J Range Manage. 1970;23:300–302. http://dx.doi.org/10.2307/3896227
21. Williams OB. A technique for studying microtine food habits. J Mammal. 1962;57:167–172. http://dx.doi.org/10.2307/1376945
22. Storr GM. Microscopic analysis of faeces, a technique for ascertaining the diet of herbivores mammals. Austral J Biol Sci. 1961;14:157–164.
23. Delaunay G. Contribution à la mise au point de méthodes de suivi des populations d'ongulés de hautes montagnes en milieu protégé: Etude sur le chamois dans le parc national des Ecrins [Contribution to the development of monitoring methods of ungulate populations of high mountains in protected areas: Study on the chamois in the Ecrins National Park] [thesis]. Rennes: Université de Rennes; 1982. p. 280. French.
24. Siegel S. Non-parametric statistics for the behavioral sciences. New York: McGraw-Hill; 1956. p. 312.
25. Cassinello J. *Ammotragus lervia*: A review on systematics, biology, ecology and distribution. Ann Zool Fenn. 1998;35:149–162.
26. Barrett RH, Beamson SL. Research needs for management of Barbary sheep in the future. In: Simpson CD, editors. Proceedings of the symposium on ecology and management of Barbary sheep; 1979 Nov 19–21; Texas, USA. Lubbock, TX: Texas Tech University Press; 1980. p.106–107.
27. Simpson CD, Krysl LJ, Hampy DB, Gray GG. The Barbary sheep: A threat to desert bighorn survival. Desert Bighorn Council Trans. 1978;22:26–31.

28. Piñero JCR, Luengo JLR. Autumn food habits of the Barbary sheep (*Ammotragus lervia* Pallas 1777) on La Palma Island (Canary Islands). Mammalia. 1992;56:385–392.

29. Hafidi MEM. Food habits and preferences of Barbary sheep *(Ammotragus lervia)* in the eastern High Atlas National Park [MSc thesis]. Morocco: International Centre for advanced Mediterranean Agronomic Studies, Mediterranean Agronomic Institute of Chania; 1996.

30. Krysl L, Simpson CD, Gray G. Dietary overlap of sympatric Barbary sheep and mule deer in Palo Duro Canyon, Texas. In: Simpson CD, editors. Symposium on ecology and management of Barbary sheep; 1979 Nov 19–21; Texas, USA. Lubbock, TX: Texas Tech University Press; 1980. p. 97–103.

31. Schaller GB. Mountain monarchs: Wild sheep and goats of the Himalaya. Chicago, IL: University of Chicago Press; 1977. p. 425.

32. Roselt/OSS. Etude de la biodiversité dans l'observatoire pilote de Haddej – Bou Hedma (TUNISIE) [Study of biodiversity in the pilot observatory Haddej - Bou Hedma (TUNISIA)]. CT no. 7. Montpellier: Collection ROSELT/OSS; 2004. p. 110. French.

33. Ogren HA. Barbary sheep. New Mexico Department of Game and Fish Bulletin 13. Santa Fe, NM: New Mexico Department of Game and Fish; 1965.

34. Ogren HA. Barbary sheep in New Mexico. New Mexico Department of Game and Fish Bulletin 11. Santa Fe, NM: New Mexico Department of Game and Fish; 1962. p. 32.

35. Ramsey CW, Anderegg MJ. Food habits of an aoudad sheep, *Ammotragus lervia* (Bovidae), in the Edwards Plateau of Texas. Southwestern Nat. 1972;16:267–280. http://dx.doi.org/10.2307/3670063

36. Evans PK. The aoudad sheep, an exotic introduced in the Palo Duro Canyon of Texas. Proceedings of the Southeastern Association of Game and Fish Commissioners Conference; 1947; New Orleans, USA. New Orleans: Southeastern Association of Game and Fish Commissioners; 1967. p. 183–188.

37. Bird W, Upham LL. Barbary sheep and mule deer food habits of Largo Canyon, New Mexico. In: Simpson CD, editor. Proceedings of the symposium on ecology and management of Barbary sheep; 1979 Nov 19–21; Texas, USA. Lubbock, TX: Texas Tech University Press; 1980. p. 92–96.

