Survey of contaminants in South Africa’s drinking water

Land-cover change in KZN: implications for biodiversity

Discovery of a new human ancestor: *Homo naledi*

Nematode pests threaten soybean production in South Africa

Food consumption changes in South Africa since 1994

SEPTEMBER/OCTOBER 2015

ISSN: 0038-2353
eISSN: 1996-7489

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

SOUTH AFRICAN Journal of Science

volume 111
number 9/10

SEPTEMBER/OCTOBER

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

SOUTH AFRICAN Journal of Science

volume 111

number 9/10

Leader

Commentary

Book Review

Review Article

Research Article

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

Disclaimer
The publisher and editors accept no responsibility for statements made by the authors.

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

Subscriptions
Subscription information can be
found at
www.sajs.co.za

What does ‘science’ mean in the title *South African Journal of Science*?

What can be said about the meaning of the word ‘science’ in the names of the *South African Journal of Science* and its publisher, the Academy of Science of South Africa (ASSAf)?

Perhaps what can be said, as a start, is that we need to have a clearer understanding of the etymology of the word and the implications that those meanings have had for the ways in which science has been practised and understood, at least in the Western world.

‘Science’ is one of hundreds of thousands of words in English that has an extraordinarily long etymological history and whose popular meaning has changed, century by century, and sometimes even more rapidly than that. Yet even amongst those words there are core meanings that have remained consistent.

In English, we have ‘science’ from Old French (meaning ‘knowledge, learning, application; a corpus of human knowledge’), where it originally entered from the Latin word ‘scientia’ meaning ‘knowledge, a knowing, expertness, or experience’. By the late 14th century, ‘science’ meant, in English, ‘collective knowledge’. But it has consistently carried the meaning of being a socially embedded activity: people seeking, systematising and sharing knowledge. Nonetheless, in the English speaking world at least, there are fierce debates about what constitutes the proper ways of defining and constituting the proper ways of undertaking research and designating ‘real knowledge’. These debates have their origins in the earliest Western universities whose intellectual context was that of the values and belief systems of the Catholic Church – and in the impact that the secularisation of universities had in later centuries.

Disciplines as we know them today only arose in the 18th and 19th centuries; and although they have changed, with new disciplines being added and some shrinking or disappearing, the debates continue about which disciplines are ‘superior’ to others, and which are undertaking ‘real’ research. This periodic ‘taunting’ of some disciplines by others is, then, hardly new.

Muller[1] captures the essence of this kind of ‘debate’ as it played out in the 1960s, in the furore generated by papers given by politician Lord CP Snow (a Cambridge trained chemist and published novelist) and Professor FR Leavis, a Cambridge literary scholar. Here is the story that Muller sets out:

> *Snow…presented a Rede Lecture at Cambridge, called provocatively ‘The Two Cultures and the Scientific Revolution’. It was at the secularised guardians of elite ‘traditional’ culture that Snow aimed his provocation. Snow characterised scientific culture as optimistic and forward looking, though regarded as shallow and philistine by the cultivated literary culture of the literary elite, who Snow considered ignorant snobs. He derided the mutual incomprehension of the two cultures: ‘The degree of incomprehension on both sides is the kind of joke which has gone sour’ and lamented the ‘sheer loss to us all’. The fault he laid squarely at the door of the literary intellectuals, calling them ‘natural Luddites’ who lacked the culture to grasp the second law of thermodynamics, a piece of general cultural knowledge he likened to knowing something about Shakespeare. …*[And] *then went on to say that industrialisation was the only hope for the poor and the Third World, and that the best the developed world could do was to produce as many engineers as it could and export them to where they were needed in the developing world.*
>
> *Despite his oversimplifications, Snow had hit a nerve. The most extreme response came from FR Leavis, doyen of the literary elite. In a lecture first given also at Cambridge,…and re-published by Leavis, Leavis heaped derision on Snow’s ‘embarrassing vulgarity of style’, on his ignorance, and on his ineptness as a novelist; he is, said Leavis, as ‘intellectually undistinguished as it is possible to be’. Leavis’ attack drew an avalanche of responses, which called it inter alia ‘bemused drivelling’ of ‘unexampled ferocity’.*

The debates may no longer be quite that ferocious, but their sounds still echo faintly through academia – more so in some countries than in others.

Yet a core of commonality is to be found: whether working within a paradigm (and remember that these too shift as research progresses) or ‘pre-paradigmatically’, three basic foundations are explicitly present. In fields as different as Genomics or Human Geography, the *raisons d’être* of ‘hard’ and ‘soft’ sciences and, of course, many of their ‘applied’ allies (Engineering, Accountancy…), are the development of new knowledge through research; advancing that knowledge; and sharing it through publication and teaching.

It is as complicated – and yet as simple – as that: the *South African Journal of Science* publishes work based on, or leading to, those foundations.[2] The Journal is about quality knowledge-producing research, not about disciplines. After all, the National Research Foundation has just made top ‘rating’ awards to scholars in widely diverse disciplines such as Epidemiology (to Quarraisha Abdool Karim); Policy Studies (Nuraan Davids); Medicine (Ntobeko Ntusi); History (Charles van Onselen) and Computational and Applied Mathematics (Daya Reddy, the President of ASSAf). That is precisely what the ‘science’ in the *South African Journal of Science* is all about, just as it is what ASSAf is about.

In fact, it is the diversity of different disciplines that enshrines the strength of the contemporary university (and the Journal) – a strength sometimes obscured by rankings which favour the ‘natural’ sciences.

In the 21st century scientific world of inter-, multi- and trans-disciplinary research, all of which are increasingly valuable approaches to discovery and innovation, what remains fundamental are the inescapable disciplinary foundations and their contributions to universities. Yet, while protecting the value of the essential, it is clear that there is an equally inescapable need for greater (and growing) mutual respect of the different ways in which knowledge is produced, and research findings reported, so that cooperation becomes more, rather than less, possible. To make the most of science, it is now more important than ever to celebrate the contributions that it makes, across the spectrum of disciplines, whether individually or collectively. It is in this way that science contributes significantly to the well-being of ourselves, the environment on which we depend, and the richness of our world: genetics, agriculture, meteorology, music, literature, and so on. How might we possibly live without the benefits that they, and their fellow disciplines, all offer?

References

1. Muller J. In search of coherence: A conceptual guide to curriculum planning for comprehensive universities. Report prepared for the SANTED Project; 2008 [unpublished report].

2. The *South African Journal of Science* adopts online publishing (Leader). S Afr J Sci. 2010;106(1/2), Art. #151, 1 page. http://dx.doi.org/10.4102/sajs.v106i1/2.15

HOW TO CITE: Butler-Adam J. What does ‘science’ mean in the title *South African Journal of Science*? S Afr J Sci. 2015;111(9/10), Art. #a0123, 1 page. http://dx.doi.org/10.17159/sajs.2015/a0123

A new star rising: Biology and 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:
Cradle of Humankind; hominin; Dinaledi Chamber; palaeoanthropology; geology; taphonomy

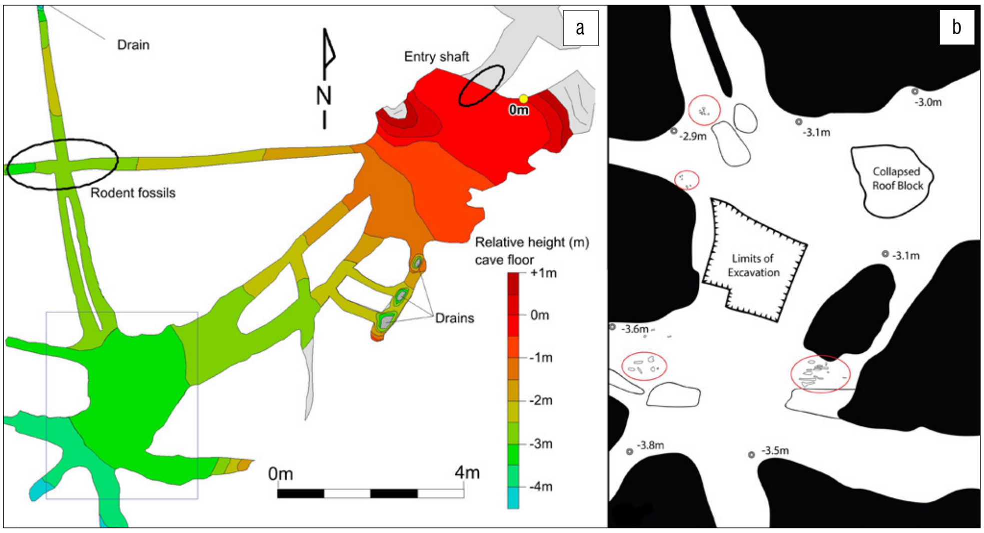
HOW TO CITE:
Randolph-Quinney PS. A new star rising: Biology and mortuary behaviour of *Homo naledi.* S Afr J Sci. 2015;111(9/10), Art. #a0122, 4 pages. http://dx.doi.org/10.17159/sajs.2015/a0122

September 2015 saw the release of two papers detailing the taxonomy[1], and geological and taphonomic[2] context of a newly identified hominin species, *Homo naledi* – *naledi* meaning 'star' in Sesotho. Whilst the naming and description of a new part of our ancestral lineage has not been an especially rare event in recent years,[3-7] the presentation of *Homo naledi* to the world is unique for two reasons. Firstly, the skeletal biology, which presents a complex mixture of primitive and derived traits, and, crucially, for which almost every part of the skeleton is represented – a first for an early hominin species*.* Secondly, and perhaps more importantly, this taxon provides evidence for ritualistic complex behaviour, involving the deliberate disposal of the dead.

The initial discovery was made in September 2013 in a cave system known as Rising Star in the Cradle of Humankind World Heritage Site, some 50 km outside of Johannesburg. Whilst amateur cavers had been periodically visiting the chamber for a number of years, the 2013 incursion was the first to formally investigate the system for the fossil remains of early hominins. The exploration team comprised Wits University scientists and volunteer cavers, and was assembled by Lee Berger of the Evolutionary Studies Institute, who advocated that volunteer cavers would use their spelunking skills in the search for new hominin-bearing fossil sites within the Cradle of Humankind. Whilst most South African field palaeoanthropologists have at one time or another had cause to venture into the caves of the Cradle, very few have received the formal training that would allow them to climb, abseil or squeeze into the deepest, most dangerous cave environments; Berger's team is unique in that they willingly went into the dark spaces of the earth that the rest of us (or at least those of us with a healthy survival instinct) balk at.

The investigation proved immediately fruitful in that undoubtedly hominin skeletal remains were recorded by the exploration team in a chamber some 30 m underground and 90 m from the cave entrance; the skeleton of a single individual was suspected, and lay exposed on the surface of soft unconsolidated sediments. Because of the sedimentary environment, recovery could proceed using methods more akin to bioarchaeology, rather than the traditional palaeontology of the Cradle in which fossils are entombed in calcified breccia. However, the structure of the cave was such that the fossil chamber (named Dinaledi Chamber) was only accessible by a near-vertical chute and crawl so narrow that only very small and slender individuals could access it. As such, a formal excavation and recovery effort was set up, involving the use of excavators who were physically suited to access the cave. The geology and context paper[2] details the unique methods of the excavation strategy, in-situ recording, and recovery of the fossils, which include 'remote' field direction and 3D scanning of the excavations rather than using conventional archaeological survey equipment – limitations imposed by the inaccessibility and small size of the Dinaledi Chamber. These imaginative strategies culminated in the Rising Star Expedition which during November 2013 and March 2014 recovered more than 1550 identifiable fossil elements; about 300 numbered bone specimens were collected from the surface of the Dinaledi Chamber and about 1250 numbered fossil specimens were recovered from a small excavation pit in the chamber (Figure 1). This discovery is the largest single fossil hominin assemblage found on the African continent to date.

The fossils, which have yet to be radiometrically dated, were derived from at least 15 individuals, a total likely to represent a small fraction of the fossils remaining in the chamber and awaiting excavation. Through the



Source: (a) Modified with permission from Paul Dirks; (b) Patrick Randolph-Quinney

Figure 1: Schematic of the Rising Star (Dinaledi) Chamber. Figure 1a shows the overall layout of the Dinaledi Chamber, side passages and crawls, and the relative drop in height from the entry shaft and the limits of accessibility within the chamber. Figure 1b (the area in 1a outlined in blue) shows the main area of excavation within the chamber, with significant concentrations of hominin bone discovered on the surface of the floor outlined in red.

course of a long and highly detailed paper, Berger and colleagues[1] lay out the argument to place the Rising Star fossils into a new species, which whilst highly primitive in terms of cranial capacity and body proportions, still warrants inclusion in the genus *Homo.* The species is characterised by a body mass and stature similar to those of small-bodied human populations, with an average stature of approximately 1.5 m and an average body mass of about 45 kg (range 40–56 kg), but has a small endocranial volume (465–560 cc) similar to that found in *Australopithecus*. The teeth of *H. naledi* are generally small, with simple cusp morphology – traits shared with *Homo habilis*. The skull morphology of *H. naledi* is unique, but shares similarities to other early *Homo* species including *H. erectus*, *H. habilis* and *H. rudolfensis,* and differs markedly from taxa such as *Paranthropus* and *Australopithecus ghari* through lack of cranial crests, and *A. afarensis, A. africanus* and *A. sediba* through (amongst others) the expression of sagittal keeling and an angular occipital torus, and in the brow region with a pronounced supraorbital torus with post-toral sulcus (a depression between the brow ridge and the rising frontal bone). The mandible is gracile, with a vertically oriented symphyseal region, overall more akin to early *Homo* than *Paranthropus* or *Australopithecus.*

Postcranially *H. naledi* is a complex mixture of primitive and derived. The hands and feet are human-like in their functional morphology, although *H. naledi* has extremely curved fingers. The hand morphology suggests the capacity for tool-using capabilities, whilst the curvature demonstrates climbing capability; together a fascinating combination. The human-like hands and feet are contrasted in the postcranial skeleton with a more primitive or australopith-like thorax, shoulder, pelvis and proximal femur.

No phylogenetic analysis is presented in the paper – and I eagerly await analyses which incorporate both the cranial *and* the postcranial skeleton. It will be interesting to see how such an organism aligns phylogenetically with the current record given the combination of derived features of the cranium (bar cranial capacity), the human-like hand and foot, and such a primitive shoulder, thorax and pelvis.

In retrospect, what is perhaps most striking about the analyses presented is that *H. naledi* as a species is defined on an entire corpus of skeletal material, with almost every single element of the body represented multiple times, across multiple individuals of differing biological ages, and which overall displays limited variability in morphology across the species range for any element. The range of variation and taphonomic associations preclude this material being a commingled assemblage made up of multiple taxa. This situation is in opposition to the usual situation which bedevils palaeoanthropology, in which a taxon is narrowly defined on the basis of a single jaw or skull, or through contextually unassociated specimens, because of the vagaries of fossil preservation and recovery.

Whilst the skeletal morphology represents a new and complex suite of evolutionary characteristics, the geological context of the chamber also presents an anomalous depositional environment in comparison to the ‘classic’ sites of the Cradle of Humankind in Gauteng – Sterkfontein, Kromdraai and Swartkrans. The latter sites are noted for fossil remains contained in lithified breccia or found in decalcified sedimentary units derived ultimately from clastic breccia.[8-13] The Dinaledi Chamber is unique in that the fossils were recovered from soft unconsolidated sediments within the karstic system, and appear never to have been breccified during their depositional history (Figure 2). The primary cave structure and stratigraphy were studied in the main by Dirks, Roberts and Kramer.[2] They identified a basic stratigraphic development of two facies (Facies 1 and 2, with 1 being the oldest) subdivided into three stratigraphic Units (1–3, from old to young). Unit 1 represents the remnants of laminated mudstones preserved as erosion remnants within the chamber. Unit 2 is a composite unit that consists of remnant outcrops of variably consolidated sediments which contains several hominin bones, including the shafts of a juvenile hominin ulna and radius. Unit 3 is the youngest stratigraphic unit, and is represented by sediment that accumulated along the floor of the chamber and is composed of largely unconsolidated sediment derived from weathering and erosion of Units 1 and 2. The majority of the hominin bones was derived from Unit 3. Dirks’ geology team have demonstrated that the clay-rich sediments making up the units were derived from in-situ weathering, and from exogenous clays and silts, which entered the chamber through fractures that prevented passage of coarser-grained materials; thus the infill of the Dinaledi Chamber is the end product of a series of filters or traps, which winnowed out all large-grained sediments or clastic material, en route to final deposition within the chamber. This winnowing process did not include the hominin fossil material, which must therefore have been transported into the chamber by a mechanism other than sedimentary accumulation processes.

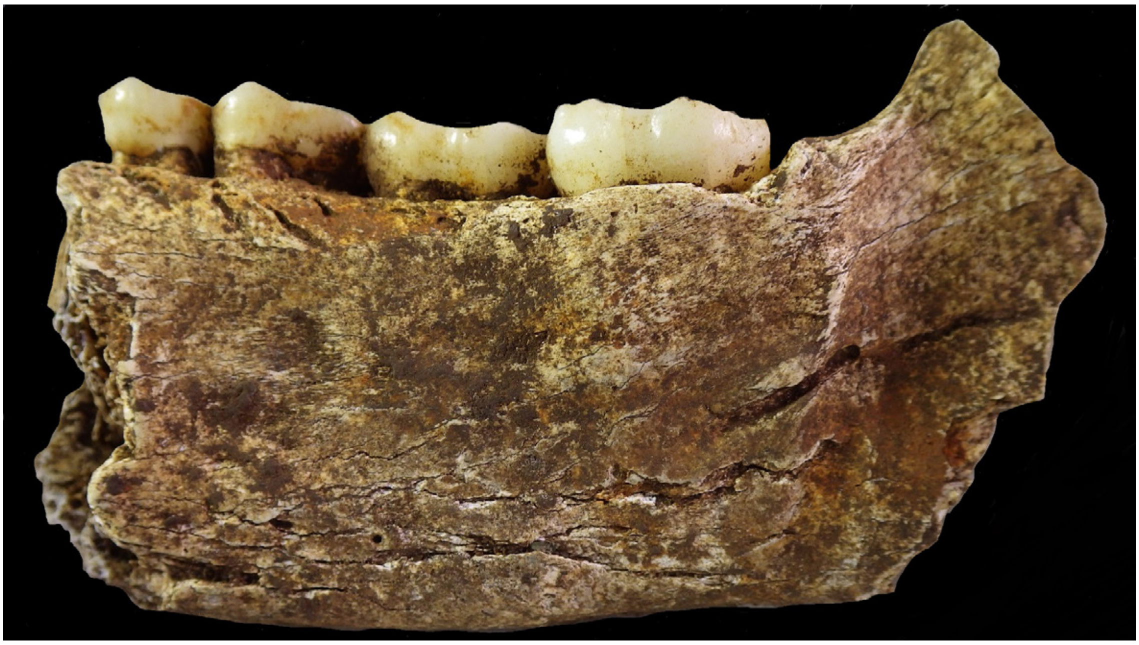


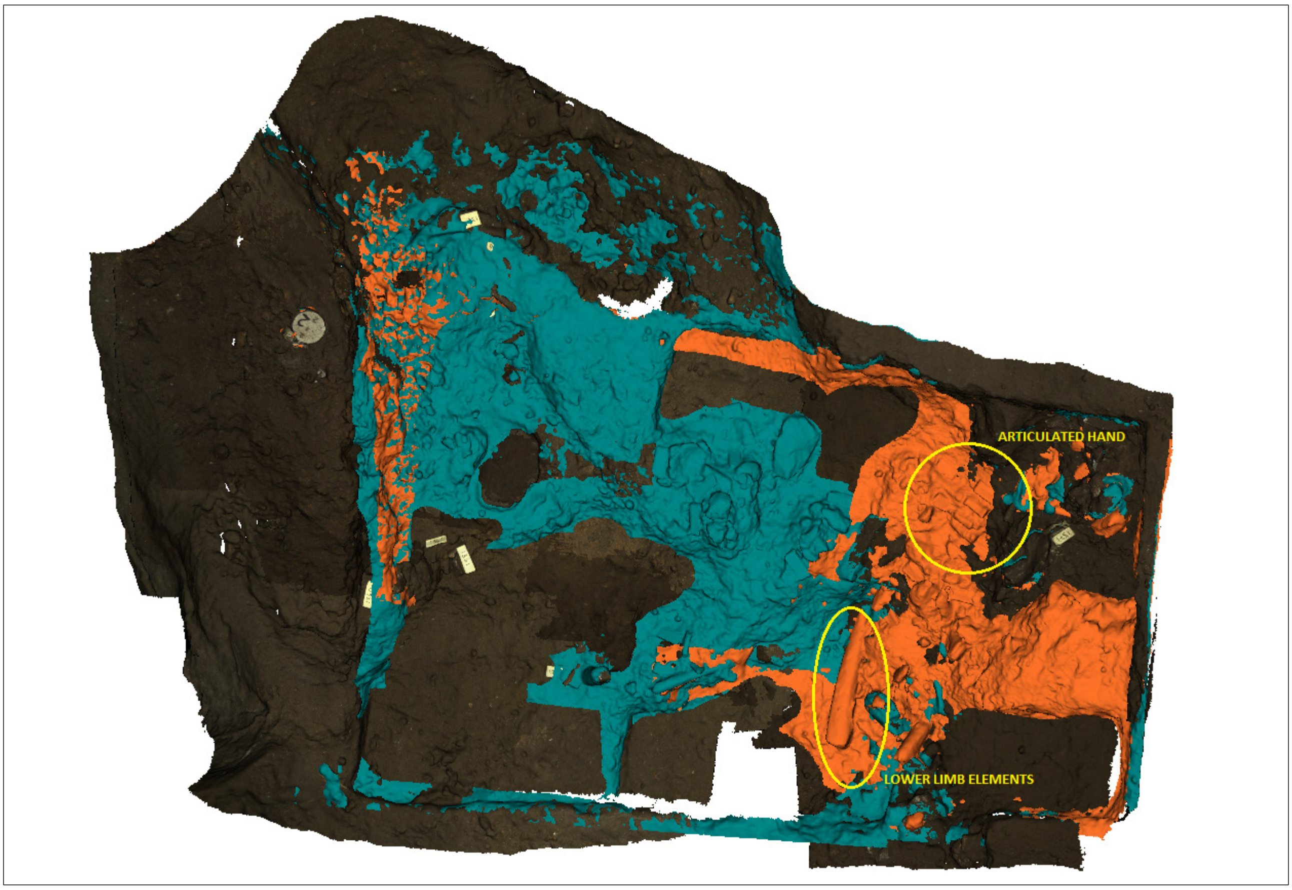
Photo: Patrick Randolph-Quinney

Figure 2: One of the mandibles of *Homo naledi* shortly after recovery from the Dinaledi Chamber. Note the excellent state of preservation and the cave sediment still adhering to the specimen. The network of cracks visible across the body and ramus of the mandible were caused by compression of the bone by sediment, and cycles of repeated wetting and drying of the bone within the chamber.

This brings us to the taphonomy of the assemblage (undertaken by Backwell, Musiba, Roberts and myself), the signatures of which represent a unique situation compared to other South African early hominin-bearing karstic cave sites for three main reasons. Firstly, the hominin remains are numerous and concentrated into a very small area (the main bone-bearing section of the chamber is only 4 m^2). Secondly, with the exception of a small number of recent intrusive rodent and bird bones, the only bones contained within the chamber are hominin. And thirdly, the assemblage is unique by what it *does not* evidence, in that it presents no evidence of perimortem breakage or trauma, no carnivore modifications (puncture marks, gnawing, etc.), no cut marks, no sub-aerial exposure or weathering, no evidence of water transportation, and no evidence of burning or charring. The assemblage was not carried in by carnivores, transported by water or mudflow, was not exposed to the elements on the land surface before being brought into the cave, and shows no evidence of the individuals having fallen into a death trap. What it does evidence is: (1) partial or near-complete articulation of anatomical elements which usually disassociate early in the decomposition sequence, such as hands and feet (Figure 3); (2) the presence of bone elements usually and easily lost to winnowing; (3) extensive modification of the surfaces of the bones by invertebrates (possibly snails or beetles); and (4) mineral staining and patterns of fracturing and cracking arising as a result of fluctuations in soil moisture content and changing water levels in the Dinaledi Chamber over time.

Taken together, this presents a formational process unlike any other fossil assemblage identified to date within the Cradle. Many of these sites represent natural death traps, or bone dens of carnivores, where hominin fossils are accumulated along with the remains of other animals that inhabited the landscape of South Africa. Dirks and his team attest that unlike other southern African cave sites in which hidden shafts and sinkholes provide access from the surface to the cave, there is no indication that a direct vertical passageway from the surface into the Dinaledi Chamber ever existed, with reconstructions of the cave environment indicating that reaching even the entrance of the Dinaledi Chamber would always have been a difficult obstacle course, particularly in the absence of artificial light. Dinaledi is unique in that only hominins are represented, and the skeletons show no evidence of transportation into the cave by the usual suspects (gravity, carnivores or flowing water) which represents a depositional scenario that deviates from all other hominin localities in the region.[14-16] Ruling out possibilities such as hominin occupation, water transport from the surface into the chamber, accumulations by predators such as hyaenas or leopards, or a mass fatality situation such as a death trap or fall, the authors are left to conclude that the fossil assemblage was produced by the hominins themselves – by the deliberate disposal of the dead. In this scenario, bodies of the individuals found in the cave would either been carried into the chamber or dropped through an entrance similar to, if not the same as, the one presently used to enter the Dinaledi Chamber. It is important to stress that nowhere in the paper is it suggested that deliberate burial is being practised. Disposal is not the same as burial, and may be purely functional as a mechanism of, say, predator avoidance, or may carry some deeper significance in terms of the primate grief response.

Given the primitive morphology and small brain size of this hominin, and the current lack of a date for the species, this interpretation raises a remarkable series of questions and 'what ifs'. After all, body disposal is a behaviour previously thought to be unique to humans or near-humans[17]



Source: Patrick Randolph-Quinney and Ashley Kruger

Figure 3: As access to the cave was limited (by the 180-mm diameter chute), conventional archaeological survey equipment could not be used during excavation. Instead, 3D high-resolution scans were used to record the position of each bone as uncovered. This image is made up of scans taken at four separate stages during the excavation process. The brown base layers show the limits of the excavation pit, whilst the blue and orange layers show the fossils being excavated on two separate days in March 2014. The articulated elements of a fossilised hand and part of a lower limb can clearly be seen. The scans were produced in the cave with an Artec Eva photogrammetric scanner and aligned and assembled in the lab using Artec Studio software.

– the possibility of a form of ritualised behaviour (in this context, 'ritualised' refers to a repeated or habituated pattern of behaviour rather than the notion of symbolic thought) in such a primitive-looking species such as *H. naledi* is bound to meet with resistance. Whilst body disposal by hominins is known from sites in the Middle and Upper Pleistocene such as Atapuerca Sima de los Huesos and many Neanderthal sites, these behaviours are ostensibly practised by hominins with a relatively large brain, which look relatively 'human', and which display other archaeological aspects of complex cognitive behaviours.[18-25] The situation is compounded by the fact that *H. naledi* remains, as yet, undated. If the species turns out to be Plio-Pleistocene in date, perhaps in the order of 2 Mya, it would represent the earliest appearance of *Homo* that is based on more than just isolated fragments of bone, together with the early adoption of ritualised behaviour coupled with primitive morphology. On the other hand, if *H. naledi* is young, less than 1 Mya, it would demonstrate that several different types of ancient humans coexisted at a similar time in southern Africa, including an especially small-brained form like *H. naledi*. Given its primitive skeletal adaptations, this might have profound implications for the production of the African archaeological record. It would also have profound implications for our understanding of the origins of complex behaviours previously thought to arise only with the origins of hominins biologically and archaeologically similar to our own species. Resolution will require firm dates, and I think willingness for archaeologists and palaeoanthropologists to throw out historical notions of cognitive evolution proceeding hand in hand with derived biological morphology.

As a footnote, the Rising Star Expedition is perhaps the first early hominin project to have been open to public scrutiny from excavation to analysis through the avenues of social media and the Internet. Throughout the excavation, the team shared expedition progress with a large public audience, from schoolchildren to fellow scientists, through blogs and video diaries, and the (often terrifying) immediacy of Twitter and Facebook; the strategy behind this approach was developed and implemented by Hawks in conjunction with Berger. This public and intellectual openness extended to the analysis phase, where the fossils were studied in a unique workshop in May 2014 funded by the South African DST/NRF, Wits University and National Geographic. More than 50 experienced scientists and early-career researchers came together over the space of a month to analyse the fossils and begin reporting on them. As a consequence, *H. naledi* progressed from first discovery to scientific dissemination in under 2 years, an almost unheard of feat in South Africa, where hominin fossils can remain undescribed and unpublished for decades. We eagerly await (presumably) forthcoming individual papers on the discrete anatomical regions of the *H. naledi* skeleton, and analyses of sexual dimorphism, growth and development, and the phylogenetic relationship between *H. naledi* and other hominin taxa.

*The papers '*Homo naledi*, a new species of the genus* Homo *from the Dinaledi Chamber, South Africa' and 'Geological and taphonomic context for the new hominin species* Homo naledi *from the Dinaledi Chamber, South Africa' can be freely accessed online at http://dx.doi.org/10.7554/eLife.09560 and http://dx.doi.org/10.7554/eLife.09561*

Patrick Randolph-Quinney is a forensic anthropologist and taphonomist and is one of the co-authors of the Dinaledi context paper, in which he contributed to the excavation and body recovery protocols, and analyses of weathering, skeletal damage patterns and other taphonomic processes in the deposition and formation of the assemblage.

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. 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. http://dx.doi.org/10.7554/eLife.09561
3. 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
4. Haile-Selassie Y, Gibert L, Melillo SM, Ryan TM, Alene M, Deino A, et al. New species from Ethiopia further expands Middle Pliocene hominin diversity. Nature. 2015;521(7553):483–488. http://dx.doi.org/10.1038/nature14448
5. Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, et al. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature. 2010;464(7290):894–897. http://dx.doi.org/10.1038/nature08976
6. Brown P, Sutikna T, Morwood MJ, Soejono RP, Jatmiko, Wayhu Saptomo E, et al. A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. Nature. 2004;431(7012):1055–1061. http://dx.doi.org/10.1038/nature02999
7. Asfaw B, White TD, Lovejoy CO, Latimer B, Simpson S, Suwa G. *Australopithecus ghari*: A new species of early hominid from Ethiopia. 1999;284:629–635.
8. Clarke RJ. On some new interpretations of Sterkfontein stratigraphy. S Afr J Sci. 1994;90:211–214.
9. Partridge TC. Re-appraisal of lithostratigraphy of Sterkfontein hominid site. Nature. 1978;275:282–287. http://dx.doi.org/10.1038/275282a0
10. Berger LR, Menter CG, Thackeray JF. The renewal of excavation activities at Kromdraai, South Africa. S Afr J Sci. 1994;90:209–210.
11. Kuman K, Field AS, Thackeray JF. Discovery of new artefacts at Kromdraai. S Afr J Sci. 1997;93:187–193.
12. Brain CK. Structure and stratigraphy of the Swartkrans Cave in the light of the new excavations. In: Brain CK, editor. Swartkrans. Transvaal Museum Monograph 8. Pretoria: Transvaal Museum; 1993. p. 23–34.
13. De Ruiter DJ. Revised faunal lists for Members 1–3 of Swartkrans, South Africa. Ann Trans Mus. 2003;40:29–41.
14. Brain CK. A taphonomic overview of the Swartkrans fossil assemblages. In: Brain CK, editor. Swartkrans. Transvaal Museum Monograph 8. Pretoria: Transvaal Museum; 1993. p. 257–264.
15. Watson V. Composition of the Swartkrans bone accumulations, in terms of skeletal parts and animals represented. In: Brain CK, editor. Swartkrans. Transvaal Museum Monograph 8. Pretoria: Transvaal Museum; 1993. p. 35–74.
16. Pickering TR, Clarke RJ, Moggi-Cecchi J. Role of carnivores in the accumulation of the Sterkfontein Member 4 hominid assemblage: A taphonomic reassessment of the complete hominid fossil sample (1936–1999). Am J Phys Anthropol. 2004;125(1):1–15. http://dx.doi.org/10.1002/ajpa.10278
17. Parker Pearson M. The archaeology of death and burial. Thrupp: Sutton Publishing; 1999.
18. Gargett RH. A response to Hovers, Kimbel and Rak's argument for the purposeful burial of Amud 7. J Hum Evol. 2000;39:261–266. http://dx.doi.org/10.1006/jhev.2000.0419
19. Hovers E, Kimbel WH, Rak Y. The Amud 7 skeleton – Still a burial. A response to Gargett. J Hum Evol. 2000;39:253–260. http://dx.doi.org/10.1006/jhev.1999.0406
20. Suzuki H, Takai F, editors. The Amud man and his cave site. Tokyo: Keigaku Publishing Co. Ltd; 1970.
21. Gargett RH. Grave shortcomings: The evidence for Neanderthal burial. Curr Anthropol. 1989;30(2):157–190. http://dx.doi.org/10.1086/203725
22. Tattersall I. Neanderthal burials: Excavations of the Dederryeh Cave, Afrin, Syria. Am J Phys Anthropol. 2005;128(1):239–240. http://dx.doi.org/10.1002/ajpa.20118
23. Minugh-Purvis N, Radovcic J, Smith FH. Krapina 1: A juvenile Neandertal from the early late Pleistocene of Croatia. Am J Phys Anthropol. 2000;111(3):393–424. http://dx.doi.org/10.1002/(SICI)1096-8644(200003)111:3<393::AID-AJPA7>3.0.CO;2-U
24. Bocquet-Appel J, Arsuaga J. Age distributions of hominid samples at Atapuerca (SH) and Krapina could indicate accumulation by catastrophe. J Archaeol Sci. 1999;26(3):327–338. http://dx.doi.org/10.1006/jasc.1998.0370
25. Carbonell E, Bermudez de Castro JM, Arsuaga JL, Diez JC, Rosas A, Cuenca-Bescos G, et al. Lower Pleistocene hominids and artifacts from Atapuerca-TD6 (Spain). Science. 1995;269:826–829. http://dx.doi.org/10.1126/science.7638598

South African scholars make Thomson Reuters 'Highly Cited Researchers 2014'

AUTHOR:
Makia L. Diko[1]

AFFILIATION:
[1]School of Physical and Mineral Sciences, University of Limpopo, Polokwane, South Africa

CORRESPONDENCE TO:
Makia Diko

EMAIL:
dikom73@gmail.com

POSTAL ADDRESS:
School of Physical and Mineral Sciences, University of Limpopo, Private Bag X1106, Sovenga 0727, South Africa

KEYWORDS:
research quality; impact; relevance; policy implications; South Africa

HOW TO CITE:
Diko ML. South African scholars make Thomson Reuters 'Highly Cited Researchers 2014'. S Afr J Sci. 2015;111(9/10), Art. #a0121, 4 pages. http://dx.doi.org/10.17159/sajs.2015/a0121

Evaluating individual and institutional scientific performance is an essential component of research assessment, and outcomes of such evaluations play a key role in institutional research strategies, including funding schemes, staffing and international recognition.[1] In a recent communication by Slippers et al.[2] entitled 'Global trends and opportunities for development of African research universities', featured in the January/February 2015 edition of the *South African Journal of Science*, they observed that:

> *The demand for demonstrating the relevance and impact of research at higher education institutions is increasing at the same time, particularly in developing nations in which funders are becoming impatient with a perceived lack of results.*

In light of the above assertion, one may be forced to ask what constitutes relevant research or how is research impact assessed? Providing a succinct response to these questions seems to be a tall order. Indeed, the quest for comprehensive criteria to assess scholarly outputs such as research publications continues to dominate academic discourse across the globe.[1,3-6] From the affluent 'North' to the developing 'South', educational systems of several countries[1,3-6], including South Africa[7-10] continue to grapple with the issues around quality of research output. Efforts at evaluating productivity, scientific impact and research quality are compounded by the seeming lack of a general consensus on an acceptable metric system by all related stakeholders (e.g. governments, academic institutions and publication agencies) within and across educational systems. In addition, variability between metric systems (inclusive of pros and cons)[6-8] and continuous emergence of new indices, coupled with inherent differences in socioeconomic and resource potentials between the 'North' and 'South', makes the conceptualisation of a globally acceptable definition of 'relevance' and 'impact' of research even more complex. Despite the above constraints, academic institutions from the 'South' (especially African universities) in search of best practice in research, are encouraged to benchmark with trends that they obtain from the 'North'.[11-13]

In South Africa, research publications are one of the instruments used to monitor the performance of institutions of higher learning. As an incentive towards increasing research output, the Department of Higher Education and Training (DHET) – through the 'Policy for Measurement of Research Outputs of Public Higher Education Institutions (2003)' – awards subsidy to higher education institutions whose members publish in an approved list of South African journals. Amongst this list is a significant cross-section of journals included on the Thomson Reuters Web of Science index (formerly known as ISI). In December 2014, the Web of Science published a list of over 3000 researchers from across the world with the most cited publications over an 11-year period (2002–2012).[14] This list featured ten researchers with affiliation to a South African university. The aim of this article is to celebrate the South African scholars whom, together with fellow listees, have been deemed the 'most influential scientific minds' by Web of Science[14], on the basis of peer recognition through citations. This study also elucidates the global distribution of these researchers as well as South Africa's performance within the global context. It concludes by interrogating the implications of this 'South African achievement' on the National Research Foundation (NRF) researcher rating system and current DHET publication subsidy policy.

The selection procedure for highly cited researchers

The list of highly cited researchers was drawn from highly cited articles and reviews in science and social sciences journals indexed in the Web of Science Core Collection during the 11-year period 2002–2012. Highly cited papers were defined as those that rank in the top 1% by citations per field based on data derived from Essential Sciences Indicators® (ESI). (For more detailed information on the analytical and selection procedures, consult the official website: www.highlycited.com)

With respect to the current study, data analysis and interpretation were based on the total number of individual researchers (F1) as well as those listed in more than one ESI category or country (F2).[15] In terms of percentage distribution of researchers per country, weightings have been computed based on F1 (3073) and F2 (3215).

Synopsis of findings

The 3073 individually cited researchers are drawn from 47 countries. A distribution of these researchers by country is presented in Table 1. Results reveal that the bulk of influential researchers are based in institutions from the United States of America (USA) with a total of 1667 researchers, followed by the United Kingdom (UK) with 360 and Germany (271) as distant second and third respectively. Saudi Arabia (174), China (161), France (140) and Japan (102) are close fourth, fifth, sixth and seventh respectively, while Canada (89), Netherlands (82) and Switzerland (68) complete the top ten countdown. South Africa is the sole representative from Africa, ranking 25th out of 47 countries. Based on F2 criteria, North America contributed 48.8% of the most influential researchers, followed by Europe (46.89%), Asia (15.17%), Oceania (1.11%), Africa (0.3%) and 0.25% for South America. In terms of performance in relation to the BRICS countries (Brazil, Russia, India, China and South Africa), South Africa with 11 highly cited researchers ranked third, closely behind India (12) and China (161) whereas, Russia and Brazil had 7 and 5, respectively.

Table 1: Ranking of Thomson Reuters Highly Cited Researchers 2014 by country, frequency and percentage distribution

Rank	Country	No. of researchers†	% Distribution based on F1	% Distribution based on F2
1	USA	1667	54.25	46.33
2	UK	360	11.71	10.01
3	Germany	271	8.82	7.53
4	Saudi Arabia	174	5.66	4.84
5	China	161	5.24	4.47
6	France	140	4.56	3.89
7	Japan	102	3.32	2.83
8	Canada	89	2.90	2.47
9	Netherlands	82	2.67	2.28
10	Switzerland	68	2.21	1.89
11	Italy	53	1.72	1.47
12	Spain	49	1.59	1.36
13	Australia	35	1.14	0.97
14	Belgium	34	1.11	0.94
15	Denmark	33	1.07	0.92
16	Sweden	32	1.04	0.89
17	South Korea	22	0.72	0.61
18	Austria	21	0.68	0.58
19	Singapore	18	0.59	0.50
20	Finland	17	0.55	0.47
21	Ireland	14	0.46	0.39
22	Israel	13	0.42	0.36
23	Iran	13	0.42	0.36
24	India	12	0.39	0.33
25	South Africa	11	0.36	0.31
26	Taiwan	11	0.36	0.31
27	Iceland	11	0.36	0.31
28	Turkey	10	0.33	0.28
29	Norway	9	0.29	0.25
30	Greece	7	0.23	0.19
31	Russia	7	0.23	0.19
32	New Zealand	5	0.16	0.14
33	Brazil	5	0.16	0.14
34	Poland	4	0.13	0.11
35	Malaysia	4	0.13	0.11
36	Jordan	4	0.13	0.11
37	Indonesia	4	0.13	0.11
38	Portugal	3	0.10	0.08
39	Hungary	3	0.10	0.08
40	Czech	2	0.07	0.06
41	Argentina	2	0.07	0.06
42	Serbia	2	0.07	0.06
43	Chile	1	0.03	0.03
44	Lithuania	1	0.03	0.03
45	UAE	1	0.03	0.03
46	Slovakia	1	0.03	0.03
47	Colombia	1	0.03	0.03

F1, the actual 3073 individual researchers listed; F2, total number of researchers from all 47 countries (inclusive of those listed in more than one country or ESI category).

†Figures generated in accordance with F2.

South Africa's most influential scientific minds

In terms of individual researchers, there are actually 10 researchers with South African affiliation (one of them being listed in two ESI Categories). They comprise six South African based scholars and four others with secondary affiliation to a South African university (Table 2). The following section briefly presents the short list of six, comprising three listees in the Environment/Ecology category, two in Social Sciences – General, and one in the Biology and Biochemistry category. More elaborate bibliographical information is hosted by their respective institutional websites.

David M. Richardson is an A1 NRF-rated researcher[16] and leading international scholar in the field of invasion biology. He is currently Director of the Department of Science and Technology (DST)/NRF Centre of Excellence in Invasion Biology at Stellenbosch University. He has published over 307 peer-reviewed articles in scientific journals and books, including chapters in 40 edited books. According to the Web of Science, his works have been cited 11 230 times.

Guy Midgley is a B1 NRF-rated researcher[16] and internationally acknowledged expert in the field of biodiversity and global change science. He has published more than 160 articles and papers, of which four have been in the top academic journals *Nature*, *Science,* and *Nature Climate Change.* His academic works have been cited more than 12 000 times.

William J. Bond is a B3 NRF-rated researcher[16] and Emeritus Professor in the Department of Botany at the University of Cape Town. His niche areas includes: processes influencing vegetation change in the past and present, including fire, vertebrate herbivory, climate extremes, atmospheric (CO_2) and habitat fragmentation; plant-animal interactions; plant form and function; and biomes.

Lyn Wadley is an A2 NRF-rated researcher[16] and Honorary Professor of Archaeology, affiliated jointly with the Archaeology Department (University of the Witwatersrand) and the Institute for Human Evolution. She is also the Director of Ancient Cognition and Culture in the Africa Research Unit at the University of the Witwatersrand. The group's research focuses on issues of cognition and culture in the Middle Stone Age of southern Africa.

Rachel K. Jewkes is an A2 NRF-rated researcher[16] and Director of the Medical Research Council's Gender and Health Research Unit. She is Honorary Professor in the Faculty of Health Sciences, School of Public Health (University of the Witwatersrand). Her research focuses on the interface of gender inequity and gender-based violence and health, particularly HIV. She has authored well over 100 articles in peer-reviewed journals and over 20 book chapters.

Nicola J. Mulder is a B3 NRF-rated researcher[16] and Head of the Computational Biology Group at the University of Cape Town. Her main research interests lie in the areas of infectious diseases and human genetics, with particular emphasis on the molecular biology of the pathogen *Mycobacterium tuberculosis*. Under her leadership, the Computational Biology Group generated over 30 publications in 2011.

Implications for South Africa

From a South African perspective, the implications of this recent achievement on NRF ratings and DHET publication subsidy policy are interrogated. Beyond right or wrong, this article seeks to generate questions that may encourage and sustain more rigorous debate around these topical issues. It is anticipated that this paper may serve as an impetus towards more in-depth appraisal by all related stakeholders.

Implications on National Research Foundation rating

The six South African based researchers have NRF ratings between A1 and B3.[16] The NRF rating system is an international benchmarking process through which individuals that exemplify the highest standards of research, as well as those demonstrating strong potential as researchers, are identified by an extensive network of South African and international peer reviewers. Ratings are based on the quality and impact of recent research outputs (over an 8-year period).[16] Taking into consideration the

Table 2: Highly cited South African researchers for the period 2002–2012

Names	Essential Science Indicators® (ESI) category	Primary affiliation	Additional affiliation/s	NRF rating
Bruce D. Walker[†]	Microbiology/ Immunology	Howard Hughes Medical Institute, USA	Ragon Institute of MGH, MIT and Harvard, USA; University of KwaZulu-Natal, South Africa	N/A
Christopher S. Henshilwood	Social Sciences, General	University of Bergen, Norway	University of the Witwatersrand, South Africa	N/A
David M. Richardson	Environment/Ecology	Stellenbosch University, South Africa		A1
Guy F. Midgley[††]	Environment/Ecology	South African National Biodiversity Institute	University of KwaZulu-Natal, South Africa	B1
Lyn Wadley	Social Sciences, General	University of the Witwatersrand, South Africa		A2
Matthias Egger	Clinical Medicine	University of Bern, Switzerland	University of Cape Town, South Africa	N/A
Nicola J. Mulder	Biology & Biochemistry	University of Cape Town, South Africa		B3
Rachel K. Jewkes	Social Sciences, General	University of the Witwatersrand, South Africa		A2
William J. Bond	Environment/Ecology	University of Cape Town, South Africa		B3
Yves van de Peer	Plant & Animal Science	Ghent University, Belgium	Flanders Institute for Biotechnology, Belgium; University of Pretoria, South Africa	N/A

Source: Thomson Reuters Highly Cited Researchers, December 2014.[14]

[†] *Bruce D. Walker was cited under two ESI categories, namely Immunology and Microbiology.*

[††] *The above data have been presented as is, pending updates by the Web of Science (Guy F. Midgley's current affiliation is Department of Botany and Zoology at Stellenbosch University).*

recent achievement of the six aforementioned researchers, the obvious questions would be:

- Are their NRF ratings a true reflection of the quality and impact of their research outputs?
- Is it possible for a researcher in the top 1% of their field globally to be classified as B rated?
- How consistent is the rating system in terms of meeting its objective – to identify, encourage and celebrate research excellence through quality and impact of research output?
- How significantly different are the phrases: 'all reviewers', 'overwhelming majority of reviewers' and 'most reviewers', as applied in the description of NRF categories?
- Alternatively, does the NRF have a comparably more rigorous and reliable evaluation scheme that needs to be projected and adopted by the rest of the world?

Implications on the publication subsidy policy

What really matters: quality or quantity? Publication subsidy is an invaluable source of institutional support from government. In addition, the financial benefits to individual researchers cannot be overemphasised. However, beyond these direct benefits, several studies have suggested negative impacts of the current DHET publication subsidy policy on the quality of research output[7-10] (notably with regard to journal articles). Despite inclusion of top-ranking high impact factor journals as part of its accredited list, researchers are not compelled to publish in them. This is further exacerbated by the option of low impact journals, often characterised by a comparatively less rigorous review process and shorter turnaround period (i.e. from initial submission to publication).

In terms of the current DHET remuneration policy, emphasis is placed on units of publication – somewhat synonymous to quantity. For example, a journal article published by a single researcher affiliated to a South African higher educational institution is worth one unit. Irrespective of the type of journal (high or low impact, local or international), the researcher is entitled to one unit and the associated financial gains. Where two or more authors with affiliations to the same or different South African institutions are involved, the single unit is shared among them. Similarly, where two South African based researchers co-publish with two other researchers without affiliation to a South African university, the former are only entitled to 0.5 units (i.e. 0.25 each). Based on the above provision, Woodiwiss[10] argued that international collaboration is seriously discouraged. Furthermore, Jeenah and Pouris[8] posited that the quest for financial gain tends to encourage quantity at the expense of quality. According to Valerie Mizrahi[17], in a lecture on 'The practice of research and publication in the South African context' during the University of Cape Town's Library Research Week (12 May 2014), the current DHET policy 'penalises collaboration' and is 'open to abuse as a numbers game'. In light of the above challenges, the sole question one may pose is: Is it not time to introduce another variable, for example the Quality Index, to the calculation of research units and corresponding financial reward? Such a variable may take into consideration key quality control elements such as the journal impact factor, journal ranking or number of article non-self-citations over the conventional 2-year cycle prior to the release of funds by DHET.

In a nutshell, as we ponder on these issues and more, let us continue to reflect on the sentiments of Salmi[13], as captured in his book, *The challenge of developing world class universities:*

> *…institutions will inevitably, from here on out, be increasingly subject to comparisons and rankings, and those deemed to be the best in these rankings of research universities will continue to be considered the best in the world.*

Acknowledgement

Data on Highly Cited Researchers 2014 are freely available on the website www.highlycited.com. Excerpts of bibliographic data presented herein are courtesy of the researcher's institutional websites.

References

1. Sahel J. Quality versus quantity: Assessing individual research performance. Sci Transl Med. 2011;3(84cm13):1–4.
2. Slippers B, Vogel C, Fioramonti L. Global trends and opportunities for development of African research universities. S Afr J Sci. 2015;111(1/2), Art. #a0093, 4 pages. http://dx.doi.org/10.17159/sajs.2015/a0093
3. Paul JR. Measuring research quality: The United Kingdom government's research assessment exercise. Eur J Inf Syst. 2008;17:324–329. http://dx.doi.org/10.1057/ejis.2008.31

4. Schreiber M. Twenty Hirsch index variants and other indicators giving more or less preference to highly cited papers. Ann Phys. 2010;522:536–554. http://dx.doi.org/10.1002/andp.201000046

5. Derrick GE, Haynes A, Chapman S, Hall WD. The Association between four citation metrics and peer rankings of research influence of Australian researchers in six fields of public health. PLoS ONE. 2011;6(4):e18521.

6. Dodson MV, Duarte M, Dias LA. SP-index: The measure of the scientific production of researchers. Biochem Biophys Res Commun. 2012;425:701–702. http://dx.doi.org/10.1016/j.bbrc.2012.07.161

7. Jacobs D. Analysis of scientific research in selected institutions in South Africa: A bibliometric study. SA Jnl Libs Info Sci. 2006;72(1):72–77.

8. Jeenah M, Pouris A. South African research in the context of Africa and globally. S Afr J Sci. 2008;104(9/10):351–354.

9. Schulze S. Academic research at a South African higher education institution: Quality issues. S Afr J Higher Educ. 2008;22(3):629–643.

10. Woodiwiss AJ. Publication subsidies: Challenges and dilemmas facing South African researchers. Cardiovasc J Afr. 2012;23(8):421–427.

11. Teferra D, Altbach PG. African higher education: Challenges for the 21st century. High Educ. 2004;47:21–50. http://dx.doi.org/10.1023/B:HIGH.0000009822.49980.30

12. Sawyerr A. African universities and the challenge of research capacity development. J Higher Educ Afr. 2004;2(1):211–240.

13. Salmi J. Directions in development-human development: The challenge of establishing world-class universities. Washington DC: The World Bank; 2009.

14. Thomson Reuters. Welcome to Highly Cited Researchers [webpage on the Internet]. c2014 [cited 2015 Mar 25]. Available from: http://highlycited.com

15. Bornmann L, Bauer J. Which of the world's institutions employ the most highly cited researchers? An analysis of the data from highlycited.com. J Assoc Inf Sci Technol. 2015 January 08. http://dx.doi.org/10.1002/asi.23396

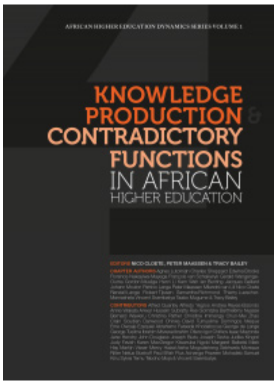
16. National Research Foundation. NRF Rating. National Research Foundation [webpage on the Internet]. c2015 [cited 2015 Mar 25]. Available from: http://www.nrf.ac.za/rating

17. Mizrahi V. The practice of research and publication in the South African context. Lecture by Prof. Valerie Mizrahi, University of Cape Town, 2014 May 12. Available from: http://researchcommonsblog.uct.ac.za/wp-content/uploads/2014/05/Practice-research-publication-SA-context_12-05-14_mizrahi.pdf

An assessment of eight African universities: Contradictory functions, knowledge production and pacts

BOOK TITLE:
Knowledge production contradictory functions in African higher education

BOOK COVER:


AFRICAN HIGHER EDUCATION DYNAMICS SERIES VOLUME 1
KNOWLEDGE PRODUCTION & CONTRADICTORY FUNCTIONS IN AFRICAN HIGHER EDUCATION
EDITORS NICO CLOETE, PETER MAASSEN & TRACY BAILEY

EDITORS:
Nico Cloete, Peter Maassen and Tracy Bailey

ISBN:
9781920677855

PUBLISHER:
African Minds, Somerset West; ZAR200

PUBLISHED:
2015

REVIEW TITLE:
An assessment of eight African universities: Contradictory functions, knowledge production and pacts

REVIEWER:
Fred M. Hayward

EMAIL:
haywardfred@hotmail.com

AFFILIATION:
Senior Higher Education Advisor, University of Massachusetts, Amherst, Massachusetts, USA

POSTAL ADDRESS:
3628 Van Ness St. NW,
Washington DC 20008, USA

HOW TO CITE:
Hayward FM. An assessment of eight African universities: Contradictory functions, knowledge production and pacts. S Afr J Sci. 2015;111(9/10), Art. #a0120, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0120

This study of higher education in eight sub-Saharan nations in Africa was built around a research project of the Higher Education Research and Advocacy Network in Africa (HERANA). It was initiated by the Centre for Higher Education Transformation (CHET) in 2007 and includes data on these institutions from two different surveys, the second completed in 2013/2014. The study shows a great deal of thought and care in its preparation, a difficult task in its undertaking, data gathering, and presentation, and is an impressive and useful study which will be of interest to most scholars of Africa. It is an important addition to the literature on higher education in Africa. The effort to assess the relationship between higher education and development, economics and democracy in Africa is timely and well-conceived, and provides a wealth of very useful information on higher education in Africa. There has been evidence of this research by Cloete et al. over the last few years and it was with great anticipation that I read the final result of their efforts. It was well worth the wait. The eight higher education institutions that are the focus of this research are in Botswana, Ghana, Kenya, Mauritius, Mozambique, South Africa, Tanzania and Uganda and provide an excellent representation of the status of higher education in sub-Saharan Africa.

This book is an attempt to provide a data driven analysis using performance indicators of eight universities picked as comparable ‘flagship’ institutions to assess their success as knowledge-producing and research-intensive institutions in the age of the knowledge economy. The study involved an assessment of the core functions of the university and the success of these African universities in those areas. The authors see knowledge production and technological innovation as the most important productive forces of higher education institutions (p.5). All this is in a context of years of declining funding for African higher education, tremendous growth in enrolments, staff shortages, and a variety of economic, political and social problems within the region as a whole. The authors call for the ‘revitalisation’ of African higher education as a result of their study.

The authors assess the eight institutions based on a set of eight measurable goals and targets: enrolments in science and technology, strong master’s and doctoral enrolments but with a majority of undergraduates, a high proportion of permanent academic staff in senior ranks, well-qualified senior staff, low student to academics ratios, high outputs of graduates in SET fields, high outputs of master’s and doctoral degrees, and high levels of new knowledge production (p.39). The institutions examined were the Universities of Botswana, Cape Town, Dar es Salaam, Eduardo Mondlane, Ghana, Makerere, Mauritius and Nairobi. Only the University of Cape Town meets all the targets; the Universities of Mauritius, Dar es Salaam and Makerere come close. At the same time the authors explore, as the book title suggests, the contradictory functions and pressures on these institutions – on the one hand, the pressure to produce knowledge useful to national development, and on the other hand, the contradictory pressures of fundraising, public service, outside jobs, growing enrolments and thus pressure on teaching loads, lack of research funding, the growth of fee-paying students, etc. That in and of itself is an especially interesting and telling discussion which plays out throughout the book and demonstrates how difficult it is to move in the direction of the targets set out for flagship institutions by the project and to establish universities which truly make a contribution to national development.

The first chapter by Nico Cloete and Peter Maassen on the ‘Roles of Universities and the African Context’ provides an excellent introduction to the study, drawing heavily on Manuel Castells, who was very influential in the project, as well as Clark Kerr and others who place this work in the proper higher education intellectual context. They spell out four major functions of higher education: production of values and social legitimisation; selection of elites; training of the labour force; and production of scientific knowledge. They then trace the development and complexity of these functions, historically placing Africa in context – one which has not been a picture of great success. The next chapters lay out data collected for the project in each of the project institutions, where possible, with most chapters focusing on two or three of the institutions. What is impressive, as someone who has done multi-country research can attest, is the care taken to make the data comparable, and the straightforwardness of the authors in noting the problems, inconsistencies and differences in the various comparisons they present.

Chapter two, on research universities, makes the case for the importance of research universities – but notes that for the most part the research universities examined are not strengthening their knowledge-generating capacity and are failing to make a substantive contribution to new knowledge generation. Nonetheless, they note that the universities are virtually the only producers of knowledge in Africa. Of the eight universities studied, only the University of Cape Town attains the requirements of a ‘high quality research and scholarship delivering knowledge producer’, with Makerere University somewhat behind (p.29). The authors go on to note that higher education remains the best, and in most cases, the only institution capable of knowledge production in Africa, in spite of its many weaknesses, and emphasise the importance of efforts to revitalise higher education, especially the research and knowledge production functions.

Chapter three, on the flagship universities, presents what seems to me to be too glowing a picture of these institutions. Much of the information presents a sorry picture of the situation of research at these universities, but the data gathered is very important in placing research and knowledge production in its current unsatisfactory context. The case of Makerere, with its low level of research, is a telling one – the discussion is of a variety of factors that hinder research, plus the negative consequences of the admission of self-funded students on research prospects. The discussion is an excellent demonstration of the contractions noted in the title of the book – both in statements of intent to carry out research and the lack of actual support and facilitation thereof.