Insecticide resistance in the malaria vector *Anopheles arabiensis* in Mamfene, KwaZulu-Natal

AUTHORS:
Basil D. Brooke[1,2]
Leanne Robertson[1]
Maria L. Kaiser[1,2]
Eric Raswiswi[3]
Givemore Munhenga[1,2]
Nelius Venter[1,2]
Oliver R. Wood[1,2]
Lizette L. Koekemoer[1,2]

AFFILIATIONS:
[1]Centre for Opportunistic, Tropical & Hospital Infections, National Institute for Communicable Diseases, Johannesburg, South Africa
[2]Wits Research Institute for Malaria, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[3]KwaZulu-Natal Department of Health, KwaZulu-Natal Provincial Government, Jozini, South Africa

CORRESPONDENCE TO:
Basil Brooke

EMAIL:
basilb@nicd.ac.za

POSTAL ADDRESS:
Vector Control Reference Laboratory, National Institute for Communicable Diseases, Private Bag X4, Sandringham 2131, South Africa

KEYWORDS:
malaria vector control; malaria elimination; vector surveillance; South Africa; resistance management

HOW TO CITE:
Brooke BD, Robertson L, Kaiser ML, Raswiswi E, Munhenga G, Venter N, et al. Insecticide resistance in the malaria vector *Anopheles arabiensis* in Mamfene, KwaZulu-Natal. S Afr J Sci. 2015;111(11/12), Art. #2015-0261, 3 pages. http://dx.doi.org/10.17159/sajs.2015/20150261

The control of malaria vector mosquitoes in South Africa's affected provinces is primarily based on indoor spraying of long-lasting residual insecticides. The primary vectors in South Africa are *Anopheles arabiensis* and *An. funestus*. South Africa's National Malaria Control Programme has adopted a malaria elimination agenda and has scaled up vector control activities accordingly. However, despite these plans, local transmission continues and is most likely because of outdoor feeding by populations of *An. arabiensis*. An outdoor *Anopheles* surveillance system has been set up in three sections of the Mamfene district in northern KwaZulu-Natal in order to assess the extent of outdoor resting *An. arabiensis* in Mamfene and to assess the current insecticide susceptibility status of this population. According to WHO criteria, the *An. arabiensis* samples tested showed evidence of resistance to deltamethrin (pyrethroid), DDT (organochlorine) and bendiocarb (carbamate), and full susceptibility to the organophosphates pirimiphos-methyl and fenitrothion. Pre-exposure to piperonyl butoxide completely nullified the deltamethrin resistance otherwise evident in these samples, supporting previous studies implicating monooxygenase-based detoxification as the primary mechanism of pyrethroid resistance. The data presented here affirm the presence of pyrethroid and DDT resistance previously detected in this population and also indicate the comparatively recent emergence of resistance to the carbamate insecticide bendiocarb. These data show that special attention and commitment needs to be given to the principles of insecticide resistance management as well as to investigations into alternative control techniques designed to target outdoor-resting *An. arabiensis* in northern KwaZulu-Natal.

Introduction

The control of malaria vector mosquitoes in South Africa's affected provinces is primarily based on indoor spraying of long-lasting residual insecticides.[1] The indoor residual spraying (IRS) method has been the mainstay of malaria vector control in South Africa since the late 1940s and has remained effective owing to carefully co-ordinated IRS programmes in South Africa's Limpopo, Mpumalanga and KwaZulu-Natal Provinces.[2]

Only *Anopheles* mosquitoes can transmit human malaria parasites and the primary vectors in South Africa are *Anopheles arabiensis* and *An. funestus*.[1] Of these, *An. funestus* is almost entirely anthropophilic (human biting), endophagic (indoor feeding) and endophilic (indoor resting).[3] These characteristics make this species especially susceptible to control by IRS, assuming that the insecticide employed for this purpose is effective against the target *An. funestus* population. Control by IRS means that the mosquitoes must retain complete or near complete susceptibility to the insecticide class being used, which can only be ascertained by regular monitoring and surveillance. The malaria epidemic experienced in South Africa during the period 1996–2000 was largely the result of the development of resistance to pyrethroid insecticides in populations of this species in northern KwaZulu-Natal and Mpumalanga which led to vector control failure.[1,2] Control was re-established using a mosaic resistance management system which was later drafted into a World Health Organization (WHO) document –- the Global Plan for Insecticide Resistance Management (GPIRM).[4] South Africa currently subscribes to the principles outlined in GPIRM as part of its malaria elimination agenda.[5] However, despite these plans and the scaling up of vector control activities in South Africa, local transmission continues, most likely because of outdoor transmission by populations of *An. arabiensis*. Unlike *An. funestus*, *An. arabiensis* has evolved substantial behavioural plasticity and will feed and rest indoors and outdoors, and will feed on humans as well as livestock, especially bovines.[3] *Anopheles arabiensis* is therefore substantially less susceptible to control by IRS.

Recently, a project was launched to assess the feasibility of the sterile insect technique for malaria vector control in South Africa, with special emphasis on controlling outdoor transmission by *An. arabiensis*.[6] As part of the baseline survey linked to this project, an outdoor *Anopheles* surveillance system has been set up in three sections of the Mamfene district in northern KwaZulu-Natal. This surveillance system has enabled recent assessments of insecticide resistance in outdoor-resting *An. arabiensis* in Mamfene as a follow-up to the discovery of pyrethroid resistance in this region in 2005.[7]

Methods

In order to assess the current insecticide susceptibility status of outdoor resting *An. arabiensis* in Mamfene, adult *Anopheles* mosquitoes were collected from outdoor-placed ceramic pots[8] and modified plastic buckets deployed in 20 households in Mamfene Sections 2, 8 and 9 during March and April 2015. These collections were transported live to the Botha De Meillon insectary facility at the National Institute for Communicable Diseases (NICD) in Johannesburg. Blood-fed female specimens were individually placed in egg-laying vials so that eggs could be harvested and reared by family. All wild-caught female individuals, including those that produced eggs, were identified to species group using morphological keys[9,10] and to species by standard PCR[11]. A total of 35 families identified as *An. arabiensis* was pooled and the F1 progeny were reared to adults under standard insectary conditions of 25 °C and 80% relative humidity.[12]

Samples of 2–5-day-old female adult F1 progeny were assessed for their susceptibility to diagnostic concentrations of a range of insecticides according to the standard WHO bioassay method.[13] Controls included samples of 2–4-day-old F1 male adults exposed to untreated papers. In addition, a subset of samples was used to assess the effect of pre-exposure to the insecticide synergist piperonyl butoxide (PBO) on the expression of pyrethroid resistance according to a method previously described by Brooke et al.[14]

Results and discussion

According to WHO criteria,[13] the F1 *An. arabiensis* samples tested showed evidence of resistance to deltamethrin (pyrethroid), DDT (organochlorine) and bendiocarb (carbamate), and full susceptibility to the organophosphates pirimiphos-methyl and fenitrothion (Table 1). Pre-exposure to PBO completely nullified the deltamethrin resistance otherwise evident in these samples (paired sample *t*-test: d.f. = 1, $t = 15.65$, $p = 0.04$) (Table 2).

The first assessments of resistance in *An. arabiensis* at Mamfene were conducted in 1996 and no resistance phenotypes were recorded.[15] However, subsequent samples collected in 2002 indicated the emergence of resistance to DDT[16] which was again recorded in 2005 together with the first indication of pyrethroid resistance[7]. The 2015 data presented here affirm the presence of pyrethroid and DDT resistance in this population, albeit at a low frequency, and also indicate the comparatively recent emergence of resistance to the carbamate insecticide bendiocarb. The PBO exposure data support previous analyses implicating monooxygenase-mediated detoxification as the primary mode of pyrethroid resistance in *An. arabiensis* at Mamfene,[7,17,18] because PBO enhances insecticide toxicity by providing an alternative substrate for monooxygenases. PBO pre-exposure assays can therefore be used to elucidate monooxygenase-based resistance mechanisms.

Conclusion

South Africa's 1996 to 2000 malaria epidemic illustrates the effect that a single insecticide resistance phenotype (pyrethroid resistance in *An. funestus*) can have on an IRS-based vector control programme.[2,19,20] The occurrence of multiple vector species carrying multiple resistance mechanisms coupled to ongoing outdoor transmission in northern KwaZulu-Natal means that special attention and commitment needs to be given to the principles of insecticide resistance management as outlined in the GPIRM document[4] as well as to investigations into alternative control techniques designed to target outdoor-resting *An. arabiensis*.

Acknowledgements

We thank Mr Sifiso Ngxongo and Mr Jabulani Zikhali for their assistance with the collection of specimens and Prof. Maureen Coetzee for her comments on the manuscript. This work was funded by the Nuclear Technologies in Medicine and the Bioscience Initiatives (NTeMBI), a national platform developed and managed by the South African Nuclear Energy Corporation and supported by the Department of Science and Technology. Funding was also provided in part by the National Research Foundation and by the International Atomic Energy Agency ((Contract no SAF16780 (under the G34002)) as well as by a CDC/GDD (Global Diseases Detection programme) grant to B.D.B. (U19GH000622-01 MAL01).