In Chapter six, the authors stress '…the potential contribution of academic research to African societies cannot be overstated' (p.110). I heartily agree with them. It is encouraging that both research and knowledge production are so strongly emphasised and are not written off as expectations for other parts of the world, but not Africa. The authors go on to point out that almost all sub-Saharan African universities are struggling to improve their academic research productivity with little success.

Chapter seven, on 'Academic Incentives for Knowledge Production', is a particularly poignant piece contrasting the incentives in South Africa for publications with the situations in Mozambique, Kenya and, to some extent, the rest of Africa. It is an excellent description of the commercialisation of the university and the ways in which a combination of low salaries, lack of support, and donor foci have 'undermined the possibility of establishing a research culture'. The authors note that other than in South Africa there is little by way of financial incentive for research – that the major incentive is the drive for knowledge production (p.129). This section is excellent.

The findings emphasise the lack of connection between research institutions, even those involved in this and other projects. Indeed, they point out that the stronger connections are with research partners abroad. This is not too surprising given the disparities of research support, but it is disheartening given the years of efforts to promote research linkages between African institutions by foundations, the World Bank and other funders. As they point out, most of the linkages are individual and not institutional. Most discouragingly, they point out that: 'The university in the guise of service provider to the community, does little more than import and transfer existing knowledge instead of creating new knowledge, will at best make a marginal, short-term contribution to development' (p.205). Overall, the study shows that the universities are largely not involved in creating new knowledge. All eight institutions are involved in a variety of activities and talk about research. However, the reality is that most are not involved in knowledge production. New pressures, limited funding, changes in the workplace, new technologies, dilution of both the academic culture and common purposes of higher education in recent years, have weakened academic commitments to institutions.

The two student surveys carried out at Makerere University and the University of Cape Town are very interesting and useful. For reasons that are clearly described, the surveys at the two institutions turned out to be less comparative than hoped. Nonetheless, the results are fascinating. The research and write-up here are excellent, upholding high standards of research and resisting the temptation to treat the data as more than it is. They conclude that: 'Key aspects of student experience have profound impacts on raising levels of citizen competence' (p.256). They then explore how universities develop citizen competencies though student experiences. This work is very suggestive of the potential of universities in nation-building and a demonstration of what can be done. But to do this, the universities need adequate funding, equipment, and well-trained and committed faculty to take full advantage of what can be done. This chapter alone makes the book worth reading; it reflects excellent work. The work of HERANA on how key aspects of student experience have an impact on citizenship competence, civic education and democratisation is among the best parts of this study.

Some of the chapters, such as the one on 'Governance of Higher Education Councils and Commissions' are somewhat pedestrian, providing what one would expect about their purposes and functions, but then that is important information too. And for those not familiar with them, it will be useful reading.

The final chapter emphasises the need for African higher education to catch up and invest in knowledge production for teaching and development. The authors emphasise the need for research-intensive universities – at least one in each country – pointing out that only three universities focus their plans on economic development – Makerere, Botswana and Mauritius. They also emphasise the critical importance of differentiation, for focused work and to limit costs, as well as for system level recognition of the need to develop research generally.

Overall this is an excellent publication, one that most people will want to read. It shows why the knowledge production functions were not developed historically in sub-Saharan Africa, and lays out what needs to be done to get them moving, with data based on evidence. It presents especially rich and very relevant material which I have found extremely useful, as will others. As someone who has done a great deal of quantitative analysis, including survey research, and has worked on the international collection of university data, I know how very difficult it is to collect accurate and useful data of this kind. The HERANA group and CHET are to be congratulated on the care and time they took in preparing this study, gathering and checking the data, and presenting it in this book. The study breaks new ground, is a major contribution to our understanding of higher education in sub-Saharan Africa and will significantly reward the reader's attention.

BOOK TITLE:
Insects of cultivated plants and natural pastures in southern Africa

BOOK COVER:

Insects of Cultivated Plants and Natural Pastures in Southern Africa
Editors
Gerhard L. Prinsloo • Vivienne M. Uys

EDITORS:
Gerhard L. Prinsloo and Vivienne M. Uys

ISBN:
9780620608411 (hardcover)

PUBLISHER:
Entomological Society of Southern Africa, ZAR600

PUBLISHED:
2015

REVIEW TITLE:
For the love of insects

REVIEWER:
Michael J. Samways

EMAIL:
samways@sun.ac.za

AFFILIATION:
Department of Conservation Ecology and Entomology, Stellenbosch University, Stellenbosch, South Africa

POSTAL ADDRESS:
Department of Conservation Ecology and Entomology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

HOW TO CITE:
Samways MJ. For the love of insects. S Afr J Sci. 2015;111(9/10), Art. #a0118, 1 page. http://dx.doi.org/10.17159/sajs.2015/a0118

For the love of insects

Insects of Cultivated Plants and Natural Pastures in Southern Africa really is a superb book. It radiates a love of insects and dedication to detail. The whole presentation is highly professional and user-friendly, and the editors are to be congratulated on their setting of such high standards.

As Cliff Moran points out in his Foreword, every time we purchase a plant product we are in part paying a bounty to offset the control of pests on those products. Financially, this situation is a real concern, with 15% of total plant food production lost to pests worldwide before the plant product is harvested, and another 10% lost during transport and storage after harvest. The situation is magnified when crops are exported, as is very much the case for many southern African plant products. Importers have strict tolerance limits with regard to the potential entry of pests into their countries. In the case of some of our indigenous pests, there is zero tolerance, which places high demands on producers and exporters to be rid of pests.

This huge volume of 786 pages is authored by 38 integrated pest management specialists who focused on their crops or domains of interest. The book comprehensively covers all the significant pest insects across southern Africa – which is a massive achievement. A total of 416 pest species are discussed, with a focus on their origin and distribution, identification, host plants, damage, and life history, and with brief notes on natural enemies and on management. Another 277 species are listed as minor pests. This book is made even more appealing by the numerous excellent photographs of the main subject insects and the damage that they cause.

The book is arranged into 12 sections: (1) vegetables, (2) cereals and sugarcane, (3) oil and protein-rich seed crops, (4) pastures and fodder crops, (5) turf grasses, (6) miscellaneous field crops, (7) deciduous fruit and nut trees and olive, (8) grapevine and berries, (9) citrus, (10) subtropical fruit and nut trees and coffee and tea, (11) plantation trees and, finally, (12) ornamental plants. Some of these sections are divided into individual crops and other subsections. Each section or subsection has a useful introductory/summary table listing the significant pests, with their scientific names, plant host (including the plant parts affected), and extent of descriptive treatment in the book. A glossary is also provided.

It behoves a reviewer to also criticise, so as to advance the field. There are some idiosyncrasies with this volume. Firstly, it should really have been called *Insect Pests of…* rather than *Insects of …* as it focuses on pest insects, and the treatment of natural enemies of the pests is very superficial. It is strange that mites are not included as they can often be as significant as insects in causing crop losses, and are usually included side by side with insects in most integrated pest management programmes. Perhaps the volume was large enough anyway, but at least a few major species could have been included. One would hope that a companion volume on mite pests – which meets the standards set by this book – will one day become available.

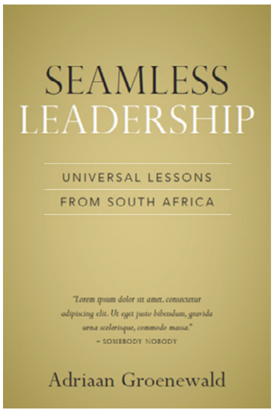
Arguably, the weakest part of the book is the referencing, with a very strong bias in some sections to the authors' own works and almost no recognition of several important and fundamental works of others in the same field. This is especially so in the case of citrus. An electronic link to an extensive literature base would have been appropriate in this age of such linkages, and perhaps can still be included, not just of the pests but also of the natural enemies.

These criticisms aside, this book is a masterful piece of work, and Gerhard Prinsloo and Vivienne Uys have succeeded without doubt in producing a fine reference work that will remain the standard for many years to come.

Mirror, mirror: The science

BOOK TITLE:
Seamless leadership: Universal lessons from South Africa

BOOK COVER:


SEAMLESS LEADERSHIP
UNIVERSAL LESSONS
FROM SOUTH AFRICA
"Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut eget justo bibendum, gravida urna scelerisque, commodo massa."
– SOMEBODY NOBODY
Adriaan Groenewald

AUTHOR:
Adriaan Groenewald

ISBN:
9781868426461 (softcover);
9781868426478 (ebook)

PUBLISHER:
Jonathan Ball, Cape Town;
ZAR250 (softcover), USD18 (ebook)

PUBLISHED:
2015

REVIEW TITLE:
Mirror, mirror: The science

REVIEWER:
Iain Edwards

EMAIL:
edwardsI@gibs.co.za

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

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

HOW TO CITE:
Edwards I. Mirror, mirror: The science. S Afr J Sci. 2015;111(9/10), Art. #a0119, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0119

Globally, it's a time of asymmetry, instability, volatility and also terror. In this age of uncertainty, much public debate and scholarly enquiry focuses on serious informed reflection on this all too complex issue. There are big books on strategy, international relations, political economy, and ultimately, power. Titles are indicative: *Capitalism in the 21st Century*, *Governance and the World Order in the 21st Century*, *The End of Certainty*, *The End of Power*, *The Locust and the Bee: Predators and Creators in Capitalism's Future* and *Why Nations Fail*. Whilst the popularity of celebrity autobiographies and exposés plummet, quality contemporary autobiographies, memoirs and biographies of properly so prominent public figures – leaders – continue to attract huge readerships. And this really is the crux of so much informed public discussion: it is all about leaders and leadership. So emerge scholarly works such as *For the Common Good: The Ethics of Leadership in the 21st Century*, *The Myth of the Strong Leader* and *Transforming Leadership*. The aspirational self-help book market is booming, particularly in its business leadership, spiritual and strategy genres. Recent titles are illustrative: *How Organizations Develop Activists*, *Radical Management*, *Servant-Leadership*, *Transformational Leadership*, *Transforming Leaders into Progress Makers*, and so forth.

Groenewald is a South African leadership analyst and consultant. *Seamless Leadership* is firmly within the latter category of book, but his wide-ranging and confidently strong and ambitious claims are such that the work inevitably ranges into fields covered by the former genres as well. The book has its origins in 2010 when Ellis Mnyandu, the editor of the Business Report (Independent Newspapers' weekday daily financial supplement) and Groenewald sought a means to 'foster a national conversation on leadership' (p.ix). Soon Groenewald was interviewing leaders, largely from within big business but also politicians and sportspeople. Groenewald's interviews became the core of his biweekly Business Report features on leadership. As Mnyandu writes in his Foreword, their mutual understanding was that 'South Africa's legacy is about nothing else but leadership', and in the interviews – and now in this book – Groenewald provides 'a mirror through which the current generation of South African leaders can look at themselves…' (p. ix and x). The book comprises 18 thematically organised chapters, each with insights and sometimes brief edited transcripts of interviews with one or more South African leaders, plus brief pieces on two US leaders. Chapter titles begin with the 'Seamless Leadership' catchphrase, for example, Chapter 16 is entitled 'Seamless leader attitude to difficulties and leading in difficult times with discussion on Hlengani Mathebula, Gary Crittenden and Gary Player'.

Based on his long experience, and most particularly through these interviews, Groenewald makes a singularly bold statement, which is the key to this book:

> *So what is the destiny that South Africa must and will fulfil? When destiny and timing collide there is no stopping the movement that will explode. Nelson Mandela is a fitting example of this principle. The collision of his destiny and timing caused an explosion that resulted in a global phenomenon.* ***Our destiny is to bring to the attention of the world that brand of leadership it so badly craves and needs*** [my emphasis]. *Collectively leaders of this and other generations have failed the people of this planet. Followers are disillusioned – and when a beacon of hope for better leadership is held up we gravitate towards it as the entire globe gravitates towards Mandela or to the memory and feelings his leadership brought. The level of gravitation towards Nelson Mandela is a direct reflection of the hunger the world has for better leadership, and we are clearly very hungry!* (p.8–9)

To back up this assertion, Groenewald spends much effort in defining his science of seamless leadership and seamless society (p.89). There appears to be four key components, which are developed as the chapters progress. First is the SiPCOM Experience, for which it is important to quote Groenewald in full summary:

> *…the SiPCOM Experience…captures the truth that all visitors to this planet continually experience Situations (experiences); People (relationships); Choices (decisions); Obstacles (or challenges); and Movement (growth, development). We cannot escape the SiPCOM experience! And the 'COM' in SiPCOM reminds us that all people have this experience in COMMON.* (p.15)

Second is the 'Destiny Chain', which 'is about successfully activating and managing the universal law of movement: "All movement is a *process* governed by integration of *motivation, direction and structure*." This is what Seamless Leaders do, often subconsciously.' (p.54) [emphasis in the original]. Then there is LIM, the Leadership Impact Model, in which all seamless leaders move 'through at least four stages before making a real impact in leaving a lasting legacy'. These four stages, in sequence, are 'understanding and acceptance', 'credible leadership for consistent performance', 'leadership multiplication of legacy' and 'successful handover' (p.138–142). Finally there is what Groenewald considers his 'contribution'. This is the 'universal law of movement' which states that 'all movement is governed by the integration of motivation, direction and structure: nothing moves without the effective blending of these three components'. He proceeds to cite Edward Kieswetter, CEO of Alexander Forbes, as a successful leader able to 'activate this universal law' (p.163).

There are very many basic problems in understanding this perspective. None of Groenewald's nostrums have any known place in science or contemporary social theory. Where Groenewald does introduce and quote from chosen authorities, he all too often fails to introduce them to readers. He quotes William James writing that 'the greatest revolution of our generation (being) the discovery that human beings, by changing the inner attitudes of their minds, can change the outer aspects of their lives' (p.36). Groenewald describes James as 'a pioneering American psychologist and philosopher', but fails to mention that James' lifetime was from 1842 to 1910. Likewise with

Dr Gary Hamel (p.67) –we are not told that he is a long-time member of faculty at the prestigious London Business School and ranked by the *Wall Street Journal* as a world-leading business strategist. These problems are compounded by further issues, none of which are just technical. The book lacks vital scholarly but accessible appendages: footnotes or end notes, citations and references and even an index. And for a work which makes very big claims to global relevance, why is there no evidence of the very large corpus of scholarly work on these issues? From a reputable publisher this is inexcusable. Is this a celebrity book?

There is great dissonance between Groenewald's overall claims and his interview material. Groenewald never makes his choice of interviewees clear. Rarely is it disclosed or even discernible when interviews occurred – which is important. Are the interview transcripts publicly accessible, and if so where? Nor are chapters linked to dated *Business Report* articles. Of the 38 interviewees mentioned in this book, 25 are white. The majority are corporate private sector figures. Only four are women: Gill Marcus, Mardia van der Walt-Korsten, Terry Volkwyn and Helen Zille – all of whom are white. As he does not explain his choices, it is difficult to suggest other leading figures. But Groenewald's claims regarding South African leaders, leadership and agency, and a global relevance in wider public affairs must surely imply that his interviewee selection must extend far, far further than private sector corporate white men. So there are glaring omissions. Where are the public figures – in the buzzword terms 'thought leaders' or 'shape-shifters': faith leaders, political analysts, public intellectuals, or 'Struggle' leaders turned business people – and in Cyril Ramaphosa's case – turned politicians again? Why are women so under-represented?

Some of the interview material is interesting, but little is significant, and much is very disappointing. All too often Groenewald seems to be imposing intrusive, trite and pseudo-scientific views onto a near *Who's Who* entry. The respective sections on President Zuma and Brian Dames, neither of which includes any transcript material, are very disappointing. The interview transcript of Helen Zillie is authentic, but hardly revelatory. The section on Amplats CEO Chris Griffiths – who surely holds one of the toughest corporate private sector positions, even in South Africa's tough mining sector – has potential, but is ultimately a lost opportunity to delve into leadership in real adversity. The same can be said for the sections on Sizwe Nxasana, Bheki Sibiya and Herman Mashaba.

Groenewald never analyses Mandela's leadership style, although there is plenty of primary material available. The section on Roelf Meyer barely spans two pages and is inexplicably banal at best. Despite Groenewald's continual stressing of the defining moment of political negotiations, when 'Nelson Mandela rose to the occasion', his treatment of Mandela and Meyer raises serious doubts over the authenticity of the entire project.

It is generally accepted that South Africa has and continues to produce very powerful and globally influential private sector companies. Some of these companies had their roots in the apartheid years, but many others have a later birth and continue this tradition of being well-led and well-run companies with global reach. The story of why this is so remains to be properly analysed and told. Despite Groenewald's heavy bias towards private sector corporates, readers rarely get even hints as to key features of this larger picture and its global significance. But it is also accepted that such company executives have long shown a marked reluctance in making considered public analysis and comments on wider social, moral and political issues, either concerning South Africa or more globally. This was glaringly apparent during the apartheid years – even during the insurrectionary times of the 1980s – and it remains so to this day. In the contemporary period – amongst the more prominent business leaders to assert their presence in wide public affairs, together with people such as now outgoing Nedbank chairperson Ruel Khoza and businessman and Free Market Foundation chairperson Herman Mashaba – is Australian-born Mark Cutifani. As Anglo American's head, Cutifani is consistently making considered and trenchant public critiques of the South African government's economic and developmental policies and decision-making and managerial competencies. Cutifani provides vital glimpses into his leadership style: no front office door or designated parking places, no ties and no 'Sir' or 'Meneer'. All this is so refreshingly different from the vanity and peremptoriness which all too easily and quickly afflicts many in high private or public leadership positions. Yet the section on Cutifani is as weak as that on Mashaba. Groenewald is just not looking bravely enough.

Finally, this book is all about successes, often despite tremendous adversity. Where is the discussion and deep 'mirror, mirror' reflections on moral courage, where suffering is an implicit consequence of principled and public bravery? We can think of M.K. Gandhi, Nelson Mandela, and Rev. Beyers Naudé. But where is leadership and moral courage now? Where are the stories of failed leadership, and what lessons can we learn from these? Groenewald has much to say on morality, and tells us often how failure can be the forge on which character is deepened. But leaving matters at the level of aphorism trivialises the importance of the point. Both Gill Marcus, who when interviewed was Governor of the SA Reserve Bank, and Kieswetter raise and stress the important, but still unrecognised, issue of a greater common good in South Africa. Groenewald hardly sees it.

Asking big questions is not only vital for interviewers, but is also an essential responsibility of leaders. As South Africans celebrated various 20th and 25th anniversaries (of the events throughout February 1990 and in 1994) and commemorated the death of Mandela, various influential voices – many with impeccable anti-apartheid and pro-democracy backgrounds – sounded notes of caution and warning. In essence their view is that South Africans have yet to ask and address key and difficult questions – including confronting the Mandela cult. Failure to ask difficult questions to which there are no easy answers is the defining point of contemporary South African leadership. South Africans have something to export to a 'hungry world'. But this does not involve turning hardworking, determined and courageous South Africans into celebrity name-drops in pursuit of making an unsustainable global point. This must surely be a new pinnacle in delusionary special-case exceptionalism that all too many South Africans, of all political persuasions, are prone to vainly and delightedly bask in, which is precisely that which must be jettisoned. When will South Africans take a good look at themselves? If there are export quality 'universal lessons from South Africa', these are neither evangelical nor messianic messages of inspiration. South Africa is a primary example of how not to develop a post Cold War post-conflict late developing constitutional democracy. *Seamless Leadership* is hocus-pocus. The mirror is cracked.

A review of carbon dioxide as a refrigerant in refrigeration technology

AUTHORS:
Paul Maina[1,2]
Zhongjie Huan[1]

AFFILIATIONS:
[1]Department of Mechanical Engineering, Tshwane University of Technology, Pretoria, South Africa
[2]Department of Mechanical Engineering, Moi University, Eldoret, Rift Valley, Kenya

CORRESPONDENCE TO:
Paul Maina

EMAIL:
mainap@tut.ac.za

POSTAL ADDRESS:
Department of Mechanical Engineering, Tshwane University of Technology, Private Bag X680, Pretoria 0001, South Africa

DATES:
Received: 31 July 2014
Revised: 13 Nov. 2014
Accepted: 19 Jan. 2015

KEYWORDS:
environment; safety; heat pump; energy efficiency; transcritical

HOW TO CITE:
Maina P, Huan Z. A review of carbon dioxide as a refrigerant in refrigeration technology. S Afr J Sci. 2015;111(9/10), Art. #2014-0258, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20140258

Tough environmental laws and stringent government policies have revolutionised the refrigeration sector, especially concerning the cycle fluid known as the refrigerant. It has been observed that only natural refrigerants are environmentally benign. When other refrigerant qualities are considered, especially those relating to toxicity and flammability, carbon dioxide emerges as the best among the natural refrigerants. However, carbon dioxide based refrigerants are not without drawbacks. Even though the use of R744 – a carbon dioxide based refrigerant gas – has solved the direct effect of emissions on the environment, studies to investigate the indirect effects of these systems are needed. Improvement in existing technical solutions and the formulation of additional solutions to existing R744 refrigeration problems is paramount if this technology is to be accepted by all, especially in areas with warm climates. National policies geared to green technologies are important to clear the way and provide support for these technologies. It is clear that carbon dioxide is one of the best refrigerants and as environmental regulations become more intense, it will be the ultimate refrigerant of the future.

Introduction

Most refrigerators use a liquefiable vapour to transfer heat. This fluid is known as the refrigerant. Refrigerant selection is a key design decision that influences the mechanical design of the refrigeration equipment. Factors that must be considered in refrigerant selection include performance, safety, reliability, environmental acceptability and cost. However, the primary requirements are safety, reliability and, nowadays, environmental friendliness (in terms of ozone depletion and global warming potential). Table 1 summarises the properties of some refrigerants and indicates that no progress has been made in terms of global warming potential (GWP) when switching from hydrochlorofluorocarbons (HCFCs) to the hydrofluorocarbon (HFC) family. When securely contained in a properly operating system, refrigerants do not impact climate change; however, system leaks and improper recovery of refrigerants during repairs or at end of life result in these harmful gases entering the atmosphere. Furthermore, during production of refrigerants, toxic and harmful wastes are released into the environment, which cause air, water and land pollution in addition to releasing greenhouse gases. An alternative to HFCs is to apply naturally occurring and ecologically safe substances, the so-called natural working fluids. The most important substances in this category are hydrocarbons, ammonia and carbon dioxide, although when safety concerns are raised (toxicity and flammability), R744, a carbon dioxide based refrigerant gas, becomes the best substitute.

Carbon dioxide (CO_2) is a clear gas (at atmospheric conditions) without a particular smell when the concentration is below suffocation level. When the concentration reaches toxic levels, it has a slightly pungent smell and somewhat acidic taste. It has a higher density than air, which has its own advantages with respect to refrigeration and disadvantages with respect to safety. CO_2 is made both naturally and artificially – artificially through the burning of fuel and other industrial processes.[1,3] Approximately 0.04% of atmospheric air is CO_2, thus CO_2 is at a concentration of approximately 380 parts per million (ppm) in air. Exhaled air from the body has a CO_2 concentration of about 4%.

History of R744 as a refrigerant

Since the invention of the vapour-compression cycle by Evans and Perkins in 1834, R744 has been a candidate for a refrigerant. Documented studies state that Alexander Twining was the first to propose R744 refrigeration using a steam compression system in his British patent of 1850. However, Thaddeus Lowe was the first to actually build a refrigerator running on R744 for ice production in 1866 after discovering its potential while using it in military balloons. Carl Linde followed suit and built a better refrigerator running on R744 in 1881, just after Windhausen had built the first R744 compressor in 1880. In 1884, W Raydt built a R744 refrigeration system for making ice using a vapour compression mechanism while, at the same time, J Harrison was the first person to build a device for manufacturing R744 purely for refrigeration use. The British company J and E Hall built the first R744 marine refrigerator in 1890 using Windhausen's compressor designs, while in the USA, continuous production of these refrigerators was started in 1897, mainly by Kroeschell Bros. Ice Making Company. Owing to its safety aspects when compared to other refrigerants during this period, R744 refrigerators grew in number, especially in the marine sector. At the same time, its technology was improving. For example, in 1889, J and E Hall created a two-stage R744 compressor which was more efficient, and in 1905, Voorhees created a flash chamber which was very similar to a liquid-vapour separator.[4-7]

Calcium chloride solution was used in most refrigerators as a secondary fluid. The salt solution was cooled to around -10 ºC (evaporation temperature of -15 ºC). Originally, the evaporator and condensers used galvanised steel pipes, 32 mm in diameter for small refrigerators and 51 mm in diameter for large cold rooms. Tank and coil heat exchangers were the first to be used, before tube in tube (double pipe) technology was introduced in 1902. The shell and tube type were invented in the early 1930s and fin technology in the 1920s. Copper replaced steel pipes during this decade too, with pipe diameters being reduced to 13 mm because of the increased heat transfer offered by the fins and copper. Air circulating fans were introduced around this time for improved cooling, especially in cold rooms. R744 used to cost around 9 cents per kg but the price increased to 12 cents per kg in the late 1920s.

Table 1: Properties of some refrigerants[1,2]

Refrigerant	Critical temperature (ºC)	Critical pressure (bar)	Ozone depletion potential	Global warming potential (100 years)	Flammable or explosive	Toxicity
CFCs and HCFCs						
R12	100.9	40.6	0.9	8100	No	No
R22	96.2	49.8	0.055	1500	No	No
Pure HFCs						
R32	78.4	58.3	0	650	Yes	No
R134a	101.1	40.7	0	1200	No	No
R152a	113.5	45.2	0	140	Yes	No
HFC mixtures						
R404A	72.1	37.4	0	3300	No	No
R407C	86.8	46.0	0	1600	No	No
R410A	72.5	49.6	0	1900	No	No
Natural refrigerants						
Propane (R290)	96.8	42.5	0	3	Yes	No
Isobutane (R600a)	135.0	36.5	0	3	Yes	No
Ammonia (R717)	132.2	113.5	0	0	Yes	Yes
Carbon dioxide (R744)	31.0	73.8	0	1	No	No

CFC, chlorofluorocarbon; HCFC, hydrochlorofluorocarbon; HFC, hydrofluorocarbon

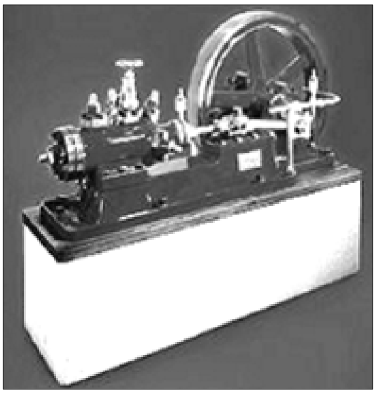


Figure 1: An old single cylinder R744 compressor.[3]

With the invention of compressors, vertical, cylinder-type compressors of up to 42 kW (325 rpm) were first used but were later replaced by horizontal compressors of up to 176 kW (120 rpm) in size. Both these constructions were similar to the steam engine design. R744 required heavy duty parts for valves, fittings, compressors and heat exchangers as a result of its associated pressure. Refrigerator size increased up to 704 kW by 1916.[7]

The use of R744 air conditioners for comfort cooling began in the 1900s. Because of the toxicity and/or flammability of NH_3 and SO_2, R744 gained in popularity, especially in food-related industries (food markets and eateries) and human comfort applications, e.g. in theatres, bars, hospitals, ships and hotels. In 1900, only 25% of all ships were using R744 as the refrigerant, but by the 1930s, this proportion had increased to 80%. Although the ships used old technology, R744 equipment still worked, albeit inefficiently, especially because they used a convectional subcritical refrigeration cycle. In addition, there were sealing and capacity loss problems related to R744 high pressures. These disadvantages encouraged a search for safe and efficient (especially at high discharge temperatures in warm climates) refrigerants which ended with the discovery of chlorofluorocarbons (CFCs) in the 1930s. The invention of CFCs, coupled with a lack of technological improvements from the R744 refrigeration industry, caused the decline in the use of R744. The last large R744 refrigeration system was installed in 1935 for Commonwealth Edison Company headquarters. The system was replaced by CFC refrigerators 15 years later. CFCs eliminated the problems encountered when using R744, such as the need for high pressure sealing, capacity and efficiency loss and the high cost of components. Eventually by the 1950s, R744 refrigeration was completely phased out.[4-7]

After the discovery of the adverse effects of synthetic refrigerants in the late 1980s, there was a renewed interest in R744. Professor Gustav Lorentzen was the pioneer of the revival of R744 refrigeration in the early 1990s, with many studies and ideas dedicated to its improvement. He suggested that, because of the properties of R744, motor vehicle air-conditioning systems (the leading sector in refrigerant leakage – 60% of all leakages[8]) and water heat pumps are best suited for R744 refrigeration.[9] His idea was positively acknowledged with many motor vehicle and water heat pump companies investing in R744 research. Leading car manufacturers, such as Nissan, Bavarian Motor Works AG (BMW) and DaimlerChrysler, have installed R744 air-conditioning systems in new cars.[4,10] Many R744 domestic heat pumps are manufactured and marketed in Asia and Europe.[11-13] Recently, there has been a keen interest in using R744 in supermarkets and other commercial refrigeration applications.[14,15] Leading beverage companies like Coca-Cola and PepsiCo have embarked on converting their vending machines to use R744 as the refrigerant of choice.[12] Also, R744 refrigeration has been applied in residential and commercial buildings' air-conditioning systems with great promise.[16] In short, there are currently many application prospects for R744 refrigeration under investigation, most of which are quite promising.

R744 properties

General introduction

As indicated in Table 1, R744 is a non-toxic, non-flammable natural refrigerant with an ozone depletion potential of zero and a GWP of 1. Furthermore, it is widely and cheaply available as a constituent of the atmosphere and as a result of industrial processes, especially the ones involving fuel. Its critical temperature, however, is 31.1 ºC. Low critical temperature means that R744 cannot be used effectively in a convectional (subcritical) refrigeration cycle simply because the condenser will not transfer heat above the critical temperature. Therefore, the condenser will be ineffective and many losses may occur. Furthermore, at temperatures that are near the critical temperature but less than it, there is a drastic reduction on the vaporisation enthalpy which leads to a reduction in heating capacity and reduced system performance. Therefore, as Lorentzen has suggested, R744 can be effectively used only in a transcritical cycle.[9]

Transcritical cycle

A transcritical cycle (Figure 2) is not limited by the critical temperature because heat output is through a temperature glide. The condenser in a transcritical cycle is replaced by a gas cooler because there is no condensation taking place, but rather a gas cooling process. This temperature glide is advantageous especially in applications such as water heating and air heating (e.g. drying processes) because of the associated efficiencies. The temperature range in which R744 refrigeration can operate in transcritical operation is the highest when compared to other convectional refrigerants, i.e. -50 ºC to 120 ºC.[4] The only drawback with the transcritical cycle for R744 is the high pressure. The critical pressure for R744 is 73.7 bars. If R744 is operated through a transcritical cycle, then its high pressure will be above 73.7 bar, which is quite high. This necessitates an equipment design that can handle such a high pressure. The high pressure has its own advantages (e.g. compact equipment and design) and disadvantages (costly equipment and safety issues). However, with current technological advances, this pressure is not a big concern.

Thermophysical properties

The high latent heat of vaporisation and volumetric heat transfer caused by the high pressures involved means that R744 equipment components can be designed in smaller size. Coupled with the fact that R744 has very low viscosity, Reynolds number is reached even with a low flow rate, which means that most flows are turbulent in nature and the heat transfer rate is high. Apart from the flow nature, R744 has a small liquid to vapour density ratio (especially near the critical point) which leads to uniform and homogeneous distribution of the refrigerant in channels. This homogeneity also adds to the high heat transfer rate. Furthermore, proximity to the critical point and less pressure loss, especially in the gas cooler, contributes to improved convective heat transfer. When the heat transfer rate is high, the size of the heat exchangers can be drastically reduced for the same amount of heat transfer to occur.[6]

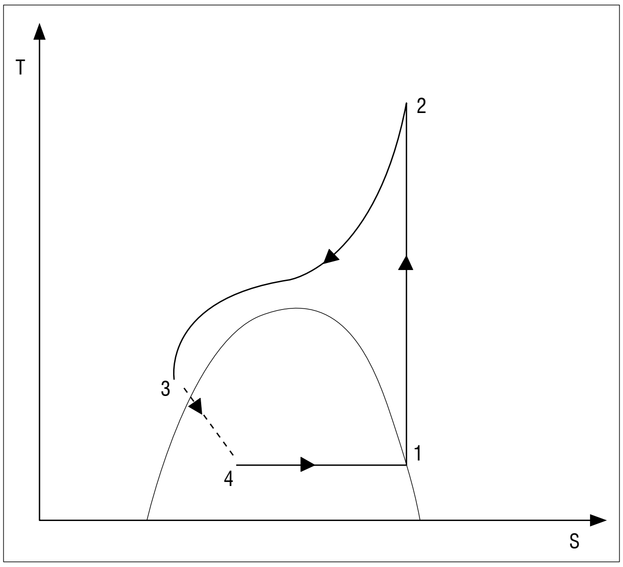


Figure 2: A theoretical transcritical cycle characterised by isentropic compression (process 1–2), isobaric heat output (process 2–3), isenthalpic expansion (process 3–4) and isobaric heat intake (process 4–1).

The compressor size can also be reduced because of the associated compression ratio involved. Even though the R744 transcritical cycle deals with high operational pressures, the ratio between the high pressure and low pressure is low when compared to other refrigerants. The ratio is also closely associated with the high adiabatic index of R744 (which is approximately 1.3). These factors lead to a compact, smaller and more efficient compressor, ideal for applications for which space is limited, e.g. in mobile air conditioners in cars.[3] The R744 compressor efficiency increase is also brought about by the low effect of valve pressure drops and the re-expansion ratio it experiences. Furthermore, internal leakages and piston blow-by losses are negligible when compared to those of other compressors. Even though the compressor walls must be thicker because of the pressures involved, R744 volumetric capacity dictates small parts in the compressor and therefore the overall size of the compressor is smaller when compared to compressors of convectional refrigerants of the same capacity.[5,6]

Properties of R744 at a supercritical state are always in between those at liquid state and those at gaseous state. The critical point, defined as the point at which no liquefaction occurs above the critical pressure and no gas is formed above the critical temperature, is peculiar in nature because, near the critical point, there is always a sudden variation in the properties. Specific heat, thermal conductivity, enthalpy, entropy, density and viscosity undergo a major change as the critical point is approached.[17] Figure 3 shows the variation of specific heat at constant pressure (c_p) against temperature for R744 at several pressures.[18] As can be seen from Figure 3, the highest value of c_p occurs at a pseudo-critical temperature for that pressure. This property can be incorporated into the design of gas coolers to maximise on their output.

R744's surface tension is smaller than that of other refrigerants. The significance of this observation is that surface tension affects the wetting characteristics, flow characteristics and evaporation characteristics of a fluid. Low surface tension might be positive in that the temperature required to initiate and maintain nucleate boiling is reduced, but it can also be negative because of drop formation and entrainment, especially when there is reduced surface stability. On the other hand, the thermal conductivity of R744 is considerably higher than that of most synthetic refrigerants, while its viscosity is lower. A high thermal conductivity means a higher heat transfer rate, while viscosity affects flow properties, which in turn affects heat transfer and pressure drop. Therefore, the thermophysical characteristics of R744 are favourable and encourage its use as a refrigerant.[5] Still, being relatively inert, R744 is compatible with most lubricants and equipment materials, as documented in numerous studies (even ones not related to refrigeration).

High (gas cooler) pressure

At supercritical state, the temperature and pressure of R744 are independent of each other and thus can be regulated independently to optimise output. Still, compressor input power is proportional to the high pressure. As shown in Figure 2, if the gas cooling process occurs at a constant pressure (process 2–3), the magnitude of the pressure will affect the specific enthalpy. This pressure is not controlled by the cooling fluid conditions (temperature and flow rates) as is the case with convectional refrigerants; it is mostly controlled by the amount of refrigerant charge present.[8] As the pressure increases, there will be an initial increase in heat output with a moderate increase in compressor power input, thus there will be an overall increase in system efficiency. As the pressure is increased further, it will reach a point at which the additional work input is more than the additional heat output, and thus the efficiency of the

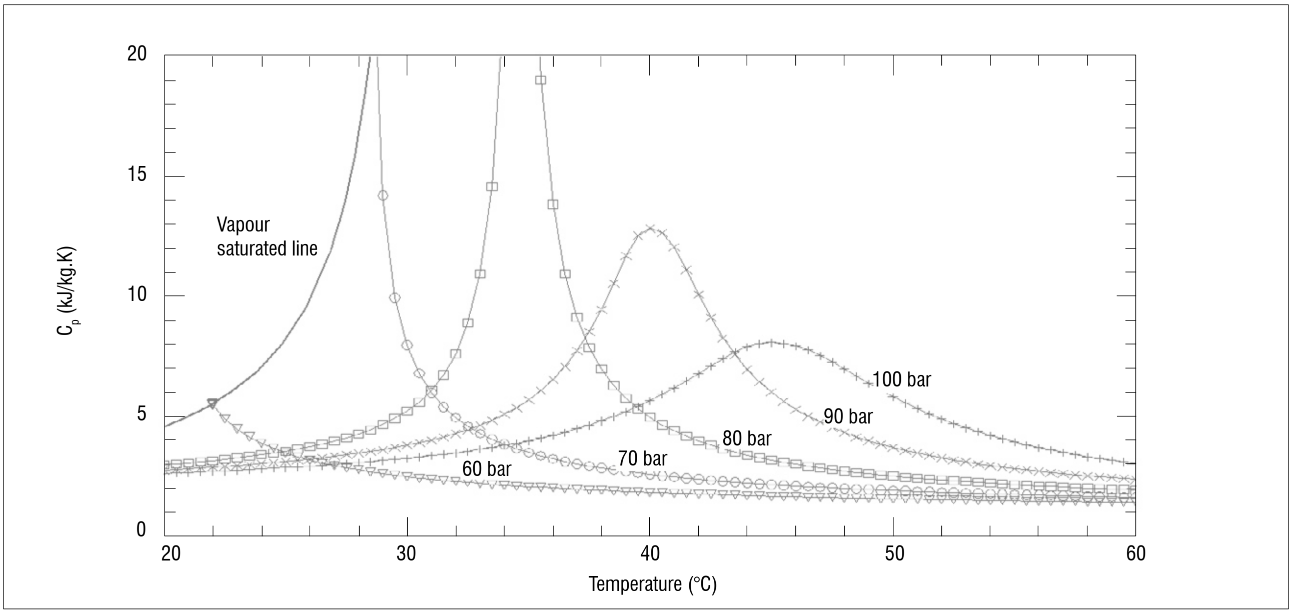


Figure 3: Variation of specific heat with temperature.

refrigerator will begin to decrease. This behaviour is attributed to the shape of isotherms (which affect the heat output) and isentropes (which affect the compressor power) at pressures above the critical point. There is always an optimum high pressure value which corresponds to certain operational conditions. The optimum pressure depends on the gas cooler outlet temperature, the evaporation temperature, the compressor isentropic efficiency and the amount of refrigerant.[19,20] Therefore, it is paramount that there is a means of controlling the high pressure. The optimum pressure can approximately be given by:

$$P_{opt} = 2.6T_{exit} + 8 \qquad \text{Equation 1}$$

where P_{opt} is expressed in bars and T_{exit} is the gas cooler exit temperature in °C.

Another important characteristic of R744 is the relationship between its pressure and temperature. The vapour pressure of R744, apart from being higher than other refrigerants (Table 2), also has a greater variation per unit temperature change, especially near the critical point. The slope of change in vapour pressure to change in temperature is much steeper for R744 than for other refrigerants. This means that for every unit change in pressure, there is a lesser change in temperature with R744 as compared to other refrigerants. Therefore, the effect of pressure loss is less severe and more tolerable in R744 than in other refrigerants. The high vapour pressure also causes high vapour density (Clausiu–Clapeyron's relation). Because volumetric heat transfer is the product of vapour density and latent heat of evaporation, a high vapour density results in a high volumetric heat transfer.[5] Furthermore, a high vapour density results in a low velocity of R744 in pipes, thus resulting in less of a pressure drop. In addition to high thermal conductivity of R744 (Table 2), high vapour density also allows the use of small components, e.g. tubes, which results in lower radiation losses.[21] Phase separation characteristics between vapour and liquid phases are directly controlled by phase density differences. R744 has a low density ratio, which is necessary for a homogenous two-phase flow.[4]

Even though R744 has very favourable properties, especially when used as a heat pump, its cycle is still affected by many losses. Given an evaporation temperature and minimum heat rejection temperature, the transcritical cycle is affected by higher throttling losses when compared to the convectional subcritical cycle. These losses increase the theoretical work done on the transcritical cycle. The throttling losses in refrigeration are normally caused by the temperature difference in the throttle device and the refrigerant properties. With the temperature being set at a specific value, R744's unique properties, especially near the critical point (i.e. high liquid specific heat and low evaporation enthalpy), increase the throttling losses and thus increase the compressor power required. Therefore, even though compressor losses are lower in R744 machines, other factors tend to increase power consumption. Thus, it is paramount to reduce the losses as much as possible if R744 technology is to be fully embraced.[5]

Table 2: Properties of several common refrigerants[21]

Refrigerant	Evaporator temperature = -30 °C			Evaporator temperature = 0 °C		
	Saturation pressure (bar)	Liquid thermal conductivity ($Wm^{-1}K^{-1}$)	Vapour density (kg/m^3)	Saturation pressure (bar)	Liquid thermal conductivity ($Wm^{-1}K^{-1}$)	Vapour density (kg/m^3)
R22	1.64	0.1084	7.38	4.98	0.0947	21.23
R407C	1.62	0.1187	7.21	5.13	0.102	21.88
R134a	0.84	0.1058	4.43	2.93	0.092	14.43
R410a	2.79	0.1293	10.57	7.99	0.1099	30.63
R404a	2.05	0.0862	10.69	6.05	0.074	30.72
R744	14.28	0.1469	37.1	34.85	0.1104	97.65

Safety

Finally, R744 is considered non-toxic, although at concentrations above 2% (about 5 kg of R744 in a 120-m^3 room) it can start to become harmful. It is colourless and odourless so it cannot easily be detected; therefore, in installations for which a high amount of charge is used (25 kg by European standards) and the ventilation is poor, R744 detectors should be installed. Furthermore, having a higher density than air (R744 relative density is about 1.53 at atmospheric conditions), it will tend to occupy lower areas so can prove to be more deadly when spilled in a non-ventilated room. Therefore, the detector should be installed near the floor. Table 3 gives the effects of R744 at various concentrations in the air. To prevent an accumulation of R744 in an enclosed space with a big installation:

- sufficient ventilation must be provided in the equipment room
- R744 containers must always be in an upright position, especially when charging the system

In the event of a large R744 release, the equipment room should be avoided until the concentration in the room is within allowable limits (measured either through sensors or other means). In case of exposure, the person should be taken into open air immediately.

Apart from toxicity issues, explosions as a result of the operational pressures experienced in R744 equipment is also a safety concern. The shock and flying fragments caused by a blast might cause injury and harm. Apart from explosions caused by the pressure in R744 systems, there is another severe type of explosion known as boiling liquid expanding vapour explosion (or BLEVE), which usually occurs when a vessel containing pressurised saturated liquid is rapidly depressurised as a result of a crack or rupture and can cause a blast which is more severe than the pressure blast. The rapid depressurisation leads to explosive vaporisation and a sudden overpressure in the tank which might blow the vessel.[21] To prevent these catastrophes, the safety design of the R744 equipment should include the following[5]:

- The equipment should have an over-pressure release valve on both the high pressure side and low pressure side.
- Components should be pressure tested at twice (or more) the amount of normal operation pressure and temperature.
- The pressure test should incorporate the other operational effects like fatigue due to pressure cycles, creep, vibration and corrosion (especially if there is water present).
- All components should be designed with the highest pressure and temperature (including a safety factor) the system may encounter even in standstill mode.

When considering the size of the system versus the pressures involved, the relative explosion energy in R744 systems is approximately the same as in convectional systems. This is mainly because R744 systems tend to have a smaller system size and less charge for the same capacity.[8]

R744 applications

Apart from being a refrigerant, R744 is used in many processes and applications. However, in refrigeration, there are numerous applications for R744, some of which have been commercialised.

Water heat pumps

Production of hot water is the best application for R744 heating refrigerators (heat pumps) as the temperature slide in the transcritical cycle suits the thermodynamic properties of water well.[5] Very efficient heat transfer and very high water temperatures are achieved with water heating applications, especially when using a counterflow gas cooler.[13] With heat pumps being the preferred water heating method when compared to electricity or fuel-fired systems, governments and other energy efficiency conscious bodies are encouraging their use. This promotion comes after it was realised that water heat pumps with an average coefficient of performance (COP) of 3 can reduce energy usage by 67% when compared to electric heating, and by even more when compared to fuel heating. The percentage energy saving of a heat pump (ΔE) when compared to another heating system with an efficiency of η is given by[21]:

$$\Delta E = \left(\frac{1}{\eta} - \frac{1}{COP} \right) \qquad \text{Equation 2}$$

where η for electric heaters is approximately 1 while η for fuel-fired heaters varies from 0.5 to 0.95, depending on the equipment and the type of fuel.

In addition, environmentalists are encouraging the use of heat pumps which are environmentally friendly, such as R744 heat pumps.[22] This is one of the applications that the ‘father’ of R744 refrigeration re-invention, Prof. Lorentzen, suggested for R744 as a refrigerant.[9] In fact, it is this sector that has seen major commercialisation of R744 refrigerators. Since 2001, the production of commercial R744 heat pump units has taken place under the general name of ECO CUTE in Japan (Figure 4). These units, which are marketed both in Asia and Europe, have now surpassed an annual production rate of 1 million, a growth encouraged both by their high efficiency and by incentives from government and environmental bodies.[3,11,23] Other manufacturers have introduced similar systems for residential, commercial and industrial use. The possibility of efficiently producing hot water at 90 ºC and above is encouraging the

Table 3: Effect on humans of R744 at various concentrations[1,21]

Concentration (%)	Effects on humans
0.1	Human comfort limit
0.5	8 h per day exposure limit
2	50% increase in breathing rate
3	100% increase in breathing rate; 10 min short-term exposure limit
5	300% increase in breathing rate; headache and sweating may begin after about an hour. This is the immediate danger to life and health concentration. An escape within 30 min will avoid irreversible health effects.
8–10	Headache after 10 to 15 min; dizziness, buzzing in the ears, blood pressure increase (because of high blood R744 content and lowered pH), high pulse rate, excitation and nausea. Pungent smell and irritant to both nose and throat. This is the lowest lethal concentration.
10–18	Cramps similar to epileptic fits, loss of consciousness and shock after exposure of a few minutes
18–20	Stroke symptoms. Death can easily occur.
30	Rapid unconsciousness and convulsions

industrial and commercial use of R744 heat pumps, especially in hotels, hospitals and the food industry.[5]

With water heating being one of the highest energy uses, especially in the residential sector (approximately 20%), coupled by the environmental and energy efficiency advantages, R744 heat pump technology has great future prospects.[22] Lorentzen has described the possibility of using R744 heat pumps for both heating and cooling applications simultaneously. These systems have high overall system efficiencies and can find application in places where both refrigeration and hot water are needed, for example, in hospitals, supermarkets, hotels, and the food processing and chemical industries.[9]

Mobile air conditioning

Mobile air conditioning is the second application that Lorentzen envisioned for R744 refrigeration systems. Compact R744 systems coupled with good heat transfer characteristics between air and R744 encourage the use of these systems in a sector in which equipment space and weight is limited while energy efficiency is paramount.[9] Mobile air conditioning is the largest consumer of refrigerants in the world, followed by commercial refrigeration. Mobile air conditioning consumes 31% of the world's refrigerant, which adds up to more than 150 000 t/year.[15,24] On the same note, mobile air conditioning has the highest leakage rate.[21] This means a complete change to R744 will be highly advantageous. Recent studies demonstrated the superiority of R744 systems when compared to commonly used HFC systems.[25] With the phasing out of HFC in progress, especially in European countries, R744 is the best alternative for mobile air conditioning. R744 air conditioning will work especially well with fuel-efficient hybrid or electrical cars with little waste heat available. With electrical cars specifically, if the air-conditioning system is efficient enough, more energy will be used to drive the car and thus more travelling distance will be covered before the electricity runs out.[8]

The greatest disadvantage of R744 in mobile air-conditioning applications is its high heat rejection temperature. However, in the new technologically advanced vehicles, in which there is little or no excess heat generation, this high temperature heat output can be effectively used, especially in cold season, not only for human comfort but also for heating essential vehicle parts like engine fluids. Common HFCs used in mobile air conditioning do not perform well as heat pumps because of their thermophysical properties.

As mobile air conditioning systems are prone to more maintenance problems than stationary systems, logistical concerns arise as R744 systems are fairly new and are different from convectional systems. There is a need for qualified technicians to handle the new technology, which is made more complex in vehicles because of the limited weight and space requirements coupled with integration of the air-conditioning system and other electrical and mechanical systems of the car. These, among other minor economic issues, need to be addressed before R744 mobile air conditioning is fully embraced. Furthermore, frost accumulation on evaporators presents a complication which has not yet been solved effectively, especially in a mobile application. Despite these challenges, R744 is proving to be an ideal refrigerant, capable of providing high temperature heat instantly while requiring less air to convey the heat. This makes it a hot research topic for complex mobile environment control with improved efficiency.[5]

Commercial refrigeration

Commercial refrigeration is the equipment used by retail outlets to display, hold or prepare food and beverages that customers purchase. This equipment includes refrigerated display counters in supermarkets, refrigerated vending machines, water coolers/heaters and ice generating machines. Commercial refrigeration consumes about 28% of worldwide refrigerants, thus is the second largest user of refrigerants.[15,24] This makes it one of the largest emitters of refrigerants into the environment and accounts for approximately 37% of worldwide emissions. In 2002, commercial refrigeration was responsible for more than 185 000 t of leaked refrigerant into the atmosphere.[26] Furthermore, the energy utilisation in this sector is usually very high, necessitating a need for efficient refrigeration systems.

Ironically, until the year 2000, R744 applications in commercial refrigeration were not considered viable. The perception has since changed with its use either as a heat transfer fluid, in a cascade system, or on its own in either a transcritical cycle or a subcritical cycle, depending on the environmental temperature. External factors like safety requirements, extra tax on HFC systems and limitations on the maximum amount of HFC charge that can be used on a single system were the main reasons for R744 acceptability in commercial refrigeration. It was first purely used indirectly as a heat transfer fluid, then in cascade systems in conjunction with HFC at a reduced charge or with hydrocarbons (HCs). With time, more skills and knowledge were acquired and the cascade systems were replaced with fully transcritical R744 systems. With the possibility of heat recovery (for space heating or tap water heating), R744 commercial refrigeration has a great potential. As the world accepts the use of R744 in supermarkets, studies show that its associated costs and energy consumption are comparable to

Figure 4: Examples of EcoCute heat pumps.[1]

other convectional refrigeration systems. Actually, it has been suggested that R744 systems are the dominant technology of the future because of their good thermophysical and safety properties.[8]

R744 technology is even being utilised in stand-alone bottle coolers and both hot and cold vending machines. Major investment in the light commercial sector is directed to R744 technology, with more than 85 000 units in operation worldwide.[8] Most of these vending machines utilise the transcritical cycle.

Other applications

Residential air conditioning

R744 air in air refrigerators has been the focus of investigations by both research institutions and industry because of the high demand for such equipment and the requirement for HFC alternatives. The annual demand for residential air-conditioning units is more than 40 million units and further market growth is expected.[21] Air conditioning is the second largest consumer of energy after water heating in most residential areas. Environmental concerns in this application are more focused on the indirect impacts of emissions due to energy use, than on the direct impacts of refrigerant leaking. Therefore, energy efficiency is paramount.[5]

There have been promising results with the application of R744 in stationary air-conditioning systems. Units with a one-way refrigerant circuit are working reliably and are more efficient than those with complicated switching devices for redirecting refrigerant flow in different seasons. Redirecting refrigerant flow involves additional valves and fittings which increase the capital costs and efficiency losses through leakages and pressure loss. One-way refrigerant flows (with the air flow redirected in different seasons) are simple, flexible, compact and cost efficient.

Heat pump dryers

Compared to convectional dryers, heat pumps may reduce energy consumption of food dryers by up to 80%. Other products which need drying include wood, laundry and sewage sludge. The product quality can be optimised more easily by using a heat pump because of the availability of greater control options for different drying chamber conditions. Water vapour from the products is absorbed by warm air, which is heated by the gas cooler prior to the drying chamber. As this air passes through the evaporator, it is dehumidified and cooled down before returning to the gas cooler to be reheated. This closed air cycle provides remarkable energy conservation, contributes tremendously high energy efficiency (by participating in the heat transfer in both the gas cooler and evaporator) and reduces environmental contaminants and the unpleasant odour experienced in some drying processes, e.g. drying of sewage sludge.[21] Owing to the gliding temperature and better temperature adaptation of heat exchangers, R744 heat pumps can achieve substantial energy savings when used as a dryer. Higher air temperatures can be easily and efficiently achieved in these systems, thus enhancing the moisture extraction rate.[27] As more efficient R744 equipment parts are produced (e.g. compressors and heat exchangers), commercial heat pump dryers using R744 as the refrigerant are becoming a possibility.[21]

Transport refrigeration

The efficient, reliable and compact characteristics of R744 equipment encourage their application in the transport sector.[8] R744 refrigeration systems are considered as a replacement for HFC refrigeration systems, both in public and goods (especially perishable) transport. These transport modes include perishable goods trucks, public and goods trains and ships. The relatively high density and capacity of R744 is an added advantage in this sector. Furthermore, because of the global nature of transport refrigeration, an environmentally benign alternative is required which is available everywhere (even in rural areas where the refrigerated trucks operate) and acceptable to all (some countries in which the refrigerated ships dock have stringent environmental rules which need to be strictly adhered to[5]).

Environment control units

Military operations usually require space conditioning for their temporary shelters, command modules and vehicles, which should be able to withstand the unique operational environment. The compactness of R744 equipment in addition to its availability globally has led to an increased interest in R744 space-conditioning systems for the military.[5,28]

Future applications for R744

The potential of R744 refrigeration is far wider than the applications discussed above. Experts assert that R744 as a refrigerant shows promise of capturing more markets, even outside the refrigeration industry, although currently, it is the natural refrigerant with the widest range of use.[21] Some future potential uses of R744 as a refrigerant are information technology (IT) equipment cooling, industrial heat pumps and industrial waste heat recovery.

Information technology equipment cooling

The necessity of high performance data centres is increasing as the need for information to be made available at anytime from anywhere to anyone grows. For this to be possible, high-density data storage and processing environments are required, and must be coupled with an efficient heat absorption system which ensures that working environments are not overheated by processes and that a smooth operation and flow of information can exist. A high-efficiency cooling system with high COP and automatic control is therefore paramount for both energy-saving purposes and smooth operation of the data centre. R744 cooling systems can provide the heat absorption process efficiently and reliably without interfering with the IT equipment because R744 is electrically benign when compared to traditional water-based systems. Another advantage of R744 cooling systems is their compactness, which eases their integration with the IT equipment and surrounding structures. Furthermore, it is a form of waste heat recovery with the possibility of simultaneous heating and cooling.[21]

Industrial heat pumps

R744 can be used for recovery of useful waste heat while at the same time providing low temperature heat (up to 130 ºC) for industrial processes. The application of R744 heat pumps in this sector therefore has potential while providing energy savings. Heat pump drying is an example of such an application already introduced into the market. Other sectors with future potential are[21]:

- Washing processes: Warm to hot water is required in some industrial processes, for example, textile washing, washing of food and cosmetics production facilities. This hot water can easily and efficiently be provided by R744 heat pumps.
- Process water: Warm process water is required in certain industries, for example, the production of starch and other viscous chemicals.
- Process air: Some industries use warm to hot air instead of water, for example, in the production of flake boards and some plastic containers.
- Steaming processes: Steam is required in most manufacturing processes as a heating media or just for cleansing purposes. A good example where R744's simultaneous heating and cooling can be effectively applied is in the regeneration of activated carbon filters in order to recover solvents. In this process, steam generated by the R744 gas cooler vaporises the activated carbons which are loaded with solvents. Subsequently, the steam absorbs the solvent, which is condensed in the evaporator and extracted. Even though it might be difficult to produce steam at the required temperature and pressure with the current R744 heat pump technology, studies are being conducted to investigate this possibility. Still, R744 can be used to preheat feed water to the boilers, thus reducing the amount of energy used in the boiler while improving overall system efficiency.

Industrial waste heat recovery

Many industrial processes produce heat that is released to the environment as waste because it cannot be reused effectively. This waste heat can be used efficiently by a R744 heat pump to produce useful heat, either for air conditioning or for tap water applications. This waste heat recovery will reduce the energy costs of generating hot tap water and/or building air. In addition, the industry will reduce its waste product treatment by meeting temperature regulations concerning waste products released into the environment and will reduce expenditure on waste products while providing useful heat.

R744 applications in South Africa

South Africa is ranked 16th in the world in terms of total primary energy consumption. This makes it the highest consumer of commercial energy per capita in Africa and thus a relatively energy intensive country. However, in terms of energy efficiency, South Africa performs poorly. In fact, it was ranked among the bottom 50 of the 150 countries compared in a study.[29] Therefore, energy costs form a large part of total production costs because the efficiency of utilisation is low. South Africa produces about 2% of the world's carbon emissions and thus requires a check in its carbon emissions.[30] With the threat of less energy created than the current energy demand, an energy crisis is looming in South Africa and energy efficiency is essential through incentives from the government and carbon tax adoption. South Africa committed itself at the Copenhagen Accord in 2009 that by 2020 it would reduce its greenhouse gas emissions to 34% below its projected emission value and by 2025 by 42%. By the look of its energy, environmental and industrial policies, South Africa is trying to achieve this commitment even though it is not legally binding, simply because of the numerous economic and resource advantages that can be achieved by being a low carbon economy.[31]

Refrigeration equipment (i.e. industrial, commercial and residential equipment) accounts for a sizeable chunk of national energy consumption. Unfortunately, in addition to the indirect effect of this equipment on the environment, most of the equipment in South Africa, and Africa in general, also still requires synthetic refrigerants. Ozone depleting and global warming refrigerants like R22 are still common in most refrigerators, while new equipment uses high GWP refrigerants like R134a and R404a.[32] Between 2005 and 2009 in South Africa, HCFCs such as R22 had the highest consumption of 25 759 t (81.4%), HFCs such as R134a had a consumption of 3.439 t (10.9%), HFC blends such as R404a of 1089 t (3.4%), methyl bromide of 747 t (2.4%) and bromochloromethane of 624 t (2%).[33] With high refrigerant leakage rates reported in the literature (between 10% and 15%),[26,34] it can be assumed that system leaks are also relatively high in South Africa. Therefore, direct emission effects are high too. South Africa, being a signatory to the Montreal Protocol, needs to reduce its HCFC consumption to 90% of baseline (2010 amount) by 2015, to 65% by 2020, to 32.5% by 2025, to 2.5% by 2030 and be completely phased out by 2040, while its methyl bromide consumption is supposed to be completely phased out by 2015.[33] To reduce the environmental effects due to refrigerant leakages, it is paramount that these synthetic refrigerants are replaced by more efficient natural and environmentally friendly ones. While these environmentally friendly refrigerants are getting the required attention and acceptance in Europe and Asia, application is still at the infancy stage in South Africa and Africa in general.[32]

Currently in South Africa, there are approximately 30 industrial and commercial installations of R744 refrigeration equipment and a negligible number of R744 residential and transport refrigeration installations. In a country where industrial and commercial refrigeration installations exceed 2000 and there are millions of residential and commercial refrigerators, it is clear how far behind we are in terms of green technology. Still, compared to the rest of Africa, South Africa leads with this green technology, therefore emphasising how the continent is lagging behind. The country's main power generating company (Eskom) has also encouraged the adoption of heat pumps and environmentally friendly refrigerators. This was brought about by its agreement to the carbon tax regime of reducing carbon emissions by 20% by 2025. The Department of Energy of South Africa forecast that for this target to be easily and smoothly achieved, alternative and efficient technologies need to be adopted.[35] To achieve carbon emission targets, Eskom has encouraged all stakeholders to reduce electricity usage by 40% by 2015 while for ozone depletion substances, the government introduced a policy of eliminating the use of R22 in all new commercial and industrial refrigerators.[36]

As per Eskom estimates, for every 1 kWh of electricity produced at the power station, 1.4 L of water and 530 g of coal are consumed. The pollutants emitted from the generation of 1 kWh of electricity include: 7.75 g of SO_2, 4.18 g of NO_x, 990 g of CO_2 and 157 g of ash. These estimates do not include the pollutants emitted while mining the coal, i.e. directly from the mine (methane which has a GWP of 20 is normally released during mining), indirectly from the mining equipment, from transporting the coal to the power station, and from establishing and maintaining the power station and mine infrastructure. By Eskom estimates, if a single household converts from electrical geysers to a normal water heat pump with a COP of about 3, approximately 355 kWh of electricity will be saved in 1 month. If the above pollutant estimates are used, it is possible to protect the environment from a large amount of pollutants. This becomes clearer if the estimated 5.4 million households which use electric geysers in South Africa all convert to heat pumps. Still, this is a conservative estimate of pollutants saved because it only considers power station production. In addition, if heat pumps with higher efficiencies are used, fewer pollutants will be released.[37]

In addition to being environmentally friendlier, the efficiency of operation of R744 refrigerators is also comparable (if not better) to conventional systems, thus they are also competitive in terms of indirect emissions, as reported in the literature.[32] In the first R744 refrigeration supermarket in South Africa (Woolworths), a 35% reduction of electricity consumption was achieved, resulting in much fewer pollutants being released.[36] The only setback is that most studies concerning R744 refrigeration and other environmentally benign alternatives have been done in countries with cold climates. Studies in warm tropical climates like in South Africa and Africa in general are scarce, especially in the open literature. It is therefore paramount that more studies are conducted in this field so as to ascertain the advantages of these systems in warm regions. With the existing R744 installations, the commercial and industrial owners have reported satisfactory performance to date and are motivated to install more of these refrigerators. Still, because of the perceived low efficiencies of R744 refrigerators in warm climates owing to their low critical point, additional studies are required in order to further improve the performance of these refrigerators and make them even more attractive.

Therefore, even though the use of R744 has solved the direct effect of emissions on the environment, if there are no studies to investigate the indirect effects of these systems, we might end up with inefficient systems consuming much energy, thus still affecting the environment.[38] Improvement of existing technical solutions and the formulation of more solutions to existing R744 refrigeration problems is vital if this technology is to be accepted by all, especially in areas with warm climates. Their installation and operating costs should be lower than that for conventional systems. Theoretical and experimental studies should be conducted on existing and new R744 systems in order to perfect this technology. System optimisation and modification are paramount if this technology is to completely replace conventional synthetic refrigerants. Also, national policies geared to encourage R744 refrigeration and other green technologies are important so as to clear the way and provide support for these technologies. In addition to research and industrial input, other stakeholders like the government and other policy organisations are important in facilitating the widespread use of these technologies.

Conclusion

Carbon dioxide as a refrigerant was explored from its historical background to specific properties which affect its performance in the refrigeration industry. As a result of its superior properties, especially concerning refrigeration, we believe R744 will be a dominant refrigerant in many applications of the refrigeration technology in the future.

Table 4: Some existing practical examples of NH_3 refrigeration[1,23,39-42]

Country	Application	Organisation and/or manufacturer
Global		
All	Commercial refrigeration	Coca-Cola vending machines since 2000
	Industrial refrigeration	Green and cool chillers Nestle new food processors since 2012
	Heat pumps	EcoCute
Europe		
All	Commercial refrigeration	Tesco supermarket new facilities since 2009
United Kingdom	Commercial refrigeration	Marks and Spencer’s retail market (White City) Harrods supermarkets (plus heat pump) Asda supermarkets new facilities since 2002 Sainsbury’s supermarkets new facilities since 2010 Booths supermarkets new facilities since 2010
	Industrial refrigeration	ABN Amro bank data systems coolers (London) Netto fresh meat warehouse
Denmark	Commercial refrigeration	Metro supermarket new facilities Fakta supermarket new facilities since 2011
	Industrial refrigeration	Netto fresh meat warehouses Netto central cold store (Arhus)
Switzerland	Commercial refrigeration	COOP supermarket new and rebuilt facilities since 2009
	Heat pumps	Le Locle hospital
Germany	Air-conditioning system	Konvekta buses since 1996 Berliner Verkehrsbetriebe (BVG) buses since 1996 (Berlin)
	Commercial refrigeration	Tengelmann Group supermarket (plus heat pump in Mulheim an der Ruh), and all the new facilities since 2008 Tegut…gute Lebensmittel supermarket (Lorsch), and all the new and rebuilt facilities since 2010 Edeka supermarket (North) Metro supermarket new facilities REWE supermarket new facilities since 2008 Aldi Sud food discount chain new facilities since 2010
	Industrial refrigeration	Carrier transicold NaturaLINE container since 2010 Hapag–Lloyd carrier container since 2010
Sweden	Commercial refrigeration	City gross supermarket (Rosengard, Malmo) ICA Kvantum supermarket (Varberg) and all other new and rebuilt facilities since 2010
Turkey	Commercial refrigeration	Carrefour hypermarket (Izmir and Istanbul)
Switzerland	Commercial refrigeration	Prodega Cash and Carry supermarket (plus heat pumps in St Blaise) Migros supermarket new facilities since 2002
Austria	Commercial refrigeration	Eurospar supermarket refrigerators (Klangenfurt and St. Gilgen)
France	Commercial refrigeration	Carrefour supermarket (Beaurans-les-Arras)
Asia		
All	Commercial refrigeration	Tesco supermarket new facilities since 2009
China	Air-conditioning system	2008 Olympic buses (Beijing)
	Commercial refrigeration	AEON retail markets new facilities since 2012
	Industrial refrigeration	Zhangzi sea food processing centre (Dalian)
	Heat pumps	Bumade railway station Wuhan University
Malaysia	Commercial refrigeration	AEON retail markets new facilities since 2012
Jordan	Industrial refrigeration	Jordan poultry plant (Amman)
Hong Kong	Commercial refrigeration	AEON retail markets new facilities since 2012
United States of America		
All	Commercial refrigeration	Tesco supermarkets new facilities since 2009
Philadelphia	Commercial refrigeration	Star Market supermarket (Chestnut Hill)
California	Commercial refrigeration	Supervalu supermarket (Albertsons, Carpinteria)
Maryland	Commercial refrigeration	Wegmans supermarket (Woodmore)
Canada		
Quebec	Commercial refrigeration	Sobeys supermarket new and rebuilt facilities (since 2006)
South America		
Brazil	Commercial refrigeration	Condor hypermarket (Curitiba) Verdemar supermarket and food distribution (Nova Lima)
	Domestic refrigeration	Metalfrio Solutions plug-ins
Oceania		
Australia	Commercial refrigeration	Drakes supermarket (Angle Vale and North Adelaide) Woolworths supermarket (Sydney, Melbourne and all the new stores) Coles supermarket (Ropes Crossing) Foodland IGA store (Adelaide)
New Zealand	Commercial refrigeration	Countdown Auckland supermarket Warehouse supermarket
Africa		
South Africa	Commercial refrigeration	Woolworths supermarket (Grey Owl, Midrand, 2009 and Claremont, Cape Town, 2010 and all the new stores and rebuilt stores) Pick n’ Pay supermarket (Johannesburg and Cape Town) Makro supermarket, Polokwane and Vaal (plus heat pump) and all the other new stores since 2011

Acknowledgement

We acknowledge the financial and logistical assistance of Eskom and the Tshwane University of Technology.

Authors' contributions

P.M. conducted the literature review under the guidance of Z.H. who was also the project leader. Both authors wrote the paper.

References

1. Bensafi A, Thonon B. Transcritical R744 (CO_2) heat pumps. Report no. 2414173. Villeurbanne: Centre Technique Des Industries; 2007.

2. Mohanraj M, Jayaraj S, Muraleedharan C. Environment friendly alternatives to halogenated refrigerants – A review. Int J Greenh Gas Con. 2009;3(1):108–119. http://dx.doi.org/10.1016/j.ijggc.2008.07.003

3. Hua T, Zhao Y, MinXia L, YiTai M. Research and application of CO_2 refrigeration and heat pump cycle. Sci China Ser E-Technol Sci. 2009;52(6):1563–1575. http://dx.doi.org/10.1007/s11431-009-0175-4

4. Sarkar J. Transcritical carbon dioxide heat pumps for simultaneous cooling and heating. Kharagpur: Indian Institute of Technology; 2005.

5. Kim MH, Pettersen J, Bullard CW. Fundamental process and system design issues in CO_2 vapor compression systems. Prog Energ Combust. 2004;30(2):119–174. http://dx.doi.org/10.1016/j.pecs.2003.09.002

6. Ma Y, Liu Z, Tian H. A review of transcritical carbon dioxide heat pump and refrigeration cycles. Energy. 2013;55:156–172. http://dx.doi.org/10.1016/j.energy.2013.03.030

7. Bodinus WS. The rise and fall of carbon dioxide systems. ASHRAE J. 1999;41(4):37–42.

8. Neksa P, Walnum HT, Hafner A. CO_2 – A refrigerant from the past with prospects of being one of the main refrigerants in the future. Paper presented at: The 9th IIR Gustav Lorentzen Conference; 2010 Apr 2–14; Sydney, Australia.

9. Lorentzen G. Revival of carbon dioxide as a refrigerant. Int J Refrig. 1994;17(5):292–301. http://dx.doi.org/10.1016/0140-7007(94)90059-0

10. Hoffmann G, Plehn W. Natural refrigerants for mobile air conditioning in passenger cars. Dessau: German Federal Environment Agency, Office GFEAP; 2010.

11. Hashimoto K. Technology and market development of CO_2 heat pump water heaters (ECO CUTE) in Japan. Borås, Sweden: IEA Heat Pump Centre; 2006.

12. Kolke GV. Natural refrigerants: Sustainable ozone- and climate-friendly alternatives to HCFCs. Eschborn: Deutsche Gesellschaft für Technische Zusammenarbeit GmbH (GTZ); 2008.

13. Nekså P, Rekstad H, Zakeri GR, Schiefloe PA. CO_2 heat pump water heater: Characteristics, system design and experimental results. Int J Refrig. 1998;21(3):172–179. http://dx.doi.org/10.1016/S0140-7007(98)00017-6

14. Girotto S, Minetto S, Neksa P. Commercial refrigeration system using CO_2 as the refrigerant. Int J Refrig. 2004;27(7):717–723. http://dx.doi.org/10.1016/j.ijrefrig.2004.07.004

15. Sawalha S. Carbon dioxide in supermarket refrigeration. Stockholm: Royal Institute of Technology; 2008.

16. Stene J. Residential CO_2 heat pump system for combined space heating and hot water heating. Trondheim: Norwegian University of Science and Technology; 2004.

17. Hwang Y, Radermacher R. Theoretical evaluation of carbon dioxide refrigeration cycle. HVAC&R Res. 1998;4(3):245–263. http://dx.doi.org/10.1080/10789669.1998.10391403

18. NIST. REFPROP V.6.0. NIST thermodynamic and transport properties of refrigerants and refrigerant mixtures database [database on the Internet]. No date [cited 2015 Aug 31]. Available from: http://www.boulder.nist.gov/div838/theory/refprop.htm.

19. Stene J. Integrated CO_2 heat pump systems for space heating and hot water heating in low-energy houses and passive houses. International Energy Agency (IEA) Heat Pump Programme – Annex 32 – Workshop; 2007 Dec 06; Kyoto, Japan. Paris: International Energy Agency; 2007. p. 1–14.

20. Liao SM, Zhao TS, Jakobsen A. A correlation of optimal heat rejection pressures in transcritical carbon dioxide cycles. Appl Therm Eng. 2000;20(9):831–841. http://dx.doi.org/10.1016/S1359-4311(99)00070-8

21. Reulens W. Natural refrigerant CO_2. Diepenbeek: Katholieke Hogeschool Limburg, Diepenbeek C; 2009.

22. Nekså P. CO_2 heat pump systems. Int J Refrig. 2002;25(4):421–427. http://dx.doi.org/10.1016/S0140-7007(01)00033-0

23. Shecco. 50 examples of natural refrigerant stories in article 5 countries [document on the Internet]. No date [cited 2015 Aug 31]. Available from: http://conf.montreal-protocol.org/meeting/oewg/31oewg/ngo-publications/Observer%20Publications/50%20Examples-Part1-of%20Natural%20Refrigerants%20Stories%20in%20Article%205%20Countries.pdf

24. Kullheim J. Field measurements and evaluation of CO_2 refrigeration systems for supermarkets. Stockholm: KTH School of Industrial Engineering and Management; 2011.

25. Hafner A, Nekså P, editors. Global environmental and economic benefits of introducing R744 mobile air conditioning. Paper presented at: The 2nd International Workshop on Mobile Air Conditioning and Auxiliary Systems; 2007 Nov 29–30; Orbassano, Italy.

26. Freléchox D. Field measurements and simulations of supermarkets with CO_2 refrigeration systems [MSc dissertation]. Stockholm: KTH Royal Institute of Technology; 2009.

27. Steimle F. CO_2 drying heat pumps. Essen: Institut fuer Angewandte Thermodynamik und Klimatechnik, Universitaet Esse; 1998.

28. Manzione JA, Neksa P, Halozan H. Development of carbon dioxide environmental control unit for the US Army. Paris: Institut International du Froid; 1998.

29. Department of Environmental Affairs and Tourism (DEAT). How energy generation causes environmental change in South Africa. Pretoria: DEAT; 2001.

30. Fawkes H. Energy efficiency in South African industry. J Energy South Afr. 2005;16(4):18–25.

31. Covary T. Development of 1st draft of a national energy efficiency action plan (NEEAP) for the Republic of South Africa. Johannesburg: Unlimited Energy; 2013.

32. Siegele B. Conversion of supermarket refrigeration systems from F-gases to natural refrigerants. Eschborn: Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, Proklima P; 2008.

33. United Nations. Goal 7: Ensure environmental sustainability. In: The Millenium Development Goals report. New York: United Nations; 2010. p. 52–64. Available from: http://www.un.org/millenniumgoals/pdf/MDG%20Report%202010%20En%20r15%20-low%20res%2020100615%20-.pdf

34. Likitthammanit M. Experimental investigations of NH_3/CO_2 cascade and transcritical CO_2 refrigeration systems in supermarkets. Stockholm: KTH School of Energy and Environmental Technology; 2007.

35. Blackharvest Trading. Energy efficiency CO_2 systems [document on the Internet]. c2010 [cited 2015 Aug 31]. Available from: http://www.doe-irp.co.za/hearing1/BLACKHARVEST_Carbon_Efficiency.pdf

36. Smith J. Good business journey: Retail sector. African Utility Week: Delivering Beyond Tomorrow 2013.

37. Eskom. Annual report. Johannesburg: Eskom; 2011.

38. Cottineau V. Calculation and comparison of different supermarket refrigeration systems. Stockholm: KTH School of Industrial Engineering and Management; 2010.

39. Riffat SB, Afonso CF, Oliveira AC, Reay DA. Natural refrigerants for refrigeration and air-conditioning systems. Appl Therm Eng. 1997;17(1):33–42. http://dx.doi.org/10.1016/1359-4311(96)00030-0

40. Heaney C, Swinard R, Pang A, West S. Natural refrigerants case studies. Melbourne: Australian Institute of Refrigeration, Airconditioning and Heating; 2007.

41. Kristensen AM. Natural refrigerants for new applications. Oslo: Nordic Chemicals Group (NKG), Kuldebrukere FF; 2008.

42. Mate J, Papathanasopoulos C, Latif S. Cool technologies: Working without HFCs. Amsterdam: Greenpeace; 2012.

Review of the stability of biodiesel produced from less common vegetable oils of African origin

AUTHORS:
Thomas Kivevele[1]
Zhongjie Huan[1]

AFFILIATION:
[1]Department of Mechanical Engineering, Mechatronics and Industrial Design, Tshwane University of Technology, Pretoria, South Africa

CORRESPONDENCE TO:
Thomas Kivevele

EMAIL:
kivevelethomas@gmail.com

POSTAL ADDRESS:
Department of Mechanical Engineering, Mechatronics and Industrial Design, Tshwane University of Technology, Private Bag X680, Pretoria 0001, South Africa

DATES:
Received: 04 Dec. 2014
Revised: 03 Feb. 2015
Accepted: 06 Feb. 2015

KEYWORDS:
oxidation stability; thermal stability; storage stability; antioxidants; fuel properties

HOW TO CITE:
Kivevele T, Zhongjie H. Review of the stability of biodiesel produced from less common vegetable oils of African origin. S Afr J Sci. 2015;111(9/10), Art. #2014-0434, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140434

The stability of biodiesel is dependent on storage conditions such as contact with ambient air and metals, exposure to sunlight and high temperature conditions which accelerate oxidation reactions. In addition, biodiesels are more susceptible to degradation when compared to fossil diesel because of the presence of unsaturated fatty acid chains which are prone to oxidation. The stability of biodiesel is categorised according to oxidation stability, storage stability and thermal stability. Oxidation instability can led to the formation of oxidation products such as aldehydes, alcohols, shorter chain carboxylic acids, insolubles, gums and sediments in the biodiesel. Thermal instability is concerned with the increased rate of oxidation at higher temperature, which in turn increases the weight of oil and fat due to the formation of insolubles. Storage stability is the ability of liquid fuel to resist change to its physical and chemical characteristics brought about by its interaction with its storage environment, such as contamination with metals. These fuel instabilities give rise to the formation of undesirable substances in biodiesel beyond acceptable limits as per global biodiesel standards such as those of the American Society for Testing and Materials (ASTM D6751) and European Standards (EN 14214). When such fuel is used in the engine, it impairs engine performance through fuel filter plugging, injector fouling, and deposit formation in the engine combustion chamber and various components of the fuel system. We review the stability of biodiesel made from less common vegetable oils of African origin and synthetic antioxidants used in improving the stability of produced biodiesels.

Introduction

Diesel fuel has been widely used in industry and in automobiles for over a century.[1] However, as the petroleum prices continue to rise, the diesel supply is becoming scarce and unreliable.[2,3] Also, because of environmental issues concerning the use of petro-diesel, the search for cleaner environmental fuels has increased in the past few decades.[4] Thus, the mono-alkyl esters of long chain fatty acids derived from renewable lipid feedstock, such as vegetable oils, animal fats and used cooking oils, also known as biodiesel, are well positioned to replace mineral diesel.[5-8] Biodiesel is a biodegradable, non-toxic biofuel, which possesses inherent lubricity. It reduces most regulated exhaust emissions and has a relatively high flash point in comparison to petroleum based diesel, making it safer than other fuels during transportation, storage and handling. In addition, the use of biodiesel reduces dependence on imported fossil fuels, which continue to decrease in availability and affordability.[1,9]

However, despite the advantages, biodiesel's chemical nature makes it more susceptible to oxidation in comparison to mineral diesel during long-term storage.[10] The sensitivity to oxidation varies depending on the fatty acid composition of the raw materials or feedstocks used for production of biodiesel; the presence of naturally occurring antioxidants; and the storage conditions, such as exposure to atmospheric oxygen, daylight, high temperatures and metals that have a catalytic effect and expedite the oxidation reaction. Poor oxidation stability of biodiesel is the central problem associated with its commercial acceptance.[10,11] Therefore, to enhance the practical feasibility of biodiesel, antioxidants are added to increase its storage stability. However, it is quite possible that these additives may also affect other basic fuel related properties of biodiesel.[12]

We therefore review the work done on the oxidation, thermal and storage stability of biodiesel produced from less common vegetable oils of African origin such as those from *Croton megalocarpus*, *Moringa oleifera*, *Jatropha*, manketti seeds, marula nuts and rubber seeds as well as neem oils. In addition, we provide the background and chemistry of various synthetic antioxidants used in improving the stability of biodiesel made from these vegetable oils.

Stability of biodiesel

The stability of biodiesel depends on the fatty acid profile of the parent feedstock, with the biodiesels with high unsaturated fatty acids content such as linoleic and linolenic acids being unstable compared to the ones containing saturated fatty acids.[11,13] The oxidative degradation of biodiesel affects some basic properties such as kinetic viscosity, cetane number and acid value. This fuel instability through oxidation can give rise to sediments and gum formation and fuel darkening.[10] As previously reported in the literature, the neat biodiesels are more prone to oxidation than the feedstocks or straight vegetable oils. The oxidised biodiesels can develop a wide variety of alcohols, aldehydes, peroxide, insolubles, gums and sediments which are formed during transport and long-term storage, causing acidity in the biodiesel.[10,14,15] The use of such oxidised biodiesel in engines can impair the performance of the engine because of possible fuel filter plugging, injector fouling and deposit formation in the engine combustion chamber and various components of the fuel system.[8,14] The decrease in stability of biodiesel is recognised by the increased iodine value, peroxide value and total acid number of either straight vegetable oils or methyl esters/biodiesels.[8,10] The lower stability of biodiesel when compared to that of straight vegetable oils is possibly because the antioxidants naturally present in the vegetable oils are either deactivated during the transesterification process or removed during the subsequent purification or separation procedures. Therefore, addition of synthetic antioxidants is imperative to increase oxidation stability of biodiesels for longer storage.[11] Generally, the stability of biodiesel is categorised according to oxidation, thermal and storage stability.[10] The specifications related to oxidation stability of biodiesels in the global biodiesel standards such as ASTM D6751 and EN14214 are summarised in Table 1.

Table 1: Specifications related to oxidative stability in biodiesel standards[16]

Specification	Method	ASTM[†] D6751	EN[‡] 14213	EN[‡] 14214
Oxidative stability (110 °C)	EN 14112	3 h (minimum)	4 h (minimum)	6 h (minimum)
Content of FAME≥ 4 double bonds (%m/m)		–	1 (maximum)	1 (maximum)
Linolenic acid content (%m/m)	EN 14103	–	–	12 (maximum)
Iodine value (g iodine/100 g)	EN 14111	–	130 (maximum)	120 (maximum)
Kinematic viscosity (mm^2/s)	D445; ISO 3104/3105	1.9–6.0	3.5–5.0	3.5–5.0
Acid value	D664; EN 14104	0.50 (maximum)	0.50 (maximum)	0.50 (maximum)

[†]American Society for Testing and Materials; [‡]European Standard

Measurement of oxidation stability

The oxidation stability of biodiesels without and with different dosages of antioxidants and its blends with mineral diesel are measured using Rancimat equipment as per EN 14112 specification of biodiesel oxidation stability. The working principle of the Rancimat instrument is illustrated in Figures 1 and 2. The biodiesel sample (10 mL) kept at constant temperature (110 °C) in the Rancimat is induced by passing a stream of purified air at a flow rate of 10 L/h over it. The vapours released during the oxidation process together with the air are passed through the flask containing distilled water, which contains an electrode for measuring the conductivity. The electrode is connected to a measuring and recording device which indicates the end of the induction period, when the conductivity of water begins to increase rapidly. The acceleration of conductivity is caused by the dissociation of volatile carboxylic acids produced during the oxidation process of biodiesel. These volatile organic acids are absorbed by the water. When the conductivity of this solution is recorded continuously, an oxidation curve is obtained (Figure 2) whose point of inflection is known as the induction period. This provides a good parametric value for oxidation stability.

Oxidative degradation chemistry

Figure 3 depicts the typical oxidation reactions of biodiesel.[10,17] Oxidation of biodiesel starts with the removal of hydrogen from a carbon atom to produce a carbon free radical. If diatomic oxygen is present, the subsequent reaction to form a peroxy radical is extremely fast. The peroxy free radical is not as reactive as the carbon free radical, but is sufficiently reactive to quickly abstract hydrogen from a carbon to form another carbon radical and a hydroperoxide (ROOH). The new carbon free radical can then react with diatomic oxygen to continue the propagation cycle. This chain reaction terminates when two free radicals react with each other to yield stable products like aldehydes, shorter chain carboxylic acids, insolubles, gum and sediments. As previously discussed, when biodiesel containing these oxidation products is used in the engine, it impairs engine performance.[10]

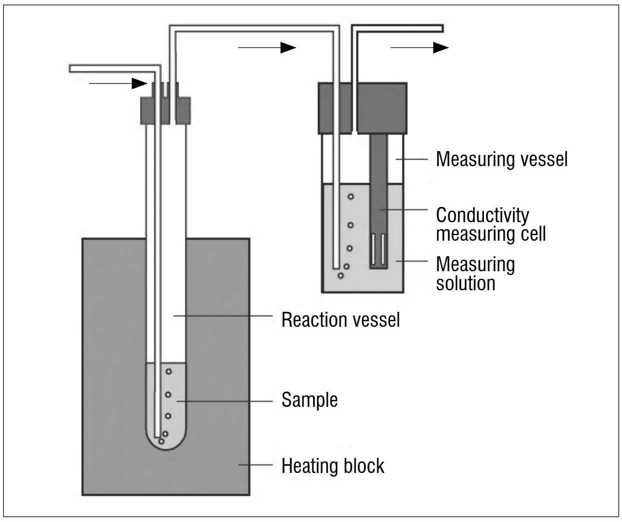


Figure 1: Principle of the Rancimat instrument.[11]

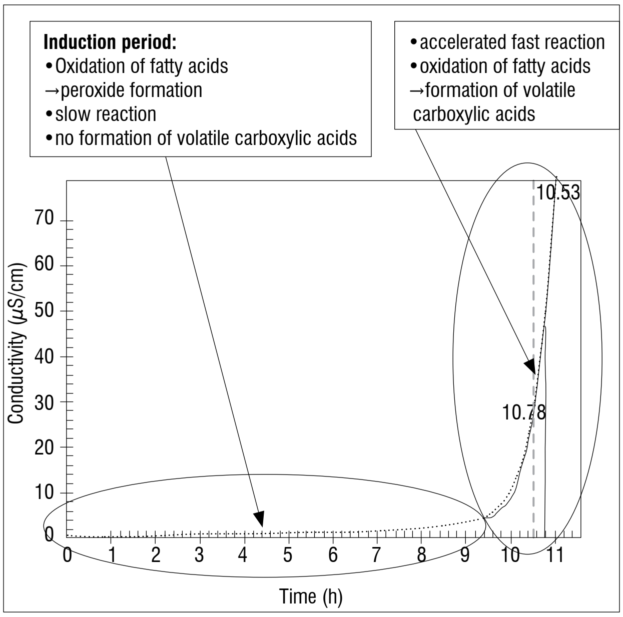


Figure 2: Typical curve obtained with a Rancimat instrument.

Initiation: $RH + I \rightarrow R^- + IH$

Propagation: $R^- + O_2 \rightarrow ROO^-$

$ROO^- + RH \rightarrow ROOH + R^-$

Termination: $R^- + R^- \rightarrow R\text{-}R$

$ROO^- + ROO^- \rightarrow$ stable products

Figure 3: Typical oxidation reaction of biodiesel.[10,11]

Thermal stability

Thermal stability of biodiesel fuel refers to the resistance of biodiesel fuel to oxidation when exposed to high temperatures. The increase in temperature significantly increases biodiesel oxidation, which consequently increases the oil weight due to the formation of insolubles.[18-20] Sarin et al.[21] measured the induction period of various biodiesels, including jatropha, by using Rancimat equipment at 100, 110 and 120 °C. They found that the oxidation stability of biodiesel from all sources decreased as the temperature increased, but no difference in relative stability was noticed. However, the oxidation stability of biodiesel from the oil which contained only a small fraction of unsaturated fatty acids and a large fraction of saturated fatty acids was found to be even better than that of the biodiesel from oil sources which contained a large fraction of unsaturated fatty acids, for example, jatropha and Karanja oils, which showed less oxidation stability.

Freire et al.[22] conducted a thermal investigation of oil and biodiesel produced from *Jatropha curcas* L. The jatropha biodiesel was synthesised by the transesterification reaction with ethanol, using oils from seeds of different crops and a homogeneous catalyst (KOH). Samples were named 2005/2006, 2006/2007, 2007/A and 2007/B, terms which were related to the harvest period and storage conditions. The results indicated that physic nut oil and ethyl biodiesel produced from different crops were thermally stable until 203 °C (2007/B) and 108.9 °C (2007A). The higher volatility of biodiesel, indicated by a lower initial decomposition temperature, certifies the quality of physic nut biodiesel as biofuel. The oil and biodiesel produced from the 2005/2006 harvest were less stable than the others because of the higher water content in the seeds. In general, the oxidative induction time values were 13 min for both oils and biodiesels, except for the 2007/B sample, which was 33 min. The quality of *Jatropha curcas* oils was dependent on how the seeds were dried, treated and stored.[23]

Measurement of thermal stability using thermogravimetric analysis

Thermogravimetric analysis is a test that is performed on a sample to determine changes in weight in relation to change in temperature. Such analysis relies on a high degree of precision in three measurements: weight, temperature and temperature change. Analysis is carried out by raising the temperature gradually and plotting the weight against the temperature. A derivative weight loss curve can be used to infer the point at which weight loss is most apparent (Figure 4). For example, a sample of biodiesel with or without antioxidant of approximately 5 mg is placed on a partially sealed pan and positioned into a platinum pan beam attached to the instrument. The air is purged at a rate of 130 mL/min. The temperature can be programmed to increase from 30 °C to 500 °C at a ramp rate of 10 °C/min. When oxidation takes place, the removal of secondary oxidation products causes a sudden weight loss in the sample. The onset temperature of oxidation can be obtained by the intersection of the extrapolated baseline and the tangent line of the curve, as shown in Figure 4 for the thermogravimetric measurement of croton oil methyl ester, which was found to be thermally stable up to 211.40 °C. It should be noted that, practically, a biodiesel which maintains its stability up to 150 °C can be regarded as thermally stable.[11]

Storage stability

Storage stability is defined as the relative resistance of a liquid fuel to physical and chemical changes brought about by interaction with its environment.[10,24] Storage instability occurs when liquid fuel or biodiesels interact with contaminants, light, factors causing sediment formation and other stress factors that accelerate the oxidation degradability of the fuel and reduce the cleanliness of the fuel.[10] The resistance of biodiesel to oxidation degradation during storage is an important issue for the viability and sustainability of an alternative fuel. Several studies related to the storage stability of biodiesel derived from less common tree-borne non-edible oil seeds under different conditions have been reported in the literature.[25-28] Sarin et al.[21] studied the influence of metal contaminants on the oxidation stability of jatropha biodiesel. Different metals were put into contact with jatropha biodiesel for a period of 6 months. The oxidation stability results indicated that copper contamination had the strongest detrimental and catalytic effect on the oxidation stability of biodiesel, where even a small concentration thereof showed nearly the same influence on the oxidation stability as that of the large quantities.

Das et al.[27] investigated the long-term storage stability of biodiesel produced from Karanja oil for 180 days under various conditions and reported that the oxidative stability of Karanja oil methyl ester (KOME) decreased; that is, the peroxide value and viscosity increased with the increase in storage time of the biodiesel. KOME samples were stored in different storage conditions such as dark or sunlight exposure, with air or without air exposure and with or without antioxidant additives, to assess the effect of storage conditions on oxidation stability and the most appropriate conditions for biodiesel storage. The samples stored under the condition of being 'open to air inside the room' had a high peroxide value and viscosity compared to those stored in other conditions, because the presence of air enhanced oxidation degradation. It was concluded that the long-term storage study gave a better understanding of the effect of the different storage conditions on the stability of biodiesel. This suggests that it is necessary to take special precautions during the storage of biodiesel, for example, limiting access to oxygen and exposure to light, metal and moisture.

Antioxidant chemistry

An antioxidant is a chemical that delays the start or slows the rate of the oxidation reaction.[29] It inhibits the formation of free radicals or interrupts the propagation of free radicals and hence contributes to the stabilisation of the biodiesel.[30] The two most common types of antioxidants are chain breakers and hydroperoxide decomposers. The most frequently used antioxidants at present are the chain breakers, which include phenolic types and aminic types.[31,32] The antioxidant contains a highly labile hydrogen that is more easily abstracted by a peroxy radical than fatty oil or ester hydrogen. The resulting antioxidant free radical is either stable or further reacts to form a stable molecule that does not contribute to the chain oxidation process. In this way, the chain breaking antioxidants interrupt the oxidation chain reaction.[11,21,31] Most of the previous studies on the stability of fatty acids and esters investigated applications of the phenolic type of antioxidants.[11,12,17,21] In esters and fatty acids, two common sources of antioxidants are natural antioxidants (α, β, γ and δ tocopherols) and the synthetic antioxidants.[21,30]

Table 2 depicts three widely used effective synthetic antioxidants for improving the oxidation stability of biodiesels derived from non-edible oils, with their chemical structures, as reported in the literature.[8,11,17,26,28] It should be noted that most of the less common vegetable oils of African origin used for biodiesel production are non-edible oils. In most studies, pyrogallol (PY) and propyl gallate (PG) were more effective than butylated hydroxyanisole (BHA) because they possess three hydroxyl (−OH) groups in their aromatic rings as shown in Table 2, while BHA has only one −OH group in its molecular structure. The −OH group of the antioxidant is very active so the hydrogen is abstracted from −OH and donated to the

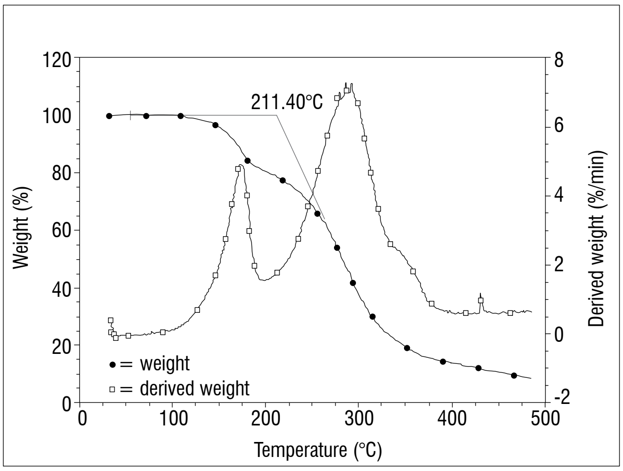
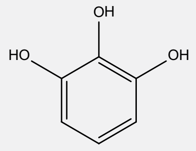
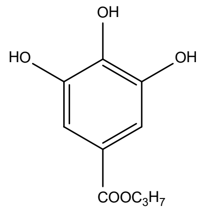
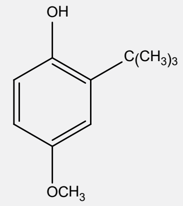


Figure 4: Typical thermogram for croton oil methyl ester.[11]

oxidised free radical to inhibit the rate of oxidation in methyl esters. The resulting antioxidant is a stable radical that can react with other fatty acid free radicals and further contribute to oxidation inhibition.[11]

Table 2: Synthetic antioxidants

Antioxidant name	Molecular structure
Pyrogallol: 1,2,3 tri-hydroxy benzene, 98%	
Propyl gallate: 3,4,5 tri-hydroxy benzoic acid, 99%	
Butylated hydroxyanisole: 2-tert butyl-4-methoxy phenol, 96%	

Stability of biodiesels from vegetable oils of African origin

Most of the less common vegetable oils of African origin are non-edible oils. They are derived from non-food feedstocks such as *Jatropha*, *Croton megalocarpus*, neem, *Moringa oleifera*, and rubber seed oils. Biodiesel produced from these non-edible oils (tree-borne non-edible oil seeds) is more economical compared to that produced from edible oils, which are more expensive than conventional diesel fuel, and eliminates the fuel versus food conflict.[11]

The challenge with regard to most non-edible oils is that they contain a high proportion of free fatty acids (FFAs) which, when they react with alkaline catalysts during the transesterification process, result in foam soap which prohibits the separation of biodiesel and glycerol. The soaps formed by the FFAs cause foaming in aqueous media which results in an increase in biodiesel viscosity.[33] The best method for reducing FFAs in non-edible oils is acid esterification; this is mainly a pre-treatment process for reducing the FFAs. The process converts FFAs to esters using an acid catalyst (H_2SO_4). The acid esterification process reduces the FFA concentration below 2% in the oil which is then recommended for the application of the one step alkaline transesterification method for biodiesel production in which the oil reacts with an alcohol (e.g. methanol) in the presence of a catalyst (e.g. KOH and NaOH).[34] The end products of the reaction are the fatty acid alkyl ester (biodiesel) and glycerine.[11] It should be noted that most of the biodiesel synthesised from non-edible oils of African origin are rich in unsaturated fatty acids which are prone to oxidation. It is therefore imperative that they are doped with antioxidants for longer storage.[21]

There have been several studies related to the production of biodiesel from less common vegetable oils of Africa origin (edible and non-edible oils) and their oxidation stability. The background of these feedstocks, the oil content of the seeds and the oxidation stability of synthesised biodiesels are discussed in the sections to follow.

Jatropha oil

Jatropha oil has been widely used for biodiesel production. It is mostly found in developing countries, especially in Africa and Asia. Jatropha plants have a high seed yield which can be continuously produced for 30–40 years. The oil content in the jatropha seeds is approximately 30–40% by weight.[21,23] Most researchers have found that jatropha oil can be used in biodiesel production as an alternative fuel in diesel engines and does not require major engine modification.[35] Free fatty acid content in jatropha oils is very high (approximately 14%) compared to those of other feedstocks.[34] Therefore, it requires two steps for biodiesel production (esterification followed by transesterification). The concern with biodiesel derived from jatropha is that it is rich in unsaturated fatty acid methyl esters[12,34] which are prone to oxidation.

Kivevele and Huan[12] and Kivevele et. al[17] studied the oxidation stability of biodiesel synthesised from jatropha oil of African origin. The fatty acid composition of produced jatropha oil methyl ester (JOME) was rich in unsaturated fatty acid (77.5%) with only 22.4% saturated fatty acid methyl esters. The neat JOME recorded oxidation stability of 5.85 h, meeting the minimum requirement of ASTM D6751 of 3 h. It is possible that the presence of naturally occurring antioxidants in the produced JOME favoured this reasonable oxidation stability. However, it did not meet the minimum requirement prescribed in the EN 14214 of 6 h as displayed in Table 1. Therefore, it was necessary to add antioxidants to increase oxidation stability for longer storage of JOME.

Amongst the antioxidants investigated were PY, PG and BHA. The overall performance of these antioxidants were in the order of PY>PG>BHA as previously discussed. It required only 200 ppm of PY and PG for JOME to meet the minimum requirement of ASTM D6751 (3 h) and EN 14112 (6 h). The South African standard for minimum requirement of oxidation stability of biodiesel is also 6 h (SANS 1935). Therefore, it can be seen that it is impossible to store JOME without antioxidant additives. JOME was observed to be thermally stable, displaying an onset temperature of 242.30 °C.

Moringa oleifera *oil*

Moringa oleifera is indigenous to the sub-Himalayan regions of northwest India, Africa, Arabia, Southeast Asia, the Pacific and Caribbean Islands and South America. It thrives best in a tropical insular climate and is plentiful near the sandy beds of rivers and streams. Moringa seeds contain between 33% and 41% (w/w) vegetable oil.[5,20] The potential of moringa oil of African origin as a feedstock for preparing biodiesel has been discussed previously in the literature.[36,37] However, few studies have reported on the stability of biodiesel derived from *Moringa oleifera* oil of African origin.

Kivevele and Huan[12] reported on oxidation stability of biodiesel synthesised from *Moringa oleifera* oil of African origin which was obtained from Tanzania and Kenya. Moringa oil methyl ester (MOME) is rich in 68.5% oleic acid methyl esters (C18:1) and 13.5% palmitic acid methyl esters (C16:0) with lower polyunsaturated fatty acid methyl esters (2.5%). Pure MOME recorded oxidation stability of 5.07 h, less than what is recommended in EN 14214 and the South African standard on oxidation stability of biodiesel (SANS 1935) of 6 h. The antioxidants were doped to increase the oxidation stability of MOME. Among the antioxidants investigated, PY and PG were more effective than BHA. Similar findings were previously reported in the literature on the oxidation stability of MOME.[17] It is also important to note that the most conspicuous property of MOME is its high cetane number of about 62.25, which is reported to be amongst the highest cetane numbers for a biodiesel fuel.[5] This is attributed to high saturated fatty acid methyl esters in its composition (26.5%). In addition, MOME was observed to be thermally stable, recording the onset temperature of 237.05 °C.

Rubber seed oil

Rubber trees have been widely used as a natural source of rubber, but now are regarded as one of the perfect biodiesel feedstocks in Nigeria as a result of their oil rich seeds.[38] Although there are variations in the oil content of the seed from different countries, the average oil yield has been reported to be around 40%. Rubber seed oil (RSO) contains 17–20% saturated fatty acids (myristic, palmitic, stearic, arachidic and behenic) and 77–82% unsaturated fatty acids.[38,39] It can be noted that RSO is rich in unsaturated fatty acids which are prone to oxidation. Therefore, for longer storage of biodiesel derived from RSO, it needs to be doped with antioxidants. Njoku et al.[38] reported that RSO recorded higher oxidation stability than rubber oil methyl ester (ROME). The oxidative stability of ROME was reduced possibly because of the trans-methylation method used in its production in which the natural occurring antioxidants in the RSO were either deactivated during trans-methylation process or removed during separation and purification procedures. However, it was concluded that rubber oils can be used to produce biodiesel fuel with similar properties to those of conventional diesel fuel and can be used directly in a diesel engine without major engine modifications.

Neem seed oil

Neem oil is a non-vegetable oil pressed from the fruits and seeds of neem trees (*Azardirachta indica*). The neem plant is a fast-growing and long-living tree, native to Myanmar and India, but now grown all around the world. In Africa, it is mainly found in West Africa (Nigeria). It is an evergreen tree growing in tropical and semi-tropical regions.[40,41] Neem oil comprises mainly triglycerides and large amounts of triterpenoid compounds, which are responsible for the bitter taste. Neem leaves in the form of powder are used as a herbal supplement in health care and in bio-pesticides in agriculture.[40] A mature neem tree produces 30 to 50 kg of fruits every year and has a productive life span of 150 to 200 years.[40] Neem seed has been reported to have a high oil content of about 39.7–60% by weight, which is a high yield desirable for a potential feedstock for biodiesel production.[40,41] Aransiola et al.[40] investigated biodiesel derived from neem nut oil and observed that neem seed oil exhibits a high FFA content (acid value of 32.538 mg KOH/g) which required two steps for biodiesel production (esterification followed by the transesterification process). The produced neem oil methyl ester (NOME) had a high percentage (44.5%) of monounsaturated fatty acids (C18:1); polyunsaturated acids (C18:2, C18:3) at 18.3% and 0.2%, respectively, which are prone to oxidation; and a controlled amount of saturated fatty acids (C16:0, C18:0) at 18.1% each. Although oxidation stability of NOME was not investigated in their study, the reported fatty acid profile indicates the instability of NOME, especially during long-time storage. Therefore, it is imperative NOME be doped with antioxidants.

Croton megalocarpus oil

Croton megalocarpus plants are indigenous to East Africa, and are widely found in the mountains of Tanzania, Kenya and Uganda.[42-44] They are used to make a good living fence while the leaves are used for mulch and green manure and the oil mostly in medicinal activities.[42] In recent years, it has been discovered that the oil from *Croton megalocarpus* seeds is a potential source for biodiesel production.[42] *Croton megalocarpus* seeds contain approximately 40–45% of oil on mass basis when extracted mechanically using a hydraulic press.[42]

There are several reports of biodiesel production from croton oil in the literature.[43-46] In most of the studies, it has been reported that croton oil methyl ester (COME) is rich in unsaturated fatty acid methyl esters. Kafuku and Mbarawa[44] reported that COME has 72.7% linoleic fatty acids and that because it is rich in unsaturated fatty acids, COME has remarkably cold flow properties. It yielded a cloud and pour point of −4 °C and −9 °C, respectively. These superior cold flow properties displayed by COME indicate that it is viable for use in cold regions.

Kivevele et al.[11] investigated the impact of various antioxidants on the oxidation stability of COME. The neat COME recorded an oxidation stability of 4.04 h, which did not meet the minimum requirement of oxidation stability prescribed in EN 14214 and SANS 1935 of 6 h. The presence of polyunsaturated fatty acid methyl esters of about 78.5% in total was the reason behind COME recording lower oxidation stability. To improve the oxidation stability of COME for longer storage, the effectiveness of various antioxidants was investigated. Among the antioxidants investigated, PY and PG were still the most effective antioxidants in improving the oxidation stability of COME. Only 200 ppm of PY and PG was required to increase the oxidation stability of COME above the minimum requirement prescribed in EN 14214 and SANS 1935. Similar observations on the oxidation stability of COME were reported in other studies.[8,12,13,17]

Manketti seed oils

The manketti tree (*Schinziophyton rautanenii*) occurs naturally in southern and western Zambia, where it is locally known as mungongo and is called manketti in Angola, Namibia, Botswana, western Zimbabwe and northern Mozambique. The edible oil extracted from manketti tree seeds is used locally in cooking, food preparation and personal care products. In addition, the seed oil has applications in modern cosmetic and personal care products, such as a body rub during dry winter months or as a skin cleanser and moisturiser because of its healing and nurturing properties. The land where the manketti trees are indigenous is not suitable for agricultural exploitation and all of the nuts are collected from the wild. The development of additional uses and external markets for this under-recognised oil seed could benefit the rural communities, provide a new export product for Africa, a new ingredient for the global cosmetic industry and an alternative fuel (biodiesel). There are few studies in the literature discussing manketti seeds oil as a possible source for biodiesel production. Juliani et al.[47] studied mungongo cold-pressed oil as a new natural product with potential cosmetic applications. They discovered that this oil is rich in unsaturated fatty acids – 25% linolenic acid (C18:3), 37% linoleic acid (C18:2), 15% oleic acid (C18:1), 8% palmitic acid (C16:0) and 9% stearic (C18:0) acid.

Kivevele and Huan[48] produced biodiesel from manketti seeds oil and evaluated its physical and chemical properties. The FFAs of manketti seeds oil was 1.57% which is below the 2% required for a one step transesterification process to produce manketti seeds oil methyl ester (MAME). Most of the fuel related properties of the produced MAME fulfilled the minimum requirement for biodiesel standards such as ASTM D6751, EN 14214 and SANS 1935. MAME showed slightly higher oxidation stability (4.75 h) which fulfilled the minimum requirement of ASTM D6751 (3 h), but did not meet the minimum requirement of EN 14214 and SANS 1935 of 6 h, due to their high percentage of methyl linoleate (45.6%) and methyl linolenate (20.3%) which are prone to oxidation. To improve the oxidation stability of MAME, the antioxidants were doped at 200, 500 and 1000 ppm dosage and tested in the Rancimat to observe effectiveness. It was found that oxidation stability increased with the increase in dosage of these antioxidants. Amongst the antioxidants used, PY and PG were found to be more effective than BHA at all dosages.

Marula nut oil

The marula tree (*Sclerocarya birrea)* is indigenous to most parts of southern Africa. The tree grows in warm and dry climatic conditions and produces oval fruits that turn pale yellow when ripe. The fruit consists of a hard woody seed covered by pulp and juice that make up the fleshy part of the fruit. The hard seed contains mostly two oil rich nuts (kernel) which can be eaten as a snack. However, small groups of rural communities in some parts of southern Africa are currently using the nut oil to produce cosmetic ointments.[49] There is now a worldwide trend to explore wild plants for oil to supplement the already existing sources of oil. The fact that the marula tree grows in drier parts where common oil seeds cannot thrive has stirred interest in marula nut oil as a valuable renewable source of energy. Studies exploring the potential use of marula nut oil as a potential source for biodiesel production are scarce. Gandure and Ketlogetswe[49] investigated the crude marula nut oil of Botswana’s climatic conditions as a possible source of biodiesel. The oil content of the marula nut was reported to be about 59%. This is a relatively high yield desirable as a potential feedstock for biodiesel production. The FFA content of the crude marula oil was 0.7%, which thus required a one step transesterification process during biodiesel production. The oil was rich in oleic fatty acid (about 70%); however, the oxidation stability of marula nut oil was not reported.

Mariod et al.[50] studied the synthesis of alkyl esters from three unconventional Sudanese oils for their use as biodiesel. Amongst the oils investigated was the Sudanese marula oil. It was observed that biodiesel synthesised from marula oil recorded remarkably high oxidation stability (27.1 h), meeting both global biodiesel standards (ASMT D6751 of 3 h and EN 14214 of 6 h). The higher oxidation stability recorded by Sudanese marula methyl ester possibly was because of the presence of naturally occurring antioxidants in crude marula oil and also reasonably high saturated fatty acids (25.2%) and lower polyunsaturated fatty acids (6.1%), which are more prone to oxidation than are monounsaturated fatty acids.

Conclusion

In the present study, we reviewed the stability of biodiesel synthesised from less common vegetable oils of African origin such as *Jatropha*, *Moringa oleifera*, *Croton megalocarpus*, rubber seed, manketti seed, neem and marula nut oils. Most of the biodiesels synthesised from these vegetable oils were observed to meet the minimum requirements prescribed in the global biodiesel standards such as ASTM D6751 and EN 14214. However, the outstanding issue is the oxidation stability. Most of the biodiesels are rich in unsaturated fatty acids which are prone to oxidation. The biodiesels recorded oxidation stabilities below the minimum requirements prescribed in EN 14214 and SANS 1935 of 6 h, except biodiesel made from marula nut oil. Therefore, it can be concluded that for longer storage it is imperative that these biodiesels be doped with antioxidants to increase oxidation stability. Among the antioxidants investigated in various studies on improving the oxidation stability of biodiesel derived from the reviewed vegetable oils of African origin were PY and PG, which were found to be the most effective.

Acknowledgement

We acknowledge funding from the National Research Foundation (South Africa) and Tshwane University of Technology.

Authors' contributions

T.K developed the concept, wrote the manuscript and performed the data analysis. Z.H was the project supervisor.

References

1. Demirbas A. Progress and recent trends in biofuels. Prog Energ Combust. 2007;33(1):1–18. http://dx.doi.org/10.1016/j.pecs.2006.06.001

2. Demirbas A. Biofuels sources, biofuel policy, biofuel economy and global biofuel projections. Energ Convers Manage. 2008;49(8):2106–2116. http://dx.doi.org/10.1016/j.enconman.2008.02.020

3. Demirbas A. Progress and recent trends in biodiesel fuels. Energ Convers Manage. 2009;50(1):14–34. http://dx.doi.org/10.1016/j.enconman.2008.09.001

4. Escobar JC, Lora ES, Venturini OJ, Yá-ez EE, Castillo EF, Almazan O. Biofuels: Environment, technology and food security. Renew Sust Energ Rev. 2009;13(6/7):1275–1287. http://dx.doi.org/10.1016/j.rser.2008.08.014

5. Rashid U, Anwar F, Moser, BR, Knothe G. *Moringa oleifera* oil: A possible source of biodiesel. Bioresour Technol. 2008;99(17):8175–8179. http://dx.doi.org/10.1016/j.biortech.2008.03.066

6. Monteiro MR, Ambrozin ARP, Lião LM, Ferreira AG. Critical review on analytical methods for biodiesel characterization. Talanta. 2008;77(2):593–605. http://dx.doi.org/10.1016/j.talanta.2008.07.001

7. Nabi MN, Rahman MM, Akhter MS. Biodiesel from cotton seed oil and its effect on engine performance and exhaust emissions. Appl Therm Eng. 2009;29(11/12):2265–2270. http://dx.doi.org/10.1016/j.applthermaleng.2008.11.009

8. Kivevele TT, Kristof L, Bereczky A, Mbarawa MM. Engine performance, exhaust emissions and combustion characteristics of a CI engine fuelled with *Croton megalocarpus* methyl ester with antioxidant. Fuel. 2011;90:2782–2789. http://dx.doi.org/10.1016/j.fuel.2011.03.048

9. Balat M, Balat H. A critical review of bio-diesel as a vehicular fuel. Energ Convers Manage. 2008;49:2727–2741. http://dx.doi.org/10.1016/j.enconman.2008.03.016

10. Jain S, Sharma MP. Stability of biodiesel and its blends: A review. Renew Sust Energ Rev. 2010;14(2):667–678. http://dx.doi.org/10.1016/j.rser.2009.10.011

11. Kivevele TT, Mbarawa MM, Bereczky A, Laza T, Madarasz J. Impact of antioxidant additives on the oxidation stability of biodiesel produced from *Croton megalocarpus* oil. Fuel Proc Technol. 2011;92:1244–1248. http://dx.doi.org/10.1016/j.fuproc.2011.02.009

12. Kivevele T, Huan Z. Effects of antioxidants on the cetane number, viscosity, oxidation stability, and thermal properties of biodiesel produced from nonedible oils. Energ Technol 2013;1:537–543. http://dx.doi.org/10.1002/ente.201300072

13. Kivevele TT, Mbarawa MM. Comprehensive analysis of fuel properties of biodiesel from *Croton megalocarpus* oil. Energ Fuel. 2010;24(11):6151–6155. http://dx.doi.org/10.1021/ef100880g

14. Monyem AH, Van Gerpen J. The effect of biodiesel oxidation on engine performance and emissions. Biomass Bioenerg. 2001;20(4):317–325. http://dx.doi.org/10.1016/S0961-9534(00)00095-7

15. Monyem A, Vangerpen JH, Canakci M. The effect of timing and oxidation on emissions from biodiesel-fueled engines. T ASAE. 2001;44(1):35–42. http://dx.doi.org/10.13031/2013.2301

16. Knothe G. Some aspects of biodiesel oxidative stability. Fuel Process Technol. 2007;88(7):669–677. http://dx.doi.org/10.1016/j.fuproc.2007.01.005

17. Kivevele TT, Agarwal AK, Gupta T, Mbarawa MM. Oxidation stability of biodiesel produced from non-edible oils of African origin. SAE Technical Paper 2011-01-1202; 2011. http://dx.doi.org/10.4271/2011-01-1202 http://dx.doi.org/10.4271/2011-01-1202

18. Dunn R. Effect of oxidation under accelerated conditions on fuel properties of methyl soyate (biodiesel). J Am Oil Chem Soc. 2002;79(9):915–920. http://dx.doi.org/10.1007/s11746-002-0579-2

19. Dunn RO. Effect of temperature on the oil stability index (OSI) of biodiesel. Energ Fuel. 2008;22(1):657–662. http://dx.doi.org/10.1021/ef700412c

20. Sharma B, Rashid U, Anwar F, Erhan S. Lubricant properties of moringa oil using thermal and tribological techniques. J Therm Anal Calorim. 2009;96(3):999–1008. http://dx.doi.org/10.1007/s10973-009-0066-8

21. Sarin R, Sharma M, Sinharay S, Malhotra RK. Jatropha-palm biodiesel blends: An optimum mix for Asia. Fuel. 2007;86(10/11):1365–1371. http://dx.doi.org/10.1016/j.fuel.2006.11.040

22. Freire L, Bicudo T, Rosenhaim R, Sinfrônio F, Botelho J, Carvalho Filho J, et al. Thermal investigation of oil and biodiesel from *Jatropha curcas* L. J Therm Anal Calorim. 2009;96(3):1029–1033. http://dx.doi.org/10.1007/s10973-009-0055-y

23. Berchmans HJ, Hirata S. Biodiesel production from crude *Jatropha curcas* L. seed oil with a high content of free fatty acids. Bioresour Technol. 2008;99(6):1716–1721. http://dx.doi.org/10.1016/j.biortech.2007.03.051

24. Bouaid A, Martinez M, Aracil J. Long storage stability of biodiesel from vegetable and used frying oils. Fuel. 2007;86(16):2596–2602. http://dx.doi.org/10.1016/j.fuel.2007.02.014

25. Sarin A, Arora R, Singh NP, Sharma M, Malhotra RK. Influence of metal contaminants on oxidation stability of jatropha biodiesel. Energy. 2009;34(9):1271–1275. http://dx.doi.org/10.1016/j.energy.2009.05.018

26. Jain S, Sharma MP. Long term storage stability of *Jatropha curcas* biodiesel. Energy. 2011;36:5409–5415. http://dx.doi.org/10.1016/j.energy.2011.06.055

27. Das LM, Bora DK, Pradhan S, Naik MK, Naik SN. Long-term storage stability of biodiesel produced from Karanja oil. Fuel. 2009;88(11):2315–2318. http://dx.doi.org/10.1016/j.fuel.2009.05.005

28. Jain S, Sharma MP . Effect of metal contaminants and antioxidants on the storage stability of *Jatropha curcas* biodiesel. Fuel. 2013;109:379–383. http://dx.doi.org/10.1016/j.fuel.2013.03.050

29. Rizwanul Fattah IM, Masjuki HH, Kalam MA, Hazrat MA, Masum BM, Imtenan S, et al. Effect of antioxidants on oxidation stability of biodiesel derived from vegetable and animal based feedstocks. Renew Sust Energ Rev. 2014;30:356–370. http://dx.doi.org/10.1016/j.rser.2013.10.026

30. Liang YC, May CY, Foon CS, Ngan MA, Hock CC, Basiron Y. The effect of natural and synthetic antioxidants on the oxidative stability of palm diesel. Fuel. 2006;85(5/6):867–870. http://dx.doi.org/10.1016/j.fuel.2005.09.003

31. Jain S, Sharma MP. Prospects of biodiesel from jatropha in India: A review. Renew Sust Energ Rev. 2010;14(2):763–771. http://dx.doi.org/10.1016/j.rser.2009.10.005

32. Liang YC, May CY, Foon CS, Ngan MA, Hock CC, Basiron Y. The effect of natural and synthetic antioxidants on the oxidative stability of palm diesel. Fuel. 2006;85(5/6):867–870. http://dx.doi.org/10.1016/j.fuel.2005.09.003

33. Demirbas A. Biodiesel fuels from vegetable oils via catalytic and non-catalytic supercritical alcohol transesterifications and other methods: A survey. Energ Convers Manage. 2003;44(13):2093–2109. http://dx.doi.org/10.1016/S0196-8904(02)00234-0

34. Kumar TA, Kumar A, Raheman H. Biodiesel production from jatropha oil (*Jatropha curcas*) with high free fatty acids: An optimized process. Biomass Bioenerg. 2007;31(8):569–575. http://dx.doi.org/10.1016/j.biombioe.2007.03.003

35. Sahoo PK, Das LM. Combustion analysis of jatropha, Karanja and polanga based biodiesel as fuel in a diesel engine. Fuel. 2009;88(6):994–999. http://dx.doi.org/10.1016/j.fuel.2008.11.012

36. Kafuku G, Lam MK, Kansedo J, Lee KT, Mbarawa M. Heterogeneous catalyzed biodiesel production from *Moringa oleifera* oil. Fuel Process Technol. 2010;91:1525–1529. http://dx.doi.org/10.1016/j.fuproc.2010.05.032

37. Kafuku G, Mbarawa M. Alkaline catalyzed biodiesel production from *Moringa oleifera* oil with optimized production parameters. Appl Energ. 2010;87(8):2561–2565. http://dx.doi.org/10.1016/j.apenergy.2010.02.026

38. Njoku OU, Ononogbu IC, Owusu JY. An investigation of oil of rubber (*Hevea bransiliensis*). Rubber Res Inst Sri Lanka. 1996;78:52–59.