Authors' contributions

B.D.B. assisted with data analysis and interpretation and produced the manuscript. L.R., N.V. and O.R.W. assisted with specimen collection, species identification and experimental procedures. M.L.K. and G.M. assisted with the experimental procedures and data analysis. E.R. assisted with data interpretation and the drafting of the manuscript. L.L.K. conceived the project, assisted with experimental procedures, data analysis and interpretation and the drafting of the manuscript. All authors read and approved the final draft of the manuscript.

Table 1: Mean percentage mortalities recorded for samples of 2–5-day-old F1 *Anopheles arabiensis* female adults following exposure to listed insecticides by class. Mortalities were recorded 24 h post exposure. The number of replicates, sample size (*n*) and standard error (s.e.) are given for each insecticide treatment. Samples by treatment are categorised as resistant (R) or susceptible (S) according to standard criteria.[13] The control was exposure of 2–4-day-old F1 *Anopheles arabiensis* male adults to untreated papers.

Insecticide (concentration)	Insecticide class	Number of replicates	*n*	Mean % mortality	s.e.	Resistance or susceptibility
Deltamethrin (0.05%)	Pyrethroid	8	191	87.21	4.63	R
DDT (4%)	Organochlorine	7	140	83.85	7.32	R
Bendiocarb (0.1%)	Carbamate	7	145	94.1	2.8	R
Pirimiphos-methyl (0.25%)	Organophosphate	2	45	100	–	S
Fenitrothion (1%)	Organophosphate	2	44	100	–	S
Control	–	10	234	2.74	1.6	–

All exposures were of 1 h duration except for fenitrothion for which there was a 2-h exposure

Table 2: Mean percentage mortalities recorded for samples of 2–6-day-old F1 *Anopheles arabiensis* male adults (M) or female adults (F) following exposure to either deltamethrin (0.05%), piperonyl butoxide (PBO) (4%), deltamethrin + PBO or untreated control papers. Mortalities were recorded 24 h post exposure. The number of replicates, sample size (*n*) and standard error (s.e.) are given for each treatment. Mortalities following exposure to PBO and control papers were negligible. There is a significant difference in mean mortality between the deltamethrin and deltamethrin + PBO treatments 24 h post exposure (paired sample *t*-test: d.f.=1, $t=15.65$, $p=0.04$).

Treatment	Sex	Number of replicates	*n*	Mean % mortality	s.e.
Deltamethrin	F	2	47	78.64	1.37
PBO	M	2	46	6.44	1.89
Deltamethrin + PBO	F	2	49	100	–
Control	M	4	90	5.49	3.35