39. Ikwuagwu OE, Ononogbu IC, Njoku OU. Production of biodiesel using rubber [*Hevea brasiliensis* (Kunth. Muell.)] seed oil. Ind Crop Prod. 2000;12(1):57–62. http://dx.doi.org/10.1016/S0926-6690(99)00068-0

40. Aransiola EF, Betiku E, Ikhuomoregbe DIO, Ojumu TV. Production of biodiesel from crude neem oil feedstock and its emissions from internal combustion engines. Afr J Biotechnol. 2012;11(22):6178–6186. http://dx.doi.org/10.5897/AJB11.2301

41. Ragit SS, Mohapatra SK, Kundu K, Gill P. Optimization of neem methyl ester from transesterification process and fuel characterization as a diesel substitute. Biomass Bioenerg. 2011;35(3):1138–1144. http://dx.doi.org/10.1016/j.biombioe.2010.12.004

42. Aliyu B, Agnew B, Douglas S. *Croton megalocarpus* (Musine) seeds as a potential source of bio-diesel. Biomass Bioenerg. 2010;34(10):1495–1499. http://dx.doi.org/10.1016/j.biombioe.2010.04.026

43. Aliyu B, Shitanda D, Walker S, Agnew B, Masheiti S, Atan R. Performance and exhaust emissions of a diesel engine fuelled with *Croton megalocarpus* (Musine) methyl ester. Appl Therm Eng. 2010;31:36–41. http://dx.doi.org/10.1016/j.applthermaleng.2010.07.034

44. Kafuku G, Mbarawa M. Biodiesel production from *Croton megalocarpus* oil and its process optimization. Fuel. 2010;89(9):2556–2560. http://dx.doi.org/10.1016/j.fuel.2010.03.039

45. Kafuku G, Lam MK, Kansedo J, Lee KT, Mbarawa M. *Croton megalocarpus* oil: A feasible non-edible oil source for biodiesel production. Bioresour Technol. 2010;101(18):7000–7004. http://dx.doi.org/10.1016/j.biortech.2010.03.144

46. Kafuku G, Tan KT, Lee KT, Mbarawa M. Noncatalytic biodiesel fuel production from *Croton megalocarpus* oil. Chem Eng Technol. 2011;34(11):1827–1834. http://dx.doi.org/10.1002/ceat.201100204

47. Juliani HR, Koroch AR, Simon JE, Wamulwange C. Mungongo cold pressed oil (*Schinziophyton rautanenii*): A new natural product with potential cosmetic applications. Acta Hort (ISHS). 2007;756:407–412. http://dx.doi.org/10.17660/actahortic.2007.756.43

48. Kivevele T, Huan Z. Mungongo seeds oil (*Schinziophyton rautanenii*) as a potential source of bio-diesel. Appl Mech Mater. 2014;472:780–784. http://dx.doi.org/10.4028/www.scientific.net/AMM.472.780

49. Gandure J, Ketlogetswe C. Chemical extraction and property analyses of marula nut oil for biodiesel production. Adv Chem Eng Sci. 2011;1:96–101. http://dx.doi.org/10.4236/aces.2011.13016

50. Mariod A, Klupsch S, Hussein IH, Ondruschka B. Synthesis of alkyl esters from three unconventional Sudanese oils for their use as biodiesel. Energ Fuel. 2006;20:2249–2252. http://dx.doi.org/10.1021/ef060039a

Systematic land-cover change in KwaZulu-Natal, South Africa: Implications for biodiversity

AUTHORS:
Debbie Jewitt[1,2]
Peter S. Goodman[1,2]
Barend F.N. Erasmus[2]
Timothy G. O'Connor[2,3]
Ed T.F. Witkowski[2]

AFFILIATIONS:
[1]Biodiversity Research, Ezemvelo KZN Wildlife, Pietermaritzburg, South Africa
[2]School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa
[3]South African Environmental Observation Network, Pretoria, South Africa

CORRESPONDENCE TO:
Debbie Jewitt

EMAIL:
Debbie.Jewitt@kznwildlife.com

POSTAL ADDRESS:
Biodiversity Research, Ezemvelo KZN Wildlife, PO Box 13053, Cascades 3202, South Africa

DATES:
Received: 20 Jan. 2015
Revised: 31 Mar. 2015
Accepted: 25 Apr. 2015

KEYWORDS:
conservation planning; habitat loss; intensity analysis; legislation; satellite imagery

HOW TO CITE:
Jewitt D, Goodman PS, Erasmus BFN, O'Connor TG, Witkowski ETF. Systematic land-cover change in KwaZulu-Natal, South Africa: Implications for biodiversity. S Afr J Sci. 2015;111(9/10), Art. #2015-0019, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20150019

Land-cover change and habitat loss are widely recognised as the major drivers of biodiversity loss in the world. Land-cover maps derived from satellite imagery provide useful tools for monitoring land-use and land-cover change. KwaZulu-Natal, a populous yet biodiversity-rich province in South Africa, is one of the first provinces to produce a set of three directly comparable land-cover maps (2005, 2008 and 2011). These maps were used to investigate systematic land-cover changes occurring in the province with a focus on biodiversity conservation. The Intensity Analysis framework was used for the analysis as this quantitative hierarchical method addresses shortcomings of other established land-cover change analyses. In only 6 years (2005–2011), a massive 7.6% of the natural habitat of the province was lost to anthropogenic transformation of the landscape. The major drivers of habitat loss were agriculture, timber plantations, the built environment, dams and mines. Categorical swapping formed a significant part of landscape change, including a return from anthropogenic categories to secondary vegetation, which we suggest should be tracked in analyses. Longer-term rates of habitat loss were determined using additional land-cover maps (1994, 2000). An average of 1.2% of the natural landscape has been transformed per annum since 1994. Apart from the direct loss of natural habitat, the anthropogenically transformed land covers all pose additional negative impacts for biodiversity remaining in these or surrounding areas. A target of no more than 50% of habitat loss should be adopted to adequately conserve biodiversity in the province. Our analysis provides the first provincial assessment of the rate of loss of natural habitat and may be used to fulfil incomplete criteria used in the identification of Threatened Terrestrial Ecosystems, and to report on the Convention on Biological Diversity targets on rates of natural habitat loss.

Introduction

Land-cover change and habitat loss are widely recognised as the major drivers of biodiversity loss in the world.[1-3] These changes not only fragment the landscape but alter biogeochemical cycles, climate, ecosystem processes and ecosystem resilience, thereby changing the nature of ecosystem services provision and human dependancies.[4-6] These losses and changes pose significant challenges for meeting biodiversity conservation goals and targets.

KwaZulu-Natal (KZN), a province situated on the eastern seaboard of South Africa, has a complex landscape, both in terms of its physical and biological diversity,[7] and the varied use and ownership of the landscape. The KZN landscape ranges from mountain climes of the Drakensberg escarpment of over 3000 m in the west to the subtropical climes of the Indian Ocean in the east (Figure 1) in an area of 93 307 km^2. KZN is the wettest of South Africa's provinces with a mean annual precipitation of 837 mm.[8] Consequently, agriculture – consisting primarily of sugar cane, orchards, commercial and subsistence crops, and timber plantations (agro-forestry) – represents major features of the landscape. The species-rich natural vegetation consists of mesic grasslands, savannas, forests and wetlands, and contains portions of the Maputaland-Pondoland-Albany biodiversity hotspot and the Midlands, Maputaland, Pondoland and Drakensberg Alpine centres of endemism.[9]

KwaZulu-Natal is the second most populous province in the country[10] with a mid-year population estimate of approximately 10.8 million people in 2011[11] (0.9 people per hectare). The province is experiencing a loss of natural habitat,[12] which has profound ecological consequences for this species-rich area. Similarly, the loss of natural capital and environmental degradation has socio-economic consequences for the many, mainly rural, inhabitants reliant on natural resources for fuel, fibre, food and medicine.[13] Many rural communities live on communally owned land, for which the drivers of change may differ from those on privately or state-owned land. It is thus important to quantify and understand the processes driving land-use and land-cover change (LULCC) in the province, and across different land tenure systems.

The availability of remotely sensed imagery has facilitated the monitoring of LULCC worldwide. In South Africa, two national land-cover (NLC) maps have been developed from satellite imagery based on *circa* 1994 (NLC 94)[14] and 2000 (NLC 2000)[15] conditions, but they are not directly comparable. Ezemvelo KZN Wildlife, the provincial conservation authority, has facilitated the development of three KZN land-cover maps based on 2005,[16,17] 2008[18,19] and 2011[20,21] conditions as part of its biodiversity monitoring mandate. These provincial data sets are valuable because they were developed using similar methodology, have similar legend categories and are mapped at the same resolution (20 m), making temporal comparisons more precise than less standardised land-cover maps. This series of five land-cover maps offers a valuable long-term period of 17 years within which to analyse land-cover change and rates of habitat loss within the province.

Understanding the patterns, processes and impacts of LULCC is essential in order to plan effectively for biodiversity conservation, especially in the face of other agents of global change such as climate.[22] Common methods of analysing land-cover change involve computing transition matrices between two points in time.[23] However, this method does not adequately account for category persistence, which tends to dominate the landscape. Failure to account for category persistence may mask important signals of land change.[24] The static state of the landscape between two time periods means that the signal of change is small in light of the overwhelming signal of persistence. Similarly, a lack of net change in a traditional analysis does not necessarily mean a lack of change on the landscape, because there could be location changes or swapping among categories. Thus an analysis that considers transitions of categories in terms of gains, losses, net change and swapping is insightful about patterns and processes of landscape change. Pontius et al.[24] developed a framework to account for these deficiencies. Further improvements to this method of analysis were developed in the Intensity Analysis framework[23] which was designed to analyse several points in time for the same study area. For each time interval, the method investigates the extent and speed of change and categorical gains and losses, whilst specifically considering the size and intensity of those changes and determining the intensity and variation of land-cover transitions from the categories available for the transitions. The framework thus identifies the underlying processes of the landscape transformations.

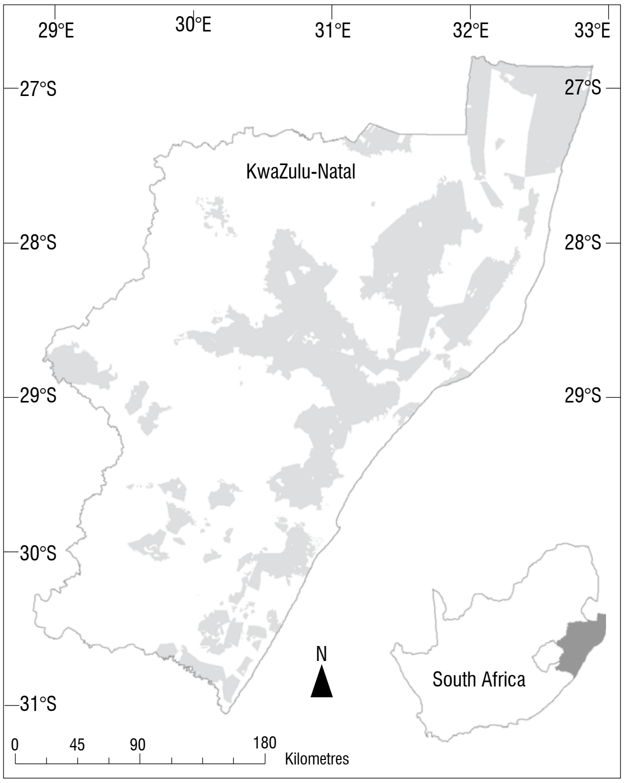


Figure 1: The location of KwaZulu-Natal, South Africa. The Ingonyama Trust Board administered areas are shown in grey, and were used as a proxy for communal areas.

Given the complex nature of KZN, it is essential to understand the drivers, patterns and processes of change for biodiversity conservation as the nature of the changes will have different management and policy implications. Using a quantitative method that addresses known inadequacies of conventional LULCC analyses and specifically assists in identifying the underlying processes of landscape change, and using the unique land-cover data set now available, should markedly improve our understanding of LULCC in KZN for conservation planning. Consequently, we have used the Intensity Analysis framework to characterise the systematic land-cover changes occurring in KZN using the three provincial land-cover maps (2005, 2008 and 2011). Differences in the pattern, rates and intensities of change are compared between land tenure systems (communal versus private and state-owned areas). In addition, the extent and rate of natural habitat loss are determined (from 1994 to 2011).

Methods

Communally owned lands

The historical legacy of the country has created three major land tenure systems in the province: communal, private and state-owned properties. The Ingonyama Trust was established in 1994 (*KwaZulu Ingonyama Trust Act No. 3 of 1994*) to hold land in title for members of communities in the province. The Ingonyama Trust Board (*KwaZulu-Natal Ingonyama Trust Amendment Act No. 9 of 1997*) administers the affairs of the Trust and the trust land and oversees the development of approximately 2.8 million hectares of communally owned land. The Ingonyama Trust Board (ITB) jurisdictional area was used as a proxy for communally owned land. Land-cover change differences were investigated between the ITB areas and the other land tenure systems (non-ITB).

Land-cover maps

Five land-cover maps were used in the analysis of land-cover change (Table 1). The details of methods used to develop these maps are dealt with in their associated documentation.[14-16,18,20] The methods in brief are as follows:

- The 1994 land-cover map was manually digitised from hard copy Landsat Thematic Mapper imagery based on 1994–1995 conditions, at a 1:250 000 scale but incorporating smaller features wherever feasible.[14]

Table 1: Land-cover map accuracy statistics, minimum mapping unit and number of classes for the national, KwaZulu-Natal (KZN) and aggregated class KZN land-cover maps. Initially nine categories were used in the analysis using the aggregated class maps, whereafter an ‘abandoned’ category was added and used in the Intensity Analysis.

Land-cover map	Overall map accuracy (%)	90% Confidence limits		Kappa index	Minimum mapping unit (ha)	Number of classes
		Low	High			
National 1994[12]	79.40	78.50	80.40	74.80	25	31
National 2000[13]	65.80	65.10	66.52	57.00	1–2	49
KZN 2005[4,5]	83.06	81.26	84.86	81.55	0.25	43
KZN 2008[6,7]	78.92	77.24	80.60	78.14	0.25	47
KZN 2011[8,9]	83.51	81.95	85.07	82.92	0.25	47
Aggregated classes KZN 2005	92.18	90.86	93.50	–	0.25	9;10
Aggregated classes KZN 2008	92.43	91.32	93.55	–	0.25	9;10
Aggregated classes KZN 2011	89.39	88.05	90.73	–	0.25	9;10

- The 2000 land-cover map was classified from multi-temporal Landsat 7 Enhanced Thematic Mapper imagery based on 2000–2003 conditions, although KZN formed part of phase 1 which used the earlier dated imagery.[15]
- The 2005 KZN land-cover map was developed from SPOT 2/4 imagery.[16,17] Certain post-classification modifications were made to improve the map, including the use of externally sourced data and expert edits.
- The 2008[18,19] and 2011[20,21] KZN land-cover maps were developed from SPOT 5 imagery. These maps represented temporal updates to the 2005 land-cover map.

The map accuracies ranged from 65.8% for the NLC 2000 to 83.5% for the provincial 2011 land-cover map (Table 1). Aggregating the classes of the provincial maps used for the change analysis significantly improved the map accuracies – up to 92.43% for the aggregated 2008 KZN land-cover map. The Kappa index for the provincial maps was high with the strength of classification agreements deemed ‘substantial’ and ‘almost perfect’ as per accepted benchmarks.[25] The aggregation of all five maps into only two categories for the long-term rate of habitat loss analysis would similarly significantly improve the accuracy statistics. Thus confidence was placed in accurately detecting change rather than error in this analysis. The imagery used to develop the provincial land-cover maps was provided as part of the South Africa Government/SANSA/SPOT IMAGE Agreement to supply annual SPOT imagery for the country.[26]

Data analysis

Detailed provincial analysis 2005–2011

The three provincial maps were analysed for land-cover change between 2005 and 2011 using IDRISI Selva.[27] The maps excluded the highly dynamic coastal sand and rock category and were standardised to the 2008 vegetation extent of the seashore line. The provincial boundary for this analysis includes the currently disputed Matatiele region in the southwest which is currently administered by the Eastern Cape but which was previously administered by KZN, and is included here for planning purposes only. Minor corrections were made to known dam and mine category errors. The maps were reclassified into 9 aggregated categories (Table 2; Supplementary figure 1 online) from the initial 43–47 land-cover categories and the associated aggregated accuracy statistics calculated from the accuracy assessment contingency tables (Table 3). The users’ accuracy exceeded 91% in almost all cases, but some categories had lower statistics in specific years. Aggregation of the categories served to improve the accuracy of the maps by eliminating errors among the more detailed land-cover categories (Table 1).

Based on initial analyses that detected changes, swapping and persistence[24] of categories in the landscape, an additional ‘abandoned’ category was created that specifically tracked changes of non-natural vegetation classes back into a semi-natural state at a future time point. It is imperative for conservation planning that these changes be tracked as this category does not hold the same biodiversity value as primary natural vegetation. Hence 10 categories were used in the Intensity Analysis.

The land-cover changes were examined using the modified transition (cross-tabulation) matrix and the hierarchical Intensity Analysis framework which uses statistical methods to identify the most important transitions and the signals of systematic processes related to the patterns of land change.[23,24] The relevant papers detail the methods used, hence they are not repeated in this paper but the equations and notation used are provided in the online supplementary material for ease of reference. These matrices were used to calculate the extent of gains, losses and swapping between categories. The Intensity Analysis considers the size of the category concerned and analyses the data at three levels of analysis, namely time interval, category gains and losses and transition intensities across available categories. The interval analysis determines the annual rate of change compared with a uniform change level across the temporal period of the analysis, and may be classed as slow or fast in comparison to the uniform change level. The category analysis investigates each time interval’s intensities of gains and losses per category and the categorical changes can be classed as dormant or active changes in comparison with the uniform intensity level. The transition analysis investigates transitions between particular gaining and correspondingly losing categories and vice versa to examine how the size and intensity of the transition varies. The transitions can be classed as targeted or avoided by comparing the observed intensity of each transition with a uniform intensity level.

Longer-term analysis of the rates of habitat loss

In addition to the provincial maps, the earlier national land-cover maps were used to investigate the amounts and rates of habitat transformation since 1994. In order to render the legend categories congruent, two categories were created across all five land-cover maps, namely untransformed (natural features and vegetation) and transformed (anthropogenically altered landscapes such as built infrastructure, cropland, plantations, mining and dams). Once an area had become transformed it was not permitted to become a natural category again at a future time point, effectively excluding the ‘abandoned’ category and thereby identifying primary natural areas best suited for biodiversity conservation. Data was resampled to a 500-m pixel size associated with the minimum mapping unit of 25 ha of the NLC 94, the coarsest level of mapping detail. A logarithmic regression curve was fitted to the temporal sequence of estimated remaining natural habitat in an attempt to best describe past pattern, and forecast the most likely state in 2050.

Table 2: The aggregated land-cover categories and a description of the categories included in the aggregated class

Aggregated land-cover category	Description
Water	Natural open water occurring in pans, rivers, wetlands, mangroves and estuaries
Plantations	Agro-forestry including clear-felled timber and rehabilitated plantation areas
Agriculture	Irrigated and dryland agriculture including permanent orchards, pineapples, sugar cane, subsistence agriculture, commercial annual crops and old cultivated fields
Mines	Major surface-based mineral, rock and sand excavation and dumping sites including rehabilitated mine areas
Built	All major urban and built-up areas, rural or low density dwellings, sports fields and race tracks, smallholdings, national, main and district roads, railways and airfields
Natural vegetation	Natural vegetation including forests, dense bush, bushland, woodland, bush clumps, grasslands, Alpine heath and degraded natural vegetation
Sand or rock	Naturally occurring exposed bare rock and sand, excluding coastal rock and sand
Erosion	Non-vegetated areas resulting from primarily gully erosion processes
Dams	Artificially impounded water
Abandoned	Secondary vegetation areas arising from abandoned non-natural categories, e.g. abandoned agricultural fields. From a biodiversity conservation perspective this category is tracked and separated in analyses because once abandoned, biodiversity value is never restored to its original state

Table 3: Aggregated class accuracy statistics for the KwaZulu-Natal 2005, 2008 and 2011 land-cover maps

Year	Category	Users' accuracy (%)	Producers' accuracy (%)	90% Confidence limits		Omission error	Commission error
				Low	High		
2005	Water	95.3	89.0	84.7	93.3	0.1	0.0
	Plantations	93.1	97.1	94.3	99.9	0.0	0.1
	Agriculture	92.9	91.7	88.7	94.6	0.1	0.1
	Mines	100.0	75.0	46.9	100.0	0.3	0.0
	Built	93.0	76.8	70.2	83.4	0.2	0.1
	Natural vegetation	91.2	95.7	94.2	97.2	0.0	0.1
	Sand or rock	75.0	75.0	46.9	100.0	0.3	0.3
	Erosion	100.0	100.0	97.3	100.0	0.0	0.0
	Dams	100.0	100.0	–	100.0	0.0	0.0
2008	Water	96.2	91.5	87.4	95.5	0.1	0.0
	Plantations	84.9	96.1	93.0	99.1	0.0	0.2
	Agriculture	93.2	91.8	89.7	94.0	0.1	0.1
	Mines	91.4	86.5	79.1	93.8	0.1	0.1
	Built	94.6	91.6	87.8	95.3	0.1	0.1
	Natural vegetation	94.6	95.7	94.2	97.3	0.0	0.1
	Sand or rock	92.3	60.0	45.8	74.2	0.4	0.1
	Erosion	80.6	80.6	71.4	89.9	0.2	0.2
	Dams	84.4	97.4	93.9	100.0	0.0	0.2
2011	Water	93.7	83.1	78.0	88.3	0.2	0.1
	Plantations	95.2	89.9	85.7	94.1	0.1	0.0
	Agriculture	95.3	93.1	90.8	95.4	0.1	0.0
	Mines	100.0	68.6	58.4	78.7	0.3	0.0
	Built	93.8	88.3	84.9	91.7	0.1	0.1
	Natural vegetation	78.8	97.3	95.8	98.7	0.0	0.2
	Sand or rock	100.0	28.0	16.3	39.7	0.7	0.0
	Erosion	91.3	84.0	74.4	93.6	0.2	0.1
	Dams	100.0	100.0	98.2	100.0	0.0	0.0

Results

Detailed provincial analysis 2005–2011

The percentage landscape change in KZN was 7.74% between 2005 and 2008, but slowed to 2.69% between 2008 and 2011 (Table 4, Figure 2). The greatest losses occurred in natural vegetation with a net loss of 721 733 ha (7.6%) since 2005. The greatest gains were made by agriculture with a net gain of 496 152 ha (5.2%) over the analysis period. Natural vegetation and agriculture were involved in the largest changes in the landscape in part because they accounted for a large part of the landscape. Importantly the agriculture and natural vegetation categories displayed high levels of swapping in the landscape (1.28% and 0.99%, respectively, between 2005 and 2008), i.e. changing to or from various categories over time. Commercial agriculture increased from 7.7% to 9.0% of KZN, driven primarily by dryland cropping, whilst subsistence agriculture increased from 3.3% to 7.4% in extent over the analysis period. The built environment had a net gain of 111 485 ha (1.2%) followed by plantations with 46 157 ha (0.5%).

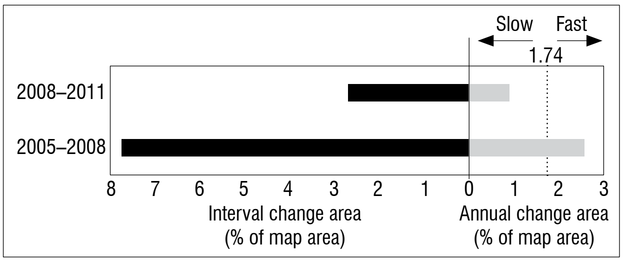


Figure 2: The landscape interval change occurring across KwaZulu-Natal between 2005–2008 and 2008–2011. The bars to the left (black) indicate the percentage area change occurring in the province in each interval, whilst the bars to the right (grey) represent the intensity of annual area of change within each time interval. Grey bars extending to the right or left of the vertical dashed line indicate a fast or slow change, respectively, relative to a uniform change across the analysis period.

Between 2005 and 2008 agriculture gained the most, followed by the built environment and plantations. Similarly, in the second time period (2008–2011), the major gains were made by agriculture and the built environment, but the gain in plantations slowed significantly. The natural vegetation category always showed the greatest losses. Figures 3 and 4 depict the annual size of the gain or loss, respectively, of a category on the left-hand side of the graph whilst the right-hand side indicates the intensity of the category gain and loss percentages relative to uniform change intensity across the landscape in general across the analysis period. Examining the intensity of the category gains and losses, which also considers the size of the category concerned, reveals that dams, mines and erosion were actively gaining categories in both time periods. The number of dams in the province increased from approximately 14 455 in 2005 to over 20 980 in 2011, representing a 45% increase in the number and a 26% increase in the extent of dams. Mining extent increased by 90% and erosion by 44%. In terms of losses, plantations was consistently a dormant category. The water and sand/rock categories were dynamic in nature.

Table 4: Percentage change in the aggregated land-cover categories in the KwaZulu-Natal landscape for 2005–2008 and 2008–2011. The gain in semi-natural vegetation (the change of non-natural vegetation classes back into a semi-natural state at a future time point) was tracked by the ‘abandoned’ category in the Intensity Analysis.

Category	2005–2008					2008–2011				
	Gain	Loss	Total change	Swap	Absolute value of net change	Gain	Loss	Total change	Swap	Absolute value of net change
Water	0.35	0.09	0.44	0.18	0.27	0.11	0.11	0.22	0.21	0.00
Plantations	0.71	0.22	0.92	0.44	0.49	0.18	0.18	0.36	0.36	0.00
Agriculture	4.92	0.64	5.56	1.28	4.28	1.26	0.30	1.56	0.60	0.96
Mines	0.04	0.00	0.04	0.01	0.03	0.01	0.00	0.02	0.01	0.01
Built	0.82	0.17	0.98	0.33	0.65	0.67	0.14	0.81	0.28	0.52
Natural vegetation	0.49	6.52	7.02	0.99	6.03	0.26	1.85	2.11	0.51	1.60
Sand or rock	0.01	0.03	0.04	0.03	0.02	0.02	0.06	0.08	0.04	0.04
Erosion	0.25	0.06	0.30	0.12	0.19	0.15	0.03	0.18	0.06	0.12
Dams	0.15	0.01	0.16	0.02	0.14	0.03	0.02	0.05	0.04	0.01
Total	7.74	7.74	7.74	1.69	6.04	2.69	2.69	2.69	1.05	1.63

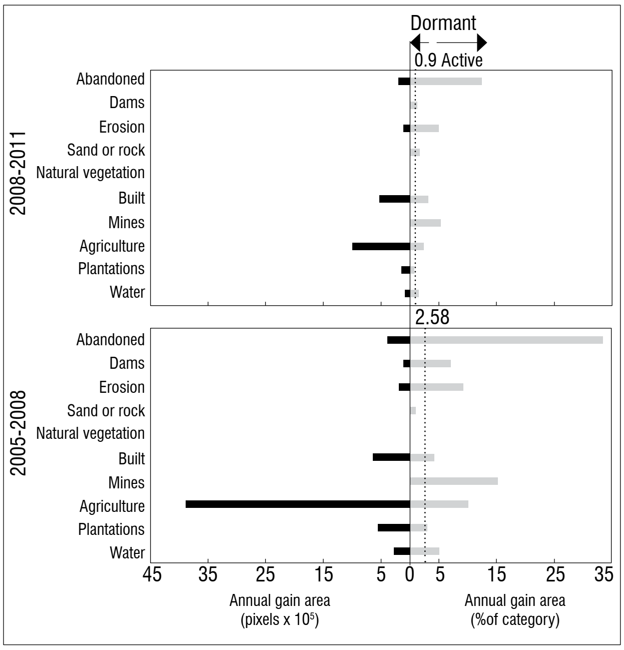


Figure 3: The gains per category for the 2005–2008 and 2008–2011 time intervals. The bars to the left (black) indicate the gross annual area gains per category. The bars to the right (grey) represent the intensity of the annual gains. Grey bars extending to the right or left of the vertical dashed line indicate active or dormant changes, respectively, relative to a uniform intensity across each analysis period.

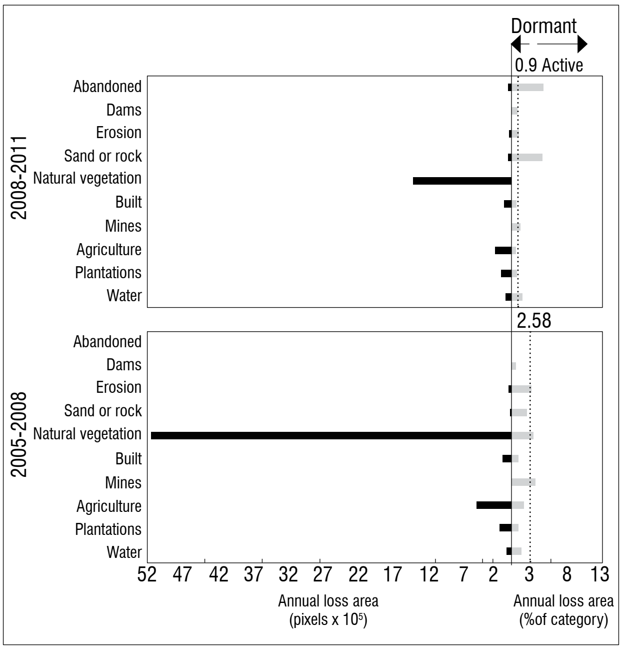


Figure 4: The losses per category for the 2005–2008 and 2008–2011 time intervals. The bars to the left (black) indicate the gross annual area losses per category. The bars to the right (grey) represent the intensity of the annual losses. Grey bars extending to the right or left of the vertical dashed line indicate active or dormant changes, respectively, relative to a uniform intensity across each analysis period.

Examination of the category transitions (Supplementary tables 2 and 3) reveals the abandoned category targeted agriculture, mines, built and dynamic natural categories such as erosion, sand and water. Agriculture targeted natural vegetation and erosion initially, but thereafter the abandoned, water, built and mine categories. The built areas targeted agricultural areas, but despite claiming an average of 7247 ha per annum of natural vegetation, cannot be said to have actively targeted this category, because of the large size and relative persistence of natural vegetation in the landscape. Erosion consistently targeted natural and abandoned vegetation and mines. Dams consistently targeted water, mines and erosion. Mines targeted natural and abandoned vegetation and dams.

The patterns of change in the communal (ITB) and non-communal (non-ITB) areas of the province followed similar patterns to those of KZN in that the rate of change slowed significantly in the second time period for both land tenure areas (Supplementary figures 2–7). A greater portion of the landscape changed in the ITB areas (10.84% and 3.41%) than in the non-ITB areas (6.45% and 2.39%) for both the first and second time periods, respectively. The major landscape differences between the ITB and non-ITB areas were that the communal areas practised a far greater degree of subsistence agriculture than commercial agriculture (30:1 versus 1:3 in non-ITB areas in 2008 with an increasing trend of subsistence agriculture). The ITB areas had a threefold higher proportion of low density settlements than high density settlements compared with non-ITB areas and the proportion of degraded natural vegetation was 50% higher in ITB areas than in non-ITB areas. The rate of increase in the built category was similar for both land tenure areas.

The ITB areas consistently gained in the abandoned, mining, agriculture, erosion and plantation categories, and the major losses stemmed from natural vegetation. The amount of swapping in the ITB landscape was 1.06% and 1.11%, respectively, in the first and second time periods. The non-ITB areas consistently gained in the agriculture, abandoned, built, dams, mining and erosion categories, and major losses stemmed from natural vegetation. The amount of swapping in the non-ITB landscape was 1.26% and 0.74%, respectively, in the first and second time periods.

Longer-term analysis of the extents and rates of habitat loss

In 1994, 73% of KZN was in a natural state. By 2011 this portion had decreased to 53% (Figure 5). The annual change percentage of the landscape decreased in each successive time period: 1.88% for 1994–2000, 1.05% for 2000–2005, 0.82% for 2005–2008 and 0.24% for 2008–2011. The average rate of habitat loss was 1.2% per annum between 1994 and 2011. A logarithmic regression function fitted the data well (adjusted $R^2=0.98$). Assuming habitat transformation continues in the same manner, it is estimated that by 2050, 45% of the landscape will remain in a natural state (Figure 6). Initially, the ITB areas had relatively more natural habitat than non-ITB areas; however, given the higher rate of change in the ITB areas, they are predicted to have less natural habitat remaining by 2050 than non-ITB areas.

Discussion

Biodiversity implications of land-cover changes

Landscape changes

The main drivers of change in the landscape were agriculture, timber plantations, built environments, mines and dams. Apart from the direct loss of natural habitat, these land covers all pose additional negative impacts for biodiversity remaining in these or surrounding areas. These effects may be direct (e.g. loss of habitat or extraction of water), indirect (e.g. pollution transported downstream), induced (e.g. associated industries and settlement) or cumulative (e.g. collective impacts on water quality and quantity).[28]

Land-cover change dynamics differed between ITB and non-ITB areas. The communal areas of the province experienced a proportionately greater degree of landscape change and development than private and state-owned areas. The ITB areas are predicted to have less natural habitat remaining than non-ITB areas in future which is problematic in that these communities rely heavily on natural resource use. Given the reliance on natural resources, the opportunity exists to promote the use of indigenous food and medicinal crops, which would benefit both biodiversity and lower the dependence on expensive agricultural inputs such as fertilisers. Low density settlement actively increased in ITB areas, posing challenges for service provision. Plantations actively increased in these areas compared with those in privately owned areas. There was a proportionally greater increase in the number of dams in the privately owned areas of the province.

Extensive land-use swapping occurred in the landscape, in particular between the agricultural and natural categories. Likely reasons for this swapping include agricultural field rotation common in subsistence

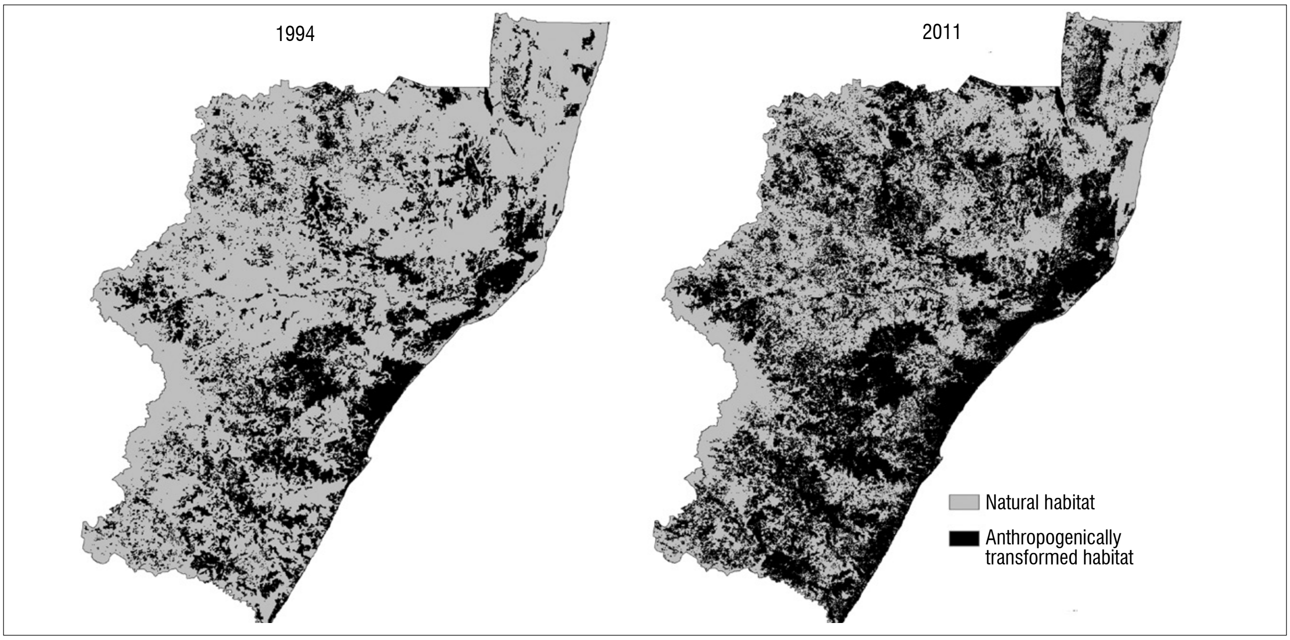


Figure 5: Accumulated transformation in KwaZulu-Natal from 1994 to 2011. The black areas represent anthropogenically transformed areas whilst the grey areas represent natural habitat.

farming and the abandonment of lands because of urbanisation, farm security issues, soil nutrient depletion, agricultural pests and diseases, invasive alien plants and economic factors. The transition analysis elucidated interesting change dynamics in the landscape, including, for example, swapping from built environments back to a natural environment, which is initially counter-intuitive. However, for diverse reasons, dwellings are often abandoned or become vandalised to the extent that vegetation overgrows the building foundations and it appears natural in later satellite images (Supplementary figures 8–22).

Agriculture

The average cultivated area per person in South Africa in 1960 was approximately 0.55 ha but this figure decreased to 0.3 ha by 1993.[29] In KZN in 2011, the commercial and subsistence agriculture equated to 0.14 ha per person, representing a significant decline over time despite a significant increase in agricultural extent. Increasing human populations will lower this ratio. Higher yields are possible from improved cultivars, irrigation, pesticide, herbicide and fertiliser use,[30] which may explain the smaller area required per person, but these inputs have negative impacts for biodiversity.

Agricultural expansion was pronounced prior to the 1960s. Policy instruments such as agricultural subsidies and minimum selling prices, were thought to have encouraged cultivation on marginal lands[29] which were later abandoned when subsidies were withdrawn. These old cultivated lands are still evident in the province but they are declining in extent, reverting primarily to agricultural use, in particular to subsistence and dryland cropping. However, these marginal areas are more prone to crop failure. The old croplands have altered soil structure, organic matter content and differing soil nutrient levels[31] and lack the full complement of native species, especially geophytic plants and those plants which rely on soil mycorrhizal associations, for example terrestrial orchids. It is not known how long it takes previously cultivated fields to return to a compositionally complete rangeland equivalent to primary rangelands, but it is estimated to be in excess of several decades.[32] This topic is worthy of further research.

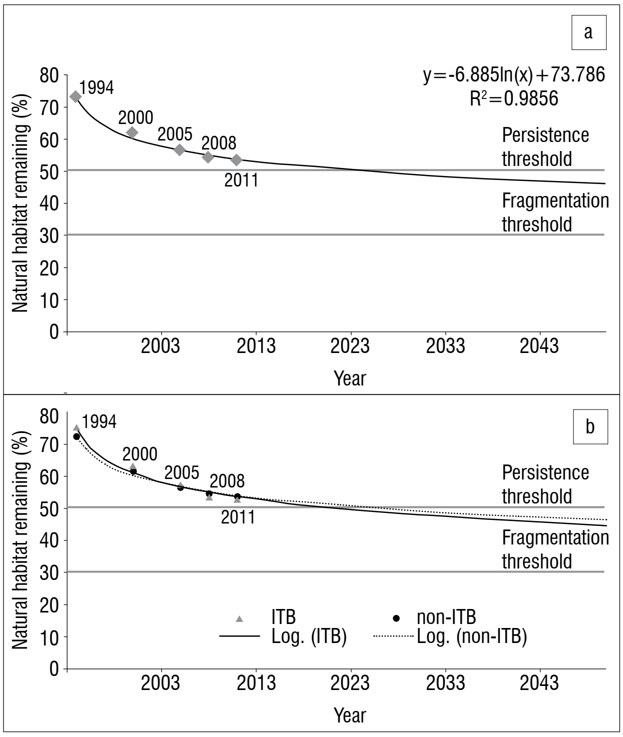


Figure 6: (a) Extrapolated rates of habitat loss in KwaZulu-Natal, assuming a business-as-usual scenario. The persistence threshold is reached once 50% of natural habitat is lost, beyond which there is a rapid decline in the probability of landscapes supporting viable populations. The fragmentation threshold is reached once 70% of natural habitat is lost, whereafter the spatial configuration of habitat patches becomes important for the persistence of remaining species.[43] (b) Extrapolated rates of habitat loss in the Ingonyama Trust Board (ITB) and non-ITB areas.

Plantations

Timber plantations occur primarily in the grassland and Indian Ocean Coastal Belt biomes. The extent of plantations has not increased significantly with a 46 000 ha increase in the first time period and a stabilisation in the second time period. The slowdown in the expansion of timber plantations in the second time period is most likely as a result of a reduction in the allocation of licences from the Department of Water Affairs and Forestry in terms of the *National Water Act No. 36 of 1998*, or because of economic factors associated with the industry. Indeed, certain catchments have been closed to new applications and a moratorium has been placed on others, pending further investigations on associated run-off reductions (Thambu D 2014, written communication, August 7). Plantations create acidic soils and an increase in available nitrate,[33] a situation for which many indigenous plant species are not adapted. Shading effects may promote shade-loving or forest species, but these species will be lost during rotational harvesting practices.[34]

Built environment

The built environment increased by 1.2% in KZN between 2005 and 2011. In particular there was an increase of the built environment in rural areas. Much of the province's biodiversity resides outside of protected areas in the rural landscape. Hence expansion in these areas poses threats for the remaining biodiversity. Sprawling urbanisation should be contained by the encouragement of higher-density settlements and the definition of an urban edge. An increase in the number of roads in rural areas is promoting development in remote areas, facilitating greater natural resource extraction and enhancing landscape fragmentation effects. Development in these areas reduces the opportunity for conserving large open spaces – which is one of the criteria used in protected area expansion plans.

Hydrological implications

The massive increase in the number of dams in the province is of significant concern for aquatic biodiversity and river health. The cumulative impacts of small dams reduces discharge, increases dissolved salts and alters macro-invertebrate indices.[35] Flow levels are reduced during dry periods, which causes hydromorphic grasslands to dry out and large trees to die back.[36] Larger dams and inter-basin transfer schemes significantly alter flow regimes and may lead to a dominance of livestock pest species.[37] Water extraction and pollution further negatively impact the ecosystem services that rivers and wetlands provide.

Mining

Mining extent almost doubled during the analysis period. The dominant form of mining affecting the change dynamics in the landscape is dune mining of titanium, iron, rutile and zircon, which occurs along the coast. The mobile nature of this form of mining creates a 'snail-trail' along the dune corridors with associated erosion, dam and abandoned category swapping. Mining impacts biodiversity principally via habitat loss, the alteration of ecological processes, pollution and the introduction of alien invasive species.[28]

Soil conservation

The extent of erosion is increasing in the province and creating degraded landscapes. Differing land-use practices may alter soil chemical and physical properties.[33] Ploughing, heavy grazing and burning deplete soil organic matter which affects soil water infiltration, retention and nutrient supply.[38] Dryland cropping, which is increasing in KZN, results in significant losses of soil organic matter. Lower soil organic matter results in lower water stable aggregates which are required to prevent soil erosion.[39] Future climate predictions suggest greater intensity of rainfall events and longer intervals between events.[40] Concomitant with the steep topography of KZN, soil erosion is thus likely to be exacerbated. Soil erosion has implications for biodiversity conservation, food security, soil conservation and water quality in terms of sedimentation and suspended sediment concentrations. Initiatives to prevent further soil erosion and degradation of natural vegetation are urgently required.

The implications of habitat loss

Extents and thresholds of habitat loss

This study highlighted the extensive loss of natural habitat occurring in the province, a massive 7.6% in only 6 years, which is of concern for biodiversity conservation and raises the question of whether this level of habitat loss is sustainable. At a national level the extent of transformed land in 2005 was 15.7%.[41] In KZN the picture is entirely different with 43% of land transformed in 2005, increasing to 46.4% by 2011. The changes in land cover, loss of habitat and the resulting fragmentation of the landscape have resulted in the loss of biodiversity and species population declines.[42] These losses will continue as more of the landscape is transformed by anthropogenic use as habitat is a finite resource, thus conservation efforts should focus on habitat preservation. As more natural habitat is lost, the opportunity costs associated with adding to the protected area estate increase. Certain areas in the landscape are unlikely to be transformed from their natural state because of, for instance, steep topography, protected areas or legislated development exclusion areas. Thus protected area expansion should focus on the areas most likely to experience a change to an anthropogenic category.

Flather and Bevers[43] identified a persistence threshold that exists once natural habitat is reduced below 50% of the total landscape for low degrees of patch aggregation. Beyond this level of transformation there is a rapid decline in the probability of landscapes supporting viable populations of organisms and a decline in habitat connectivity. The amount of natural habitat remaining in KZN is rapidly approaching this threshold. A target of no more than 50% of habit loss should be adopted to adequately conserve biodiversity in the province. Loss of habitat leads to a loss of ecological resilience and habitat specialist species.[44] This loss is of particular concern in this species-rich province and in the face of climate change for which ecological resilience is of paramount importance. A fragmentation threshold exists once 20–30% of natural habitat remains, whereafter the spatial configuration of habitat patches becomes important for maintaining population persistence.

Transformation of the landscape is creating ‘islands’ of protected areas in a matrix of anthropogenically transformed areas. This transformation is despite the province having good systematic conservation plans and data, which demonstrate that much of the biodiversity resides outside of protected areas. This situation calls into question the effectiveness of current conservation strategies and processes related to environmental authorisations. In light of extensive calls for further development in the province, a major rethink is required in order to determine how this development should be implemented. Effective management of the matrix is critical for the persistence of a vast majority of species that utilise these areas for breeding, foraging or migration.[45] New efforts to mainstream biodiversity into various land-use sectors are to be encouraged and supported.[46]

The science–policy interface

The Convention on Biological Diversity, to which South Africa is a signatory, has a target of halving (or where feasible bringing close to zero) by 2020, the rate of loss of natural habitat and significantly reducing degradation and fragmentation. The rates of habitat loss have declined in the province, but, given the recent global economic recession, caution needs to be exercised in interpreting the slowing rates as an achievement towards this target. Drivers such as changes in the economy, legislation, technological advances or even social resistance are likely to alter the rate of habitat transformation.

Some of the criteria used in the identification of Threatened Terrestrial Ecosystems (according to the *National Environmental Management: Biodiversity Act No. 10 of 2004* and the *National List of Ecosystems that are Threatened and in Need of Protection Act No. 1002 of 2011*) are incomplete because of data constraints. Specifically, criteria B – which examines the rate of loss of natural habitat – is not defined. Our data set and method of analysis provides the first provincial assessment of the rate of loss of natural habitat, and together with the Convention on Biological Diversity targets, could be used to fulfil this criterion of the legislation.

The main drivers of land-cover change are human responses to economic opportunities which are mediated by institutional factors.[6] Markets and policies constrain or encourage land-use change. Thus it is essential that decision- and policymakers are cognisant of the full implications that decisions and policy development may have on the rates of habitat loss. It is critical that a longer-term decision and planning framework, that is cognisant of constitutional and international agreements, be adopted. It is essential that conservation officials actively lead the way in biodiversity conservation, as stated eloquently by Noss et al.[47]:

> *The pro-growth norms of global society foster timidity among conservation professionals, steering them toward conformity with the global economic agenda and away from acknowledging what is ultimately needed to sustain life on Earth.*

Conclusions

The development of three directly comparable land-cover maps by Ezemvelo KZN Wildlife has permitted the first time-series analysis of LULCC in the province. The analysis elucidated the drivers, patterns and processes of land-cover change in KZN. The Intensity Analysis framework explicitly revealed change dynamics that other LULCC approaches would not have been able to do, by examining change at different levels of detail and considering category sizes and intensity of changes. This framework allowed a deeper understanding of systematic transitions in the province.

The challenge of conserving biodiversity in KZN is becoming increasingly difficult because natural habitat continues to be lost and the associated negative impacts and habitat degradation related to the identified land-cover drivers further threaten biodiversity. The provincial trends in habitat loss unfortunately follow global trends, but should this province, and South Africa, wish to lead the way in biodiversity conservation, some very important and difficult policy and legislative decisions need to be made now.

Acknowledgements

We are grateful to the anonymous reviewers for their valuable comments and suggestions. Ezemvelo KZN Wildlife is thanked for providing a study bursary. B.F.N.E. is supported by Exxaro. B. Naidoo and H. Snyman are thanked for their assistance with the graphs and maps, and J. Craigie and G. Nanni are thanked for their assistance with the accuracy assessment flights.

Authors’ contributions

D.J. was the project leader, researcher and wrote the manuscript; P.S.G., B.F.N.E., T.G.O.C. and E.T.F.W. made conceptual and editorial contributions.

References

1. Vitousek PM. Beyond global warming: Ecology and global change. Ecology. 1994;75(7):1861–1876. http://dx.doi.org/10.2307/1941591

2. Millenium Ecosystem Assessment. Ecosystems and human well-being: Biodiversity synthesis. Washington DC: World Resources Institute; 2005.

3. Jetz W, Wilcove DS, Dobson AP. Projected impacts of climate and land-use change on the global diversity of birds. PLoS Biol. 2007;5(6):1211–1219. http://dx.doi.org/10.1371/journal.pbio.0050157

4. Verburg PH, Neumann K, Nol L. Challenges in using land use and land cover data for global change studies. Glob Chang Biol. 2011;17:974–989. http://dx.doi.org/10.1111/j.1365-2486.2010.02307.x

5. Chapin FS, Zavaleta ES, Eviner VT, Naylor RL, Vitousek PM, Reynolds HL, et al. Consequences of changing biodiversity. Nature. 2000;405:234–242. http://dx.doi.org/10.1038/35012241

6. Lambin EF, Turner BL, Geist HJ, Agbola SB, Angelsen A, Bruce JW, et al. The causes of land-use and land-cover change: Moving beyond the myths. Glob Environ Change. 2001;11:261–269. http://dx.doi.org/10.1016/S0959-3780(01)00007-3

7. Jewitt D, Goodman PS, O'Connor TG, Witkowski ETF. Floristic composition in relation to environmental gradients across KwaZulu-Natal, South Africa. Austral Ecol. 2015;40(3):287–299. http://dx.doi.org/10.1111/aec.12213

8. Schulze RE, Lynch SD, Maharaj M. Annual precipitation. In: Schulze RE, editor. South African atlas of climatology and agrohydrology. Water Research Commission report 1489/1/06, Section 6.2. Pretoria: Water Research Commission; 2006.

9. Mucina L, Rutherford MC. The vegetation of South Africa, Lesotho and Swaziland. Pretoria: South African National Biodiversity Institute; 2006.

10. Statistics South Africa. Census 2011: Provinces at a glance. Report no. 03-01-43. Pretoria: Statistics South Africa; 2012.

11. Statistics South Africa. Mid-year population estimates 2011. Statistical release P0302. Pretoria: Statistics South Africa; 2011.

12. Jewitt D. Land cover change in KwaZulu-Natal. Environment. 2012;10:12–13.

13. Shackleton CM, Shackleton SE, Cousins B. The role of land-based strategies in rural livelihoods: The contribution of arable production, animal husbandry and natural resource harvesting in communal areas in South Africa. Dev South Afr. 2001;18(5):581–604. http://dx.doi.org/10.1080/03768350120097441

14. Fairbanks DHK, Thompson MW, Vink DE, Newby TS, Van den Berg HM, Everard DA. The South African land-cover characteristics database: A synopsis of the landscape. S Afr J Sci. 2000;96(2):69–82.

15. Van den Berg EC, Plarre C, Van den Berg HM, Thompson MW. The South African national land-cover 2000. Report no. GW/A/2008/86. Pretoria: Agricultural Research Council – Institute for Soil, Climate and Water; 2008 [unpublished report].

16. GeoTerraImage. KZN Province land-cover mapping (from SPOT2/4 satellite imagery circa 2005–06): Data users report and metadata (version 2). Pietermaritzburg: Ezemvelo KZN Wildlife; 2008 [unpublished report].

17. Ezemvelo KZN Wildlife. KwaZulu-Natal land cover 2005 v3.1 (clp_KZN_2005_LC_v3_1_grid_w31.zip) [GIS coverage]. Pietermaritzburg: Biodiversity Conservation Planning Division, Ezemvelo KZN Wildlife; 2011.

18. GeoTerraImage. 2008 KZN Province land-cover mapping (from SPOT5 satellite imagery circa 2008): Data users report and metadata (version 1). Pietermaritzburg: Ezemvelo KZN Wildlife; 2010 [unpublished report].

19. Ezemvelo KZN Wildlife. KwaZulu-Natal land cover 2008 v2 (clp_KZN_2008_LC_v2_grid_w31.zip) [GIS coverage]. Pietermaritzburg: Biodiversity Research and Assessment, Ezemvelo KZN Wildlife; 2013.

20. Ezemvelo KZN Wildlife, GeoTerraImage. 2011 KZN Province land-cover mapping (from SPOT5 satellite imagery circa 2011): Data users report and metadata (version 1d). Pietermaritzburg: Ezemvelo KZN Wildlife; 2013 [unpublished report].

21. Ezemvelo KZN Wildlife. KwaZulu-Natal land cover 2011 v1 (clp_KZN_2011_LC_v1_grid_w31.zip) [GIS coverage]. Pietermaritzburg: Biodiversity Research and Assessment, Ezemvelo KZN Wildlife; 2013.

22. Heller NE, Zavaleta ES. Biodiversity management in the face of climate change: A review of 22 years of recommendations. Biol Conserv. 2009;142:14–32. http://dx.doi.org/10.1016/j.biocon.2008.10.006

23. Aldwaik SZ, Pontius RG. Intensity analysis to unify measurements of size and stationarity of land changes by interval, category, and transition. Landscape Urban Plan. 2012;106(1):103–114. http://dx.doi.org/10.1016/j.landurbplan.2012.02.010

24. Pontius RG, Shusas E, McEachern M. Detecting important categorical land changes while accounting for persistence. Agric Ecosyst Environ. 2004;101:251–268. http://dx.doi.org/10.1016/j.agee.2003.09.008

25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174. http://dx.doi.org/10.2307/2529310

26. Van den Dool R. Exploring options for the national mosaic. PositionIT. 2011;(July):53–55.

27. Eastman JR. IDRISI Selva [computer program]. Worcester, MA: Clark University; 2012.

28. Mining and biodiversity guideline: Mainstreaming biodiversity into the mining sector. Pretoria: Department of Environmental Affairs, Department of Mineral Resources, Chamber of Mines, South African Mining and Biodiversity Forum, South African National Biodiversity Institute; 2013.

29. Biggs R, Scholes RJ. Land-cover changes in South Africa 1911–1993. S Afr J Sci. 2002;98:420–424.

30. Fresco LO. Some thoughts about the future of food and agriculture. S Afr J Sci. 2014;110(5/6), Art. #a0066, 2 pages. http://dx.doi.org/10.1590/sajs.2014/a0066

31. Cramer V, Hobbs R, Standish R. What's new about old fields? Land abandonment and ecosystem assembly. Trends Ecol Evol. 2008;23(2):104–112. http://dx.doi.org/10.1016/j.tree.2007.10.005

32. Roux ER. Plant succession on Iron Age I sites at Melville Koppies (Johannesburg). S Afr J Sci. 1970;66:48–50.

33. Mills AJ, Fey MV. Declining soil quality in South Africa: Effects of land use on soil organic matter and surface crusting. S Afr J Sci. 2003;99:429–436.

34. O'Connor TG. Influence of land use on plant community composition and diversity in Highland Sourveld grassland in the southern Drakensberg, South Africa. J Appl Ecol. 2005;42:975–988. http://dx.doi.org/10.1111/j.1365-2664.2005.01065.x

35. Mantel SK, Hughes DA, Muller NWJ. Ecological impacts of small dams on South African rivers. Part 1: Drivers of change – Water quantity and quality. Water SA. 2010;36(3):351–360.

36. O'Connor TG. Effect of small catchment dams on downstream vegetation of a seasonal river in semi-arid African savannah. J Appl Ecol. 2001;38:1314–1325. http://dx.doi.org/10.1046/j.0021-8901.2001.00680.x

37. Mantel SK, Muller NWJ, Hughes DA. Ecological impacts of small dams on South African rivers. Part 2: Biotic response – Abundance and composition of macroinvertebrate communities. Water SA. 2010;36(3):361–370.

38. Du Preez CC, Van Huyssteen CW, Mnkeni PNS. Land use and soil organic matter in South Africa 1: A review on spatial variability and the influence of rangeland stock production. S Afr J Sci. 2011;107(5/6), Art. #354, 8 pages. http://dx.doi.org/10.4102/sajs.v107i5/6.354

39. Du Preez CC, Van Huyssteen CW, Mnkeni PNS. Land use and soil organic matter in South Africa 2: A review on the influence of arable crop production. S Afr J Sci. 2011;107(5/6), Art. #358, 8 pages. http://dx.doi.org/10.4102/sajs.v107i5/6.358

40. Intergovernmental Panel on Climate Change (IPCC). Summary for policymakers. 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/New York: Cambridge University Press; 2007. Available from: http://www.ipcc.ch/publications_and_data/publications_ipcc_fourth_assessment_report_wg1_report_the_physical_science_basis.htm

41. Schoeman F, Newby TS, Thompson MW, Van den Berg EC. South African national land-cover change map. S Afr J Geomatics. 2013;2:94–105.

42. O'Connor TG, Kuyler P. Impact of land use on the biodiversity integrity of the moist sub-biome of the grassland biome, South Africa. J Environ Manage. 2009;90:384–395. http://dx.doi.org/10.1016/j.jenvman.2007.10.012

43. Flather CH, Bevers M. Patchy reaction-diffusion and population abundance: The relative importance of habitat amount and arrangement. Am Nat. 2002;159(1):40–56. http://dx.doi.org/10.1086/324120

44. Pardini R, Bueno AdA, Gardener TA, Prado PI, Metzger JP. Beyond the Fragmentation Threshold Hypothesis: Regime shifts in biodiversity across fragmented landscapes. PLoS ONE. 2010;5(10):e13666.

45. Franklin JF, Lindenmayer DB. Importance of matrix habitats in maintaining biological diversity. Proc Natl Acad Sci USA. 2009;106(2):349–350. http://dx.doi.org/10.1073/pnas.0812016105

46. Huntley BJ. Good news from the South: Biodiversity mainstreaming – A paradigm shift in conservation? S Afr J Sci. 2014;110(9/10), Art. #a0080, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0080

47. Noss RF, Dobson AP, Baldwin R, Beier P, Davis CR, Dellasala DA, et al. Bolder thinking for conservation. Conserv Biol. 2012;26(1):1–4. http://dx.doi.org/10.1111/j.1523-1739.2011.01738.x

Note: This article is supplemented with online only material.

Studies on CO variation and trends over South Africa and the Indian Ocean using TES satellite data

AUTHORS:
Abdoulwahab M. Toihir[1,2]
Sivakumar Venkataraman[2]
Nkanyiso Mbatha[3]
Sivakumar K. Sangeetha[4]
Hassan Bencherif[1]
Ernst-Günther Brunke[3]
Casper Labuschagne[3]

AFFILIATIONS:
[1]Atmosphere and Cyclones Laboratory, University of Réunion Island, Saint Denis, Réunion Island, France
[2]School of Chemistry and Physics, University of KwaZulu-Natal, Durban, South Africa
[3]South African Weather Service, Stellenbosch, South Africa
[4]School of Geography and Environmental Sciences, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Abdoulwahab Toihir

EMAIL:
fahardinetoihr@gmail.com

POSTAL ADDRESS:
Laboratoire de l'Atmosphère et des Cyclones UMR, 8105, 15 Avenue René Cassin, Université de la Réunion, CS 92003, 97744 Saint-Denis, Cedex, Réunion, France

DATES:
Received: 28 May 2014
Revised: 10 Dec. 2014
Accepted: 25 Apr. 2015

KEYWORDS:
Aura satellite; remote sensing; carbon monoxide; vertical profile; annual trend

HOW TO CITE:
Toihir AM, Sivakumar V, Mbatha N, Sangeetha SK, Bencherif H, Brunke EG, et al. Studies on CO variation and trends over South Africa and the Indian Ocean using TES satellite data. S Afr J Sci. 2015;111(9/10), Art. #2014-0174, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140174

In this study, we used measurements from the tropospheric emission spectrometer aboard the Earth Observing System's Aura satellite over South Africa, Madagascar and Reunion Island to investigate variations and trends in tropospheric carbon monoxide (CO) over 5 years, from 2005 to 2009, and at 47 pressure levels from 1000 hPa to 10 hPa. We believe that the study is the first of its kind to address the use of space-borne data for CO distribution over southern Africa. Maximum CO was recorded during spring and minimum during summer. Positive anomalies were identified in 2005 and 2007 during the spring and negative anomalies in the beginning of the year (especially in 2006, 2008 and 2009). The estimated trends based on a linear regression method on inter-annual distribution predicted a decreasing rate of 2.1% per year over South Africa, 1.8% per year over Madagascar and 1.7% per year over Reunion Island. The surface CO measurements made at Cape Point station (34.35°S, 18.48°E) showed an average decline of 0.1 ppb per month, which corresponded to 2.4% of the average annual mean for the studied period. The observed decrease in CO was linked to the La Niña event which occurred in 2006 and 2008 and a declining rate of biomass burning activity in the southern hemisphere over the observation period. TES measurements are in agreement with ground-based measurements and can be used with confidence to complement CO measurements for future analyses over the southern tropics and middle latitude.

Introduction

Carbon monoxide (CO) is the subject of our study and it is produced by incomplete combustion of hydrocarbons like fossil fuel, biofuel and methane (CH_4) with insufficient oxygen, which prevents complete oxidation to carbon dioxide (CO_2). CO is a reactive gas that regulates the concentrations of several greenhouse gases in the atmosphere (e.g. ozone, methane and carbon dioxide) and plays an important role in tropospheric photochemical processes. Its chemistry is relatively simple and lifetime relatively long (from weeks to months, depending on the region and on the season), enough for one to follow pollutant movements and dispersions on a regional and global scale.[1] Thus, the present study on tropospheric distribution and variability of CO is important for monitoring and evaluating photochemical processes and transport of tropospheric traces gases.

There are different remote sensing instruments currently in use for retrieving CO information from various atmospheric layers. Among the means of remote sensing, observations include measuring devices from ground-based instruments on board commercial aircraft and high technology instruments aboard satellites. Within the framework of observations by satellite, Earth Observing System (EOS) is the biggest world programme for earth monitoring and observation. This programme coordinates several polar-orbiting and low inclination satellites measuring atmospheric parameters and particles at local and global scale. Among the EOS satellites, there is EOS Aura, which contains the tropospheric emission spectrometer (TES) instrument measuring carbon monoxide in the atmosphere.[2,3] With good spatial coverage, satellite measurements offer the best method for providing CO measurements over the globe; however, their height and temporal resolution is poor in comparison to most ground-based instruments. It is therefore necessary to compare the satellite measurements with ground-based and/or mathematical numerical models.

A previous study by Kopacz et al.[4] from February to April 2001, used the atmospheric chemical transport model in order to validate and compare Asian CO sources detected by Measurements of Pollution in the Troposphere (MOPITT). Similarly, Luo et al.[5] used a simulated concentration of CO from the Global Earth Observation System (GEOS) chemistry global three dimensional models as a time varying synthetic atmosphere in order to demonstrate and assess the capabilities of TES nadir observations. It is important to consider that two instruments from the same programme may not provide identical concentrations of tropospheric CO due to their different characteristics and modes of observation. A previous comparison study between TES and MOPITT CO measurements showed significant differences in the lower and upper troposphere, which led to a review of the instruments and to adopt adjustment and correction methods in order to improve the agreement between them.[6] Heald et al.[7] illustrated that MOPITT and Transport and Chemical Evolution over the Pacific aircraft programme (TRACE-P) CO observations are independently consistent to provide information about the Asian CO emission sources, with the exception of Southeast Asia. Richards et al.[8] performed a study on TES CO measurements that was assimilated into the global chemistry transport model. Their work has shown that the assimilated data have a good agreement with MOPITT observations, especially at higher pressure levels in the southern tropics region. In some cases, CO measurements from different instruments are combined in order to estimate a global emission with accurate quantification. This is the case with an adjoining inversion method of multiple satellites (MOPITT, Atmospheric Infrared Sounder (AIRS), Scanning Imaging Absorption spectrometer for Atmospheric Chartography (SCIMACHY) and TES) used by Kopacz et al.[9] to investigate global carbon monoxide distribution sources with a high spatial resolution (4°×5°**)** and a monthly temporal resolution. A comparative study of CO total column from three satellites' sensors – AIRS, MOPITT and Infrared Atmospheric Sounding Interferometer (IASI) – and two ground-based spectrometers, one in central Moscow and the other in its suburb Zvenigorod, was made during summer when forest fires are expected; the results revealed that all of the instruments were in good agreement during fires and when there were no forest fires.[10]

Surveys on air pollution have previously been made in the Indian Ocean region using remote sensing tools such as the AIRS instrument, which investigated carbon monoxide variation over the Malaysian peninsula that had originated from Indonesian forest fires.[11] However, no similar scientific studies on carbon monoxide investigations using TES measurements have been performed. Previous studies have shown that the west southern Indian Ocean region seasonally receives CO from Africa and South America[9,12], south western Asia[12] and India[9,13]. However, the African and southern American contributions were observed below 11 km, while the south western Asia contribution was observed in the upper troposphere.[12] In addition, these previous studies have reported maximum CO in Africa and the Indian Ocean region during the biomass burning period from July to November. Khalil and Rasmussen[14] reported that from 1988 to 1992, the global CO began to decrease at a rate of 2.6±0.8% per year, probably due to reduction in tropical biomass burning. Similarly, Novelli et al.[15] confirmed that CO emissions have significantly reduced from 1990, especially in the southern hemisphere, whereas in the northern hemisphere, CO decreased at an average rate of 6.1% per year from June 1990 to June 1993. Zeng et al.[16] reported that the trend of CO in the southern hemisphere from 1997 to 2009 was negative due to a decline in industrial emissions.

The present study is focused on using TES data in order to investigate the inter-annual CO total column distribution and trends, as well as to explain monthly and seasonal CO vertical profiles over South Africa and the Indian Ocean (Madagascar and Reunion Island) region. These two regions have different climatology but have several similarities due to their geographical proximity. The present work does not only provide insight into CO distributions within the region under investigation, but also explores the capability of the TES instrument to complement the CO measurement in the geographic zones**.** We also validate the TES data by performing a comparison between the TES measurements recorded at ~1000 hPa when the satellite overpasses closest to the Cape Point Global Atmosphere Watch (GAW) station (34.35°S, 18.48°E). The ambient surface CO measured at Cape Point during the TES overpasses are used for comparative purposes.

Data sources

TES instrument

The tropospheric emission spectrometer (TES) is an imaging infrared Fourier transform spectrometer in polar sun-synchronous orbit aboard the EOS Aura satellite, launched on 15 July 2004. TES is dedicated to measure from nadir and limb, viewing the infrared radiance emitted by the earth's surface, atmospheric gases and particles from space. TES measurements are routinely performed from 650 to 3050 cm^{-1} spectral range (3.3–15.4 μm).[2] The apodised spectral resolution for standard TES observation is 0.1 cm^{-1}; however, finer spectral resolution (0.025 cm^{-1}) is used during special observations. In this present work, we are interested only in CO observation from nadir viewing in which the CO information is retrieved from space on the 2000–2200 cm^{-1} absorption band.[3] Nadir observations have a footprint of 5 km × 8 km, averaged over detectors in which each pixel covers a spatial resolution of 0.5 km × 5.3 km. TES uses both the natural thermal effects issued by the atmosphere and sunlight reflected from ground, allowing night and day observations. The routine observations performed by TES are referred to as 'a global survey'. A global survey is run every other day on a predefined schedule and collects 16 orbits (~26 h) of continuous data. A series of repetitive units, referred to as a sequence, is performed in each orbital track. The possible maximum number of sequences per run is 72 for 26 h of observation. Thus, 1152 profiles are provided at global scale during these 26 h of observation. For the purpose of the present study, only profiles recorded for overpasses over South Africa, Madagascar and Reunion Island have been selected and used. The geographic locations of these three selected sites are indicated in Figure 1. (For further information on the TES instrument, the reader may visit the web page http://tes.jpl.nasa.gov.)

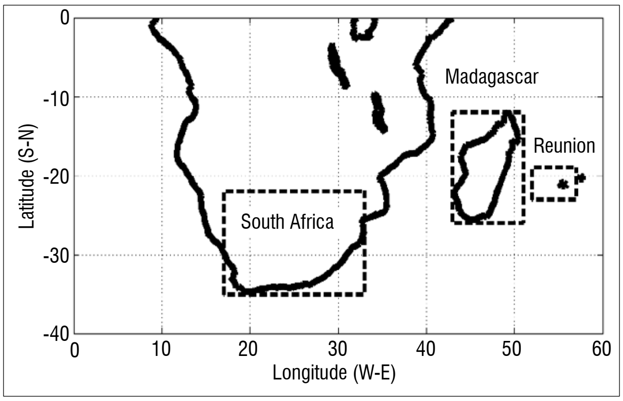


Figure 1: Geographical locations of the three study regions.

TES data

The data retrieval processing was performed in the selected areas for each day when a global survey was made. The number of profiles retrieved during a global survey varied between 0 and 16 profiles, depending on the extension of the observed area with respect to the satellite location during the observation period. It is worth noting that nadir observation covers a field of view of 5 km × 8 km. The considered geographical delimitations for the selected sites are shown in Figure 1 and Table 1. However, during observation time, there is a possibility that no observations were obtained in one or two of the selected areas (South Africa, Reunion and Madagascar).

Table 1: Geographic coordinates of the rectangle delineating the three study areas for tropospheric emission spectrometer data retrieval

Region	Longitude (°E)		Latitude (°S)	
	Maximum	Minimum	Minimum	Maximum
South Africa	33°	17°	22°	35°
Madagascar	51°	43°	12°	26°
Reunion	57°	52°	19°	23°

Figure 2 shows the number days of observation recorded in each area with respect to the month during the study period (01 January 2005–31 December 2009). Figure 2 indicates that South Africa and Madagascar have more observations, probably because of their larger size when compared to Reunion Island. Nevertheless, the number of observations over Reunion Island was found to be acceptable for the purpose of this study. It is worth noting that the number of days recorded at Reunion Island for a specific month (e.g. January 2007) was between 2 and 10, whilst in South Africa and Madagascar, the frequency was between 3 and 16 days (except in June 2005 when no CO measurements from TES were received).

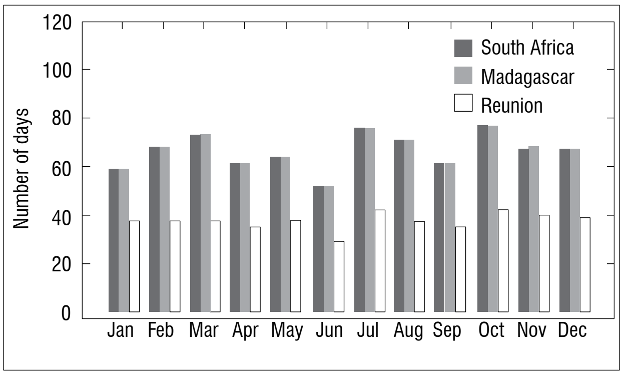


Figure 2: The number of satellite overpass days during each month for the three study areas.

Each TES mean daily vertical profile is composed of 47 pressure levels between 1000 hPa and 10 hPa and deducted by averaging daily vertical footprints recorded over the site during the observation. Figure 3 shows an example of a vertical TES CO profile recorded for 03 January 2007 during a close satellite overpass of Cape Point (34.35°S, 18.48°E) in South Africa. It illustrates CO mixing ratio values of ~59.5 ppb at ground level ~1000 hPa and the maximum is observed in between the 500 and 400 hPa pressure levels. The CO mixing ratio continues to decrease above 400 hPa and reaches a minimum value of ~14.8 ppb at around 51 hPa. The decrease is significant especially above 100 hPa (above the troposphere layer).

Trace analytical RGA-3 instrument

The trace analytical RGA-3 is an analyser used to measure CO. The different chemical components of an atmospheric sample are separated by gas chromatography (GC) utilising a column filled with a molecular sieve that is specific to CO retention. Upon heating, the eluted species are directed to a detector, which contains mercuric oxide (HgO). The RGA-3 instrument characteristics and operator mode are well documented by Hammer[17]. In this study, a trace analytical RGA-3 instrument was used for surface CO measurements at Cape Point in South Africa, which is one of South African Weather Service's long-term trace gas measurement stations.[18]

Cape Point station and CO measurements

The Cape Point global atmospheric watch laboratory (34.35°S, 18.48°E) is situated approximately 60 km directly south of Cape Town, and is located on a coastal cliff 230 metres above sea level within a nature reserve. It experiences moderate temperatures – dry summers impacted by occasional biomass burning episodes from the surrounding areas and increased precipitation during the austral winter months. Various wind systems (from both the marine and continental sectors) constitute the annual climatology. The observatory has been operational since 1978 and continues to provide long-term trace gas data – including CO. Cape Point station is managed by the South African Weather Service and coordinated by the GAW network through the World Meteorological Organization.

The station is predominantly exposed to maritime air masses that are derived from the southwestern and southeastern Atlantic Ocean regions, which constitute background conditions.[18,19] The daily average Cape Point CO mole fractions are calculated from 12-min values made by a RGA-3 trace analytical instrument and compiled to yield 30-min averages. Figure 4 presents the Cape Point CO measurements recorded on 03 January 2007 during such a specific overpass. In addition, Radon-222 (^{222}Rn), which is an excellent tracer for air mass origin, was used to characterise the air type on the day under scrutiny. Generally, ^{222}Rn >1200 mBq/m^3 is associated with continental air masses, while ^{222}Rn <250 mBq/m^3 is representative of baseline conditions with the oceanic air component dominating.[19] It can be seen from Figure 4 that on this date, CO mole fractions (solid line) varied between 46 and 60 ppb (right axis), whilst ^{222}Rn (dash line) fluctuated between 372 and 1485 mBq m^3 at the same time. These conditions primarily reflect continental and mixed air masses[19] with a small marine component after 20:30 (local time). CO and ^{222}Rn also correlate very well (Figure 4) revealing common source regions on this occasion. A strong continental air signal was evident during the early morning hours (04:00 to 07:30) when ^{222}Rn exceeded 1200 mBq/m^3, while later in the day (10:00 till midnight), ^{222}Rn levels fluctuated between 400 and 600 mBq/m^{3,} typically reflecting mixed air mass conditions (marine air with some entrainment of continental air) in response to a varying wind regime. The ^{222}Rn is used in this work to characterise the observation day in which the surface CO recorded at Cape Point station was contaminated by continental air mass. The present study uses only the daily average data from Cape Point uncontaminated by continental air mass to compare with the TES daily data recorded at 1000 hPa overpass over Cape Point.

Data analysis

The CO daily profile obtained from TES observation recorded overpasses over South Africa, Madagascar and Reunion have been analysed in order to examine the CO regional and seasonal distribution and to estimate its inter-annual trend. Thus, the statistical descriptive method is adopted and the inter-annual trend is estimated based on linear regression analysis. Prior to analysing the climatological, seasonal and monthly variations of CO over the selected regions, we first established a comparison of TES data with the ground-based measurements.

In order to facilitate a meaningful inter-comparison between TES and Cape Point CO data, the air mass block (vertical and lateral dimension) over the Cape Point region as seen by TES should be as homogeneous as possible. If this is not the case, the likelihood of TES seeing the same CO air mixture as the Cape Point laboratory will be greatly reduced, especially during pollution episodes. However, when southerly onshore maritime air flow regimes exist, the conditions for inter-comparisons are optimal. Such maritime air flow systems are easily identified using ^{222}Rn data.[19] According to Brunke et al.[19] and in order to classify the air masses over Cape Point during TES overpasses, only CO data during which the daily average ^{222}Rn level did not exceed 250 mBq/m^3 were selected for inter-comparison.

The TES profiles selected were those recorded between 34.35±0.5°S and 18.48±0.5°E. The daily selected value of TES was that registered at ~1000 hPa (ground level) while the daily value of surface based CO was computed from higher resolution in-situ measurements made at the Cape Point station. Comparison between the TES and ground-based values were performed each day when TES and filtered ground-based data were available. A total of 184 TES profiles were recorded over Cape Point during the period of observation (January 2005–December 2010). However,

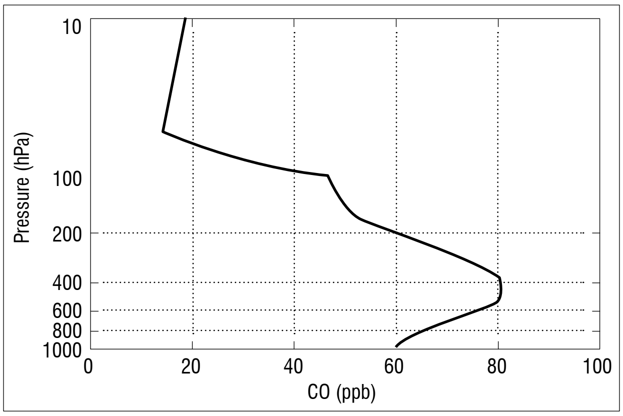


Figure 3: Height profile of CO measured by a tropospheric emission spectrometer overpass over Cape Point in South Africa on 03 January 2007.

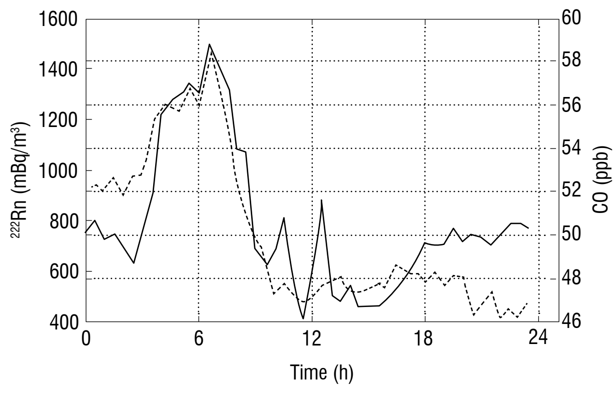


Figure 4: Diurnal distribution of CO (solid line) and ^{222}Rn (dashed line) observed at the Cape Point Global Atmosphere Watch station on 03 January 2007.

only 52 profiles were recorded on days in which ground-based data was not contaminated by continental air mass. Finally, the comparison was achieved on these 52 days of observation. The comparison indexes used were the relative differences, mean bias error, absolute root mean square and correlation coefficient R^2 between TES and ground-based observations. The CO relative difference (RD) between TES and Cape Point observed for a given day *d* was calculated with the following expression:

$$RD_d = 100x\,\frac{TES_d\text{-}Gb_d}{Gb_d} \qquad \text{Equation 1}$$

where Gb_d and TES_d represent the daily values of ground-based and TES data, respectively. The mean bias error was taken as the average value of relative difference. The mathematical expression of the mean bias error (MBE) is:

$$MBE = \frac{1}{N}\sum_{d=1}^{N} RD_d \qquad \text{Equation 2}$$

where *N* represents the number of day-pairs between the TES and Cape Point measurements. The correlative coefficient and absolute root mean square between the two observations was assessed using the linear regression method.

Results and discussion

Comparison between ground-based measurements and TES overpasses

The comparison between TES satellite and ground-based observations were performed in order to validate the satellite measurements over Cape Point station. Methodology used for data processing were given in the precedent sections while the results are presented here. Figure 5a shows temporal variations of CO measured at the Cape Point station (black crosses) and from TES (grey circles) for the time period from 01 January 2005 to 31 December 2009. In Figure 5b, the figure is superimposed with the relative difference between the two observations with respect to the ground-based measurement. It is apparent that the seasonal and inter-annual variation between the surface-based and satellite observations are found to be consistent and the obtained relative difference is within ±40%. Correlation between the two observations is given in Figure 5c. The correlation coefficient R^2 obtained between TES and ground-based measurements was evaluated and found to be around 0.61. Furthermore, Figure 5c shows a regression linear line (grey colour line) above the unit slope line (black line), indicating an overestimation of TES measurement with respect to ground-based. However, the obtained mean bias error is assessed at around 6.6% and the absolute root mean square recorded between the two observations is estimated to be less than 8.1 ppb. This difference is in part due to the different instrumental characteristics. TES uses an infrared high-spectral resolution Fourier transform spectrometer whilst at the Cape Point Observatory, the instrument is a RGA-3 trace analytical gas chromatograph. The difference in the time of observation may also explain the different values observed (one observation per day for TES and 48 individual 30-min observations in the case of surface measurements). Even the spatial resolution of TES (0.5 km × 5.3 km) can play a role in the observed differences between TES and ground-based measurements. However, the difference in measurements has not modified the overall variations within the temporal structure for the annual distribution between the TES Aura satellite and surface observations. Hence, we propose that it is possible to use TES data to study CO distribution for other localities.

The climatology of CO observed over the three regions

The monthly climatological profile was deduced by averaging daily vertical profiles recorded during the month, irrespective of the year. The monthly variations of TES CO measurements over the three regions during 5 years (2005–2009) are plotted in Figure 6. Over the three selected regions, CO measurements show a high variability in the free troposphere between 850 hPa and 150 hPa. It was found that the CO mixing ratio varied between 55 ppb and 120 ppb from surface to ~100 hPa; thereafter, it showed linear variations, mostly decreasing over height and reaching less than 20 ppb at 50 hPa. The maximum value was observed during spring when maximum bushfires and biomass burning are reported in southern Africa and adjacent regions.[9,11-13,20] The minimum value of CO mixing ratio was observed during the summer period. This period was also marked by high humidity and high activity of UV (ultraviolet) radiation in the southern tropic and subtropics region that led to production of radical hydroxyls (OH) from ozone photolysis.[21] Reaction with OH constitutes the primary sink for CO in the troposphere and consequently, the seasonal cycle of CO is primarily driven by the OH radical.[21,22]

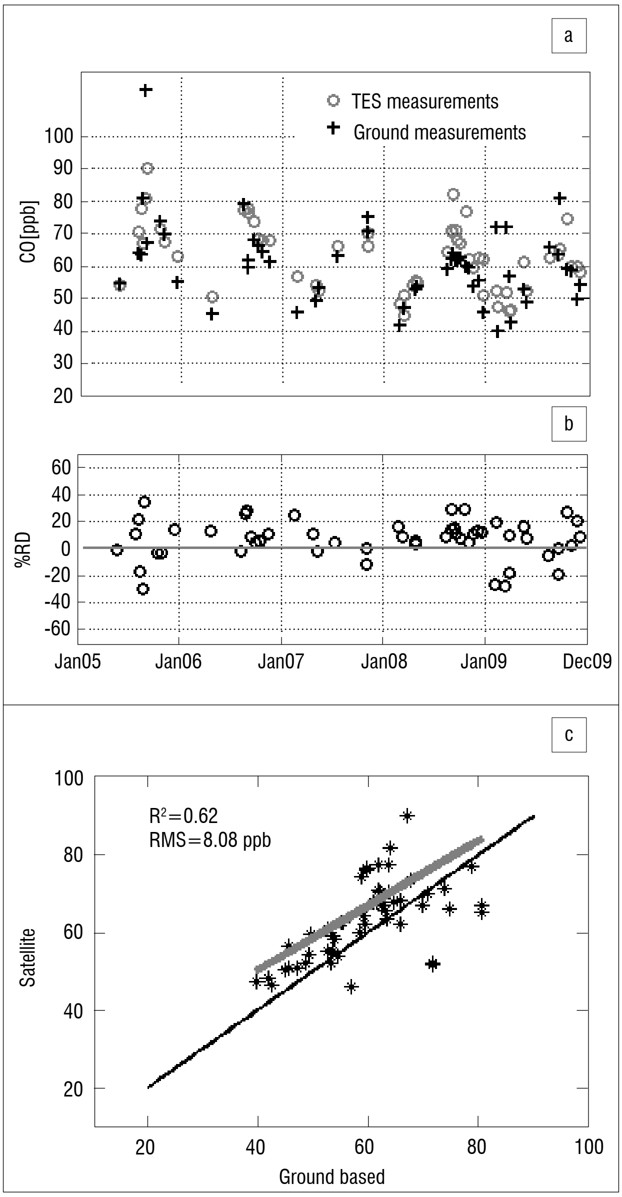


Figure 5: (a) Inter-annual variation of CO observed over Cape Point by the tropospheric emission spectrometer (TES) and from surface measurements at the Cape Point Global Atmosphere Watch station (ground) for the days on which there are both ground-based and TES observations under oceanic air mass conditions. (b) The relative difference (RD) observed between the two observations. (c) Scatter plot between TES satellite and ground-based observations over Cape Point; the black line represents the zero bias line while the bold grey line is the regression line between the two observations.

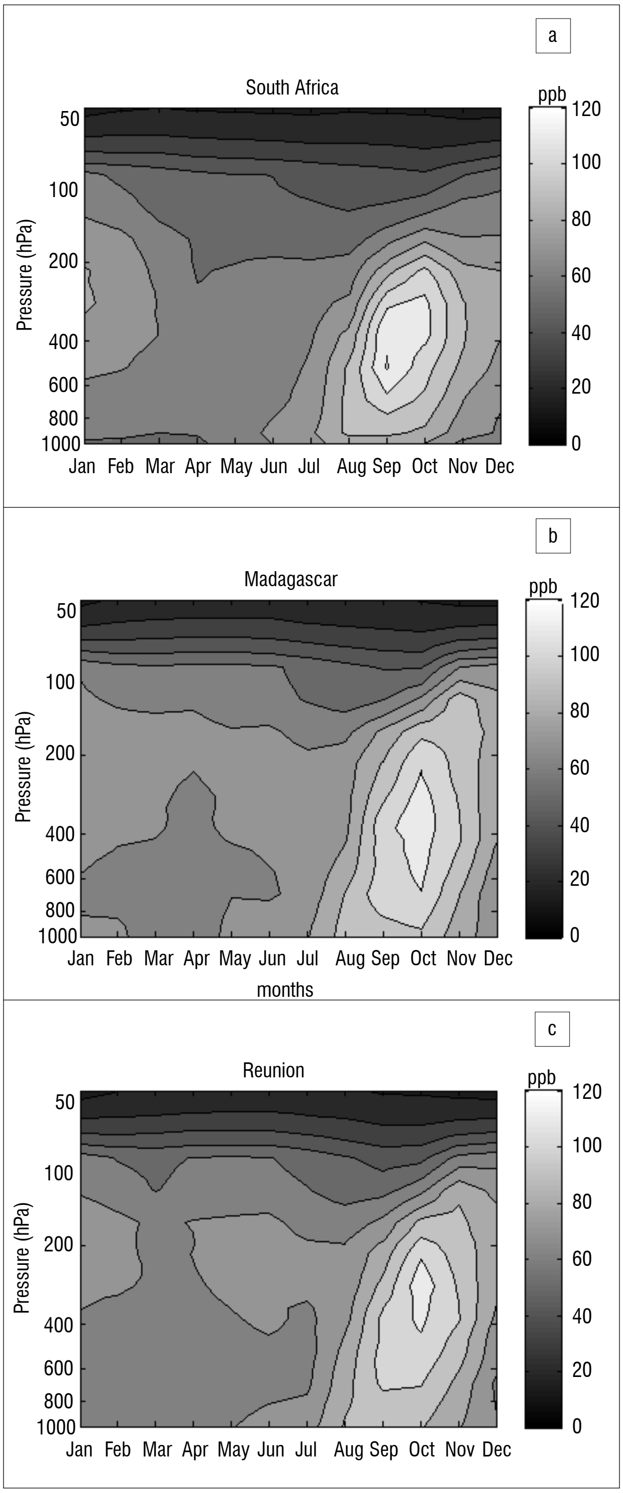


Figure 6: Seasonal distribution of CO concentration in ppb over (a) South Africa, (b) Madagascar and (c) Reunion Island.

It is worth noting that tropospheric carbon monoxide is produced essentially around the atmospheric boundary layer. The rather elevated quantity observed in the free troposphere between 850 hPa and 150 hPa is due to large-scale air mass transport and vertical convection from the free troposphere to tropopause. During this upward movement, CO reacts with OH to produce other chemical species such as CO_2 and hydrogen. Alternatively, it can also be oxidised to CO_2 and ozone (O_3) following specific reactions.[22] This could be one of the reasons for the observed lower values of CO mixing ratios close to the tropopause. Biomass burning activity plays a key role on the variability of the CO annual cycle. Indeed, the observed increase of CO begins in July at the three areas, in accordance with the start of the southern African savannah burning period. It is worth noting that the African burning period starts between January and February in West Africa and moves progressively southward; it reaches below the Inter-tropical Convergence Zone (ITCZ) around July, influencing CO transport in the southern troposphere. The CO mixing ratio begins to accumulate in southern Africa from July, however the maximum is observed between September and October when the CO variability is strongly pronounced. After November, the CO values were found to decrease. This could be firstly, a combination of the end of biomass burning period in Africa and South America[23] and secondly, the seasonal increase in OH levels. This study found that CO levels observed during the 4 months, from July to October represented 45% of the recorded annual CO. This high percentage shows the important role played by biomass burning on the annual CO production in southern Africa and adjacent regions.

In terms of vertical variation of CO, three principal layers are identified: the atmospheric boundary layer, the free troposphere and the lower stratosphere, representing 1000 hPa to 850 hPa, 850 hPa to 100 hPa, and 100 hPa to 10 hPa, respectively. The relative percentages of deviation with respect to the observed mean values are found in Table 2. Maximum deviation is observed in the free troposphere and minimum deviation in the lower stratosphere. The CO monthly average and its concomitant deviation observed over South Africa and Madagascar, especially in the atmospheric boundary layer, are higher than those observed over Reunion Island. These deviations are most likely a result of local contributions and also possibly due to the lower number of scans achieved by the satellite overpasses over Reunion. The satellite scanning of several regions increases the probability of achieving a variety of daily vertical profiles, which subsequently increase the standard deviation observed for the monthly vertical distributions. On the other hand, the high variability observed over Reunion compared to South Africa and Madagascar in the LS layer is most likely due to the reduced number of observation days recorded per month, most of which are spaced over several days. Hence, the concentrations recorded mainly between the middle and upper troposphere up to stratosphere, may not represent the same structures. As the CO mixing ratio does not change quite as remarkably beyond the tropopause height region, the variability was found to be low, which could be a valid reason for the observed lower deviations.

Table 2: Percentage variability recorded in the atmosphere boundary layer (ABL), free troposphere (FT) and lower stratosphere (LS) over the three selected areas

Region	ABL (%)	FT (%)	LS (%)
South Africa	4.8	7.1	2.9
Madagascar	4.4	6.2	2.7
Reunion	3.8	6.9	3.5

Seasonal vertical profile

In order to give an atmospheric dynamic explanation to the observed seasonal vertical profile distributions, seasonal vertical profiles were calculated on 47 pressure levels between 1000 and 10 hPa. The months were grouped into the South African seasons: summer (December to February), autumn (March to May), winter (June to August) and spring (September to November). The profiles and associated standard deviations obtained for each pressure level are presented in the Figure 7 for all the three regions. From Figure 7, we observe that CO values measured between 1000 hPa and 200 hPa are higher than 80 ppb during the spring having a maximum at about 400 hPa over our study location. During winter, summer and autumn, the seasonal average is still below 80 ppb with a maximum located above 300 hPa except over South Africa during summer where the maximum exceeds 80 ppb and is located at around 300 hPa. It is found that the observed quantity from the surface to approximately 600 hPa during winter exceeds that of summer and autumn because the forest fires that usually start in July. It is apparent from the figure that the maximum CO is located around 400 hPa during the spring over both South Africa and the Indian Ocean region. However, during the summer and autumn the CO mixing ratio is observed to be dominant at pressure levels from approximately 350 hPa to 150 hPa primarily over the Indian Ocean region.

Above 400 hPa, the profiles observed during winter and autumn have the same vertical variation. CO levels observed at 800 hPa over South Africa are generally always higher than the observations at ground level. This is contrary to Madagascar and Reunion Island, where the ground level has a higher CO concentration than the 800 hPa level, especially during winter, summer and autumn. This high quantity observed at 800 hPa over South Africa compared to the other areas is most likely due to various industrial emissions and stable circulation layers, amongst other reasons.

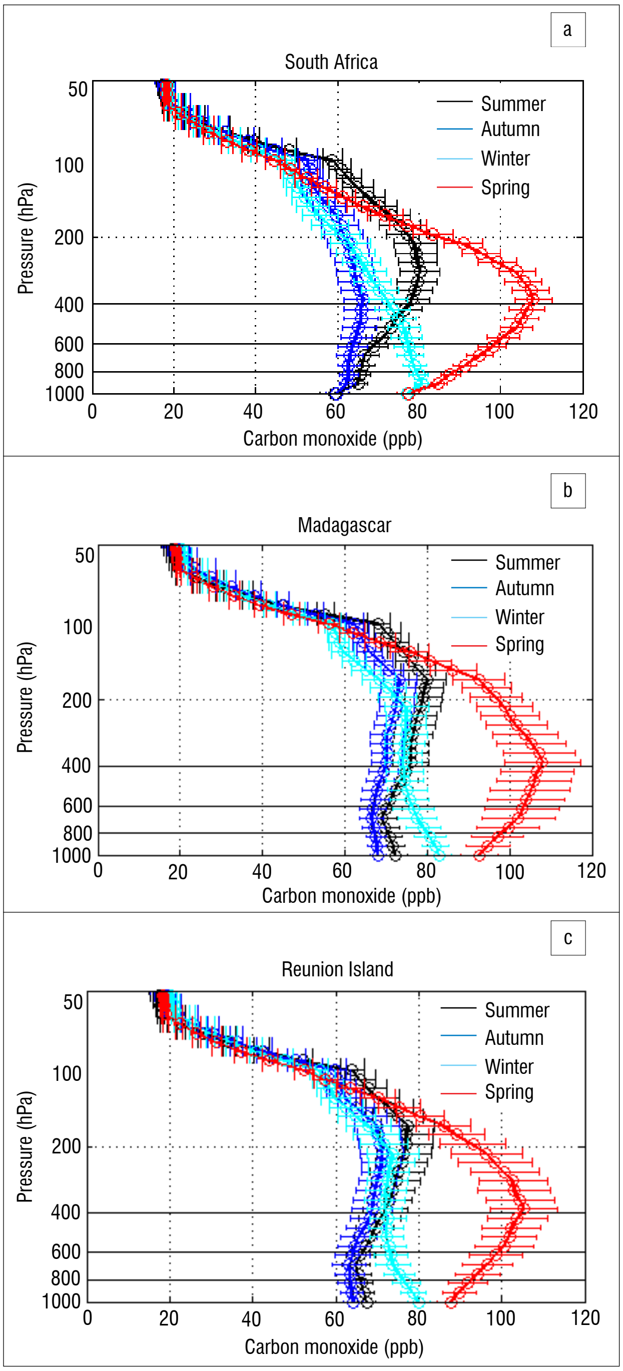


Figure 7: Seasonal vertical profile and their associated standard deviations for each pressure level over (a) South Africa, (b) Madagascar and (c) Reunion Island.

South Africa is located approximately at 30°S of latitude; where the Hadley cell from the tropical region and the Ferrell cell from the middle latitude converge to form a sub-tropical jet-stream. The jet-stream is located at around 300 hPa and extends up to the tropopause region. This tropospheric air current from the west also has a high speed (~25 m/s) and acts as a pathway for moving and transporting air masses and pollutants to the east. A previous study by Stohl et al.[26] demonstrated that emission from South America can be advected into the Atlantic to reach southern Africa, thus it can also be transported from South Africa to Australia by crossing the Indian Ocean. The CO transport over the western Indian Ocean region is also governed by the trade winds. Stohl et al.[24] has also shown that CO, with a lifetime exceeding 25 days, may reach ~9 km (~300 hPa) over Africa. The effects of trade winds are evident until approximately 6 km altitude (~450 hPa) where there is air mass and pollutants exchange (jet stream-trade wind) in the pressure range from 450 hPa to 300 hPa. This exchange results in homogeneity between the vertical profiles found within the Indian Ocean area (Reunion and Madagascar) and the subtropical zone of South Africa. Thus, because of the jet stream–trade wind exchange and the longer lifetime of CO over Africa, it is also expected that the CO concentrations observed may reach a maximum at approximately 400 hPa over southern Africa and the Indian Ocean during spring.

Deep convection within the ITCZ over the continental region in the southern summer continuously lifts the tropopause upwards by some kilometres and as such deepens the extent of the troposphere. This happens as a result of thunderstorm activity mixing tropospheric air at a moist adiabatic lapse rate, thus pollutants can be transported by vertical convection up to 200 hPa. As a result of this mixing, maximum CO is usually observed at around 200 hPa especially over Madagascar and Reunion Island. This observation also explains why the tropopause layer may, on occasion, move above 200 hPa. This can be clearly seen from 100 hPa level where the seasonal vertical profiles have identical variation. These results confirm that the tropopause is at its lowest level of the atmosphere during the spring and at its highest levels during the others periods. During summer, a maximum CO is observed at around 300 hPa over South Africa and above 200 hPa over Madagascar and Reunion Island. We assume the existence of a dynamic structure that prevents CO convection movements. This dynamic structure is probably located below 200 hPa in the mid-latitudinal region and around 150 hPa in the tropical region, especially during summer, and this may be regarded as the CO tropopause height.

Seasonal distribution of total column CO over the three sites

In this study, the annual cycle of CO distribution was obtained by combining 5 years of TES data, which included 2005, 2006, 2007, 2008 and 2009. The obtained seasonal distribution is plotted in Figure 8. The recorded quantities of CO in January, February and March were lower and are associated with small standard deviations as well. However, the recorded values during this period were higher over Madagascar compared to South Africa and Reunion. The minimum annual CO concentration over Reunion Island and Madagascar occurred during March whilst in May over South Africa, (mainly as a result of the factors discussed in the section on the climatology of CO above). During May, June and July (austral winter), the CO distribution over Reunion Island was similar to that observed over Madagascar. During the same period, the monthly CO values observed over South Africa were below 2.1 ppm. The annual maximum occurred during spring and reached its peak value in September over South Africa (2.8 ± 0.2 ppm) and a month later, in October, over Reunion Island (2.9 ± 0.2 ppm) and Madagascar (3.1 ± 0.2 ppm).

The high CO observed over Madagascar during December, January and February may be caused by emission sources from northern Africa, India and the Middle East, which are transported via western trade winds to the ITCZ (usually located above the Madagascar area during this time). The ITCZ constitutes a convective source and is an effective and significant pollution trapping zone.[13] From March onwards, the ITCZ migrates to the northern hemisphere after which Madagascar and Reunion Island receive the same climatic conditions as the rest of the study area and thus have the same characteristics which regulate the CO distribution within the mid-latitudes. This is the reason why the monthly observed CO levels over Reunion and Madagascar, especially during April, May, June and July are so similar.

During the austral winter, CO lifetime is typically longer (around a month) within the tropical region.[25] This could be a reason why more CO is observed over Madagascar and Reunion Island (tropical zone) during April to July, but is less pronounced over South Africa (mid latitude). However, the CO levels observed over the Indian Ocean region cannot adequately explain the high CO mixing ratio recorded during spring since it generalises over the three study sites and the climatology over South Africa is different from that of the other two sites. As mentioned earlier, the high CO mixing ratio observed during spring is generally associated with increased biomass burning activity over sub-equatorial Africa, Madagascar and southern America.[25,26] Also, the CO transport takes sufficient time to reach the Indian Ocean and could be a reason for the observed high background values of ambient CO during July to August at Cape Point, maximum in September over South Africa and in October over the Indian Ocean. However, a more in-depth investigation on CO-coupled air mass trajectories at a regional and continental scale is needed to validate this statement.

Trend estimates

Trend analyses were carried out using the monthly mean values of CO for the 5-year period from 2005 to 2009. The monthly values for the year are represented by integration of the quantities measured over the entire 47 pressure levels. The inter-annual distribution, which is derived from the monthly mean variation, is presented in Figure 9. Positive anomalous were evident in 2005 and 2007 and negative anomalous were observed during the summers of 2006, 2008 and 2009. In the framework of the annual emission budget, higher quantities were observed in 2005 and 2007 while lower quantities were recorded in 2006, 2008 and 2009.

Two important events were observed during the spring of 2005 and 2007 which contributed significantly to the high quantities observed in those years: (1) the El Niño events and (2) an increase in biomass burning activity. Indeed, the El Niño events that occurred in 2005 and 2007 were associated with positive anomalies of total column CO over the entire globe, except for high latitudes of the two hemispheres.[27] In addition, 2005 and 2007 were marked by an increase in biomass burning activity in the southern hemisphere, especially over the South America area, according to Torres et al.[28] and Giglio et al.[29]

As the annual cycle of CO is strongly linked with biomass burning activity in the southern hemisphere,[28,30,31] we assume that the low quantity observed in 2006 was partly due to the reduced biomass burning activity in the South American region. This happened as a result of the tri-national committee action and small farmer initiative which both aimed to reduce biomass activity in 2006.[30,31] A study by Torres et al.[28] showed strong negative anomalies of biomass burning activity during 2008 and 2009 in the southern hemisphere. Their inter-annual fire account and aerosol optical absorption distributions were shown to have similar lower peaks upon comparison with our CO inter-annual distribution results (Figure 9) over the same period. In addition, Yongxing[27] showed that the reduced CO levels observed in 2006, 2008 and 2009 were partly due to La Niña events which were observed in January and February of 2006, 2008 and 2009. These events are normally associated with heavy rainfalls, which can considerably decrease the amount of pollutants within the atmosphere while reducing the total biomass burning events themselves. Therefore, the observed inter-annual variability in Figure 9 is primarily a result of a mixture of biomass burning and El Niño Southern Oscillation (ENSO) events.

A linear regression was applied on the total column CO data for the three sites under investigation. The results obtained are also plotted in Figure 9. Overall, a negative trend was observed for all three sites. The calculated slopes of the linear trends, quantifying the total column CO decrease, amounted to 5.1 ppb per month, 4.7 ppb per month and 4.5 ppb per month for South Africa, Madagascar and Reunion respectively. Expressed as a percentage, the quantity of CO measured within the 47 pressure levels decreased by 2.1%, 1.8% and 1.7% respectively over South Africa, Madagascar and Reunion.

An interesting aspect is that the decreased trend within the CO mixing ratio, observed for our study period (2005–2009), was also reflected within the surface-based measurements of Cape Point (Figure 10). These measurements also show a declining trend of similar magnitude to that observed by the TES/Aura satellite over the three study sites. Figure 10 represents the inter-annual background CO distribution observed at Cape Point (January 2005 to December 2009), derived from mean monthly observations. A linear regression applied on the Cape Point inter-annual distribution, showed that tropospheric CO decreased by a similar negative rate, estimated at 0.1 ppb per month, which corresponds to 2.4% of the average annual mean for this period.

The concomitant observed decreases in CO mixing ratios observed by TES for the three sites studied is most likely a consequence of activities

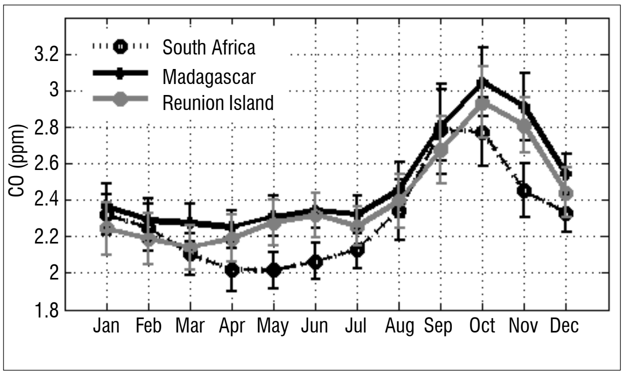


Figure 8: Seasonal distribution of CO and its variability observed from January 2005 to December 2009. The error bars in the figure represent the standard deviation of the monthly average.

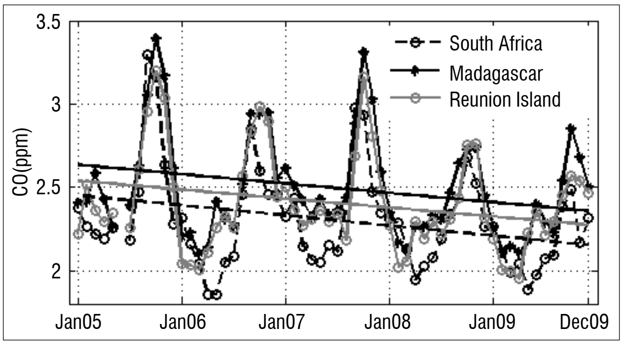


Figure 9: Monthly distribution of CO and the trends estimated for the period from January 2005 to December 2009.

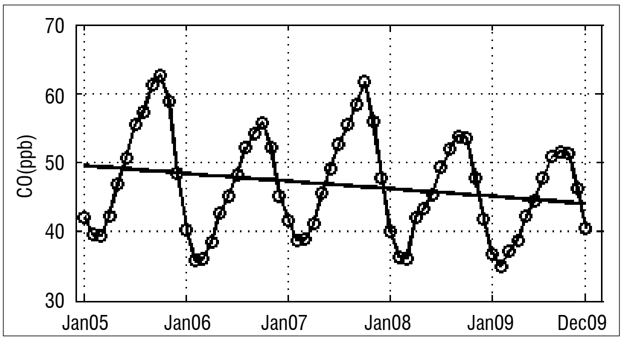


Figure 10: Cape Point Global Atmosphere Watch station monthly means (from January 2005 to December 2009) with linear trend line.

that drive, or at least can be linked to, the origins of the observed inter-annual CO distribution. These driving processes were discussed earlier in this paper and are: the reduction of biomass burning activity observed in 2006, 2008 and 2009 around South America and the La Niña events with their associated positive rainfall anomalies. The observed decrease in CO is probably also in part due to a reduction in anthropogenic carbon monoxide emissions as suggested by Zeng et al.[16] and/or an increase in global hydroxyl radical (OH),[32,33] but changes within these driving mechanisms need to be verified by further studies. A recent study by Zeng et al.[16] analysed the CO trend over two southern hemisphere sites (Lauder, New Zealand (45°S, 170°E) and Arrival Height, Antarctica (78°S, 167°E)) between 1997 and 2010. They reported a decreasing rate of CO, estimated to be around -0.94 ± 0.47% per year. They further explored the relationship between this negative rate and an overall decline in CO emissions from industrial sources, estimated to be around 26% in the southern hemisphere for the period from 1997 to 2009.

The atmospheric oxidant capacity depends directly on the amount of OH radical available. This very reactive species acts as the tropospheric detergent, by reacting with most of the emitted oxidants, and especially CO.[34] Therefore, OH distributions determine the CO lifetime in the atmosphere.[21] A study by Prinn et al.[33] has shown an increase of the global hydroxyl radical (OH) in the atmosphere, estimated to be around 1.0 ± 0.8% per year from 1979 to 1990. However, more investigations on OH distributions over our study area during the study period are needed in order to validate this statement.

Summary and conclusion

In this study, we investigated spatiotemporal distributions of CO over three different adjacent regions, namely South Africa, Madagascar and Reunion using five years (2005–2009) of data obtained from the TES Aura satellite instrument. A preliminary study was conducted to validate TES data through comparison with ground-based CO data recorded at the Cape Point GAW station. A comparison was performed between ground-based measurements uncontaminated by continental air mass and satellite data recorded at 1000 hPa over the Cape Point station and a satisfactory agreement was found between the measurements. The mean bias error observed between the two observations was 6.6% and the regression coefficient was greater than 0.6. Thereafter, data from the TES instrument was used to investigate CO variation and trends on the three selected regions. The following are salient features noted from this study:

- Monthly and seasonal vertical distributions obtained from 47 pressure levels of TES data highlighted the most prominent pressure ranges of pollutants within the free troposphere, being from 400 hPa to below 150 hPa. As a result of CO having a long lifetime within these pressure ranges, a mechanism is in place, which allows for pollutant transport on a regional and global scale, which includes vertical convective movement, the trade winds and the jet-stream.
- The bulk of the CO concentration was observed in the free troposphere, between 850 hPa and 150 hPa. Above 150 hPa, CO mixing ratio decreased linearly reaching a minimum value at approximately 50 hPa. The lowest CO mixing ratio was observed within the stratosphere.
- An investigation into the temporal variation of CO revealed a high concentration of CO during spring, driven primarily by fires and biomass burning activities over Africa and adjacent regions. The peak period was observed to be between July and October. The quantity of CO observed during this period was estimated to be approximately 45% of the annual recorded values. The minimum CO mixing ratio was observed during the southern summer, specifically during May over South Africa and March over the Indian Ocean region.
- Analyses for the inter-annual distribution of CO identified positive anomalies in the spring of 2005 and 2007 and lower quantities of CO during 2006, 2008 and 2009.
- A linear regression applied on inter-annual CO variation exhibited a negative trend in the order of 2.1% over South Africa, 1.8% over Madagascar and 1.7% over Reunion. The observed declining rate of CO in the upper atmosphere was also confirmed with an analysis made on ambient surface CO at the Cape Point GAW station which displayed a tropospheric background CO decline, estimated at approximately 2.4% per year.

From these analyses, we conclude that measurements of CO using TES are in agreement with ground-based measurements. TES can therefore be used for future analyses on the variability of CO in southern tropic and middle latitude regions, or TES measurements could be merged with correlative data to provide a high-quality and reliable long-term CO data set.

Acknowledgements

This work is supported by the GDRI ARSAIO (Atmospheric Research in Southern Africa and the Indian Ocean), a French–South African cooperative programme network supported by the French Centre National de la Recherche Scientifique (CNRS) and the South African National Research Foundation (NRF). A.M.T. acknowledges the Indian Ocean Bureau AUF (Agence Universitaire de la Francophonie) and the University of KwaZulu-Natal for travel and hosting invited student exchange on the NRF–CNRS bi-lateral research project (UID 78682). We also thank the TES team and the South Africa Weather Service for providing the CO data used in this work. S.K.S. acknowledges the SANSA Earth Observation for their financial support.

Authors' contributions

A.M.T. was the project leader; V.S. was the supervisor of the project and the principal coordinator of the ARSAIO (Atmospheric Research Southern Africa and Indian Ocean) programme; N.M., S.S. and H.B. contributed to data analysis and the review of the manuscript. E.G.B. and C.L. assisted in interpreting the results and writing the paper.

References

1. Clerbaux C, George M, Turquety S, Walker KA, Bernath P, Barret B, et al. CO measurement from the ACE-FTS satellite instrument: Data analysis and validation using ground-based, airborne and spaceborne observation. Atmos Chem Phys. 2008;8:2569–2594. http://dx.doi.org/10.5194/acp-8-2569-2008

2. Beer R. TES on the aura mission: Scientific objectives, measurement, and analysis. IEE Trans Geosci Remote Sens. 2006;44:1102–1105. http://dx.doi.org/10.1109/TGRS.2005.863716

3. Herman R, Kulawik S. Tropospheric Emission Spectrometer (TES), Level 2 (L2) data user's guide version 5. Pasadena, CA:JPL; 2011.

4. Kopacz M, Daniel JJ, Daven KH, Heald LC, David GS, Qiang Z. Comparison of adjoint and analytical Bayesian inversion method for constraining Asian source of CO, using satellite (MOPITT) measurement of CO columns. J Geophys Res. 2009;114(D04305):1–10.

5. Luo M, Beer R, Jacob DJ, Logan JA, Rodgers CD. Simulated observation of tropospheric ozone and CO with the TES satellite instrument. J Geophys Res. 2002;107(D15):ACH9 1–10.

6. Luo M, Rinsland CP, Rodgers CD, Logan JA, Worden H, Kulawik S, et al. Comparison of carbon monoxide measurements by TES and MOPITT: Influence of a priori data instrument characteristics on nadir atmospheric species retrievals. J Geophys Res. 2007;112(D9), 303, 13 pages.

7. Heald LC, Jacob DJ, Jones DBA, Palmer PI, Logan JA, Streets DG, et al. Comparative inverse analysis of satellite (MOPITT) and aircraft (TRACE-P) observation to estimate Asian source of CO. J Geophys Res. 2004;109(D23), 306, 17 pages.