References

1. Brooke B, Koekemoer L, Kruger P, Urbach J, Misiani E, Coetzee M. Malaria vector control in South Africa. S Afr Med J. 2013;103(10 Suppl 2):784–788. http://dx.doi.org/10.7196/samj.7447
2. Coetzee M, Kruger P, Hunt RH, Durrheim DN, Urbach J, Hansford CF. Malaria in South Africa: 110 years of learning to control the disease. S Afr Med J. 2013;103(10 Suppl 2):770–778. http://dx.doi.org/10.7196/samj.7446
3. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, et al. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: Occurrence data, distribution maps and bionomic précis. Parasit Vectors. 2010;3:117. http://dx.doi.org/10.1186/1756-3305-3-117
4. World Health Organization (WHO). Global Plan for insecticide resistance management. Geneva: WHO; 2012. Available from: http://www.who.int/malaria/vector_control/ivm/gpirm/en/
5. Maharaj R, Morris N, Seocharan I, Kruger P, Moonasar D, Mabuza A, et al. The feasibility of malaria elimination in South Africa. Malar J. 2012;11:423. http://dx.doi.org/10.1186/1475-2875-11-423
6. Munhenga G, Brooke BD, Chirwa TF, Hunt RH, Coetzee M, Koekemoer LL. Evaluating the potential of the sterile insect technique for malaria control: Relative fitness and mating compatibility between laboratory colonized and a wild population of *Anopheles arabiensis* from the Kruger National Park, South Africa. Parasit Vectors. 2011;4:208. http://dx.doi.org/10.1186/1756-3305-4-208
7. Mouatcho JC, Munhenga G, Hargreaves K, Brooke BD, Coetzee M, Koekemoer LL. Pyrethroid resistance in a major African malaria vector *Anopheles arabiensis* from Mamfene, northern Kwazulu-Natal, South Africa. S Afr J Sci. 2009;105:127–131.
8. Brooke B, Hargreaves K, Spillings B, Raswiswi E, Koekemoer L, Coetzee M. Novel malaria control: Can traditional clay pots be used to deliver entomopathogenic fungi to malaria vectors in northern Kwazulu-Natal? Communicable Diseases Surveillance Bulletin. 2012;10:13–15.
9. Gillies MT, De Meillon B. The Anophelinae of Africa south of the Sahara. Johannesburg: South African Institute for Medical Research; 1968.
10. Gillies MT, Coetzee M. A supplement to the Anophelinae of Africa south of the Sahara. Johannesburg: South African Institute for Medical Research; 1987.
11. Scott JA, Brogdon WG, Collins FH. Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. Am J Trop Med Hyg.1993;49:520–529.
12. Hunt RH, Brooke BD, Pillay C, Koekemoer LL, Coetzee M. Laboratory selection for and characteristics of pyrethroid resistance in the malaria vector *Anopheles funestus*. Med Vet Entomol. 2005;19:271–275. http://dx.doi.org/10.1111/j.1365-2915.2005.00574.x
13. World Health Organization (WHO). Test procedures for insecticide resistance monitoring in malaria vector mosquitoes. Geneva: WHO; 2013. http://www.who.int/malaria/publications/atoz/9789241505154/en/
14. Brooke BD, Kloke G, Hunt RH, Temu EA, Koekemoer LL, Taylor ME, et al. Bioassay and biochemical analyses of insecticide resistance in southern African *Anopheles funestus*. Bull Ent Res. 2001;91:265–272. http://dx.doi.org/10.1079/BER2001108
15. Gericke A, Govere JM, Durrheim DN. Insecticide susceptibility in the South African malaria mosquito *Anopheles arabiensis*. S Afr J Sci. 2002;98:205–208.
16. Hargreaves K, Hunt RH, Brooke BD, Mthembu J, Weeto MM, Awolola TS, et al. *Anopheles arabiensis* and *An. quadriannulatus* resistance to DDT in South Africa. Med Vet Entomol. 2003;17:417–422. http://dx.doi.org/10.1111/j.1365-2915.2003.00460.x
17. Munhenga G, Koekemoer LL. Differential expression of P450 genes in *Anopheles arabiensis* following permethrin selection. Afr J Biotechnol. 2011;10:12711–12716.
18. Nardini L, Christian RN, Coetzer N, Ranson H, Coetzee M, Koekemoer LL. Detoxification enzymes associated with insecticide resistance in laboratory strains of *Anopheles arabiensis* of different geographic origin. Parasit Vectors. 2012;5:113. http://dx.doi.org/10.1186/1756-3305-5-113
19. Maharaj R, Mthembu DJ, Sharp BL. Impact of DDT re-introduction on malaria transmission in KwaZulu-Natal. S Afr Med J. 2005;95:871–874.
20. Maharaj R, Raman J, Morris N, Moonasar D, Durrheim DN, Seocharan I, et al. Epidemiology of malaria in South Africa: From control to elimination. S Afr Med J. 2013;103(10 Suppl 2):779–783. http://dx.doi.org/10.7196/samj.7441

A sculpture by Marco Cianfanelli at the Nelson Mandela Capture Site in Howick, KwaZulu-Natal. The 50 non-linear steel columns seen individually when viewed from the side (back cover image: ©Sibusiso Manzini) align to form a two-dimensional portrait of Nelson Mandela when viewed from the front (front cover image: ©Paul Gregg/African Images/AfriPics). In an article on page 77, Manzini motivates for broadening current perspectives of innovation in South Africa, beyond science and technology.

APPLYING SCIENTIFIC THINKING IN THE SERVICE OF SOCIETY

Our vision is to be the apex organisation for science and scholarship in South Africa, internationally respected and connected, its membership simultaneously the aspiration of the country's most active scholars in all fields of scientific enquiry, and the collective resource for the professionally managed generation of evidence-based solutions to national problems.

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

T +27 12 349 6600/21/22 | F +27 86 576 9514

WWW.ASSAF.ORG.ZA