8. Richards NAD, Li Q, Bowman KW, Worden JR, Kulawik S, Osterman GB, et al. Assimilation of TES CO into a global CTM: First results. Atmos Chem Phys Discuss. 2006;6:11727–11743. http://dx.doi.org/10.5194/acpd-6-11727-2006

9. Kopacz M, Jacob DJ, Fisher JA, Logan JA, Zhang L, Megretskaia IA, et al. Global estimation of CO sources with high resolution by adjoint inversion of multiple satellite datasets (MOPITT, AIRS, SCHIAMACHY and TES). Atmos Chem Phys. 2010;10:855–876. http://dx.doi.org/10.5194/acp-10-855-2010

10. Yurganov LN, Rakitin V, Dzhola A, August T, Fokeeva E, George M, et al. Satellite- and ground-based CO total column observations over 2010 Russian fires: Accuracy of top-down estimates based on thermal IR satellite data. Atmos Chem Phys. 2011;11:7925–7942. http://dx.doi.org/10.5194/acp-11-7925-2011

11. Rajab JM , MatJafr MZ, Lim HS, Abdullahi K. Indonesia forest fires exacerbate carbon monoxide pollution over peninsular Malaysia during July to September 2005. 6th International Conference on Computer Graphics Imaging and Visualisation (ICCGIV'09); 2009 Aug 11–14; Tianjin, China. Washington DC: IEEE; 2009. p. 547–552.

12. Duflot V, Dils B, Baray JL, De Mazière M, Attié JL, Vanhaelewyn G, et al. Analysis of the origin of the distribution of CO in the subtropical southern Indian Ocean in 2007. J Geophys Res. 2010;115(D22), 106, 16 pages.

13. De Laat ATJ, Lelieveld J, Roelofs GJ, Dickerson RR, Lobert JM. Source analysis of carbon monoxide pollution during INDOEX 1999. J Geophys Res. 2001;106:28481–28495. http://dx.doi.org/10.1029/2000JD900769

14. Khalil KAM, Rasmussen RA. Global decrease in atmosphere carbon monoxide concentration. Nature. 1994;370:639–641. http://dx.doi.org/10.1038/370639a0

15. Novelli PC, Masarie KA, Tans PP, Lang PM. Recent changes in carbon monoxide. Science. 1994;263:1587–1590. http://dx.doi.org/10.1126/science.263.5153.1587

16. Zeng G, Wood SW, Morgenstern O, Jones NB, Robinson J, Smale D. Trends and variations in CO, C_2H_6, and HCN in the southern hemisphere point to the declining anthropogenic emissions of CO and C_2H_6. Atmos Chem Phys. 2012;12:7543–7555. http://dx.doi.org/10.5194/acp-12-7543-2012

17. Hammer S, Mattheier HG, Muller L, Sabasch M, Schimdt M, Schmitt S, et al. A gas chromatography system for high precision quasi-continuous atmospheric measurement of CO_2, CH_4, N_2O, SF_6, CO and H_2 [document on the Internet]. c2008 [cited 2014 Feb 20]. Available from: http://www.iup.uni-heidelberg.de/institut/forschung/groups/kk/GC_Hammer_25_SEP_2008.pdf

18. Scheel HE, Brunke EG, Sladkovic R, Seiler W. In situ CO concentrations at the sites Zugspitze (47°N, 11°E) and Cape Point (34°S, 18°E) in April and October 1994. J Geophys Res. 1998;103(D15):19295–19304. http://dx.doi.org/10.1029/96JD04010

19. Brunke EG, Labuschagne C, Parker B, Scheel HE, Whittlestone S. Baseline air mass selection at Cape Point, South Africa: Application of ^{222}Rn and other filter criteria to CO_2. Atmos Environ. 2004;38(33):5693–5702. http://dx.doi.org/10.1016/j.atmosenv.2004.04.024

20. Swap R, Annegarn HJ, Suttles TS, King MD, Platnick S, Privette JL, et al. Africa burning: A thematic analysis of the southern African regional science initiative (SAFARI 2000). J Geophys Res. 2003;108(D13):SAF 1–14.

21. Novelli PC, Masarie KA, Lang PM. Distribution and recent changes of CO in the lower troposphere. J Geophys Res. 1998;103:19015–19033. http://dx.doi.org/10.1029/98JD01366

22. Wang C, Prinn R. Impact of emissions, chemistry, and climate on atmospheric carbon monoxide: 100-year predictions from a global chemistry-climate model. Chemosphere Global Change Sci. 1999;1(1–3):73–81. http://dx.doi.org/10.1016/S1465-9972(99)00016-1

23. Brunke EG, Ebinghaus R, Kock HH, Labuschagne C, Slemr F. Emissions of mercury in southern Africa derived from long-term observations at Cape Point, South Africa. Atmos Chem Phys. 2012;12:7465–7474. http://dx.doi.org/10.5194/acp-12-7465-2012

24. Stohl A, Sabine E, Caroline F, Paul J, Nicole S. On the pathways and timescales of intercontinental air pollution transport. J Geophys Res. 2002;107(D23):ACH6-1–ACH6-17

25. Duflot V. Quantification et étude du transport des polluants dans la troposphère de l'Océan Indien [Quantification and study of pollutants transport in the stratosphere of the Indian Ocean] [thesis]. Saint Denis: LACy, University of Réunion; 2012. French.

26. Van der Werf GR, Randerson JT, Gilio L, Collatz GJ, Kasibhatla PS, Arellano AF. Inter-annual variability in global biomass burning emission from 1997 to 2004. Atmos Chem Phys. 2006;6:3423–3441. http://dx.doi.org/10.5194/acp-6-3423-2006

27. Yongxing Z. Mean global and regional distribution of MOPPIT carbon monoxide during 2000–2009 and during ENSO. Atmos Environ. 2010;45:1347–1358.

28. Torres O, Chen Z, Jethva H, Ahn C, Freitas SR, Bhartia PK. OMI and MODIS observations of the anomalous 2008–2009 southern hemisphere biomass burning seasons. Atmos Chem Phys. 2010;10:3505–3513. http://dx.doi.org/10.5194/acp-10-3505-2010

29. Giglio L, Randerson JT, Van der Werf GR, Kasibhatla PS, Collatz GJ, Morton DC, et al. Assessing variability and long-term trends in burned area by merging multiple satellite fire products. Biogeosciences. 2010;7:1171–1186. http://dx.doi.org/10.5194/bg-7-1171-2010

30. Koren I, Remer LA, Longo K. Reversal of trend of biomass burning in the Amazon. Geophys. Res. Lett. 2007;34(L20), 404, 4 pages.

31. Koren I, Remer LA, Longo K, Brown F, Lindsey R. Reply to comment by W. Schroeder et al. on ''Reversal of trend of biomass burning in the Amazon''. Geophys Res Lett. 2009;36(L03), 807. http://dx.doi.org/10.1029/2008gl036063

32. Prinn RG, Huang J. Comment on "Global OH trend inferred from methylcloroform measurements" by Maarten Krol et al. J Geophys Res. 2001;106:23151–23157. http://dx.doi.org/10.1029/2001JD900040

33. Prinn RG, Cunnold DP, Simmonds F, Alyea R, Boldi A, Crawford P. Global average concentration and trend for hydroxyl radicals deduced from ALE/GAGE trichloroethane (methyl chloroform) data for 1978–1990. J Geophys Res. 1992;97:2445–2461. http://dx.doi.org/10.1029/91JD02755

34. Duncan BN, Logan JA. Model analysis of the factors regulating the trends and variability of CO between 1988 and 1997. Atmos Chem Phys. 2008;8:7389–7404. http://dx.doi.org/10.5194/acp-8-7389-2008

Sectoral electricity elasticities in South Africa: Before and after the supply crisis of 2008

AUTHORS:
James Blignaut[1]
Roula Inglesi-Lotz[1]
Jaco P. Weideman[1]

AFFILIATION:
[1]Department of Economics, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Jaco Weideman

EMAIL:
jaco.weideman@outlook.com

POSTAL ADDRESS:
Department of Economics, Private Bag X20, Hatfield 0028, South Africa

DATES:
Received: 14 Mar. 2014
Revised: 28 Oct. 2014
Accepted: 27 Nov. 2014

KEYWORDS:
electricity tariffs; price elasticity; electricity demand; industrial sectors; structural break

HOW TO CITE:
Blignaut J, Inglesi-Lotz R, Weideman JP. Sectoral electricity elasticities in South Africa: Before and after the supply crisis of 2008. S Afr J Sci. 2015;111(9/10), Art. #2014-0093, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140093

In this paper, we estimate the price elasticity of electricity for various industrial sectors of the South African economy from 2002 to 2011. The data used include sectoral electricity consumption data and electricity tariff data, both courtesy of Eskom as well as output data based on national statistics. The most important contribution this paper makes is that it includes the period after the sharp rises in electricity tariffs in 2007/2008 following a period of load-shedding and insecurity in electricity supply. Previous studies have included data only until 2007 and, for the most part, have found statistically insignificant, positive elasticities. However, for the period post-2007, we found statistically significant and negative elasticities for 9 of the 11 sectors considered. Our results show that the majority of industrial sectors have become much more sensitive to changes in the price of electricity following 2007/2008, indicating to policymakers that tariff restructuring might influence consumer behaviour significantly.

Introduction

Price elasticity measures the sensitivity of consumer behaviour to price (or tariff) fluctuations. (It should be noted that electricity prices in South Africa are administratively set and hence are actually tariffs, but here we use the terms interchangeably.) Understanding such behavioural responses is of strategic and practical importance to policymakers and investors alike within the electricity sector when considering infrastructure development planning, the determination of future electricity tariffs, environmental policies, etc. Being able to determine the most likely behavioural responses to changes in prices in an industrial environment that is continuously becoming more electricity intensive[1] is the key to successful policy implementation.

South Africa experienced a severe electricity supply crisis during 2007/2008 with extensive blackouts or load-shedding. The damaging consequences on the economy were vast. The National Energy Regulator of South Africa (NERSA) estimated that approximately ZAR50 billion (approximately USD5 billion) was lost during this crisis.[2] Many possible reasons have been given to explain the crisis, such as the lack of capacity for generation and reticulation of electricity[3] and the lack of research on electricity topics and energy in general[4].

Eskom (the state-owned monopolistic supplier of electricity in South Africa) argued that the rate of economic and population growth in the country increased the mismatch between demand and supply of electricity,[5] and thus, only the expansion of power generation capacity would be able to alleviate the problem. Since then, the construction of two additional power plants, Medupi and Kusile, was initiated. These plants will be fully operational only from 2018 onwards, adding an extra 9600 MW (2 x 4800 MW) to the current power generation capacity.

It has also been argued[3,6] that South Africa's historically low electricity tariffs – compared with those of the rest of the world – have been a disincentive for consumers to use energy efficiently, leading to higher electricity consumption levels. Since the crisis, Eskom and NERSA have changed the electricity tariff structure resulting in increases of up to 25% per annum from 2008 to date.

Following these events, many researchers have tried to detach the behaviour of consumers and their reactions – if any – to the past and also to future changes in tariffs.[3,6-9] When these studies were published, the available data included information only up to 2007/2008 – a time when electricity tariffs were at historically low levels.[3] Hence, the data did not allow for the investigation of the possible impacts of price restructuring beyond 2007/2008 on the South African electricity market nor for any changes in elasticities.

The tariff structures in the country since 2008/2009 might have altered consumer sensitivity to price fluctuations, so an in-depth analysis of this new behaviour has become imperative. From the outset it should also be noted that, given the dynamic nature of the country's economy, it is necessary to update elasticity estimates regularly as they do change over time (see also Inglesi-Lotz[6]). Moreover, this information is also necessary to estimate the degree of consumer sensitivity to the introduction of any carbon tax or other future price restructuring.

Although we are not the first to ask these questions, this paper is unique in several ways. Firstly, it is the first to incorporate data after the electricity supply crisis in South Africa and the price restructuring of 2008/2009. Previous studies on the South African case only included information pre-energy crisis in the country and hence could not discuss the effects of the crisis and the price changes in consumer behaviour. This study is therefore not just a more updated version, but also provides more relevant and current policy suggestions.

Secondly, ours is the first study conducted for South Africa with such a detailed level of disaggregation. This analysis is not only for sectoral policy implications but will also assist in thoroughly understanding similarities and differences in the behaviour of the various sectors. This understanding allows investigations into whether there are changes and whether these changes are confined purely to certain sectors or are an economy-wide trend.

Literature review

Several studies have been conducted on the issue of electricity prices and price elasticity in South Africa, especially before 2007/2008. The local increase in the interest in price elasticity of electricity is matched by a similar increase internationally. Table 1 lists a selection of studies that deal with the influence of electricity prices on electricity consumption, both locally and internationally.

While this list is not exhaustive, the following salient facts emerge:

- International literature has focused on both developed and developing countries as electricity and its determinants are equally important for the economic growth and development of all countries.
- A large number of different methodologies was used, either in a time-series context looking at only one country at a time or in a panel context examining a group of countries at once.
- None of these studies concluded a certain trend, that is, whether the electricity demand was more or less elastic in the short run than in the long run.
- In some studies, such as Ziramba's[24] and Amusa et al.'s[25], the price was found to be a statistically insignificant factor in the determination of electricity demand and hence was excluded from the estimations.
- South African studies reported various conclusions from no effect of price on consumption (zero or statistically insignificant elasticities) to highly negative price elasticities.

Table 1: Selected international and South African studies on price elasticity of electricity demand

Author(s) (international studies)	Period	Methodology	Country	Price elasticity
Diabi[10]	1980–1992	Panel data (ordinary least squares (OLS), fixed effects (time and region), random effects)	Saudi Arabia	Range from -0.139 to 0.01
Von Hirschhausen and Andres[11]	1996–2000	Cobb–Douglas for forecasting purposes	China	By assumption -0.02
Al-Faris[12]	1970–1997	Johansen cointegration methodology	Gulf Cooperation Countries	Short-run: -0.09; long-run: -1.68 (average of GCC countries)
Kamerschen and Porter[13]	1973–2008	Flow adjustment model and 3-stage least squares	USA	Range from -0.51 to 0.02 with first method and from -0.15 to -0.13 with the second
Narayan and Smyth[14]	1969–2000	Auto-regressive distributed lag (ARDL) methodology and Granger causality	Australia	Long-run: -0.541 and short run -0.263
De Vita et al.[15]	1980–2002	ARDL methodology	Namibia	Long-run: -0.34
Atakhanova and Howie[16]	1990–2003	Panel data techniques	Kazakhstan	Price was statistically insignificant
Narayan et al.[17]	1978–2003	Panel cointegration and error correction models (ECM)	G7 countries	Long run range from -7.408 to -1.45; short-run from -1.739 to -0.0001
Amarawickrama and Hunt[18]	1970–2003	Various models (such as Engle-Granger, Johansen, fully modified OLS)	Sri Lanka	Long-run: range from -0.63 to 0; short run: 0
Narayan et al.[19]	1978-2003	OLS and dynamic OLS Panel cointegration	G7 Nations	Model 1: LR: 1.45(OLS) 1.56 (DOLS) SR:-0.1 Model 2: LR 0.35 (OLS) 0.37 (DOLS) SR: not significant
Bianco et al.[20]	1975-2008	Cointegrating OLS	Romania	Non-residential: SR: 0.136 to -0.076 LR: 0.469 to -0.247
Narayan et al.[21] (interpreted from Narayan and Smyth[22])	1974–2002	Granger causality	Six Middle Eastern countries	Long run: 0.04, but not statistically significant

Author(s) (South African studies)	Time period	Methodology	Notes	Price elasticity
Pouris[23]	1950–1983	Unconstrained distributed lag model	Aggregate electricity demand	Long run: -0.9; short run: NA
Blignaut and de Wet[7]	1976–1996	Calculation on a year-on-year basis	26 economic sectors	Average over period; varies from -0.306 to 0.760
Ziramba[24]	1978–2005	ARDL methodology	Residential electricity demand	Price was statistically insignificant
Amusa et al.[25]	1960–2007	ARDL approach	Aggregate electricity demand	Price was statistically insignificant
Inglesi[3]	1980–2005	Engle-Granger cointegration model with ECM	Aggregate electricity demand	Long run:-0.56 and short run: statistically not significant
Inglesi and Pouris[26]	1980–2005	Engle-Granger cointegration model with ECM	Aggregate electricity demand	Long run:-0.56 and short run: statistically not significant
Inglesi-Lotz and Blignaut[9]	1993–2006	Panel data techniques: seemingly unrelated regressions	Five economic sectors	Industrial sector: -0.869; Transport, Commercial, Mining and Agriculture: statistically not significant
Inglesi-Lotz[6]	1980–2005	Kalman filter methodology	Aggregate electricity demand	Varying from -1 in the mid 1980s to close to zero in the mid 2000s
Inglesi and Pouris[26]	1970–2007	Kalman filter methodology	Industrial electricity demand	Varying from -1 in 1980 to -0.953 in 1990 and then stabilised at about -0.95
Kohler[27]	1989–2009	ARDL methodology	Industrial and total electricity demand	Total energy: -0.939 whilst industrial sectors vary from not significant to significant and negative

From these observations, it can be inferred that the sensitivity of consumers to tariff changes is a dynamic process: different among countries and variable depending on the specific conditions of the country's electricity market and economy in its entirety. Beyond the anthropogenic effects, climate also has an impact on electricity demand[28] and may affect sectors differently. For these reasons, we aimed to re-examine the sectoral elasticities of electricity demand for South Africa since the tariff increases of 2008 because consumer behaviour towards electricity usage may have shifted.

The direction and magnitude of the behavioural change and adaptation of any particular sector varies highly. However, these changes can be broadly categorised into two outcomes: the industry does not adapt and output is affected, or the industry adapts and output is 'minimally' affected. In terms of what can be done with regard to price changes, a neat case study by Thollander et al.[29] exists for the Swedish iron foundry industry. Thollander et al.'s[29] findings indicate that electricity price rises could lead to the industry employing 'energy efficiency' measures. The extent of these measures varies, but in another case, also in Sweden, it was found that in the face of price hikes in a chemical wood pulp mill, consumption of electricity could be reduced by up to 50% through efficiency measures.[29]

If this was the case in South Africa, we would expect to see elasticity remain insignificant and not constrain output for the particular industry under investigation. Industries that may exhibit such qualities to some extent, based on the work of Inglesi-Lotz and Blignaut[9], include the agricultural as well as non-ferrous metal sectors. These sectors have shown considerable gains in energy efficiency over the period 1993–2006. Given this historical trend towards efficiency in these two sectors, the price increases of 2008 may ultimately serve as catalysts to accelerate this effect. On the other hand, a sector that may be particularly affected is the iron and steel sector, which showed a large decrease in energy efficiency over the same period.[10] Electricity price changes might therefore be very damaging to the output of this sector.

Research methodology

Theoretical approach

In a perfect empirical world in which all information is available and quantifiable, a rather straightforward empirical description of the factors affecting the demand for electricity (q_{it} in Equation 1) exists. The main factors influencing electricity consumption are: (1) sector-level electricity demand characteristics, denoted by *X*, (2) availability of electricity (supply side factors), denoted by *Z* and (3) the price of electricity (p). Throughout this study, a panel approach is followed, with the subscript $i=1,\dots N$ referring to the individual sector under investigation whilst the subscript $t=1,\dots T$ refers to the time period under investigation.

$$q_{it} = F(X_{it};\, Z_{it};\, p_{it}) \qquad \text{Equation 1}$$

Typically, one would expect to estimate the supply side to arrive at an estimation of the supply price, and use the estimate of the supply price in the demand equation. In that structure, *Z* would not appear directly in the demand estimate, it would rather appear indirectly. However, in a monopolistic set-up such as the South African electricity supply sector, the supply curve is actually the marginal revenue curve, although only that portion thereof that lies above the marginal cost curve. Presumably, this marginal cost curve is (in the short run) flat within the boundary of the capacity constraints.

Electricity tariffs are exogenously determined by the national supplier of electricity, Eskom (as approved by NERSA). The tariffs are not determined through the interaction of supply and demand, but are an administrative charge determined by a national process of consultation and then decided on by the regulator. Consequently, and combined with the fact that electricity supply in the country has a specified ceiling, electricity supply is not considered to be a factor affecting electricity demand.

Based on this scenario, a regression of the following form is estimated following the most used specification in the literature[3,6,2430]:

$$q_{it} = F\,(X_{it};\, p_{it}), \qquad \text{Equation 2}$$

where q is the annual electricity consumption per sector; p is the annual average tariff charged to each sector and X is the annual economic output per sector.

Econometric methodology

The data covered a number of sectors in the period 2000–2012, thus forming a panel database. Firstly, unit root tests are used to formally assess whether the panel data set is stationary or not. The tests proposed by Levin et al.[31] and Im et al.[32] are used to conclude whether the panel data set studied exhibits stationarity attributes. The first tests for the existence of a common unit root process while the second tests for individual unit roots.

If the results of the tests show that the series are stationary, then the analysis proceeds with the estimation of seemingly unrelated regressions (SUR). If not, Pedroni's[33,34] panel cointegration test is used to investigate whether the series are cointegrated or not. If the series are cointegrated, then we continue with the SUR estimation as our focus is primarily on the long run. The Pedroni test allows for interdependence between cross sections.

$$y_{it}=a_i+\delta_i t+\gamma_{1i}X_{1it}+\gamma_2 X_{2it}+\varepsilon_{it} \qquad \text{Equation 3}$$

The coefficients α_i and δ_i allow for sector-specific fixed effects and deterministic trends, respectively, while ε_{it} denotes the estimated residuals representing deviations from the long-run. The null hypothesis is that there is no cointegration (in other words, the residuals are non-stationary). A unit root test is conducted as follows:

$$\varepsilon_{it}=p_i\varepsilon_{it-1}+w_{it}. \qquad \text{Equation 4}$$

The panel tests proposed by Pedroni[33,34] use the following four statistics: panel ν, panel ρ, panel PP and panel ADF-statistic. These can be described as follows:

> *These statistics pool the autoregressive coefficients across different countries for the unit root tests on the estimated residuals. These statistics take into account common time factors and heterogeneity across countries. The group tests are based on the between dimension approach which includes three statistics: group r, group PP, and group ADF-statistics. These statistics are based on averages of the individual autoregressive coefficients associated with the unit root tests of the residuals for each country in the panel.34(p.657)*

Studies that used panel data techniques[9,10,16,17] usually first estimated a pooled estimation followed by a fixed effects model to capture cross-sectional heterogeneity. The pooled effects model is, however, considered to be limited for a number of applications, as it does not take into account any cross-section heterogeneity among the sectors. The pooled effects model presents a joint estimation of coefficients as follows:

$$y_{it} = \beta_0+\beta_1X_{1,it}+\beta_2X_{2,it}+\varepsilon_{it}, \qquad \text{Equation 5}$$

for $i = 1....N$ and $t = 1.....T$,

where y_{it} is the dependent variable observed for individual *i* at time *t*; $X_{1,it}$ and $X_{2,it}$ are the time-variant regressors; β_0 is the constant; β_1 and β_2 are the slope coefficients; and ε_{it} is the error term. However, 'pooling' has some specific characteristics, such as the increase of the degrees of freedom, hence the potential low standard errors on the coefficients as a result. Also, except for the same slope coefficients, Equation 5 assumes a common intercept. To be able to distinguish between different effects, Equation 5 can be rewritten as:

$$y_{it}= \beta_0+\beta_1X_{1,it}+\beta_2X_{2,it}+\alpha_i+u_{it}, \qquad \text{Equation 6}$$

for $i= 1....N$, and $t = 1.....T$,

where α_i is the unobserved individual effect and u_{it} is the error term.

There are two methods of dealing with the unobserved individual effect: the fixed effect model and the random effects model. The first assumes that α_i is not independent of $X_{1,it}$ and $X_{2,it}$ while the latter assumes that α_i is independent of $X_{1,it}$ and $X_{2,it}$ or $E(\alpha i \mid X_{1,it}, X_{2,it}) = 0$. However, these estimations assume an average price coefficient for all the cross-sections (economic sectors in this application). Thus, to describe the behaviour of each sector separately and hence capture differences among sectors, SURs are estimated. Equation 6 should be amended (by representing a different coefficient for each i in order to represent a SUR) as follows:

$$y_{it}=\beta_{0,I}+\beta_{1,i}X_{1,it}+\beta_{2,i}X_{2,it}+\varepsilon_{it}, \qquad \text{Equation 7}$$

for $i = 1....N$ and $t = 1.....T$.

For this specific exercise, Equation 7 is transformed into:

$$\text{electricity consumption}_{it}=a_{0,i}+a_{1,i}\text{price}_{it}+ a_{2,i}\text{output}_{it\,+}u_{it} \qquad \text{Equation 8}$$

where, once again, *i* is 1,….*N* for each sector; *t* is 1,….*T* is the time period; electricity consumption is the electricity consumption of each sector; price is the tariff charged by Eskom to each of the sectors; output is the economic output of each of the sectors and α_0, α_1, α_2 are the slope coefficients for each of the variables. All the variables are in their natural logarithms.

Data

Industrial sectors

The choice of the sectors to be included was dictated by the availability of data. These industrial sectors are:

- Agriculture: agriculture, forestry and fishing
- Coal mining
- Gold and platinum mining
- Rest of mining
- Iron and steel: basic iron and steel; metal products excluding machinery; and machinery and equipment
- Liquid fuels: coke and refined petroleum
- Non-ferrous metals
- Rest of chemicals: non-metallic minerals; basic chemicals; other chemicals and human-made fibres; rubber products and plastic products
- Rest of manufacturing: food, beverages and tobacco; electrical machinery and apparatus; textiles, clothing and leather; transport equipment; wood and paper; publishing and printing; glass and glass products; and furniture and other manufacturing
- Transport
- Commercial: trade, catering and accommodation services; financial intermediation, insurance, real estate and business services; and community, social and personal services

Given that the analysis is over 11 sectors ranging from 2002 to 2012, the cross-sectional dimension for the panel analysis performed here can be defined as *i* = agriculture, coal mining…to transport and commercial. In the case of the time variable, *t* represents any year from 2002 to 2012.

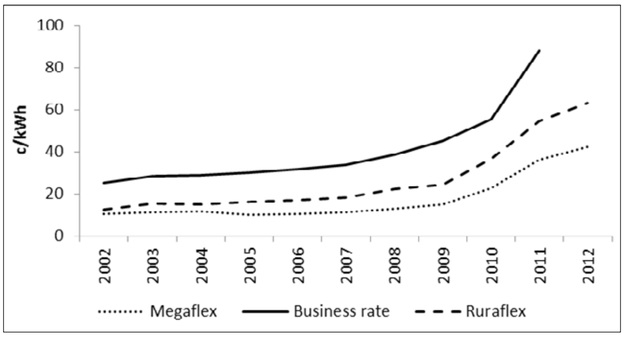
Electricity tariffs

With regard to tariffs, Eskom has a very detailed analysis of their tariffs and the time of use (TOU) structure. The issuing of tariff books started in 1995, resulting in a rich collection of tariff data available. The descriptions of the tariffs that will be used here are as follows:

- **Ruraflex**[35] [p. 44–45]: TOU electricity tariff for rural customers with a notified maximum demand (NMD) from 25 kVA.
  In our application, Sector 1 (agriculture) falls in this category. Tariffs are different between high-demand and low-demand seasons and among three categories within each season (peak/standard and off-peak). A weighted average tariff for each year was estimated using the consumption level for each TOU slot for each year.
- **Megaflex**[35] [p. 20–21]: TOU electricity tariff for urban consumers with a NMD greater than 1 MVA.
  In our application, Sectors 2–10 fall in this category. Prices are different between high-demand and low-demand seasons and among three categories within each season (peak/standard and off peak), and a weighted average was estimated for each sector based on the consumption for each TOU slot for each year.
- **Businessrate**[35] [p. 26]: Suite of electricity tariffs typically for commercial use and for high-consumption, non-commercial supplies in urban areas with a NMD of up to 100 kVA.
  In our application, Sector 11 (Commercial sector) falls in this category. Different charges for different types of business rate are described as well as an extra environmental levy charged per season of the year, with a weighted average estimated for each year based on the consumption levels.

Figure 1 shows the three weighted real averages for Megaflex, Ruraflex and Businessrate (The price series has been deflated using the consumer price index of the South African Reserve Bank.[36])

From Figure 1, it can be observed that the real electricity tariffs charged by Eskom were relatively low and stable for the years 2002–2008. As remarked earlier, in 2007/2008, the country experienced a severe electricity supply crisis resulting in prolonged periods of load-shedding with negative consequences for the entire economy. Following that, NERSA approved high increases in tariffs for the next 4 years (approximately 25% per annum).



Source: Calculated based on various Eskom tariff books[35] and reserve bank CPI[36]

Figure 1: Weighted real average annual Eskom tariffs.

Electricity consumption

The annual electricity consumption data set was provided by Eskom and includes the Eskom national sales directly to each consuming sector. Figure 2 shows the electricity consumption per sector from 2002 to 2012.

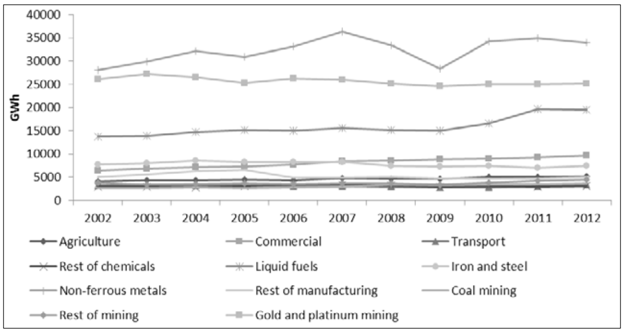
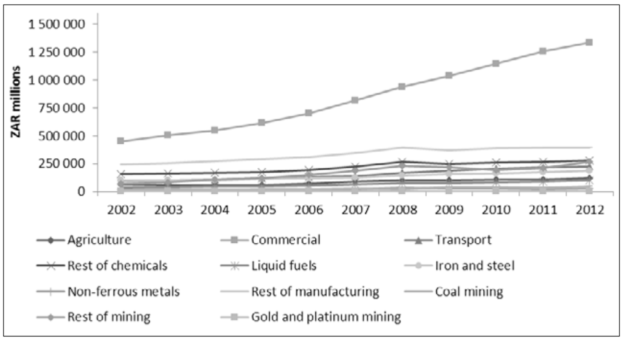


Figure 2: Sectoral electricity consumption: 2002–2012 (GWh).

Economic output

The data for the output of the various sectors were obtained from Quantec's *South African Standardized Industry Indicator Database* and are given in South African rand (ZAR) millions at 2005 constant prices. The data set extends from 1970 to 2012. Figure 3 presents the economic output per sector from 2002 to 2012.



Source: Quantec[37]

Note: USD1=ZAR9.81 (effective rate of exchange on 18 September 2013)

Figure 3: Sectoral economic output: 2002–2012 (ZAR million, constant 2005 prices).

Empirical results

As discussed in the methodology section, the first step of the exercise was to examine the stationarity properties of the variables used in the estimation. The Levin et al.[31] and Im et al.[32] tests were performed on these series. According to the conclusions for the different specifications, there is evidence that the panel data set exhibits stationarity properties (the results can be made available by the authors upon request).

As the results of the stationarity tests cannot reject the null hypothesis of stationarity, we can continue with the SUR estimation for the different economic sectors, where electricity consumption will be represented here as a function of:

- real economic output whose coefficients are expected to be positive (*a priori* expectations)
- electricity tariffs whose coefficients are expected to be negative (*a priori* expectations)

Then based on the price restructuring after 2008, illustrated in Figure 1, the following dummy was used to capture the important changes in the electricity sector since 2008:

$$\text{dum} = \begin{cases} 0, \textit{from } 2002 \textit{ to } 2007 \\ 1, \textit{from } 2008 \textit{ to } 2012 \end{cases}. \qquad \text{Equation 9}$$

The assumption is that the price restructuring might have affected the behaviour of some sectors with regard to their decision on how much electricity to consume. Thus, the final equation to be estimated was as follows (all variables in natural logs):

$$\text{electricity consumption}_t = \alpha_0 + \alpha_1 \text{price}_t + \alpha_2 \text{dum} * \text{price}_t + \alpha_3 \text{output}_t \qquad \text{Equation 10}$$

With the use of the dummy, the assumption was that the sectors may behave differently after the price increases of 2008. The *a priori* expectation is that the tariffs did not play an important role in electricity consumption before 2008, but that they did so thereafter. This assumption is derived from Inglesi-Lotz and Blignaut[9], who found price coefficients were statistically insignificant until 2006 (the last year of the sample because of data availability) and Inglesi-Lotz[6] who found the price coefficient was close to zero for the last part of the time period (last year was 2005 because of data availability). From 2002 to 2007, the price elasticity was equal to α_1 and from 2008 to 2012 was equal to $\alpha_1 + \alpha_2$ (where the coefficients are statistically significant).

Table 2 presents the results of the SUR estimation. The output coefficients are all positive and statistically significant, confirming our *a priori* expectations based on economic theory: the higher the production output of a sector, the higher the electricity usage of the sector.

In order to interpret the price elasticities, we should look at them for two separate periods: 2002–2007 and 2008–2012. As mentioned earlier, before 2008, price elasticity was represented by only the price coefficient but thereafter, it was represented by the sum of the price coefficient and the coefficient of the interaction variable of the dummy with price (for sectors for which the elasticities for both periods were statistically significant). Table 3 presents these coefficients.

For the period 2002–2007, the majority of the sectors exhibited the same pattern: the price was statistically insignificant, or in other words, did not play a role in the changes in electricity consumption. However, after the price restructuring of 2007/2008, the sectors show a significantly higher sensitivity to price changes, illustrated by the statistically significant coefficients of the interaction of price with the dummy. The lack of importance before 2008 can be attributed to the relatively low and stable level of prices.[6] Further evidence in support of this argument is that the majority of the sectors exhibited a structural break around the time that the price increase took effect (Chow test results are available upon request). Of those sectors that did exhibit a structural break, all are sectors that are heavily dependent on energy inputs and can hence be expected to react very quickly to price changes. Some sectors did not exhibit a break, possibly because of the limited number of observations available after the break dates. This limited number of observations means that it will be more difficult for the Chow test to detect more subtle structural changes. It may therefore be advantageous to conduct a more in-depth study on this issue of structural breaks once more data have become available for analysis. However, the results of the Chow test do not detract from the findings presented here, which are all statistically significant and of the *a priori* expected sign according to economic theory after the year 2008.

In the Commercial sector, the price changes did not seem to affect the electricity usage in the first period and the sector exhibited the lowest price elasticity in absolute values for the second period. That finding can be explained by the fact that the Commercial sector is not among the electricity intensive sectors of the country with the cost of electricity being a small portion of overall cost,[1] and hence, tariff fluctuations are not a factor to their small electricity consumption.

The Mining sector is another exception to the norm of the results. The Gold and Platinum Mining sector was found to have reacted negatively to price changes even for the period 2002–2007. This sensitivity dropped after the tariff increases of 2008. This drop may be explained by the platinum mining sector's electricity consumption, which is driven by large increases in output (output coefficient = 3.824, substantially higher than the rest). During the period, the world experienced a platinum 'bubble' and these figures reflect this.

There are only two sectors in which an anomaly is observed: Other Mining (which includes diamonds, quarrying, etc.) and Non-ferrous Metals. The anomaly lies in the price elasticities before the price increases of 2008 – the elasticities were 1.068 and 0.821, respectively. The anomaly did, however, disappear in the period 2008–2012 as the price elasticities during this period were found to be both negative and statistically significant.

Conclusion and discussion

South Africa has seen many studies with respect to price elasticity in the recent past. These studies were conducted either at a national or at a sectoral level and included data prior to 2008. We aimed to examine the price elasticities of electricity demand in South Africa with the purpose of testing whether consumer behaviour changed and by how much after the energy crisis of 2007/2008 and price restructuring of 2008/2009.

Table 2: Seemingly unrelated regression estimation results

	Dependent variable: Electricity consumption					
	Price coefficient	*p*-value	Dummy* price coefficient	*p*-value	Output coefficient	*p*-value
Agriculture	0.095	0.846	-0.235	0.006***	1.973	0.000***
Coal Mining	-0.201	0.652	-0.291	0.001***	2.243	0.000***
Commercial	-0.325	0.566	-0.190	0.003***	1.801	0.000***
Gold and Platinum Mining	-1.745	0.001***	-0.417	0.000***	3.824	0.000***
Iron and Steel	-0.044	0.919	-0.279	0.001***	1.977	0.000***
Liquid Fuels	0.577	0.163	-0.418	0.000***	2.027	0.000***
Non-ferrous Metals	0.821	0.037**	-0.342	0.000***	2.210	0.000***
Rest of Chemicals	0.154	0.708	-0.240	0.005***	1.771	0.000***
Rest of Manufacturing	0.319	0.439	-0.251	0.004***	1.716	0.000***
Rest of Mining	1.068	0.007**	-0.465	0.000***	1.632	0.000***
Transport	0.192	0.652	-0.346	0.000***	1.829	0.000***

R-squared = 0.89041; adjusted R-squared = 0.85056; SE of regression = 0.362046; Sum squared resid = 9.03707; Durbin–Watson stat = 1.66522.

Note:*, **, *** denote 10%, 5% and 1% level of statistical significance

Note: the terms 'price' and 'tariff' are considered interchangeable.

Table 3: Electricity price elasticities before and after 2008

	2002–2007	2008–2012
Agriculture	Non-significant	-0.235
Coal Mining	Non-significant	-0.291
Commercial	Non-significant	-0.190
Gold and Platinum Mining	-1.745	-0.417
Iron and Steel	Non-significant	-0.279
Liquid Fuels	Non-significant	-0.418
Non-ferrous Metals	0.821	-0.342
Rest of Chemicals	Non-significant	-0.240
Rest of Manufacturing	Non-significant	-0.251
Rest of Mining	1.068	-0.465
Transport	Non-significant	-0.346

The inclusion of further data beyond the energy crisis allowed for a comparison between pre- and post-energy crisis elasticities.

For the period 2002–2007, the estimates for price elasticity were statistically insignificant for the majority of the sectors, indicating that price did not play a role in the changes in electricity consumption. However, after 2008, the sectors showed a significantly higher sensitivity to price changes, illustrated by the statistically significant coefficients of the interaction of price with the dummy. The general trend we found is that those elasticity values which were insignificant before became significant, while those that were significant became more negative.

As in Inglesi-Lotz and Blignaut[9], the differences among various sectors' electricity usage to price changes are also noted here. However, the overall trend of becoming more sensitive after the 2008/2009 price restructuring should be considered by policymakers. Future applications for increases in electricity tariffs (or higher environmental levies or carbon taxes) should take into account the fact that this might result in further decreases in electricity consumption. Decreases in consumption can materialise in two ways, with opposite effects. Some consumers will aim to consume energy more efficiently while others will turn to alternative and renewable forms of energy. This improvement and substitution will have a positive impact on environmental concerns. However, with higher prices of electricity, the variable costs for many small- and medium-sized enterprises will, in some cases, be unbearable and cause them to close down, thus putting severe constraints on the economic production of the country. This possibility is echoed by E.ON[38] which highlights that using energy prices as a tool to encourage efficiency may damage growth in some sectors. A policy should be employed only for those sectors for which there is opportunity for efficiency gains.[38] Prima facie evidence for a sector that would struggle to adapt, as per the current analysis, is the Gold Mining and Platinum Sector.

From a governmental perspective, it may therefore be advantageous to apply differential pricing strategies to the various energy consuming sectors. As noted in Kohler[27], policymakers may in fact use energy pricing policy in such a manner as to discourage energy inefficiency within the South African economy by increasing the prices for energy inefficient users by more than those for energy efficient users. In so doing, the cost of inefficiency will rise and users of energy in South Africa will use energy more efficiently, reducing the demand for electricity somewhat in the long run.[39] This increased energy efficiency can be further enhanced by combining the differential pricing with subsidies specifically aimed at electricity users that adopt energy efficient technologies.[39] In fact, according to research by the Human Sciences Resource Council (HSRC),[40] there are numerous opportunities for energy efficiency (or energy savings) even in the most energy-intensive sectors, such as some mining sectors and some manufacturing sectors.[40] The HSRC furthermore identified low electricity prices as one of the impediments to the implementation of more efficient technologies.[40] Hence, a two-way approach pushing energy users away from inefficiency by making energy costly, as well as a cost-pull towards efficiency in the form of a subsidy, may be the way forward for South Africa, especially given how sensitive many sectors have become to electricity prices.

Acknowledgement

This research is based on work supported in part by the National Research Foundation of South Africa (UID: 83868).

Authors' contributions

J.P.W. collected the data and completed the literature review as well as the first draft of the manuscript; R.I-L. conducted the econometric analysis and interpreted the results together with J.B.; R.I-L. and J.B. also prepared the conclusions and policy implications of the study.

References

1. Inglesi-Lotz R, Blignaut JN. Electricity intensities of the OECD and South Africa: A comparison. Renew Sust Energ Rev. 2012;16:4491–4499. http://dx.doi.org/10.1016/j.rser.2012.04.004

2. Nersa: Power crisis costs South Africa about R50 million. Mail and Guardian [online]. 2008 August 26 [cited 2009 Mar 30]. Available from: http://mg.co.za/article/2008-08-26-nersa-power-crisis-cost-sa-about-r50bn

3. Inglesi R. Aggregate electricity demand in South Africa: Conditional forecasts to 2030. Appl Energ. 2010;87:197–204. http://dx.doi.org/10.1016/j.apenergy.2009.08.017

4. Pouris A. Energy and fuels research in South African universities: A comparative assessment. Open Inform Sci J. 2008;1:1–9. http://dx.doi.org/10.2174/1874947X00801010001

5. Bayliss K. Lessons from the South African electricity crisis. Report no. 56. Brasilia: International Policy Centre for Inclusive Growth; 2008.

6. Inglesi-Lotz R. The evolution of price elasticity of electricity demand in South Africa: A Kalman filter application. Energ Policy. 2011;39:3690–3696. http://dx.doi.org/10.1016/j.enpol.2011.03.078

7. Blignaut JN, De Wet T. Some recommendations towards reducing electricity consumption in the South African manufacturing sector. S Afr J Econ Manag S. 2001;42:359–379.

8. Odhiambo NM. Electricity consumption and economic growth in South Africa: A trivariate causality test. Energ Econ. 2009;31:635–640. http://dx.doi.org/10.1016/j.eneco.2009.01.005

9. Inglesi-Lotz R, Blignaut JN. Estimating the price elasticity of demand for electricity by sector in South Africa. S Afr J Econ Manag S. 2011;14:449–465.

10. Diabi A. The demand for electric energy in Saudi Arabia: An empirical investigation. OPEC Rev. 1998;22:13–29. http://dx.doi.org/10.1111/1468-0076.00039

11. Von Hirschhausen C, Andres M. Long-term electricity demand in China – From quantitative to qualitative growth? Energ Policy. 2000;28(4):231–241. http://dx.doi.org/10.1016/S0301-4215(00)00014-8

12. Al-Faris AR. The demand for electricity in the GCC countries. Energ Policy. 2002;30:117–124. http://dx.doi.org/10.1016/S0301-4215(01)00064-7

13. Kamerschen DR, Porter DV. The demand for residential, industrial and total electricity, 1973–1998. Energ Econ. 2004;26:87–100. http://dx.doi.org/10.1016/S0140-9883(03)00033-1

14. Narayan PK, Smyth R. The residential demand for electricity in Australia: An application of the bounds testing approach to co-integration. Energ Policy. 2005;33:467–474. http://dx.doi.org/10.1016/j.enpol.2003.08.011

15. De Vita G, Endresen K, Hunt IC. An empirical analysis of energy demand in Namibia. Energ Policy. 2006;34:3447–3463. http://dx.doi.org/10.1016/j.enpol.2005.07.016

16. Atakhanova Z, Howie P. Electricity demand in Kazakhstan. Energ Policy. 2007;35:3729–3743. http://dx.doi.org/10.1016/j.enpol.2007.01.005

17. Narayan PK, Smyth R, Prasad A. Electricity consumption in G7 countries: A panel co integration analysis of residential demand elasticities. Energ Policy. 2007;35:4485–4494. http://dx.doi.org/10.1016/j.enpol.2007.03.018

18. Amarawickrama HA, Hunt LC. Electricity demand for Sri Lanka: A time series analysis. Energy. 2008;33(5):724–739. http://dx.doi.org/10.1016/j.energy.2007.12.008

19. Narayan PK, Smyth R, Prasad A. Electricity consumption in G7 countries: A panel cointegration analysis of residential electricity demand. Energ Policy. 2007;35:4485–4494. http://dx.doi.org/10.1016/j.enpol.2007.03.018

20. Bianco V, Manca O, Nardini S, Minea MA. Analysis and forecasting of nonresidential electricity consumption in Romania. Appl Energ. 2010;87:3584–3590. http://dx.doi.org/10.1016/j.apenergy.2010.05.018

21. Narayan PK, Narayan S, Popp S. Does electricity consumption panel Granger cause GDP: A new global evidence. Appl Energ. 2010;87:3294–3298. http://dx.doi.org/10.1016/j.apenergy.2010.03.021

22. Narayan KP, Smyth R. Multivariate granger causality between electricity consumption, exports and GDP: Evidence from a panel of Middle Eastern countries. Energ Policy. 2009;37:229–236. http://dx.doi.org/10.1016/j.enpol.2008.08.020

23. Pouris A. The price elasticity of electricity demand in South Africa. Appl Econ. 1987;19:1269–1277. http://dx.doi.org/10.1080/00036848700000073

24. Ziramba E. The demand for residential electricity in South Africa. Energy Policy. 2008;36:3450–3466. http://dx.doi.org/10.1016/j.enpol.2008.05.026

25. Amusa H, Amusa K, Mabugu R. Aggregate demand for electricity in South Africa: An analysis using the bounds testing approach to cointegration. Energ Policy. 2009;37:4167–4175. http://dx.doi.org/10.1016/j.enpol.2009.05.016

26. Inglesi R, Pouris A. Forecasting electricity demand in South Africa: A critique of Eskom's projections. S Afr J Sci. 2010;106(1/2):50–54. http://dx.doi.org/10.4102/sajs.v106i1/2.16

27. Kohler M. Differential electricity pricing and energy efficiency in South Africa. Energy. 2014;64:524–532. http://dx.doi.org/10.1016/j.energy.2013.11.047

28. Ahmed T, Muttaqi KM, Agalgaonkar AP. Climate change impacts on electricity demand in the State of New South Wales, Australia. Appl Energ. 2012;98:376–383. http://dx.doi.org/10.1016/j.apenergy.2012.03.059

29. Thollander P, Karlsson M, Söderstrom M, Creutz D. Reducing industrial energy costs through energy-efficient measures in a liberalized European electricity market: Case study of a Swedish iron foundry. Appl Energ. 2005;81:115–126. http://dx.doi.org/10.1016/j.apenergy.2004.07.006

30. Inglesi-Lotz R. The sensitivity of the South African industrial sector's electricity consumption to electricity price fluctuations. Working paper no. 201215. Pretoria: Department of Economics, University of Pretoria; 2012.

31. Levin A, Lin CF, Chu C. Unit root tests in panel data: Asymptotic and finite-sample properties. J Econometrics. 2002;108:1–24. http://dx.doi.org/10.1016/S0304-4076(01)00098-7

32. Im KS, Pesaran MH, Shin Y. Testing for unit roots in heterogeneous panels. J Econometrics. 2003;115:53–74. http://dx.doi.org/10.1016/S0304-4076(03)00092-7

33. Pedroni P. Critical values for cointegration tests in heterogeneous panels with multiple regressors. Oxford B Econ Stat. 1999;61:653–670. http://dx.doi.org/10.1111/1468-0084.61.s1.14

34. Pedroni P. Panel cointegration: Asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis: New results. Economet Theor. 2004;20:597–627. http://dx.doi.org/10.1017/S0266466604203073

35. Eskom. Tariff & charges booklet 2012/13. Johannesburg: Eskom; 2013.

36. South African Reserve Bank. Quarterly bulletin time series database: Series KPB7155J. Pretoria: South African Reserve Bank; no date.

37. Quantec. South African standardized industry indicator database [database on the Internet]. No date [cited 2014 Oct 28]. Available from: www.quantec.co.za

38. Madlener R, Bernstein R, Gonzalez MAG. Econometric estimation of energy demand elasticities. E.ON Energy Research Center Series (E.ON.ERC) volume 3 issue 8. Aachen: E.ON Energy Research Center; 2011.

39. Deloitte. Economic impact of electricity price increases on various sectors of the South African economy [document on the Internet]. No date [cited 2014 Oct 28]. Available from: http://www.eskom.co.za/CustomerCare/MYPD3/Documents/Economic_Impact_of_Electrcity_Price_Increases_Document1.pdf

40. Altman M, Davies R, Mather A, Fleming D, Harris H. The impact of electricity price increases and rationing on the South African economy. Pretoria: Human Sciences Research Council Centre for Poverty Employment and Growth; 2008.

Estimation and monitoring of aboveground carbon stocks using spatial technology

AUTHORS:
Adolph Nyamugama[1,2]
Vincent Kakembo[1]

AFFILIATIONS:
[1]Department of Geosciences, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

[2]Agricultural Research Council – Institute for Soil, Water and Climate, Pretoria, South Africa

CORRESPONDENCE TO:
Adolph Nyamugama

EMAIL:
Adolph.nyamugama@gmail.com

POSTAL ADDRESS:
Agricultural Research Council – Institute for Soil, Water and Climate, Private Bag 0083, Pretoria 0001, South Africa

DATES:
Received: 26 May 2014
Revised: 30 Sep. 2014
Accepted: 04 Dec. 2014

KEYWORDS:
remote sensing; geographical information systems; GIS; subtropical thicket; model development

HOW TO CITE:
Nyamugama A, Kakembo V. Estimation and monitoring of aboveground carbon stocks using spatial technology. S Afr J Sci. 2015;111(9/10), Art. #2014-0170, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140170

Monitoring temporal changes of aboveground carbon (AGC) stocks distribution in subtropical thicket is key to understanding the role of vegetation in carbon sequestration. The main objectives of this research paper were to model and quantify the temporal changes of AGC stocks between 1972 and 2010 in the Great Fish River Nature Reserve and its environs, Eastern Cape Province, South Africa. We used a method based on the integration of remote sensing and geographical information systems to estimate AGC stocks in a time series framework. A non-linear regression model was developed using Normalised Difference Vegetation Index values generated from SPOT 5 High Resolution Geometric satellite imagery of 2010 as an independent variable and AGC stock estimates from field plots as the dependent variable. The regression model was used to estimate AGC stocks from satellite imagery for 1972 (Landsat TM), 1982 (Landsat 4 TM), 1992 (Landsat 7 ETM), 2002 (Landsat ETM+) and 2010 (SPOT 5) satellite imagery. AGC stocks for the respective years were compared by means of change detection analysis at the subtropical thicket class level. The results showed a decline of AGC stocks in all the classes from 1972 to 2010. Degraded and transformed thicket classes had the highest AGC stock losses. The decline of AGC stocks was attributed to thicket transformation and degradation, which were attributed to anthropogenic activities.

Introduction

The role of forests as a carbon source and sink has been widely explored. Recent studies have shown that forests store close to 289 919 t of carbon in trees and other vegetation.[1] However, the rate of deforestation has become a subject of major concern for many scientists. The global forest deforestation rate has been approximated at -7317% per year between 2000 and 2005.[1-3] Approximately 13 million ha of world forest was lost between 2000 and 2010,[4] implying an increase in the amount of carbon dioxide into the atmosphere.[1,5] Africa's forests are disappearing at a rate approximately four times more than that of the world average.[6] A study by the United Nations Environment Programme (UNEP) conducted in 2006 estimated that 70%, 95% and 30% of forests in West Africa, East Africa and the Democratic Republic of Congo, respectively, would be decimated by 2040. Deforestation and forest degradation contribute to atmospheric greenhouse gas emissions through the combustion of forest biomass and decomposition of the remaining plant material.[7]

The average deforestation rate in the Southern African Development Community (SADC) region is about 0.6%. The main causes of deforestation are conversion of forest land to agriculture and uncontrolled veld fires.[8] The subtropical thicket in the Eastern Cape Province of South Africa has suffered different levels of degradation, ranging from moderate to severe.[9-11] Land use/cover change activities were singled out as the major causes of degradation of the subtropical thicket. Approximately 60% of the primary subtropical thicket vegetation biome of South Africa has been severely degraded.[9,12] The disappearance of *Portulacria afra* thicket species has been observed in vast areas of the biome.[12,13]

Different methodological strategies have been employed to estimate aboveground carbon (AGC) stocks. Technological advancements have seen the introduction of remote sensing (RS) and geographical information system (GIS) techniques in this endeavour. It has been demonstrated that RS data in conjunction with ground information acquired at object level can help in the development of national and regional estimates of AGC stocks.[14-16] Using RS data, it is possible to monitor terrestrial ecosystems at various temporal and spatial scales. This approach has been widely tested for land-cover mapping and forestry applications. Previous studies have also reviewed the application of RS data for qualitative change detection of deforestation through land use/cover classification and quantification of forest AGC stock changes.[17-19]

While regression models for AGC stock estimations have been explored, most work has been dedicated to developing models that extrapolate destructive harvest data points to large scales, based on proxies measured from the ground and RS.[1,9,20,21] Different attributes are used for AGC stock estimation, depending on major factors such as climate and relief.[22-24] Most AGC stock estimations have been performed on a broad scale without analysing specific local conditions. Only a few studies[10,23,25,26] have attempted to account for variations in specific biomes. Many developed countries have at least one inventory of all their forest area that can be used as a baseline for further estimation. Conversely, very few developing countries have comprehensive national forest area inventories.[24] Against the above background, in this study we sought to estimate the changes in AGC stocks in the Great Fish River Nature Reserve (GFRNR) and its environs from 1972 to 2010. Landsat satellite imagery archives date back to 1972, providing a benchmark for estimating changes in AGC stocks within the 38-year period. A non-linear regression model was developed to estimate AGC stocks for this area using field plot AGC estimates and satellite imagery from 1972, 1982, 1992, 2002 and 2010.

The study area

The GFRNR, located in the Eastern Cape Province of South Africa (Figure 1), is divided into nature conservation areas which consist of three reserves: the Andries Vosloo Kudu Nature Reserve, the Double Drift Reserve and the Sam Knott Nature Reserve. Its environs include privately owned commercial farms and the highly populated communal villages of Glenmore, Tyefu, KwaNdlambe, Committee Drift, Chisira and Ncabasa. The study area is characterised by semi-arid conditions; both rainfall and temperature vary markedly with the seasons. The summer and winter season minimum and maximum temperatures range from 22 °C to 30 °C and -3 °C to 22 °C, respectively. While winters are dry, the mean annual rainfall of the area, which is about 450 mm, peaks in spring and autumn.[25]

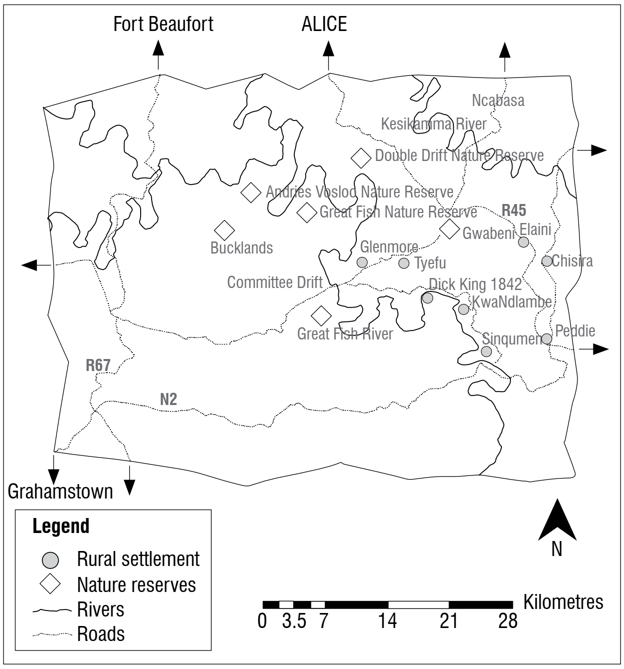


Figure 1: Map of the Great Fish River Nature Reserve and its environs – the study area.

Materials and methods

The methods – which entailed field surveys, image analysis, model development and quantifying AGC stocks – are illustrated in Figure 2. These procedures are described in the subsequent sub-sections.

Sampling design

According to the sampling design done prior to fieldwork, the study area was stratified into three subtropical thicket vegetation categories: intact, transformed and degraded thicket. This stratification was performed with the aid of a land use/cover map generated from the 2010 SPOT (HRG) imagery and was undertaken in order to obtain precise estimates of the subtropical thicket vegetation parameters. A large number of plots allowed for the measurement of the spatial variability of AGC stocks, increasing the confidence in AGC estimates. An optimum number of sample plots was selected using the formula below by Fuchs et al.[20]:

$$M = \frac{x^2 * CV^2}{AE^2}, \qquad \text{Equation 1}$$

where M is the minimum number of samples required, x is the value associated with specified probability, CV is the coefficient of variance and AE is the allowed error.

During field surveys conducted in February and March 2010, 90 plots of 30 m x 30 m were randomly allocated to the intact, transformed and degraded thicket strata. One subplot of 1 m x 1 m was also randomly selected from each of the 30 m x 30 m plots. The total number of trees was recorded for each plot and coordinates were recorded using a centimetre level precision Ashtech® ProMark2™ Global Positioning System (GPS) receiver. Within 1 m x 1 m subplots, seedlings, herbaceous material and litter were collected, dried to a constant mass and converted from biomass to carbon (tonnes) using a 0.48 conversion ratio.[13]

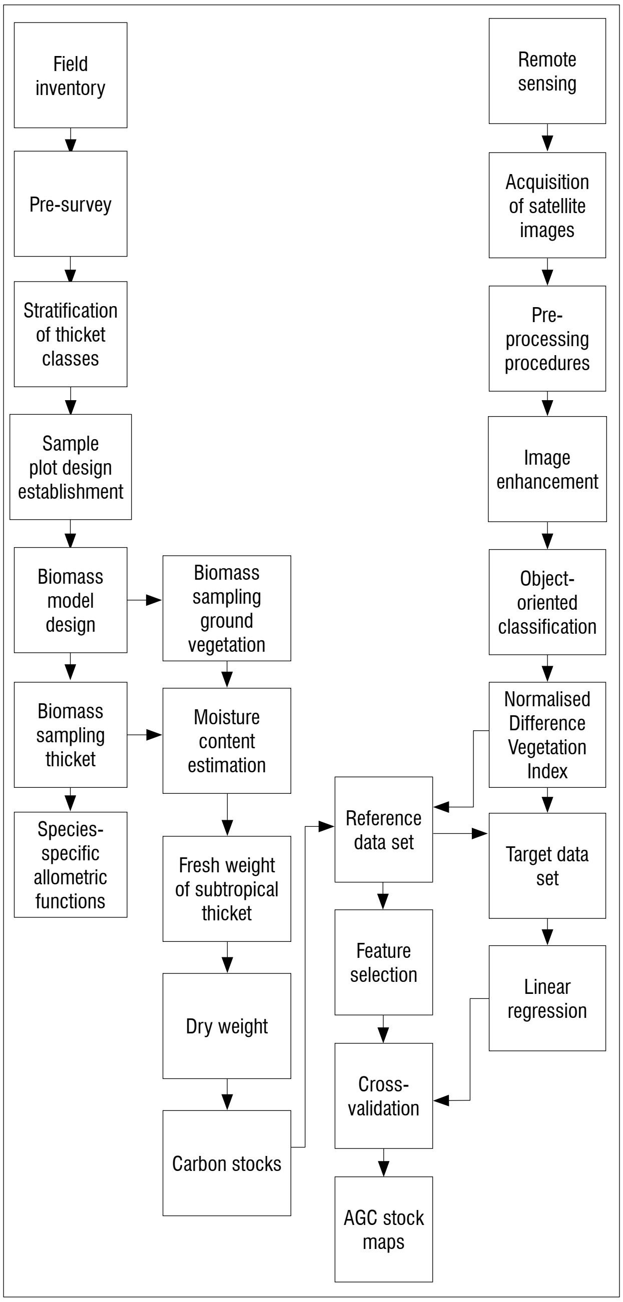

Field inventory
Pre-survey
Stratification of thicket classes
Sample plot design establishment
Biomass model design
Biomass sampling thicket
Species-specific allometric functions
Biomass sampling ground vegetation
Moisture content estimation
Fresh weight of subtropical thicket
Dry weight
Carbon stocks
Reference data set
Feature selection
Cross-validation
AGC stock maps
Remote sensing
Acquisition of satellite images
Pre-processing procedures
Image enhancement
Object-oriented classification
Normalised Difference Vegetation Index
Target data set
Linear regression

Figure 2: Framework for the integration of remote sensing and field data measurements to estimate aboveground carbon (AGC) stocks.

Aboveground biomass and AGC estimation

Samples of dominant tree species or guilds were selected in the respective strata. The trees were measured for height, canopy and basal diameter. The dominant plant species in each stratum were sampled and parts of aboveground portions (branches) were harvested and dried in the oven at a constant temperature of 80 °C. The dry mass of the branches was plotted against their canopy diameters and a power function was used to estimate the dry mass of all trees measured in each plot based on their canopy and the formula by Mills and Cowling[25] and Lu[27].

The use of allometric regression equations has proven to be a reliable and non-destructive method to estimate AGC stocks. The development of allometric equations in this study was based on the subtropical thicket vegetation characteristics present in the study area in conjunction with the ones used by Skowno[10], Patenaude et al.[12] and Myeong et al.[16] However, there are uncertainties associated with the use of allometric equations, such as the variation of vegetation at sites, soil type, climate, stand structure and genetic properties, which are associated with subtropical thicket vegetation. Therefore, in order to account for these uncertainties, only subtropical thicket vegetation guilds found in the study area were considered. A study by Lu[27] revealed that the use of general allometric equations results in errors in tree measurements, sampling and representation of plots over large areas. The use of age, leaf area index and hyperspectral RS, which are outside the scope of this study, can be used to improve the estimation of AGC stocks in future studies.

Image processing

Subset images covering the study area were extracted from the respective sets of imagery. The satellite images acquired were captured on 20 November 1972 for Landsat MSS, 20 December 1982 for Landsat 4 TM, 15 December 1992 for Landsat TM, 1 December 2002 for Landsat 7 ETM and 25 December 2010 for SPOT 5 HRG. Clearly distinguishable spectral reflectance patterns of different vegetation surfaces were therefore identifiable, as all imagery sets were captured in the summer rainfall season. Other data sets such as aerial photographs, topographic maps and SPOT 5 at a 10-m resolution were acquired for accuracy assessments. Geometric and radiometric correction were required for reliable change detection analysis using satellite imagery. Orthorectification was used to correct different angles, which are typical of multi-temporal data sets and also to ensure that images overlay perfectly with other GIS data sets. Temporal imagery (Landsat MSS, 4 TM, 7 ETM+ and SPOT 5 HRG) data sets were geo-referenced to a 2-m SPOT mosaic and projected to the Universal Transverse Mercator (UTM) system using World Geodetic Systems, zone 11, 1984 datum. A 20-m digital elevation model was used to correct relief displacement caused by local topography. At least 30 ground control points, evenly distributed on each image, were used in the geo-referencing process.[4]

Radiometric and atmospheric correction of surface reflectance of the SPOT 5 and Landsat data were conducted using the dark object subtraction option. The equations used for this procedure[21] are:

$$R\lambda = \mathrm{PI} * \mathrm{D} * (L\lambda - L\lambda.\mathrm{haze})/(\mathrm{Esun}\lambda * \mathrm{COS}(\theta)) \qquad \text{Equation 2}$$

$$L\lambda = \mathrm{DN}\lambda/A\lambda \text{ (for SPOT HRG data)} \qquad \text{Equation 3}$$

$$L\lambda = \mathrm{gain} * \mathrm{DN}\lambda + \mathrm{bias} \text{ (for Landsat TM data),} \qquad \text{Equation 4}$$

where L_λ is the apparent at-satellite radiance for spectral band λ, DN_λ is the digital number for band λ, A_λ is the calibration factor for spectral band λ of the HRG image, R_λ is the calibrated reflectance L_λ.haze is path radiance, Esun_λ is exo-atmospheric solar irradiance, D is the distance between earth and the sun and θ is the sun zenith angle.

Challenges encountered during the application of satellite imagery of different spatial resolution for change detection are outlined by Buyantuyev and Wu[26]. Studies done by Singh[14] and Im et al.[28] revealed that the post-classification approach minimises the problems caused by variation in sensors and atmospheric conditions, as well as vegetation phenology between different dates, because data from different dates are independently classified. Therefore, Landsat MSS, 4 TM, 7 ETM and SPOT 5 HRG images were independently portioned into image objects using Definiens Developer version 7 RS software.

Object-oriented classification

An object-oriented approach was used to independently segment the respective sets of imagery into three land-cover classes, namely intact, transformed and degraded. A fractal net evolution approach embedded in Definiens Developer 7 was used.[15] The segments were merged based on their level of similarity. Layer weights, scale, shape factor and compactness were set to 1, 20, 0.3 and 0.5, respectively. These variables were determined by visual interpretation of the results of image segmentation, where objects were considered to be internally homogenous, such that all pixels within a primitive object belong to one cover class.[29] Training sites were used for supervised classification in conjunction with the rule-based classification to classify each of the objects into one of the three land-cover classes.[15,30,31,32] The post-classification comparison approach was applied to detect the land-cover change classes.[15,29,31] The overall accuracies obtained were 0.91, 0.92, 0.88, 0.91 and 0.86 for the 1972, 1982, 1992, 2002 and 2010 imagery, respectively, signifying a high classification accuracy. The accuracy for 2010 SPOT imagery in particular was improved by linking training site GPS positions in the field to their corresponding positions on the satellite image.

Model development and validation

A non-linear relationship between carbon and the Normalised Difference Vegetation Index (NDVI) values was established to develop a regression model for AGC estimation from the 2010 SPOT 5 HRG image.[19,33] The 2010 SPOT image was selected because fieldwork was conducted in 2010 as well, during the months that coincided with the December summer. In this case, the NDVI was used as the independent variable, while AGC was the dependent variable.

The model was validated using sets of data obtained from field sites. A total of 20 sets of AGC stock estimates from field plots were compared with AGC stocks predicted by the model from the satellite image. AGC stocks for the GFRNR and environs were independently estimated.

The root mean square error (RMSE) was used to check the accuracy of the model using the equation by Canty et al.[18]

$$\mathrm{RMSE} = \sqrt{\sum \frac{(CT\text{-}EC)^2}{M}}\,, \qquad \text{Equation 5}$$

where RMSE is the root mean square error, CT is the carbon predicted by the model, EC is the estimated carbon and M is the number of observations.

The non-linear regression (quadratic) equation was chosen instead of a simple linear regression because it produced higher correlations: R^2=0.975 for intact thicket, R^2=0.812 for transformed thicket and R^2=0.725 for degraded thicket. These values are comparable to the results obtained by Myeong et al.[16], Hirata[21], Buyantuyev and Wu[26] and Lu[27] who established allometric relationships between thicket trees using attributes such as crown cover area, basal area and height, while Chubey et al.[31] established relationships between the crown and diameter, breast and height. It is noteworthy, however, that the natural conditions in which subtropical thicket vegetation grows do not follow linear relationships between diameter, crown area, basal area and height.[34,35]

Mapping AGC stock changes between 1972 and 2010

The validated non-linear regression model was used to map AGC stocks in subtropical classes for the entire study area in 2010. This model was also used to estimate and map the aboveground carbon stocks for the respective years. Mapping AGC stocks was then carried out for the three thicket classes (intact, transformed and degraded thicket). The non-linear regression model used for AGC stocks estimation in the study area was:

$$\mathrm{AGC} = 108.2\mathrm{e}^{(\mathrm{NDVI}10.0184)}, \qquad \text{Equation 6}$$

where AGC is the above ground carbon stocks (kg C/pixel) and NDVI is the SPOT NDVI value.[19]

After applying this model to estimate the AGC stocks, stock amounts were upscaled to t/ha.

Results

NDVI trends

The NDVI range for intact, transformed and degraded thicket between 1972 and 2010 is illustrated in Figure 3. A decreasing trend in the NDVI range for intact and degraded thicket is noticeable. The overall range for intact thicket declined from 0.53–0.77 to 0.38–0.62 between 1972 and 2010. This decline is particularly pronounced between 2002 and 2010. The range for degraded thicket also deteriorated from -0.05 – -0.04

in 1972 to -0.60 – -0.045 in 2010. These findings reflect tremendous changes in vegetation status, which have serious implications for AGC stocks, as explained in the subsequent sub-sections.

Model validation and AGC stock quantification

To validate the AGC stocks, model data sets obtained from the 90 field plots were plotted against the AGC stocks predicted from the imagery. A model validation result of R^2=0.960, which is quite significant at the 0.05 level, was obtained. This result implies a strong correlation between predicted AGC stocks and AGC stocks calculated in the field. The RMSE value of the model was 0.21, suggesting the model was highly accurate.

The validated non-linear regression model was applied to estimate the amount of AGC stocks for the entire study area, for the rest of the imagery sets. The models developed for the GFRNR and environs were used for the estimation, quantification and mapping of AGC stocks per class for the respective dates. Figure 4 illustrates the temporal changes of AGC stocks from 1972 to 2010 in the study area. AGC stock changes per subtropical thicket class in the GFRNR and environs are shown in Table 1.

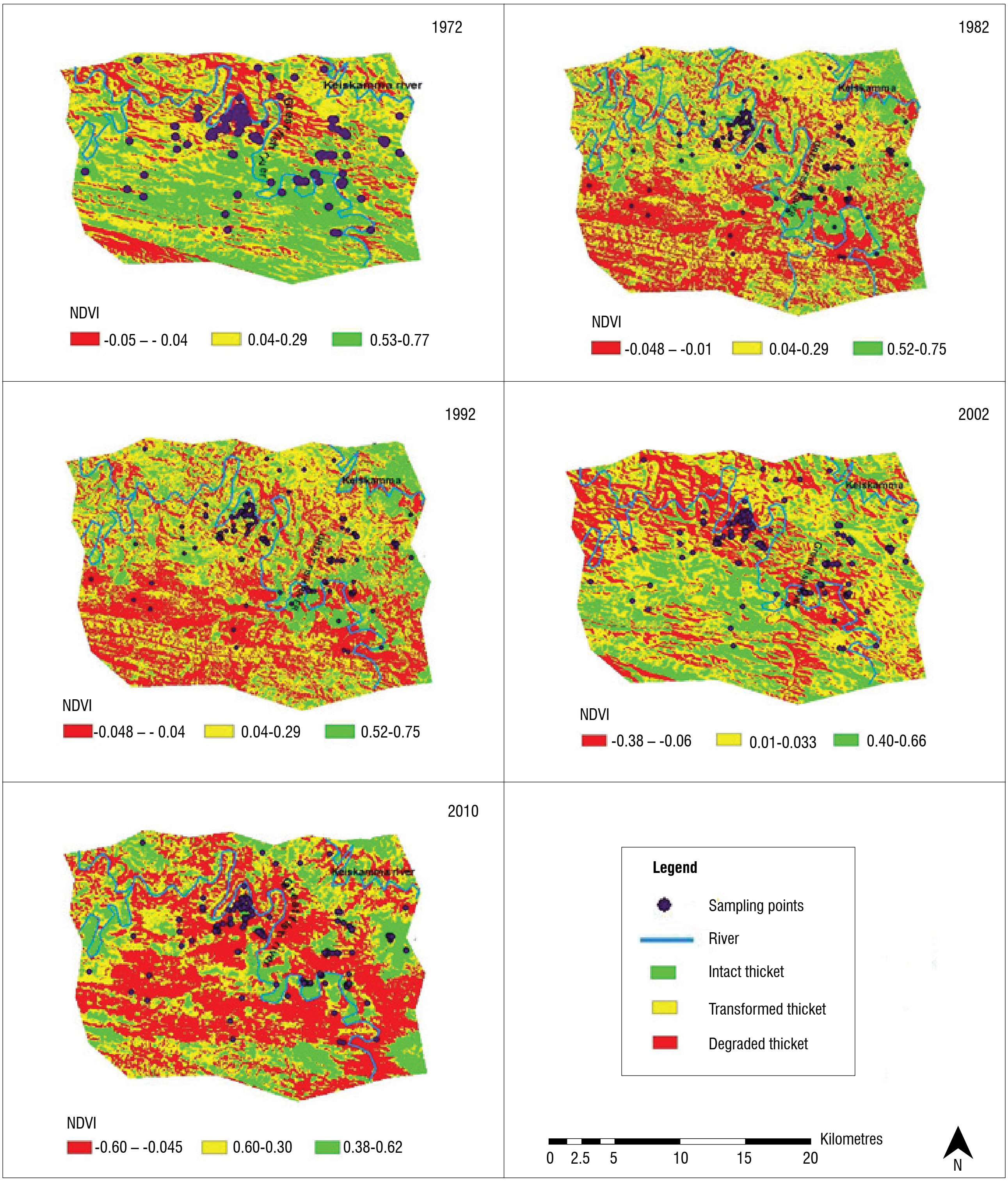


Figure 3: Normalised Difference Vegetation Index (NDVI) ranges per class from 1972 to 2010. The dots on the NDVI surfaces represent field sampling points.

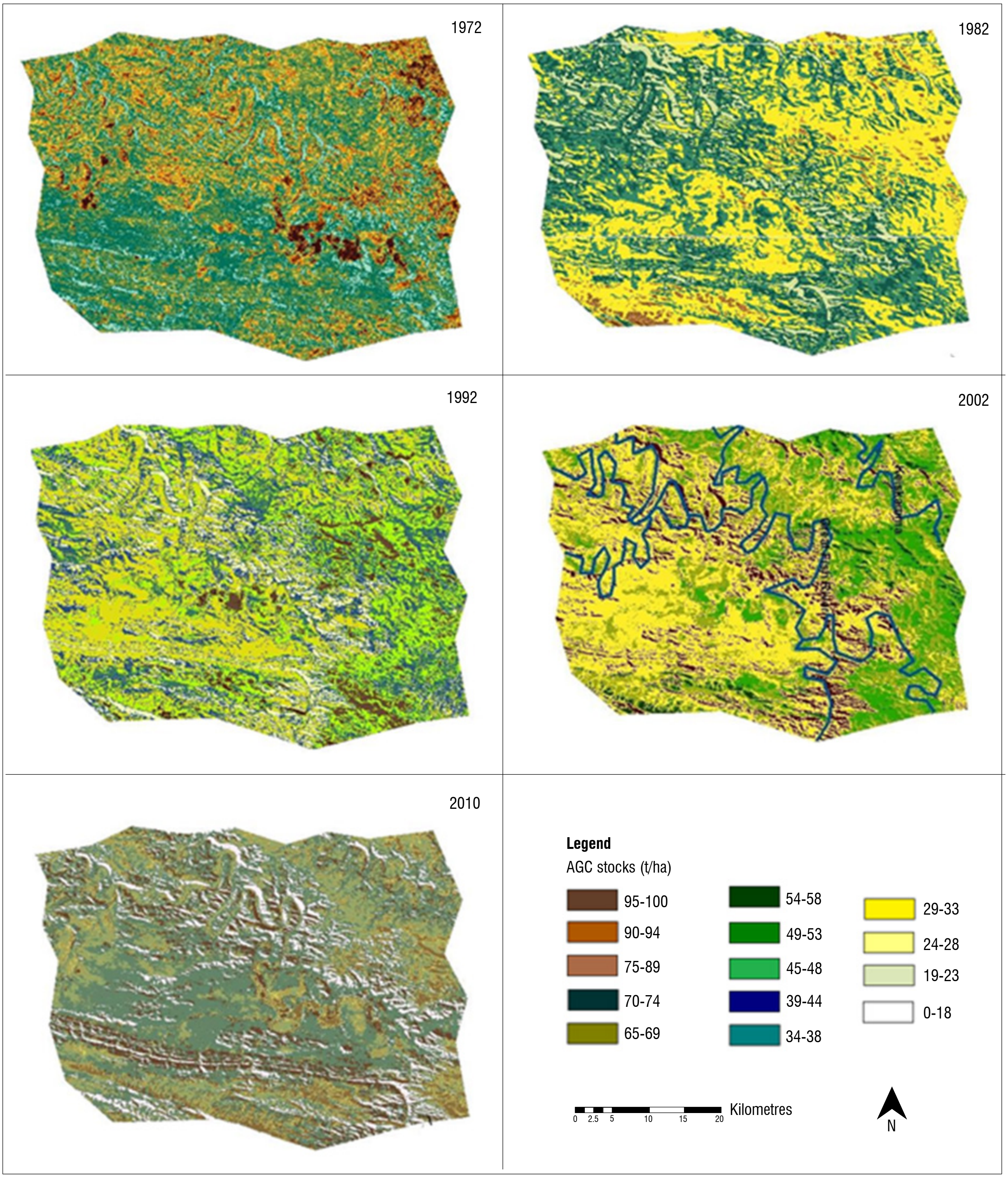


Figure 4: Aboveground carbon (AGC) stock changes between 1972 and 2010.

Temporal changes of AGC stocks

Change in mean AGC stocks in tonnes between 1972 and 1982 illustrates that intact, transformed and degraded thicket classes decreased by 15 t/ha, 15 t/ha and 5 t/ha, respectively. The decreases were quite considerable for all the classes, as reflected by the percentage change (Table 1). It can be noted that intact and transformed thicket had the highest decrease in AGC stocks (15 t/ha each). Between 1982 and 1992, the decrease in mean AGC stocks for intact and degraded thicket was tremendous, as indicated by 12.5% and 37%, respectively, while there was no change in transformed thicket stocks.

The period between 1992 and 2002 registered the lowest decrease in mean AGC stocks (Table 1). Conversely, mean AGC stock losses were highest between 2002 and 2010, coinciding with the pronounced decline in the NDVI range pointed out earlier. Overall, a high percentage of net mean AGC stock losses of 15 t/ha, 15 t/ha and 35 t/ha for intact, transformed and degraded thicket, respectively, was registered between 1972 and 2010.

Table 1: Comparison of mean aboveground carbon (AGC) stocks between 1972 and 2010

Thicket class	Average AGC (t/ha)		Change in AGC (t/ha)	Change in AGC (%)
	1972	**1982**	**1972–1982**	**1972–1982**
Intact	95	80	15	15.7
Transformed	80	65	15	18.7
Degraded	60	55	5	8
	1982	**1992**	**1982–1992**	**1982–1992**
Intact	80	70	10	12.5
Transformed	65	65	0	0
Degraded	55	40	15	37
	1992	**2002**	**1992–2002**	**1992–2002**
Intact	70	76	-6	5
Transformed	65	60	5	7.6
Degraded	40	35	5	14
	2002	**2010**	**2002–2010**	**2002–2010**
Intact thicket	76	65	11	14
Transformed	60	50	10	16.6
Degraded	35	20	15	42.8

Discussion

The object-oriented classification produced a classification of subtropical thicket vegetation classes, which enabled the accurate estimation of the area covered by each of the subtropical thicket classes.

The temporal changes that occurred within the subtropical thicket classes are a prerequisite for AGC stock change assessments.[18,23] The temporal analysis of the stocks from 1972 to 2010 shows a drastically decreasing trend in the three subtropical thicket classes. Degraded and intact thicket had the highest AGC stock losses. There were high losses in AGC between 2002 and 2010 in all three thicket classes. The considerable thicket degradation processes taking place, particularly in the communal rangelands and villages, could explain the temporal losses in AGC stocks. Thicket degradation was noted to vary with land tenure systems.[13] Communal lands are characterised by overgrazing and severe soil erosion forms related to land abandonment.

Rainfall trends between 1972 and 2010, analysed by Nyamugama[35], reveal the oscillatory nature of precipitation in the study area. For instance, despite the fact that more precipitation was received in 2010 than in 2002, a considerable deterioration in thicket condition and reduction in AGC stocks was identified. This finding suggests that precipitation had little effect with regard to vegetation and AGC stock trends. By implication, anthropogenic factors could have played an overriding role in the deterioration of thicket and AGC stocks during this period.

The AGC stock estimates for 2010 are comparable with the estimates carried out by Mills et al.[36] in the GFRNR and in the Sundays River spekboom[37]. These results are also comparable with the findings of Palmer et al.[38] in Baviaanskloof Mega Reserve, South Africa, who observed that the intact subtropical thicket class had higher AGC stocks than the transformed and degraded thicket classes. By implication, thicket transformation and degradation trends as identified in the present study translate into enormous losses of AGC. The decline in the area under intact subtropical thicket has led to high losses of AGC stocks from 1972 to 2010. The implications of this decline for climate change are tremendous.

The main challenge encountered related to similarities in spectral signatures of subtropical thicket vegetation NDVI with some other vegetation surfaces. Therefore, extensive sampling was done in those areas in order to separate subtropical thicket vegetation classes on the 2010 SPOT image. The challenge was, however, greater with the Landsat MSS, TM and ETM images. Despite these challenges, the results obtained in this research are valid, as borne out by comparison with similar studies in other areas covered by subtropical thicket. A study by Copping et al.[39] revealed the challenges encountered in long-term change detection, as a result of constraints related to vegetation phenology and variations in interannual vegetation productivity. Although the role of vegetation phenology variation is known, the post-classification techniques employed in this study compensated for the inter-date phenological variations, as each classification was done independently.[32] Each classification was then used to characterise the land use/cover changes.[15,16]

Conclusions

The results of this study confirm that GIS and RS can be reliably used to model spatial and temporal changes in AGC stocks. It is possible to estimate, quantify and map AGC stocks over space and time, with changes in various subtropical thicket classes. A general decline in AGC stocks from 1972 to 2010 was identified in all the subtropical thicket classes. Degraded and transformed subtropical thicket classes had the highest losses, while the intact subtropical thicket class showed the lowest decline. The high losses of AGC stocks in subtropical thicket classes are attributed mainly to anthropogenic activities such as overgrazing, indiscriminate wood collection for fuel and other injudicious land-use practices. The results of this study can be used to retrospectively estimate AGC losses and thus carbon emissions in the past and to predict future scenarios. The methodology used in this study can be applied to estimate AGC stocks at national and regional levels, which can assist in carbon reporting for the country as per the Kyoto Protocol requirements. The average AGC stocks for each subtropical thicket class can therefore be applied to satellite-based subtropical vegetation maps to estimate regional subtropical thicket forest emissions.

Acknowledgements

We thank the National Research Foundation (South Africa) for funding this project and the South African National Space Agency for providing satellite imagery. We are also thankful to Mr N. Ndou, Mr B. Mazeka and Ms J. Smith for their assistance during fieldwork. We acknowledge the contributions of the anonymous reviewers, who added much value to the manuscript.

Authors' contributions

A.N. was the lead researcher, performed all the experiments and wrote the manuscript. V.K. gave technical input into the field research and the writing of the manuscript.

References

1. Food and Agriculture Organization (FAO). Global forest resources assessment. Report no. 163. Report published at Biennial Meeting of the FAO Committee on Forestry and World Forest. Rome: FAO; 2010. p. 22–30.

2. Gibbs HK, Brown SO, Niles J, Foley JA. Assessing and estimating tropical forest carbon stocks: Making REDD a reality. Environ Res Lett. 2007;2(4), Art.# 045023. http://dx.doi.org/10.1088/1748-9326/2/4/045023

3. Intergovernmental Panel on Climate Change. Expert meeting on detection and attribution related to anthropogenic climate change. Report no. 94. Cambridge: Cambridge University Press; 2007. p. 540–560.

4. Gibbs HK, Brown S. Geographical distribution of woody biomass carbon stocks in tropical Africa. Oak Ridge, TN: Carbon Dioxide Information Center, Oak Ridge National Laboratory; 2007.

5. Fearnside PM, Laurance WF. Tropical deforestation and greenhouse-gas emissions. Manaus: National Institute for Research in the Amazon (INPA); 2004. p. 129–150.

6. Gibbs HK, Brown S, Niles JO, Foley JA. Monitoring and estimating tropical forest carbon stocks: Making REDD a reality. Environ Res Lett. 2007;2(4), Art. #045023. http://dx.doi.org/10.1088/1748-9326/2/4/045023

7. Ramankutty N, Gibbs HK, Achard F, DeFries R, Foley JA, Houghton RA. Challenges to estimating carbon emissions from tropical deforestation. Glob Change Biol. 2007;13:51–66. http://dx.doi.org/10.1111/j.1365-2486.2006.01272.x

8. Houghton JT, Ding Y, Griggs DG, Noguer M, Van der Linden PJ, Dai X, et al., editors. Climate change: The scientific basis. Contribution of Working Group I to the third assessment report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2001. http://dx.doi.org/10.4337/9781781950715.00018

9. Lechemere-Oertel RG, Kerley GIH, Cowling R. Patterns and implications of transformation in semi-arid succulent thicket, South Africa. J Arid Environ. 2005;62:459–474. http://dx.doi.org/10.1016/j.jaridenv.2004.11.016

10. Skowno AL. Indirect biomass estimation in subtropical thicket vegetation in the Eastern Cape Province, South Africa. Report for the conservation farming project. Cape Town: National Botanical Institute; 2003. p. 25–30.

11. Volk JHJ, Euston-Brown DIW. The patterns within and ecological processes that sustain the subtropical thicket vegetation in the planning domain of the subtropical thicket ecosystem planning (STEP) project. Report no 40. Port Elizabeth: Terrestrial Ecology Research Unit, University of Port Elizabeth; 2002. p. 20–40.

12. Patenaude G, Milne R, Dawnson PT. Review synthesis of remote sensing approaches for forest carbon estimation reporting to the Kyoto Protocol. J Environ Sci Policy. 2005; 8:161–178. http://dx.doi.org/10.1016/j.envsci.2004.12.010

13. Zhou W, Huang G, Troy A, Cadenarzo L. Object-based land cover classification of shaded areas in high spatial resolution imagery of urban areas a comparison study. Remote Sens environ. 2009;8:1769–1777. http://dx.doi.org/10.1016/j.rse.2009.04.007

14. Singh A. Digital change detection techniques using remotely-sensed data. Int J Remote Sens. 1989;10:989–1003. http://dx.doi.org/10.1080/01431168908903939

15. Sanchez-Azofeifa GA, Castro-Esau WA, Kurz A, Joyce J. Monitoring carbon stocks in the tropics and the remote sensing operational limitations: From local to regional projects. Ecol Appl. 2009;19:480–494. http://dx.doi.org/10.1890/08-1149.1

16. Myeong S, Nowak DJ, Duggin MJ. A temporal analysis of urban forest carbon storage using remote sensing. Remote Sens Environ. 2006;101:277–282. http://dx.doi.org/10.1016/j.rse.2005.12.001

17. Alhern FJ, Cahoon D, French NH, Kasischke ES, Michael JL. Forest fire monitoring and mapping: A component of global observation of forest cover. Brussels: Joint Research Centre, European Commission; 2000. p. 170–174.

18. Canty MJ, Nielsen AA, Schmidt M. Automatic radiometric normalization of multitemporal satellite imagery. Remote Sens Environ. 2004;91:4411–4451. http://dx.doi.org/10.1016/j.rse.2003.10.024

19. Chave J, Condition R, Aguilar S, Hernandez A, Lao S, Perez R. Error propagation and scaling for tropical forest biomass estimates. Philos Trans Roy Soc Lond B. 2004;359:409–420. http://dx.doi.org/10.1098/rstb.2003.1425

20. Fuchs H, Magdon P, Kleinn C, Flessa H. Estimating aboveground carbon in a catchment of the Siberian forest tundra: Combining satellite imagery and field inventory. Remote Sens Environ. 2009;113:518 – 531. http://dx.doi.org/10.1016/j.rse.2008.07.017

21. Hirata Y. Estimation of stand attributes in *Cryptomeria japonica* and *Chamaecyparis obtusa* stands using QuickBird panchromatic data. J Forest Res Jpn. 2008;13(3):147–154. http://dx.doi.org/10.1007/s10310-008-0059-7

22. Birch NE. The vegetation potential of natural rangelands in the Mid fish river valley, Eastern Cape, South Africa: Towards a sustainable and acceptable management system [unpublished PhD thesis]. Grahamstown: Rhodes University; 2000.

23. Malhi Y, Grace J. Tropical forests and atmospheric carbon dioxide. Trends Ecol Evol. 2000;15:332–337. http://dx.doi.org/10.1016/S0169-5347(00)01906-6

24. Lu D. The potential and challenges of remote sensing based biomass estimation. Int J Remote Sens. 2006;12:2509–2525. http://dx.doi.org/10.1080/01431160500486732

25. Mills AJ, Cowling RM. Rate of carbon sequestration at two thickets restoration sites in the Eastern Cape South Africa. Austral Ecol. 2006;30:797–804. http://dx.doi.org/10.1111/j.1442-9993.2005.01523.x

26. Buyantuyev A, Wu J. Effects of thematic resolution on landscape pattern analysis. Landscape Ecol. 2007;22:7–13. http://dx.doi.org/10.1007/s10980-006-9010-5

27. Lu D, Mausel P, Brondízio E, Moran E. Assessment of atmospheric correction methods for Landsat TM data applicable to Amazon basin LBA research. Int J Remote Sens. 2002;23:2651–2671. http://dx.doi.org/10.1080/01431160110109642

28. Im J, Jensen JR, Tullis JA. Object-based change detection using correlation image analysis and image segmentation. Int J Remote Sens. 2008;29(2):399–423. http://dx.doi.org/10.1080/01431160601075582

29. Benz UC, Hofmann P, Willhauck G, Lingenfelder I, Heynen M. Multi-resolution, object oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS J Photogramm. 2004;58:239–258. http://dx.doi.org/10.1016/j.isprsjprs.2003.10.002

30. Muukkonen P, Heiskanen J. Estimating biomass for boreal forests using ASTER satellite data combined with standwise forest inventory data. Remote Sens Environ. 2005;99:434–447. http://dx.doi.org/10.1016/j.rse.2005.09.011

31. Chubey MS, Franklin SE, Wulder MA. Object-based analysis of IKONOS-2 imagery for extraction of forest inventory parameters. Photogramm Eng Rem Sens. 2006;2:383–394. http://dx.doi.org/10.14358/PERS.72.4.383

32. Llyod JW, Van den Berg EC, Palmer AR. Patterns of transformation and degradation in the thicket biome, South Africa. Port Elizabeth: Terrestrial Ecology Research Unit, Nelson Mandela Metropolitan University; 2002.

33. Kakembo V. Trends in vegetation degradation in relation to land tenure, rainfall and population changes in Peddie District, Eastern Cape, South Africa. Environ Manage. 2001;1:39–46. http://dx.doi.org/10.1007/s002672001

34. Kakembo V, Rowntree KM. The relationship between land use and soil erosion in the communal lands near Peddie Town, Eastern Cape, South Africa. Land Degrad Dev. 2003;14:39–49. http://dx.doi.org/10.1002/ldr.509

35. Nyamugama A. Monitoring carbon stocks in the sub-tropical thicket biome using remote sensing and GIS techniques: The case of the Great Fish River Nature Reserve and its environs, Eastern Cape Province, South Africa [unpublished PhD thesis]. Port Elizabeth: Nelson Mandela Metropolitan University; 2013.

36. Mills AJ, Cowling RM, Fey MV, Kerley GIH, Donaldson JS, Lechmere Oertel RG, et al. Effects of goat pastoralism on ecosystem carbon storage in semi-arid thicket, Eastern Cape, South Africa. Austral Ecol. 2005;30:797–804. http://dx.doi.org/10.1111/j.1442-9993.2005.01523.x

37. Powell M. Restoration of degraded subtropical thickets in the Baviaanskloof Megareserve, South Africa [unpublished MSc thesis]. Grahamstown: Rhodes University; 2009.

38. Palmer AR, Kakembo V, Lloyd JW, Ainslie A. Degradation patterns in the succulent thicket. Proceedings of the First Step Forum, Zuurberg. South Africa. Report no. 54. Port Elizabeth: Centre for African Conservation Ecology, Nelson Mandela Metropolitan University; 2006. p .7–17.

39. Copping P, Jonckheere I, Nackaerts K, Muys B. Digital change detection methods in ecosystem assessing: A review. Int J Remote Sens. 2004;25:1565–1596. http://dx.doi.org/10.1080/0143116031000101675

Food consumption changes in South Africa since 1994

AUTHORS:
Lisa-Claire Ronquest-Ross[1]
Nick Vink[2]
Gunnar O. Sigge[1]

AFFILIATIONS:
[1]Department of Food Science, Stellenbosch University, Stellenbosch, South Africa
[2]Department of Agricultural Economics, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Lisa-Claire Ronquest-Ross

EMAIL:
lisa.ronquest@effem.com

POSTAL ADDRESS:
Department of Food Science, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

DATES:
Received: 12 Oct. 2014
Revised: 11 Dec. 2014
Accepted: 12 Dec. 2014

KEYWORDS:
food consumption data; food consumption shifts; packaged food; beverages; public health

HOW TO CITE:
Ronquest-Ross L-C, Vink N, Sigge GO. Food consumption changes in South Africa since 1994. S Afr J Sci. 2015;111(9/10), Art. #2014-0354, 12 pages. http://dx.doi.org/10.17159/sajs.2015/20140354

Food consumption patterns in South Africa have changed dramatically over the past decades and likely will continue to change over the coming decades. Various food-related studies conducted over the last few decades indicate that food consumption shifts in South Africa have been towards a more Western-orientated diet, with nutritional consequences contributing to increased obesity and other non-communicable diseases. Several sources of data may be used to examine patterns in food consumption over time. Each of these methods has its own merits depending on the desired outcome, but are difficult to compare as each measures different levels of dietary information. As a result of the lack of regular national or comparable food consumption data in South Africa, the objective of this study was to establish, through the use of databases (FAOSTAT food balance sheets and Euromonitor International© Passport), the broad food and beverage consumption shifts in South Africa since 1994. Our findings indicate that food consumption shifts have been towards an overall increase in daily kilojoules consumed, a diet of sugar-sweetened beverages, an increase in the proportion of processed and packaged food including edible vegetable oils, increased intake of animal source foods, and added caloric sweeteners, and a shift away from vegetables. The largest shifts in food consumption were observed for soft drinks, sauces, dressings and condiments, sweet and savoury snacks, meat, and fats and oils. Convenience, health and nutrition, and indulgence were the main drivers of the increase in consumption of packaged foods and beverages. These shifts in food consumption are concerning as relates to their fat, sugar and salt composition and potential effect on public health.

Introduction

Background to food consumption in South Africa

Food consumption is affected by food availability, accessibility and choice.[1] Food intake choices are influenced by factors such as geography, season, education, demography, disposable income, government and other support services, urbanisation, globalisation, marketing, religion, culture, ethnicity, social networks, time and the consumer.[1-3] In 1994 certain changes occurred in South Africa which dramatically affected food consumption patterns and will continue to do so, as a result of shifts in food availability, accessibility and choices.[4-6] Amongst others, there has been significant growth of supermarkets, which accounted for about 50–60% of retail sales, while rising urbanisation and growing per capita incomes are expected to double the demand for high-value foods such as dairy, meat, fresh fruits, vegetables, and processed, packaged and prepared foods.[5-7] Total food expenditure has increased for fruit and vegetables and processed foods such as spaghetti and oven-ready meals, while expenditure on maize and wheat flour has declined.[5,6,8] For South African women, who do most of the household grocery shopping, the most important consideration when choosing a food item is price.[9,10] Taste, health, nutrient content, safety and hygiene of the food item, and ease of preparation (in descending order) are considered after price. Furthermore, Temple et al.[11] showed that a healthier diet can cost as much as 69% more than a typical South African diet and concluded that a healthy diet is largely unaffordable for most South Africans.

Comparative food consumption information is useful in identifying trends in foods and eating patterns nationally or among various sub-populations of interest.[1] Knowledge of food consumption patterns is required for nutrient profiling purposes, costing and compiling of typical food baskets, for innovation and trend analysis relevant for the food industry, pricing policies to promote healthier food purchases, addressing food insecurity, and developing and monitoring progress against food-based dietary guidelines.[12,13]

In the largest and only national food consumption survey conducted in South Africa (in 1999), the most commonly consumed food items of children aged 1–9 years old were maize, sugar, tea, whole milk and brown bread.[14] Children's intakes of energy, calcium, iron, zinc, selenium, vitamins A, D, C and E, riboflavin, niacin, vitamin B6 and folic acid were below two-thirds of the Recommended Dietary Allowances.[14] Steyn et al.[15] concluded that South Africans, on average, primarily consume maize, wheat, vegetables, milk, potatoes and sugar. In a secondary study of dietary surveys by Steyn et al.[16], it was determined that the most commonly consumed food items by the South African adult population were maize, sugar, tea, bread (brown and white), non-dairy creamer, brick margarine, chicken meat, full-cream milk and green leaves. In the 1–5 and 6–9-year-old groups, maize porridge, sugar, tea, full-cream milk and white bread were eaten most frequently.[16] In 2009, a cross-sectional 24-hour food recall survey revealed that the most commonly consumed food groups for South Africans aged 16 and older, were cereals/roots, meat/fish, dairy and vegetables other than vitamin A-rich vegetables while eggs, legumes, and vitamin A-rich fruit and vegetables were least consumed.[17] The most recent South African National Health and Nutrition Examination Survey (SANHANES-1) study indicated a national dietary diversity score of 4.2, which is very close to the cut-off level of 4.0 for dietary adequacy.[9]

Data sources for assessing food consumption trends

Several data sources can be used to examine food consumption patterns to determine changes over time.[1,12] Such data are derived from FAOSTAT food balance sheets (FAOSTAT FBS), household budget surveys or individual dietary surveys.[1,12] Each of these methods has its own merits depending on the desired outcome,[1] but are difficult to compare as each measures different levels of dietary information.[12]

Food balance sheets

A food balance sheet provides a comprehensive picture of a country's food supply during a specified period.[18,19] For each food item, the food balance sheet indicates the availability for human consumption, which corresponds to the sources of supply and their utilisation.[18,19] When describing consumption of foods per capita of a population, FAOSTAT FBS do not represent the actual amount of food consumed and will invariably result in an overestimation in food consumption compared with individual dietary surveys.[12,15]

Household budget surveys

Household budget survey data are regularly collected in developed countries by the Organisation for Economic Co-operation and Development, Eurostat or national administrations for the population.[12] South Africa collects these data for South African households as part of the Income and Expenditure Surveys that Statistics SA coordinate.[20] Challenges with household budget surveys are that data collected on food quantities purchased are not necessarily consumed and are expressed as food categories, e.g. seed oils, and not as single food items, e.g. olive oil.[12]

Individual dietary surveys

In the developed world, many countries conduct nationwide monitoring surveys, which provide valuable data to understand long-term changes in food and nutrient intake.[1] These data can be supplemented by smaller surveys, in single locations on smaller numbers of individuals, which often do not use the same methodology.[1,12] Many countries, particularly in the developing world, do not have the resources to mount individual level nutrition surveys as these are both prohibitively expensive and labour intensive.[1] In fact, only one large-scale study has ever been conducted in South Africa – the National Food Consumption Survey conducted in 1999 for children aged 1–9 years.[14,21]

Euromonitor International© Passport

Euromonitor International© delivers market research solutions via four different channels: databases, reports, books and consulting (Hartfall R 2013, written communication, February 6). Euromonitor International© Passport packaged food data are a consensus of opinions based on data gathered from trade sources, national statistics and secondary sources. These sources could include trade press, trade association data, company published reports and store checks in a sample of various stores with the bulk collated from trade interviews with market opinion leaders such as manufacturers, retailers, distributors, packaging converters and ingredients players (Hartfall R 2013, written communication, February 6). It covers all retail channels, both formal as well as informal, independent retailers including kiosks and street stalls (Hartfall R 2013, written communication, February 6). Euromonitor International© Passport measures, amongst other items, packaged food and beverages, fresh food and ingredients used in packaged food and beverages.[22] Euromonitor International© Passport per capita food consumption data are based on total retail value or volume size in a given year divided by the population in that same year.[22]

The consequence of food consumption changes in South Africa is a transition towards a more Western-orientated diet as has been reported in various food- and nutrition-related studies conducted at a community or provincial level over the past few decades.[9,23-25] In the past 50 years, data have shown that among urban black individuals, fat intake has increased from 16.4% to 26.2% of total energy (a relative increase of 59.7%), while carbohydrate intake has decreased from 69.3% to 61.7% of total energy (a relative decrease of 10.9%).[23] Sugar intake as a percentage of total energy intake was 5.9% in rural areas and 12.3% in urban areas for adolescents and adults (aged 10 years and over).[26] For urban areas, this value is above the World Health Organization's (WHO) recommendation for the prevention of dental caries.[26] South Africans consume salt at levels of 8.1 g/day – nearly double the WHO recommendation of 4–6 g/day.[27] South Africans have also shifted to eating on the move. The SANHANES-1 study revealed that almost half (48%) of South Africans reported that they had eaten out before with 28.3% of South Africans eating out weekly.[9]

Owing to the lack of regular national food consumption surveys or comparable food consumption survey data, the objective of this study was to establish, through the use of two databases (FAOSTAT FBS and Euromonitor Passport), the broad food and beverage consumption shifts amongst South Africans since 1994.

Materials and methods

Euromonitor International© Passport databases and FAOSTAT FBS were selected for this study as they both offer accessible, comprehensive and comparable national data on food items consumed. They are both regularly tracked (annually) and the available data have comparable units of measure, i.e. kilograms per capita per annum. FAOSTAT FBS data were extracted from the FAOSTAT website for South Africa.[19] Euromonitor International© Packaged Food and Beverage Consumption (Euromonitor PFBC) was extracted from the Euromonitor International© Passport for the South African geography.[22] Both sets of exported data were converted to per capita consumption figures as this takes into account increases in population growth over time. Per capita intake is a crude estimate of consumption as it is the total amount consumed divided by the total population and does not take into account wastage, losses in storage, urban/rural distribution differences or distribution within households.[13,15] Therefore, it is comparable for daily intake values for commonly consumed foods, but may differ substantially for foods consumed by a smaller number of respondents.[13,15] Intervals of 5 years were compared, from 1994 to 2009 for FAOSTAT FBS data and from 1999 to 2012 for Euromonitor PFBC data, with specific time overlaps in 1999, 2004 and 2009. Euromonitor International© Passport data for South Africa have been recorded only from 1998 and therefore the data overlap starts only in 1999. The food items were grouped into eight food categories: (1) meat, eggs and fish; (2) cereals; (3) vegetables, fruits and nuts; (4) dairy; (5) fats; (6) sugar and stimulants (coffee and tea); (7) packaged foods; and (8) soft drinks.

Data from FAOSTAT FBS and Euromonitor PFBC were compared according to food item per 5-year interval. Each of the eight food category data sets was then compared and analysed to determine the consumption trend of that food category over the specified period. This comparative analysis was also performed to assess the integrity of the data when both FAOSTAT FBS and Euromonitor PFBC data were available. Furthermore, the Euromonitor PFBC data were used to establish consumption trends of packaged food and beverages, as FAOSTAT FBS does not measure packaged foods and beverages specifically.

Results and discussion

Meat, eggs and fish

Meat has always been an important part of the human diet and remains central to most meals in developed countries.[1] The South African Food Based Dietary Guidelines recommend that either chicken, fish, meat, milk or eggs should be eaten daily.[28] Consistent with studies indicating that meat consumption has increased in South Africa,[2,6] FAOSTAT FBS data (Table 1) indicate that South Africans consumed 18 kg of meat per annum more in 2009 than they did in 1994. This increase is mainly as a result of significant increases of 109% and 119% from 2009 to 1994 in consumption of poultry meat and pig meat respectively, with beef, mutton and goat meat consumption remaining relatively stable. However, the Euromonitor PFBC data indicate a slight decrease in meat consumption from 1999 to 2012 (Table 1). There was a 4–17 kg difference in absolute meat consumption between the FAOSTAT FBS and Euromonitor PFBC data with the FAOSTAT FBS indicating higher consumption (Table 1). The difference is because the Euromonitor PFBC meat data do not include meat consumption of processed meats such as sausages, canned meat or meat cuts with added ingredients or sauces.[29] Therefore, the FAOSTAT FBS are more representative of meat consumption in South Africa. The total data set suggests that value-added processed meat consumption has increased significantly since 1994, as the Euromonitor PFBC data set, which does not include processed meat, showed a decline of 4.5% since 1999 and the FAOSTAT FBS data set, which does include it, showed increases in consumption of processed meat of 45.8% (Table 1).

Egg consumption, consistent across both the FAOSTAT FBS and Euromonitor PFBC data, has increased by 55.8% and 24.1% from 1994 to 2009 and 1999 to 2012, respectively (Table 1). This finding is in line with marked increases in worldwide egg consumption, especially in developing countries.[1]

Fish is an important source of good quality protein and is low in fat, except for oily fish which is a good source of long-chain polyunsaturated fatty acids.[1] Fish consumption increased by more than 26% in both the FAOSTAT FBS and Euromonitor PFBC data sets (Table 1).

Cereals

According to the FAOSTAT FBS data, the largest contributors to total food and cereal consumption for South Africans in 2012 were maize (104 kg.capita/year) and wheat (60.9 kg.capita/year) (Table 2). This result is consistent with those of several food consumption surveys conducted in South Africa.[14,16] A slight decrease of 4.6% was seen for maize consumption from 1994 to 2009 (Table 2), which is expected to continue into 2020 as household incomes continue to increase and consumers move towards higher-value food items.[2,5] South African wheat consumption, ranging from 59.8 to 60.9 kg.capita/year from 1994 to 2009 (Table 2), is fairly in line with the world consumption rate of 66 kg.capita/year.[2] Euromonitor PFBC unfortunately does not measure staple foods such as maize and wheat (Johnson S 2013, written communication, July 10). Euromonitor PFBC records wheat consumption only when wheat is consumed as an ingredient in processed foods (such as baked goods like brown bread) and is therefore significantly lower than that recorded by FAOSTAT FBS, but shows that wheat consumption has increased (Table 2). Only rice and oats showed significant increases in consumption of 48% and 83.3%, respectively, since 1994 (Table 2). Rice consumption data indicated in the Euromonitor PFBC is for packaged rice of all varieties (long grain, basmati etc.) as well as various packaging formats, e.g. ready-to-eat.

Table 1: Comparison of total meat, eggs and fish consumption in South Africa assessed by FAOSTAT food balance sheets (FBS) and Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	FAOSTAT FBS (kg.capita/year)					Euromonitor PFBC (kg.capita/year)				
	1994	1999	2004	2009	% Change (1994–2009)	1999	2004	2009	2012	% Change (1999–2012)
Total meat	40.3	37.2	43.8	58.7	45.7	33.1	31.0	31.4	31.6	-4.5
Bovine meat	16.6	11.5	13.8	15.4	-7.0					
Mutton and goat meat	4.0	4.2	3.7	3.8	-5.0					
Pig meat	3.1	3.0	3.6	6.8	119.0					
Poultry meat	15.3	18.2	22.4	32	109.0					
Meat, other	1.3	0.3	0.3	0.7	-46.0					
Total offal	3.6	3.4	4.2	4.7	30.6					
Total eggs	4.3	5.6	5.7	6.7	55.8	5.8	6.3	6.9	7.2	24.1
Total fish and seafood	5.9	6.5	9.9	7.5	27.0	4.5	5.5	5.5	5.7	26.7

Table 2: Comparison of total cereals (excluding beer) consumption in South Africa assessed by FAOSTAT food balance sheets (FBS) and Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	FAOSTAT FBS (kg.capita/year)					Euromonitor PFBC (kg.capita/year)				
	1994	1999	2004	2009	% Change (1994–2009)	1999	2004	2009	2012	% Change (1999–2012)
Total cereals, excluding beer	182	182	187	182	0.4					
Wheat	59.8	60.3	56.9	60.9	1.8	23.5	21.2	23.3	25.6	8.9
Rice (milled equivalent)	10	13.3	15.7	14.8	48.0	12.3	16.7	13.3	12.6	2.4
Barley	0.1	0.1	0.1	0.1	0.0					
Maize	109	106	112	104	-4.6					
Rye	0.1	0.1	0.1	0.1	0.0					
Oats	0.6	0.6	0.8	1.1	83.3					
Millet	0.3	0.2	0.1	0.1	-66.7					
Sorghum	1.9	1.5	1.6	1.5	-21.1					
Cereals, other	0.1	0.1	0.1	0.1						

Millet and sorghum showed declines in consumption of 66.7% and 21.1%, respectively, but are not significant contributors to total cereal consumption for South Africans (Table 2).

Vegetables, fruits and nuts

Fruit and vegetable consumption plays a vital role in providing a micronutrient dense diet[30] and South Africa's Food Based Dietary Guidelines recommend eating plenty of fruits and vegetables every day and dry beans, peas, lentils and soya regularly.[28] However, the recent SANHANES-1 study revealed a low intake of fruits and vegetables (two or fewer portions per day) for 25.6% of South Africans and that people in formal urban areas appeared to consume the most fruit and vegetables.[9] This result may be linked to cost and availability.[9] Inadequate fruit and vegetable consumption is a problem worldwide[1] as well as in South Africa[30]. South Africans are eating slightly more fruit than in 1994 with both FAOSTAT FBS and Euromonitor PFBC data showing about a 6% increase (Table 3) led mainly by increases in banana, apple and grapefruit consumption. In research conducted by Steyn et al.[31] on street food consumption in South Africa, fruit was the most commonly purchased item. Unfortunately, a slight decrease in vegetable consumption can be seen in both sets of data (Table 3). However, tomato and onion consumption increased moderately. Consumption of starchy roots (mainly potatoes) increased according to both sets of data with a corresponding slight decrease in sweet potato consumption (according to FAOSTAT FBS data) (Table 3). The consumption of pulses increased as a result of a moderate increase in bean consumption of 16% with peas and other pulses consumption declining (FAOSTAT FBS data) (Table 3).

Dairy

Dairy provides at least 10 essential nutrients including protein, carbohydrates, vitamins (A, B12 and riboflavin) and minerals (calcium, phosphorus, magnesium, potassium and zinc).[32] According to the FAOSTAT FBS and Euromonitor PFBC data, dairy consumption increased by 8.4% from 1994 to 2009 and by 14.7% from 1999 to 2012 (Table 4). Some significant shifts, based on the Euromonitor PFBC data, are increases of 18.5% and 6.8% in cheese and drinking milk consumption, respectively, and a more significant increase of 73.7% in yoghurt and sour milk products consumption since 1999 (Table 4).

Cheddar is by far the most popular type of unprocessed cheese in South Africa (31%) followed by Gouda (20%) and mozzarella.[33,34] In terms of consumption of drinking milk products, cow's milk consumption increased by 7.3% with a larger increase of 16.7% in value-added flavoured milk products over the same period (Table 4). Soy beverage consumption only became evident in 2009 at 100 mL.capita/year (Table 4) as soy is seen as a healthier alternative to dairy for those individuals who are lactose intolerant or have high cholesterol.[35] A significant reduction of 42.9% in powdered milk consumption occurred over the same period (Table 4);

Table 3: Comparison of total vegetable, fruit, starchy root, pulses and tree nut consumption in South Africa assessed by FAOSTAT food balance sheets (FBS) and Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	FAOSTAT FBS (kg.capita/year)					Euromonitor PFBC (kg.capita/year)				
	1994	1999	2004	2009	% Change (1994–2009)	1999	2004	2009	2012	% Change (1999–2012)
Total vegetables	43.1	44.4	42.4	42.9	-0.5	42.0	40.4	38.5	38.7	-7.9
Tomatoes	8.7	9.0	9.3	10.4	19.5					
Onions	5.4	7.5	7.0	7.9	46.3					
Vegetables, other	29.1	27.9	26.2	24.6	-15.5					
Total fruits, excluding wine	32.8	41.3	38.4	34.8	6.1	28.1	27.0	29.0	29.9	6.4
Oranges, mandarins	10.4	12.4	8.3	6.9	-33.7					
Lemons, limes	0.9	0.8	1.7	1.2	33.3					
Grapefruit	0.8	1.4	2.1	3.7	362.5					
Citrus, other	0.1	0.1	0.1	0.2	100.0					
Bananas	2.9	7.0	5.4	7.1	144.8					
Apples	4.8	3.6	5.7	6.8	41.7					
Pineapples	1.8	2.2	2.0	1.9	5.6					
Grapes	4.4	5.2	3.3	1.3	-70.5					
Fruits, other	6.7	8.6	9.8	5.8	-13.4					
Total starchy roots	26.1	29.9	30.8	30.6	17.2	26.0	26.5	27.1	27.8	6.9
Potatoes	24.8	28.9	29.8	29.5	19.0					
Sweet potatoes	1.4	1.0	1.0	1.1	-21.4					
Total pulses	3.3	3.0	3.2	3.6	9.1	2.1	1.8	1.9	2.0	-4.8
Beans	2.5	2.5	2.5	2.9	16.0					
Peas	0.5	0.3	0.4	0.4	-20.0					
Pulses, other	0.3	0.3	0.3	0.2	-33.3					
Total tree nuts	0.1	0.3	0.4	0.3	200.0	0.5	0.6	0.7	0.7	40.0

consumers are moving away from powdered milk and towards ultrahigh temperature (or UHT) milk because it is more affordable.[35]

Yoghurt and sour milk consumption increased by a dramatic 73.7% to 6.6 kg.capita/year from 1999 to 2012 (Table 4). Danone Southern Africa led the category in 2012 with a value share of 44% through effective marketing of their leading yoghurt and sour milk brands over the last decade.[36-38]

Fats

South Africa's Food Based Dietary Guidelines recommend eating fats sparingly.[28] However, as with studies indicating increases in fats and oils consumption in South African and other developing countries,[1,23,39] both FAOSTAT FBS and Euromonitor PFBC data indicated increasing trends (Table 5). Increases were greater than 28.5% according to both FAOSTAT FBS and Euromonitor PFBC data (Table 5). These increases are significant considering the high energy that fats and oils contribute to the diet and are mainly related to increases in vegetable oils (>29.6% for both FAOSTAT FBS and Euromonitor PFBC) and oil crops consumption (108.3%, FAOSTAT FBS data), coupled with a decrease in animal fat consumption of 53.8% (FAOSTAT FBS data specifically) (Table 5). The Euromonitor PFBC data refer to butter consumption only as it relates to animal fats. FAOSTAT FBS showed a marked decline in animal fats consumption in South Africa (Table 5), which has also been observed worldwide.[1] However, butter consumption specifically has increased from 0.3 to 0.4 kg.capita/year (33.3%) from 1999 to 2012 with margarine consumption increasing by 13.6% to 2.5 kg.capita/year from 1999 to 2012 (Table 6). Olive oil consumption only became measureable in 2012 with 0.1 kg.capita/year (Table 6).

Sugar and stimulants

A positive trend in food consumption shifts since 1994 is that consumption of sugar and sweeteners as raw sugar or natural sweeteners (e.g. honey) declined by at least 7.5% according to both data sets (Table 7). However, the Euromonitor PFBC data indicated an overall increase in consumption of 7.1% as this data set includes intake of sugar and sweeteners when used as ingredients in processed foods (Table 7). There has been a large increase of 33.1% in consumption of sugar and sweeteners through consumption of processed foods which use sugar and sweeteners as ingredients e.g. confectionery and soft drinks (Table 7).

Apart from water, tea is the most consumed beverage in the world.[40] Consumption of stimulants (coffee and tea) in South Africa has increased by 54.5% and 44.4%, according to the FAOSTAT FBS and Euromonitor PFBC data, respectively (Table 7). The total South African tea market is divided into black tea (80% of the market) with, rooibos and speciality tea making up the remainder.[41] For centuries, herbs and spices have played an important role in our nutrition for both their culinary and healing value.[42,43] Consumption of spices has doubled since 1994 (Table 7).

Table 4: Comparison of total dairy consumption in South Africa assessed by FAOSTAT food balance sheets (FBS) and Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	FAOSTAT FBS (kg.capita/year)					Euromonitor PFBC (kg.capita/year)				
	1994	1999	2004	2009	% Change (1994–2009)	1999	2004	2009	2012	% Change (1999–2012)
Total dairy	51.0	47.6	52.9	55.3	8.4	37.4	37.5	38.3	42.9	14.7
Baby milk formula						0.2	0.2	0.3	0.3	50.0
Total cheese						2.7	2.9	3.0	3.2	18.5
Drinking milk products						30.7	30.0	29.5	32.8	6.8
Flavoured milk drinks						2.4	2.3	2.3	2.8	16.7
Flavoured powdered milk						0.1	0.1	0.1	0.1	0.0
Cow's milk						27.4	27.0	26.5	29.4	7.3
Powdered milk						0.7	0.6	0.5	0.4	-42.9
Soy beverages						0.0	0.0	0.1	0.1	
Yoghurt and sour milk						3.8	4.4	5.5	6.6	73.7

Table 5: Comparison of total fats and oils consumption in South Africa assessed by FAOSTAT food balance sheets (FBS) and Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	FAOSTAT FBS (kg.capita/year)					Euromonitor PFBC (kg.capita/year)				
	1994	1999	2004	2009	% Change (1994–2009)	1999	2004	2009	2012	% Change (1999–2012)
Total fats and oils	13.3	14.5	16.6	17.1	28.6	7.2	7.8	8.4	9.6	33.3
Total oil crops	1.2	2.0	1.9	2.5	108.3					
Total vegetable oils	10.8	12	13.9	14	29.6	4.2	4.6	5.1	5.9	40.5
Total animal fats	1.3	0.5	0.8	0.6	-53.8	0.3	0.3	0.3	0.4	33.3

Table 6: Consumption of packaged fats and oils in South Africa according to Euromonitor Packaged Food and Beverage Consumption (PFBC)

	Euromonitor PFBC (kg.capita/year)				
	1999	2004	2009	2012	% Change (1999–2012)
Total fats and oils	7.2	7.8	8.4	9.6	33.3
Butter	0.3	0.3	0.3	0.4	33.3
Margarine	2.2	2.4	2.4	2.5	13.6
Olive oil	0.0	0.0	0.0	0.1	
Spreadable oils and fats	0.5	0.5	0.6	0.7	40.0
Vegetable and seed oil	4.2	4.6	5.1	5.9	40.5

Packaged foods

Simple, convenient food solutions is a major global food trend and one that packaged foods address in many ways.[2] In recent years, there has been an increase in sales of almost all categories of packaged foods and beverages in South Africa,[29,44] resulting in a vibrant packaged food and beverage sector. According to the Euromonitor International© 2012 report on the South African Packaged Food sector, this sector has grown by 57% from ZAR91 billion to ZAR143 billion and by 15% in volume from 4.515 k tons to 5.202 k tons from 2007 to 2012.[29] In fact, from 2000 to 2009, the food and beverage manufacturing sector outgrew the manufacturing sector at 32.6% compared to 8.7%.[45]

The largest category in packaged food in terms of per capita consumption is the bakery sector, which contributed 43 kg.capita/year in 2012 (Table 8). Total consumption of items in this category increased slightly by 6.4% (Table 8). With respect to 'baked goods', of which bread consumption is the largest contributor, a slight increase of 4.7% was observed, mainly as a result of an increase of 27.9% in 'artisanal or unpackaged bread' consumption (Table 8). At the same time, conventional/industrial bread consumption decreased by 9.3% (Table 8). Innovation in value-added breakfast cereals and lower income consumers choosing maize as a more affordable carbohydrate are some of the reasons for conventional/industrial bread decline.[46] White bread remained the most popular bread type in 2012 at 49% value share because of its appeal with lower income consumers.[46] However, brown bread increased in popularity with its share reaching 42% in 2012.[46] A number of new speciality breads have been launched in South Africa.[46,47] Consumption of 'biscuits', both 'savoury and cracker type biscuits' as well as 'sweet biscuits' increased by more than 50% (Table 8). 'Savoury and cracker type biscuits' appeal to higher income groups who entertain at home.[48] Plain 'sweet biscuits' accounted for nearly 50% of overall value sales in 2012 as a result of lower income consumers' attraction to strong brands with lower price point offerings.[48,49] Consumption of 'breakfast cereals', both 'hot' and 'ready-to-eat', has increased by more than 42.9% since 1999 (Table 8). 'Breakfast cereal' consumption is fairly low in South Africa and price-sensitive consumers either switch out of cereals to cheaper carbohydrates such as bread or maize or between brands based on promotional prices, different pack sizes or cheaper variants.[50,51]

The next largest category in packaged food in terms of consumption was 'canned/preserved food' at 3.9 kg.capita/year in 2012 (Table 8). A decline of 13.3% in consumption from 1999 to 2012 was observed in this category (Table 8), in line with global declines in consumption of canned foods.[52] Even though canned goods are considered household essentials as they are perceived to offer value for money, consumer perception is that they do not meet requirements for fresh, quality ingredients, convenience and eco-friendliness.[52,53] Main category declines in consumption are seen in 'canned/preserved fish/seafood' (-44.4%), 'canned/preserved fruit' (-20%) and 'canned/preserved vegetables' (-12.5%) (Table 8). Consumers appear to be attracted to 'frozen processed vegetables' rather than 'canned/preserved vegetables' because of perceived freshness and lower costs.[53,54] The only two categories in canned/preserved foods that grew in consumption (by 50%) were 'canned/preserved beans' and 'canned/preserved ready meals' such as spaghetti in sauce, meat in sauce and soup (Table 8).

The third largest category at 3 kg.capita/year in 2012, increasing from 1.9 kg.capita/year in 1999, was 'sauces, dressings and condiments' (Table 8). This category experienced a significant growth of 57.9% with 'table sauces' driving most of the consumption increases at 73.3% growth from 1999 to 2012 (Table 8). Consumption of 'tomato sauce/ketchup' and 'salad dressings' has more than doubled since 1999 (Table 8) and consumption of 'mayonnaise' grew by 50%. A new category – 'spicy chilli/pepper sauces' – emerged around 2004 (Table 8). The trend towards home dining to save money as well as the 'braaing' or barbequing culture in South Africa are some of the reasons for increased consumption.[55]

The fourth largest category contributing to per capita consumption of packaged foods, at 2.8 kg.capita/year, was 'frozen processed foods' (Table 8). The 'frozen processed foods' category grew by 21.7% from 1999 to 2012, with consumption in some sub-categories, such as 'frozen ready meals', 'frozen pizza', 'frozen processed potatoes' (e.g. frozen potato chips) and 'frozen processed poultry', even doubling

Table 7: Consumption of sugar and stimulants in South Africa assessed by FAOSTAT food balance sheets (FBS) and Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	FAOSTAT FBS (kg.capita/year)					Euromonitor PFBC (kg.capita/year)				
	1994	1999	2004	2009	% Change (1994–2009)	1999	2004	2009	2012	% Change (1999–2012)
Total sugar and sweeteners	33.9	32.6	31.0	31.0	-8.6	35.4	36.1	37.1	37.9	7.1
Packaged products†						22.7	21.6	21.1	21.0	-7.5
In processed foods‡						12.7	14.5	16.0	16.9	33.1
Total stimulants	1.1	1.0	1.2	1.7	54.5	0.9	1.0	1.2	1.3	44.4
Total spices	0.2	0.3	0.5	0.4	100.0					

†Sugar and sweeteners sold as packaged products, e.g. white sugar or honey.

‡Sugar and sweeteners used as ingredients in processed foods such as confectionery.

Table 8: Consumption of packaged foods in South Africa according to Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	Euromonitor PFBC (kg.capita/year)				
	1999	2004	2009	2012	% Change (1999–2012)
Bakery	40.4	36.0	39.2	43.0	6.4
Baked goods	38.6	34.1	36.9	40.4	4.7
Packaged/industrial bread	22.6	17.9	18.2	20.5	-9.3
Unpackaged/artisanal bread	14.0	14.4	16.7	17.9	27.9
Total biscuits	0.7	0.7	0.9	1.1	57.1
Sweet biscuits	0.5	0.5	0.7	0.8	60.0
Savoury biscuits and crackers	0.2	0.2	0.3	0.3	50.0
Total breakfast cereals	1.1	1.2	1.4	1.5	36.4
Ready-to-eat cereals	0.7	0.7	0.9	1.0	42.9
Hot cereals	0.4	0.4	0.5	0.6	50.0
Canned/preserved food	4.5	3.8	3.7	3.9	-13.3
Canned/preserved fish/seafood	1.8	1.3	1.0	1.0	-44.4
Canned/preserved beans	0.6	0.6	0.8	0.9	50.0
Canned/preserved vegetables	0.8	0.7	0.7	0.7	-12.5
Canned/preserved fruit	0.5	0.4	0.4	0.4	-20.0
Canned/preserved ready meals	0.2	0.2	0.3	0.3	50.0
Canned/preserved tomatoes	0.3	0.3	0.3	0.3	0.0
Canned/preserved meat and meat products	0.2	0.2	0.2	0.2	0.0
Sauces, dressings and condiments	1.9	2.0	2.3	3.0	57.9
Table sauces	1.5	1.7	1.9	2.6	73.3
Ketchup	0.5	0.5	0.6	1.0	100.0
Mayonnaise	0.4	0.5	0.5	0.6	50.0
Salad dressings	0.1	0.1	0.2	0.3	200.0
Spicy chilli/pepper sauces	0.0	0.1	0.1	0.1	
Other table sauces	0.4	0.4	0.4	0.4	
Frozen processed food	2.3	2.3	2.6	2.8	21.7
Frozen processed vegetables	0.5	0.5	0.6	0.7	40.0
Frozen bakery	0.5	0.5	0.6	0.5	0.0
Frozen ready meals	0.2	0.2	0.3	0.4	100.0
Frozen desserts	0.2	0.2	0.2	0.2	0.0
Frozen pizza	0.1	0.1	0.2	0.2	100.0
Frozen processed fish/seafood	0.2	0.2	0.2	0.2	0.0
Frozen processed potatoes	0.1	0.2	0.2	0.2	100.0
Frozen processed poultry	0.1	0.1	0.1	0.2	100.0
Frozen processed red meat	0.1	0.1	0.1	0.1	0.0
Other frozen processed food	0.1	0.1	0.2	0.2	100.0

Continues on next page

Continues from previous page

Food item	Euromonitor PFBC (kg.capita/year)				
	1999	2004	2009	2012	% Change (1999–2012)
Confectionery	2.3	2.3	2.6	2.6	13.0
Chocolate confectionery	0.8	0.7	0.9	0.9	12.5
Countlines	0.5	0.4	0.4	0.4	-20.0
Tablets	0.2	0.3	0.4	0.4	100.0
Gum	0.3	0.3	0.3	0.3	0.0
Sugar confectionery	1.2	1.2	1.4	1.4	16.7
Boiled sweets	0.3	0.3	0.4	0.4	33.3
Liquorice	0.1	0.1	0.0	0.0	-100.0
Lollipops	0.1	0.1	0.1	0.1	0.0
Pastilles, gums, jellies and chews	0.3	0.3	0.4	0.4	33.3
Toffees, caramels and nougat	0.3	0.2	0.3	0.3	0.0
Chilled processed food	2.1	2.0	2.1	2.3	9.5
Chilled fish/seafood	0.4	0.4	0.3	0.3	-25.0
Chilled pizza	0.0	0.1	0.1	0.1	
Chilled processed meat	1.5	1.4	1.4	1.5	0.0
Chilled ready meals	0.2	0.2	0.2	0.3	50.0
Chilled/fresh pasta	0.0	0.0	0.1	0.1	
Sweet and savoury snacks	1.5	1.6	2.0	2.3	53.3
Chips/crisps	0.9	1.0	1.3	1.5	66.7
Extruded snacks	0.3	0.3	0.4	0.4	33.3
Fruit snacks	0.1	0.1	0.1	0.1	0.0
Nuts	0.1	0.1	0.1	0.1	0.0
Tortilla/corn chips	0.1	0.1	0.1	0.1	0.0
Dried processed food	1.6	1.3	1.6	2.0	25.0
Dehydrated soup	0.2	0.1	0.2	0.2	0.0
Dessert mixes	0.1	0.1	0.1	0.1	0.0
Dried pasta	0.7	0.7	0.8	1.0	42.9
Dried ready meals	0.5	0.3	0.4	0.5	0.0
Instant noodles	0.1	0.1	0.1	0.2	100.0
Ice cream	1.5	1.4	1.6	1.5	0.0
Spreads	0.4	0.4	0.4	0.5	25.0
Soup	0.2	0.2	0.2	0.2	0.0

(Table 8). Convenience is the driver of the increase[56] as the number of consumers with freezers is increasing.[54]

Total 'confectionery' consumption has grown by 13% to 2.6 kg.capita/year since 1999 (Table 8). This growth is as a result of increasing consumption of 'chocolate confectionery' (12.5%), mainly from 'tablets' (Table 8). Chocolate 'tablets' or 'slabs' are the most popular chocolate confectionery type in South Africa and hold over 50% market share.[57] Plain milk chocolate is still preferred over dark chocolate, but the ratio is changing because of the health benefits associated with dark chocolate.[58,59] Consumption increases (16.7%) in 'sugar confectionery' are mainly as a result of 'boiled sweets', 'pastilles, gums, jellies and chews' (Table 8).

'Chilled processed food' consumption grew with a modest 9.5% to 2.3 kg.capita/year when compared to growth in consumption of other packaged foods (Table 8). 'Chilled ready meals' was the only category in 'chilled processed foods' that experienced increased consumption, with new categories, 'chilled pizza' and 'chilled/fresh pasta' appearing in 2004 and 2009, respectively (Table 8).

'Sweet and savoury snacks' consumption, also at 2.3 kg.capita/year, experienced significant growth of 53.3% between 1999 and 2012 with both 'chips/crisps' and 'extruded snacks' experiencing increases in consumption of more than 33% (Table 8). Potato and corn chips dominate with innovative local flavours such as Mrs Balls Chutney and Sweet Chilli leading, rather than entirely new products.[60] Consumer drivers for this increase in consumption are busy lifestyles, snacking between meals, the launch of healthier alternatives (low in salt for example) and new flavours.[61]

Soft drinks

Consistent with the results of other studies, South Africans are increasing their consumption of soft drinks, especially in urban areas.[16,31,44] Soft drinks were second to fruit, the most commonly purchased street food item.[31] The high prevalence of soft drink consumption is concerning in terms of its association with obesity and non-communicable diseases.[31] A recent study conducted in the USA concluded that added sugar intake from sugar-sweetened soft drinks is associated with an increase in cardiovascular disease mortality and recommended that calorie intake from added sugar be limited.[62]

'Total soft drink' consumption increased by a dramatic 68.9% from 55 L.capita/year in 1999 to 92.9 L.capita/year in 2012, with all categories experiencing significant growth (Table 9). South Africa's annual per capita consumption of Coca-Cola products (including regular, low-calorie and no-calorie, sparkling beverages, ready-to-drink (RTD) juices and juice drinks, RTD coffees and teas, sports drinks, energy drinks, dairy, waters and enhanced waters) is 260 8-ounce servings (equivalent to a 237-mL serving), which is significantly higher than the worldwide average of 94 (22.3 L.capita/year).[63] This equates to 61.6 L.capita/year in 2012 for Coca-Cola products in South Africa and is a significant increase of 80% from 144 8-ounce servings or 34 L.capita/year in 1992.[63]

The largest category of soft drinks is 'carbonates', which contributes a significant 67.5 L.capita/year (Table 9). Consumption across all 'carbonates' increased (from 41.2% to 100%) between 1999 and 2012 (Table 9). 'Low calorie/kilojoule cola carbonates' consumption increased by 45% (Table 9), but unfortunately remains a small contributor to the overall 'carbonates' consumption at <4.3% of total carbonates in 1999 and 2012. Better education among consumers about nutrition is affecting consumer preferences towards healthier options such as bottled water, nectars (25–99% juice), RTD tea, low-calorie cola carbonates and non-cola carbonates.[64-66] The second largest category, contributing 9.2 L.capita/year to 'total soft drinks' consumption in 2012, is 'fruit/vegetable juice', which grew by 44% between 1999 and 2012 (Table 9). The largest growth in consumption was observed in '100% juice' offerings (Table 9), which remained the most popular juice type, followed by nectars and fruit drinks.[64]

'Bottled water' has become a substantial global business and consumption continues to increase rapidly, even in countries with available, safe potable tap water.[67,68] There are many consumer considerations for this finding, such as dissatisfaction with tap water taste, demographics, perceived quality or safety of the water source, branding and marketing influences and overall convenience.[66-68] 'Bottled water' in South Africa, similar to the rest of the world, experienced a dramatic growth of 315% after 1999 to contribute 8.3 L.capita/year in 2012 (Table 9). However, a consumption of 8.3 L.capita/year still is significantly lower than the 2003 global average of 22.7 L.capita/year.[68]

Since 1999, 'concentrates' consumption has doubled to 5.1 L.capita/year in 2012 through consumption of liquid concentrates (Table 9). Two new categories that have experienced a dramatic 600% growth from 1999 to 2012 are 'sports and energy drinks' (from 0.3 to 2.1 L.capita/year) and 'RTD tea' (from 0.1 to 0.7 L.capita/year), albeit amongst a small population (Table 9). 'RTD tea' contains approximately 30% less sugar/calories than many 'carbonates'.[40] For this reason there is an increased consumption as consumers move away from 'carbonates' to healthier alternatives.[66]

In terms of 'sports and energy drinks' consumption, even though consumption has grown significantly since 1994 to 2.1 L.capita/year, it is still low compared to that of developed countries. For example, Germany's per capita consumption of 'sports and energy drinks' is 5.3 L.capita/year and Austria's is 11.1 L.capita/year.[69]

As mentioned above per capita consumption data, as a measure, is a crude estimate and generally an overestimate, especially for food items consumed by a smaller portion of the population. For commodity food items, a data 'integrity-check' could be done to verify the reliability and accuracy of the two data sources (FAOSTAT FBS and Euromonitor PFBC). However, Euromonitor International© Passport was the only available source for packaged food and beverage data and therefore could not be verified against other sources.

Conclusions

The most significant shifts (>30% increases) in food items consumed from 1994 to 2012 (Figure 1) were observed in these food and beverage categories: meat, fats and oils, sauces, dressings and condiments, sweet and savoury snacks, and soft drinks.

A significant increase was observed in meat consumption (Figure 1), with chicken being the most consumed animal protein. South Africans are consuming significantly more fats and oils coupled with a decrease in animal fat consumption. Packaged foods and beverages have seen the most dramatic shift (>50%) in soft drinks, sauces, dressings and condiments, and sweet and savoury snacks, which is typical of an upwardly mobile consumer (Figure 1). Convenience, health and wellness, and indulgence were the main consumer drivers for the increase in consumption of packaged foods since 1994.

In conclusion, the food consumption changes in South Africa observed in this study caused a shift towards a diet of:

- sugar-sweetened beverages
- increased proportions of processed and packaged food including edible vegetable oils
- increased intake of animal source foods
- added caloric sweeteners
- decreased vegetable consumption

These shifts in food choices are concerning for public health because of selected food items contributing to fat, sugar, salt and total kilojoule intake. These consumption shifts are in keeping with various food- and nutrition-related studies conducted at community or provincial level over the past few decades.[9,23-25] As indicated in these studies, the nutritional consequence of these food consumption shifts has contributed to increased obesity and other non-communicable diseases.[23,24,70] This problem is further exacerbated by the fact that a healthy diet is largely unaffordable for most South Africans[11], especially when considering that price is the most important factor taken into consideration when selecting food items[9].

Locally, and globally, the fast food, bottled soft drink and multinational food companies are often implicated in the increase in non-communicable diseases.[24,71,72] The South African Department of Health has therefore targeted the food and beverage industry with regulations in an attempt to improve public health. Further regulations related to food composition and/or labelling as well as consumer demand for healthier, affordable products will require advances and/or the application of science and technology developments by the South African food and beverage industry.

Acknowledgements

We thank Rosalind Hartfall and Susie Johnson from Euromonitor International© Passport for their support in the use and interpretation of Euromonitor data; Johan Visser and Anee Sieberhagen at Nampak Research & Development for providing access to their extensive library; and Mars Africa for providing time and financial assistance.

Table 9: Consumption of soft drinks in South Africa according to Euromonitor Packaged Food and Beverage Consumption (PFBC)

Food item	Euromonitor PFBC (kgL.capita/year)				
	1999	2004	2009	2012	% Change (1999–2012)
Total soft drinks	55	70.8	87.3	92.9	68.9
Carbonates	43.7	54.7	65.5	67.5	54.5
Cola carbonates	24.3	30.5	34.4	34.6	42.4
Regular cola carbonates	22.3	27.6	31.4	31.7	42.2
Standard regular cola	22.3	27.2	31.4	31.7	42.2
Speciality regular cola	–	0.4	–	–	
Low calorie cola carbonates	2	2.9	3.1	2.9	45.0
Standard low calorie cola	1.7	1.9	2.4	2.4	41.2
Speciality low calorie cola	0.3	1	0.6	0.6	100.0
Non-cola carbonates	19.4	24.2	31.0	32.9	69.6
Lemonade/lime	6.1	7.7	9.9	10.1	65.6
Juice-based lemonade/lime	0.1	0.1	0.1	0.2	100.0
Non-juice-based lemonade/lime	6	7.6	9.8	9.9	65.0
Mixers	0.8	1	1.3	1.6	100.0
Ginger ale	0.1	0.1	0.2	0.2	100.0
Seltzer	0.4	0.5	0.7	0.8	100.0
Tonic water	0.3	0.4	0.4	0.6	100.0
Orange carbonates	3.5	4.1	4.4	4.4	25.7
Non-juice-based orange carbonates	3.5	4.1	4.4	4.4	25.7
Other non-cola carbonates	9	11.4	15.4	16.8	86.7
Fruit/vegetable juices	6.4	7.1	8.2	9.2	43.8
100% juice	4.4	5.1	6.2	7.2	63.6
Juice drinks (up to 24% juice)	0.3	0.2	0.2	0.2	-33.3
Bottled water	2.0	4.1	7.1	8.3	315.0
Concentrates	2.5	3.5	4.1	5.1	104.0
Liquid concentrates	2.5	3.5	4.1	5.1	104.0
RTD tea	0.1	0.3	0.6	0.7	600.0
Still RTD tea	0.1	0.3	0.6	0.7	600.0
Sports and energy drinks	0.3	1.1	1.8	2.1	600.0
Energy drinks	0.1	0.3	0.6	0.6	500.0
Sports drinks	0.2	0.7	1.3	1.4	600.0

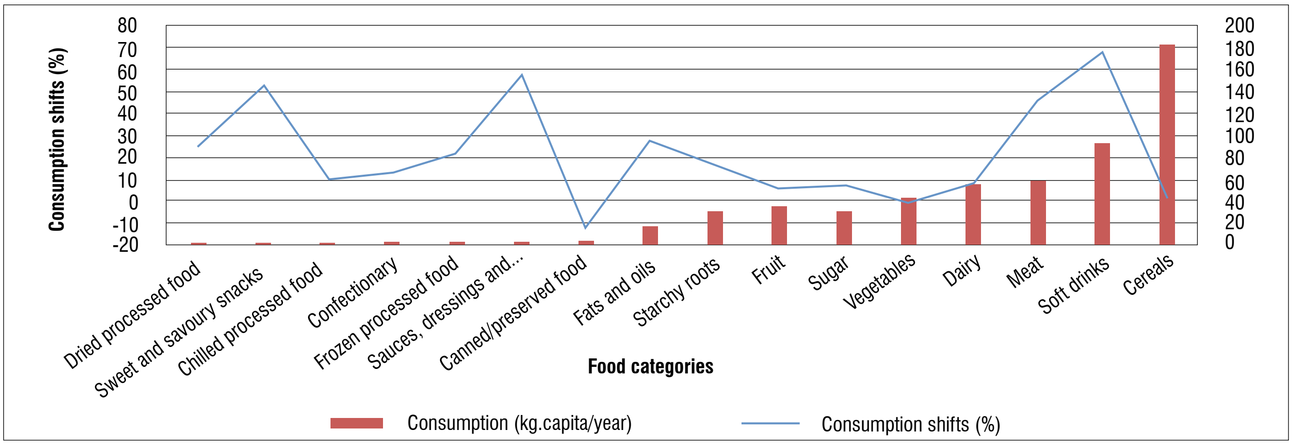


Figure 1: Per capita consumption and shifts in consumption of specific food categories in South Africa from 1994/1999 to 2009/2012 (FAOSTAT FBS & Euromonitor PFBC data sets).

Authors' contributions

L.C.R-R. was the project leader and was responsible for data collection and analysis and the write-up of the manuscript. G.O.S. was the supervisor and made editorial contributions. L.C.R-R., G.O.S. and N.V. were responsible for the project design.

References

1. Kearney J. Review: Food consumption trends and drivers. Philos Trans R Soc London B. 2010;365:2793–2807. http://dx.doi.org/10.1098/rstb.2010.0149
2. Bureau for Food and Agricultural Policy (BFAP). The South African agricultural baseline [homepage on the Internet]. c2011 [cited 2012 Feb 19]. Available from: http://www.bfap.co.za/documents/baselines/BFAP_Baseline_2011
3. Wenhold F, Annandale J, Faber M, Hart T. Water use and nutrient content of crop and animal food products for improved household security: A scoping study. WRC Report no. 2012TT 537/12. Pretoria: Water Research Commission; 2012.
4. Anonymous. Rising sophistication. Food & Beverage Reporter. 1999 Nov/Dec; 6.
5. World Wide Fund (WWF). Agriculture: Facts and trends South Africa [document on the Internet]. c2012 [cited 2012 Dec 21]. Available from: http://awsassets.wwf.org.za/downloads/facts_brochure_mockup_04_b.pdf
6. Bureau for Food and Agricultural Policy (BFAP). The South African agricultural baseline [homepage on the Internet]. c2013 [cited 2013 Nov 04]. Available from: http://www.bfap.co.za/documents/baselines/BFAP_Baseline_2013
7. Reardon T, Gulati A. The supermarket revolution in developing countries policies for "competitiveness with inclusiveness". International Food Policy Research Institute (IFPRI) Policy Brief 2. East Lansing, MI: Michigan State University; 2008.
8. Statistics South Africa. Income and expenditure of households 2005/2006. Statistical release P0100 [document on the Internet]. c2008 [cited 2014 Apr 03]. Available from: http://www.statssa.gov.za/publications/P0100/P01002011
9. Shisana O, Labadarios D, Rehle T, Simbayi L, Zuma K, Dhansay A, et al. South African National Health and Nutrition Examination Survey (SANHANES-1). Cape Town: HSRC Press; 2013.
10. Nielsen. South Africans are cautious, price-sensitive grocery shoppers [homepage on the Internet]. c2012 [cited 2014 Feb 11]. Available from: http://foodstuffsa.co.za/food-trends-mainmenu-119/food-trends-2012/2316-south-africans-are-cautious-price-sensitive-grocery-shoppers
11. Temple NJ, Steyn NP, Fourie J, De Villiers A. Price and availability of healthy food: A study in rural South Africa. Nutrition. 2011;27(1):55–58. http://dx.doi.org/10.1016/j.nut.2009.12.004
12. Serra-Majem L, MacLean D, Ribas L, Brulé D, Sekula W, Prattala R, et al. Comparative analysis of nutrition data from national, household, and individual levels: Results from a WHO-CINDI collaborative project in Canada, Finland, Poland, and Spain. J Epidemiol Commun H. 2003;57:74–80. http://dx.doi.org/10.1136/jech.57.1.74
13. Faber M, Wenhold FAM, MacIntyre UE, Wentzel-Viljoen E, Steyn NP, Oldewage-Theron WH. Presentation and interpretation of food intake data: Factors affecting comparability across studies. Nutrition. 2013;29:1286–1292. http://dx.doi.org/10.1016/j.nut.2013.03.016
14. Labadarios D, Steyn NP, Maunder E, MacIntryre M, Gericke G, Swart R, et al. The National Food Consumption Survey (NFCS): South Africa, 1999. Public Health Nutr. 2005;8(5):533–543. http://dx.doi.org/10.1079/PHN2005816
15. Steyn NP, Abercrombie R, Labadarios D. Food security – An update for health professionals. South Afr J Clin Nutr. 2001;14(3):98–102.
16. Steyn NP, Nel JH, Casey A. Secondary data analyses of dietary surveys undertaken in South Africa to determine usual food consumption of the population. Public Health Nutr. 2003;6(7):631–644. http://dx.doi.org/10.1079/PHN2003482
17. Labadarios D, Steyn NP, Nel J. How diverse is the diet of adult South Africans? Nutr J. 2011;10(33):1–11. http://dx.doi.org/10.1186/1475-2891-10-33
18. Kelly A, Becker W, Helsing E. Food balance sheets. In: Becker W, Helsing E. Food and health data: Their use in nutrition policy-making. World Health Organization Regional Publications European Series No. 34. Copenhagen: WHO Regional Office for Europe; 1991. p 39–47.
19. FAOSTAT. Food balance sheets [homepage on the Internet]. c2013 [cited 2013 Jul 26]. Available from: http://faostat3.fao.org/faostat-gateway/go/to/download/FB/*/E
20. Statistics South Africa. Statistics by theme/Living condition [homepage on the Internet]. c2014 [cited 2014 Apr 03]. Available from: www.statssa.gov.za
21. Faber M, Wenhold F. Food intake and sources of food of poor households in rural areas of South Africa. In: Water use and nutrient content of crop and animal food products for improved household food security: A scoping study. WRC Report no. TT 537/12. Pretoria: Water Research Commission; 2012. p. 24–57.
22. Euromonitor International© Passport [homepage on the Internet]. c2013 [cited 2013 Nov 05]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx
23. Bourne LT, Lambert EV, Steyn K. Where does the black population of South Africa stand on the nutrition transition? Public Health Nutr. 2002;5(1A):157– 162. http://dx.doi.org/10.1079/PHN2001288
24. Kruger HS, Puoane T, Senekal M, Van Der Merwe MT. Obesity in South Africa: Challenges for government and health professionals. Public Health Nutr. 2005;8(5):491–500. http://dx.doi.org/10.1079/PHN2005785
25. Viljoen AT, Botha P, Boonzaaier CC. Factors contributing to changes in food practices of a black South African community. J Family Ecol Consum Sci. 2005;33:46–62.
26. Steyn NP, Myburgh NG, Nel JH. Evidence to support a food-based dietary guideline on sugar consumption in South Africa. B World Health Organ. 2003;81(8):599–608.
27. Bertram MY, Steyn K, Wentzel-Viljoen E, Tollman S, Hofman KJ. Reducing the sodium content of high-salt foods: Effect on cardiovascular disease in South Africa. S Afr Med J. 2012;102(9):743–745. http://dx.doi.org/10.7196/samj.5832
28. Vorster HH, Love P, Browne C. Development of food-based dietary guidelines for South Africa – The process. S Afr J Clin Nutr. 2001;14(3):S3–S6.
29. Euromonitor International. Packaged food in South Africa [database on the Internet]. c2012 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx
30. Lombard M, Labuschagne I, Goosen C. The nutritional value of canned vegetables and fruit within a balanced diet. S Afr Food Review. 2011;38(2):24– 25.
31. Steyn NP, Labadarios D, Nel JH. Factors which influence the consumption of street foods and fast foods in South Africa – A national survey. Nutr J. 2011;10(104):1–10. http://dx.doi.org/10.1186/1475-2891-10-104
32. Nutrition Australia. Dairy food myths [homepage on the Internet]. c2009 [cited 2014 Mar 04]. Available from: http://www.nutritionaustralia.org/national/resource/dairy-food-myths
33. Rolando R. The big cheese. S Afr Food Rev. 2012;39(1):15.
34. Euromonitor International. Cheese in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx
35. Euromonitor International. Drinking milk products in South Africa. 2013 February.
36. Cochrane K. The doings of a dairy giant. S Afr Food Rev. 2008;35(11):16–20.
37. Brooks N. Danone developments. S Afr Food Rev. 2010;37(11):21–23.
38. Euromonitor International. Yoghurt and sour milk products in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx
39. Popkin BM. The nutrition transition: An overview of world patterns of change. Nutr Rev. 2004;62(7):140–143. http://dx.doi.org/10.1301/nr.2004.jul.S140-S143
40. Unilever. Why tea is an all-time favourite [homepage on the Internet]. c2013 [cited 2014 Mar 12]. Available from: http://www.unilever.co.za/brands-in-action/detail/Why-tea-is-an-all-time-favourite/292030/
41. Durham L. Stir up your hot beverage sales. Supermarket & Retailer. 2011 March; p. 37–40.
42. Rolando R. Super food, the spice of life. S Afr Food Rev. 2010;37(6):25–26.
43. Hyslop G. In good health. S Afr Food Rev. 2012;39(4):25–26.

44. Igumbor EU, Sanders D, Puoane T, Tsolekile L, Schwarz C, Purdy C, et al. "Big food," the consumer food environment, health and the policy response in South Africa. PLOS Med. 2012;9(7):1–7. http://dx.doi.org/10.1371/journal.pmed.1001253

45. Kupka J. Food and beverage industry performs well. S Afr Food Rev. 2010;37(4):18.

46. Euromonitor International. Baked goods in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

47. Hillmann J. Bread makers get creative. S Afr Food Rev. 2007;34(7):21–24.

48. Euromonitor International. Biscuits in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

49. Solomon N. Democratising the bakers brand to access sales in the informal sector. Food & Beverage Reporter. 2011 May; p. 7–8.

50. Euromonitor International. Breakfast cereals in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

51. Fast-moving Consumer Goods (FMCG). Breakfast cereals in South Africa. c2014 [cited 2014 Mar 10]. Available from: http://www.fastmoving.co.za/brand-category/breakfast-cereals-218

52. Eagle J. Canned foods industry in decline [homepage on the Internet]. c2013 [cited 2013 Jul 09]. Available from: http://www.foodproductiondaily.com/Packaging/Canned-foods-industry-in-decline

53. Euromonitor International. Canned/preserved food in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

54. Neall B. The sprouting of nature's choice. S Afr Food Rev. 2006;33(3):11–12.

55. Euromonitor International. Sauces, dressings and condiments in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

56. Euromonitor International. Frozen processed food in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

57. BMI Research. Confectionery and chocolate: SA marketing insights from BMI [homepage on the Internet]. c2011 [cited 2014 Oct 12]. Available from: http://www.foodstuffsa.co.za/food-trends-mainmenu-119/food-trends-2011/1153-confectionery-and-chocolate-sa-market-insights-from-bmi

58. Anonymous. SA's No. 1 choc supplier. S Afr Food Rev. 2009;36(10):32.

59. Euromonitor International. Chocolate confectionery in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

60. Rossouw J. The African snack market – Flavouring the world. S Afr Food Rev. 2010;37(5):21.

61. Euromonitor International. Sweet and savoury snacks in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

62. Yang Q, Zhang Z, Gregg EW, Flanders WD, Merritt R, Hu FB. Added sugar intake and cardiovascular diseases mortality among US adults. JAMA Intern Med. 2014;174(4):516–524. http://dx.doi.org/10.1001/jamainternmed.2013.13563

63. Coca-Cola company. Annual review 2012 [document on the Internet]. c2012 [cited 2014 Mar 12]. Available from: http://www.coca-colacompany.com/annual-review/2012/pdf/TCCC_2012_Annual_Review.pdf

64. BMI Research. An eye on SA beverage trends [homepage on the Internet]. c2011 [cited 2014 Mar 19]. Available from: http://www.foodstuffsa.co.za/food-trends-mainmenu-119/food-trends-2011/1391-an-eye-on-sa-beverage-trends-1

65. Hu C. Soft drinks report 2011. S Afr Food Rev. 2011;38(8):32–36.

66. Euromonitor International. Soft drinks in South Africa [database on the Internet]. c2013 [cited 2014 Oct 12]. Available from: http://www.portal.euromonitor.com/Portal/Default.aspx

67. Doria MF. Bottled water versus tap water: Understanding consumers' preferences. J Water Health. 2006;4(2):271–276.

68. Wilk R. Bottled water – The pure commodity in the age of branding. J Consum Cult. 2006;6(3):303–325. http://dx.doi.org/10.1177/1469540506068681

69. Hyslop G. Anuga 2013 in Cologne. S Afr Food Rev. 2013;40(9):45–46.

70. Popkin BM. Global nutrition dynamics: The world is shifting rapidly toward a diet linked with noncommunicable disease. Am J Clin Nutr. 2006;84:289–298.

71. Bowman, SA, Gortmaker SL, Ebbeling CB, Pereira MA, Ludwig D. Effects of fast-food consumption on energy intake and diet quality among children in a national household survey. Pediatrics. 2004;113:112–117. http://dx.doi.org/10.1542/peds.113.1.112

72. Anonymous. Food industry to blame for obesity epidemic. S Afr Food Rev. 2012;39(4):22.

Plants, people and health: Three disciplines at work in Namaqualand

AUTHORS:
Lesley Green[1]
David W. Gammon[2]
Michael T. Hoffman[3]
Joshua Cohen[4]
Amelia Hilgart[2]
Robert G. Morrell[5]
Helen Verran[6]
Nicola Wheat[7]

AFFILIATIONS:
[1]Environmental Humanities South, School of African and Gender Studies, Anthropology and Linguistics, University of Cape Town, Cape Town, South Africa
[2]Department of Chemistry, University of Cape Town, Cape Town, South Africa
[3]Department of Biological Sciences, University of Cape Town, Cape Town, South Africa
[4]School of African and Gender Studies, Anthropology and Linguistics, University of Cape Town, Cape Town, South Africa
[5]Office of the Vice-Chancellor, University of Cape Town, Cape Town, South Africa
[6]History and Philosophy of Science, University of Melbourne, Melbourne, Australia
[7]Open Box Software, Cape Town, South Africa

CORRESPONDENCE TO:
Robert Morrell

EMAIL:
Robert.morrell@uct.ac.za

POSTAL ADDRESS:
University of Cape Town –
Research Office, 2 Rhodes Ave,
Mowbray, Cape Town 7700,
South Africa

DATES:
Received: 11 Aug. 2014
Revised: 10 Nov. 2014
Accepted: 19 Dec. 2014

KEYWORDS:
interdisciplinarity; scientific knowledge; medicinal plants; natural products chemistry; phenology; indigenous knowledge

In Paulshoek, Namaqualand, three research projects focusing on medicinal plants were developed concurrently. The projects were based in the disciplines of anthropology, botany and chemistry. In this paper, we explore how these projects related to one another and describe the conversations that occurred in the process of searching for transdisciplinary knowledge. The projects ostensibly shared a common object of knowledge, but it was through working together that the medicinal plants constituted us as a community of scholars. As our insight into our respective disciplinary relationships with the plants grew, so did our understanding of the limitations of our respective disciplinary positions. The process made possible a 'reimagination' of both the object of study and our relationships to it and to one another. The research project, conceptualised in 2009, engaged current debates on indigenous knowledge and its historical erasures, and offered an approach that has potential to produce new knowledges while respecting the integrity of the disciplines. This approach requires a non-competitive attitude to research and one that acknowledges the contributions that can be made by multiple approaches.

Introduction

Historical divisions between the sciences and the humanities that reach back to the origins of modernist thought have long inhibited a productive conversation across disciplines. Different ways of establishing what counts as evidence, how it counts, and how to account for it, mean that when a botanist, a chemist, an anthropologist, a *kruiedokter* [herbal doctor] or a goatherder attend to something as apparently self-evident as plants in the veld, we notice different things, name objects differently, and put them into our respective scholarly dialogues in very different ways. Good science and reliable knowledge matters deeply to all of us, and for that reason, we are mindful of the vital role that our respective disciplinary gatekeepers play, whether they are peer reviewers, discussants in departmental seminars, or examiners of dissertations. It is not easy to write 'outside' of our disciplines, because disciplines serve literally to 'discipline' the methods of establishing understanding, making it difficult to sustain generative transdisciplinary conversations.

This paper reports on an exploration in talking in disciplinary parallel (each with our own disciplinary language) and in sharing findings. It is also about a process of developing new approaches to the objects of our enquiry and forging new relationships with them and with one another. The exploration involved 'respectively and variously meet(ing) the differing epistemic requirements and methodological obligations of knowing' in the three disciplinary scientific communities as well as among the *kruiedokters*.[1] It also involved accepting that, as research interests converged and our understandings were each enriched with new disciplinary nuggets, there was not a single knowledge of the plant waiting to be uncovered but rather many, to which we were each contributing.

This paper is consciously an exercise in transdisciplinarity. There are four elements of transdisciplinary research: a focus on life-world problems, transcending and integrating disciplinary paradigms, participatory research and the search for unity of knowledge beyond disciplines.[2] The method has an integrative effect that calls for holistic approaches:

> *The notion of the ecosystem is central to representing the world and its functioning, cycles, equilibria and dynamics. We are dealing here with coherences, balance sheets, and not with the absolute objectivation of the things that make up the world. There is no longer truth per se, about the complexity of the world, but knowledge that is more or less complete, and therefore uncertain.*[3]

Knowledge making – the 'doing of research' – does not occupy a neutral ground free of vested interests. This fact is never clearer than in colonial settings such as South Africa, where eugenics established itself as a racist form of science that justified the colonial and apartheid projects which entrenched racial (and other) inequalities.[4] In South Africa today, scholars working in a postcolonial democracy continue to face challenges, the most visible of which have been struggles over plant medicines and HIV. Both disciplinary stricture and political prescription can stand in the way of developing new knowledge.

Living and working in a democracy in which historical redress is an important issue in the conduct of research, tax-funded researchers bear a responsibility to the public to ensure that the knowledge in terms of which governments

HOW TO CITE:
Green L, Gammon DW, Hoffman MT, Cohen J, Hilgart A, Morrell RG, et al. Plants, people and health: Three disciplines at work in Namaqualand. S Afr J Sci. 2015;111(9/10), Art. #2014-0276, 12 pages. http://dx.doi.org/10.17159/sajs.2015/20140276

make decisions is not only accurate, verifiable and reliable, but also takes account of perspectives that have historically been marginalised. The intensity of the debate over indigenous knowledge and antiretrovirals in South Africa amply underscores this point.[5-9]

We began to open a dialogue about people, plants and well-being in a small village called Paulshoek in Namaqualand in 2009, in the wake of the struggle over antiretrovirals, on our respective ways of producing knowledge about the relationship between plants and people. The question we sought to pose together was what different 'knowledges' would we produce if we worked together to understand the plants and people in Paulshoek. Determined to take seriously both the benefits of disciplinary philosophies and methods, and to consider ways of bridging the gaps between the approaches, a project was conceptualised that involved three distinct studies in the same area concerning plants and plant medicine. Over time, the project became the 'ABC Project': Anthropology, Botany, Chemistry, reflecting both our respective disciplines and the challenge of 'beginning at the beginning' – to find the beginnings of a common language in which our respective studies could be brought into what Verran and Christie[10] so aptly called a 'generative dialogue'. Since that early set of conversations in 2009, one PhD has been completed, with another two PhD studies close to completion. Each study engaged with at least two of the three disciplines through a series of workshops and dialogues that included all parties – supervisors, graduates and transdisciplinary specialists.

The project builds on Timm Hoffman's extensive multi-year study of people–plant interactions in the area,[11] much of which derives from ongoing work with Paulshoek resident, Marianna Lot, with whom all the researchers in turn came to work. Questions about plant chemistry led to a dialogue with organic chemist David Gammon, and over time, both Hoffman and Gammon found themselves in conversation with a number of *kruiedokters* based in the area. At that point, a conversation opened with Lesley Green, an anthropologist interested in postcoloniality, democracy and knowledge.[12] Three graduate studies emerged in the course of this three-way conversation: the ethnobotanical, phytochemical and metabolomics of plants (undertaken by Nicola Wheat), systems biology approaches with the incorporation of the socio-biome for plant natural products chemistry (undertaken by Amelia Hilgart) and the work of *kruiedokters*, the experiences of their patients and the multi-layered setting in which plants are called upon to affect human well-being in various ways (undertaken by Joshua Cohen).

Visiting scholar Helen Verran, a trained chemist and a reader in the history and philosophy of science, who has written extensively on knowledge and culture in West Africa and Australia, became a much valued participant in discussions during her visits to South Africa. Throughout the project, Robert Morrell of the University of Cape Town (UCT) Research Office worked alongside the team to hold open the scholarly conversation in a terrain that was immensely challenging, and which has led to insights about 'the bigger picture'. Here, natural and social science knowledges took up a different kind of relationship to one another than that of the initial framing of our interdisciplinary project.

Initially after developing research questions that straddled disciplinary interests, the actual work of the various teams sought to answer these questions via strong disciplinary methodologies. The goal was to hold open conversations about our findings as they emerged. Once the respective PhD projects were close to completion, we met together regularly to discuss our findings and to try to understand where the conversation between our respective ways of working lay, and what the topics of conversation might be. The slow unfolding conversation that we had over several months included sharing understandings, working through misunderstandings and, perhaps most valuably, seeking to understand why our different questions and ways of working mattered so deeply to each of us. Working in this way, the collective began to glimpse a set of puzzles that could in some way be attended to by the particular affordances offered by each discipline, where each could be recognised and valued as contributing on its own terms.

Our particular research objects began to come into focus: the chemist's antimicrobial molecules, the anthropologist's rendering of the *bossiedokter*'s [bush doctor's] idea of *krag* (strength, energy or vitality) and the botanist's account of phenological cycles. Once these objects were clearly in focus, we could start a conversation about why and how these research objects mattered, historically, socially and scientifically and why they were objects of concern. Doing so, we argue, rendered us no longer mute in the face of one another's facts and reasonings, but rather offered possibilities for different disciplines to find their voice within our collective. Learning to explore our differences explicitly but jointly, we learned to respect medicinal plants as differentially knowable. We could glimpse how those differentiations, in being respected, suggest just and efficacious organisational approaches to managing plants.

In this article, we seek to offer an account of the kinds of conversations we have had, both in our particular studies and with each other, with a view to thinking through the challenges to transdisciplinary studies in the context of postcolonial debates over knowledge in South Africa. In our case it remains a conversation that is characterised by frankness, humour and collegiality, and it speaks directly to much wider international conversations on the problem that lack of accountability of scientific experts to publics weakens democratic debate.[13] Moreover, the extent to which all of us generate different kinds of facts makes it all the more evident that different approaches within and without the academy have contributions to make to improving our understanding of what is referred to universally simply as 'health'.

Where knowledge producers at universities have for centuries been locked in an adversarial relationship with one another over the conditions for the production of truth about nature, the recognition of the wider social and historical context of scholarship provokes a more humble relationship to 'the facts' that we produce – which in turn offers the beginnings of a more convivial conversation, akin to what Latour speaks of as 'scholarly diplomacy'.[14] Yet such a space is only just emerging and needs constantly to be tended and protected, particularly as opening up transdisciplinary conversation renders academics vulnerable to ripostes from colleagues who are more comfortable in the centre of a discipline. With these goals in mind, in this article, we set out the findings of the respective studies, presenting them each in terms that are defensible within their own disciplines, then work around and between them to open up the question of whether transdisciplinary knowledge production is worth the trouble it brings.

Three disciplinary studies

Phenology: A botanist's view of seasonal plant changes

The study begins with the work of botanist Timm Hoffman. Botanical studies of plant life in the Succulent Karoo biome describe some of the richest diversity of plant life on the planet. For this reason, Hoffman, like many other South African botanists, was drawn to the semi-desert conditions in the Northern Cape where he and his students have focused on the plants of Paulshoek and the broader Kamiesberg area for 15 years.

Hoffman gives this narrative account of his project:

> I recorded the phenology (the seasonal cycles) of a wide range of different species in Paulshoek over a long time period to, firstly, add to existing information about each species (e.g. in www.PlantZAfrica.com) and, secondly, to help with the interpretation of the livestock production system data that I had been collecting in Paulshoek over the same length of time. I wanted to develop an understanding of how each species 'behaved' in response to rainfall, temperature and grazing over long periods. I hoped to use these insights to develop a better sense of the ecological 'integrity' or health of a landscape and to add to the wider range condition assessment literature.
>
> Because the Succulent Karoo biome receives most of its rain in the form of relatively predictable frontal systems, previous work has emphasised the predictability of plant responses to low but

regular rainfall.[15,16] However, few studies have investigated the response of different growth forms (e.g. bulbs, annuals, grasses, leaf-deciduous shrubs, leaf-succulent shrubs, trees) over long time periods to describe their phenological responses. In late 1999, 70 species from different growth forms were identified and their phenological response monitored on a monthly basis. The same approximately 300-m route was usually walked each month and three or four individuals from a species, sometimes more, were used to establish a general phenological profile for each species for each month. From these qualitative observations, it was possible to document the overall response of the community of plants to drought and high rainfall periods.

Despite apparent regularity and phenological consistency, long-term observations suggest that growth within the Succulent Karoo is remarkably variable from year to year, largely in response to rainfall. For the 4 years indicated in Figure 1, rainfall varied not only in the amounts which fell each year (2003 = 123 mm, 2005 = 202 mm, 2006 = 200 mm, 2011 = 304 mm) but also in distribution. In some years, such as during the extensive drought of 2003, almost no rain fell until September, while 2005 was characterised by an abundance of early season rainfall followed by a relatively dry late winter and spring.

These results suggest that the Succulent Karoo is predictable but also that the vegetation responds to discrete rainfall events. The implications of these findings are that livestock farmers need flexible approaches to management in order to accommodate such variability.

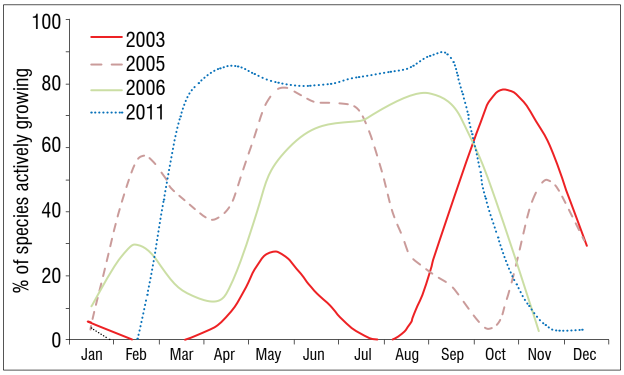


Figure 1: The percentage of species (n=70) with evidence of new shoot growth (i.e. actively growing) in each month in 4 different years in Paulshoek.

Organic chemistry: Searching for the chemistry of medicinal plants

The tentative beginning of an interdisciplinary project came about when Gammon encountered the work of Hoffman and together they began to conceive the possibility of adding new dimensions and insights to their respective endeavours. In particular, an idea explored was the extent to which insights at the molecular level, below the resolution of the human eye, aided or not by magnifying instruments, would complement or inform Hoffman's phenological observations. Gammon's preliminary formulation of the potential scope of the collaboration started from the classical definition of natural products, and noted not only the solid scientific edifice that now defines natural products science, but also the tensions associated with new technological developments and a greater awareness of the challenges at the interface of where the chemical entities are located and harvested.

Gammon gives his account of the kinds of scientific research pursued by natural products chemists:

> Natural products chemistry has classically concerned itself with small molecules present as secondary metabolites in living organisms – plants, fungi, microbes, and so on.[17] The structurally diverse secondary metabolites[18] are so called on the basis of appearing to not be directly necessary for growth and development of organisms, as opposed to primary metabolites such as lipids, nucleotides, amino acids and organic acids which are involved in essential metabolic processes. Secondary metabolites have received a great deal of attention because of the array of uses and activities that they exhibit, and, more recently, through recognition of their adaptive role and ecological functions.
>
> Historically, natural products chemistry has, through its primary focus on careful analysis of extracts of natural materials and separation of the complex mixtures in pursuit of single, pure chemical entities, led to the discovery of new substances with pharmacological or other activities. The field has been dominated by a strong drive to discover new drugs or drug leads[19] and to contribute broadly to improving healthcare, although this has not been the only motive, with research fields such as chemical ecology having significant traction.
>
> However, over the last decade or two, the discipline of natural products chemistry has been caught in a tension of introspection, on the one hand, and something of a renaissance on the other. Leading practitioners like Cordell and others[18-22] have been calling for an urgent reappraisal of the importance of natural products research. They suggest that natural products chemists should face up to the unavoidable challenges of provision of medicines and healthcare for a burgeoning world population, particularly in parts of the world where the majority of people have limited access to 'first-world' medicine, while at the same time calling for greater sensitivity to the environment and raising questions about who benefits from the research. In issuing these challenges, they however do not acknowledge the role of indigenous people living within and from the biodiversity, in terms of knowledge production or dissemination. This issue is notably taken up by Etkin and Elisabetsky in their analysis of papers published in the Journal of Ethnopharmacology over a 25-year period since the inception of the journal, where, despite the stated intentions of the journal, they conclude that:
>
> *Much of what is reported as ethnopharmacological research is comprised by decontextualised catalogues of plants and lists of phytoconstituents and/or pharmacologic properties [and] few researchers in ethnopharmacology seem to be interested in the people whose knowledge and identity are embodied in these plants. While some studies are based on plants drawn from indigenous pharmacopoeias, most of what is published as ethnopharmacology has a weak, if any, ethnographic component.*[23]

With these concerns in mind, Gammon and Wheat conceived a project which integrated contemporary approaches and considerations. The combined insights of Hoffman and anthropologist Lesley Green were considered invaluable in 'hearing' the alternative voices which several natural products chemists were seeking to bring to the fore. An additional interest for Gammon in particular was to contribute to the search for bioactive ingredients in plants, and to consider the extent to which new field technologies could be applied in ways that could contribute to teaching sciences in local schools. With this in mind, a 'field-deployable bioassay kit' was used. This kit was developed by the Global Institute of Bio-Exploration based at Rutgers University. It allows for simple, small-scale, in-the-field, preliminary assays of standardised plant extracts for broad-spectrum bioactivity, thus minimising the impact on the environment of removing quantities of plant material.[24] Wheat's study utilised these assays and validated them by comparison with more sophisticated, laboratory-based assays. In addition, the goal was to explore the scope and limitations of liquid chromatography–mass spectrometry and high-field nuclear magnetic resonance techniques, with appropriate data handling, for fingerprinting of extracts, correlation with bioactivity profiles and general preliminary assessment of plant extracts. The information from these methods of analysis enabled more efficient metabolomic profiling and offered an improved search for active constituents.

Nicola Wheat, the PhD candidate who worked with Gammon on the project, describes her entry into the work:

> My first visit to Paulshoek was in 2004 as part of a 10-day Botany Honours field trip led by Timm Hoffman. In the village we stayed in traditional reed mat houses (matjieshuise) in the camp site. That evening, the village women came to make us supper, the children sang and people played an 'action' version of dominoes that involved loudly slapping dominoes down on the table in quick succession. The next morning Timm took us to his research sites around the village and Oom ['uncle', used as a form of respect] Samuel took us on a walk through the veld, showing us important plants of the area. That evening there was a talent competition with much singing and dancing. We left the next morning. That was the beginning of my relationship with the people and plants of Paulshoek.
>
> In 2008, 5 years later, with an MSc and some work experience behind me, I was considering a PhD and visited the student advisor for chemistry for advice. He knew of David's endeavours to get a plant research team together and promptly referred me to him. David suggested a PhD on the chemistry of medicinal plants, with Paulshoek as the research site. With my background in botany, having worked with Timm and previously visited the area, all I needed was to complete a few additional chemistry courses and I was ready to start. I spent the next year taking chemistry courses, attending lectures and getting up to speed on plant and medicinal chemistry before starting my PhD in 2010. What was originally envisioned as a purely chemical analysis of medicinal plants and their constituent bioactive compounds turned out to be so much more, allowing me to work with a variety of people in a way I had not imagined.

For her doctoral project,[25] Wheat asked whether a study of medicinal plants from several different disciplinary vantage points could indeed produce an integrated approach to drug discovery from natural products. The project included broad-level ethnobotanical and anthropological studies with more focused metabolomic and phytochemical studies to better understand the pharmacological basis of culturally significant plants. Wheat set out to interrogate the widely held hypothesis that traditional knowledge, when considered from a scientific point of view, can act as a proxy for detecting bioactive molecules[26-28] and that the preferential selection of certain families for medicinal use may be used as an indicator of underlying bioactive phytochemistry.

Wheat's study involved comparing levels of biological activity in extracts from plants selected either randomly or on the basis of known medicinal or other uses. From an initial survey of over 100 plant species, she applied statistical techniques to narrow the focus to a handful of plants for further study and then used liquid chromatography–mass spectrometry and nuclear magnetic resonance to analyse whole crude extracts, with a particular focus on the extract from *Crassula brevifolia*, a common plant from Paulshoek. These data, together with results from a range of bioassays on the extracts, constitute a comprehensive profile of activity and metabolite composition – a multi-dimensional mapping of the plants (Figure 2), far richer than normally achieved from isolation of a small selection of plant constituents.[29] The study was open-ended: in principle, it was to set up the focused search for active ingredients, but it also provided a deeper insight into the hidden molecular world of the plant and a basis for a dialogue with traditional knowledge in the search for a synthesis which might accrue from the bringing together of advanced analytical techniques and lifetimes of learned experiences.

About the time that Wheat commenced her work, Amelia Hilgart approached the Botany Department at UCT to explore the possibility of studies in Botany, and conversations with Hoffman and Gammon led to the development of a PhD proposal in Chemistry.

Amelia Hilgart describes how her approach to her work evolved:

> 'Chemical stasis in a living organism equals death.' This was the opening statement of my first biochemistry lecture as an undergraduate; the lecture was on thermodynamic equilibrium in mammalian cells, and has largely shaped my perception of what a metabolite is. In a living organism, everything is constantly changing on a chemical level and involves an expenditure of energy. The compounds that we look for in plants in natural products chemistry are secondary metabolites, vaguely defined as 'the compounds produced by plants that are not directly essential for basic photosynthetic or respiratory metabolism; such compounds are known as primary metabolites.[30] I have always been impressed by the incredible diversity of compounds plants can make and particularly struck by the idea that plants use so much energy to create such a diverse array of chemicals. Unless genetically designed to do otherwise (or genetically broken), plants do not waste energy to create useless molecules. Plants make things that they can use when they need them. These were the ideas I had when I arrived in Cape Town.
>
> My first appreciation for the project came through Timm Hoffman when we took the 7-hour drive from UCT to Paulshoek to try to find something for a project that garnered my interest. Timm loves to talk about plants and my introduction to Paulshoek flora included a long discussion on the Aizoaceae family which dominates the landscape. By the end of that trip there were two ideas that featured strongly in my mind; firstly, that there was an extraordinary metabolic process – called facultative crassulacean acid metabolism – that featured in at least some plants in the Aizoaceae family, which allowed the plants to change their carbon uptake mechanisms; and secondly, that one Aizoaceae species in particular (Galenia africana) was killing goats and sheep when ingested during the summer months.

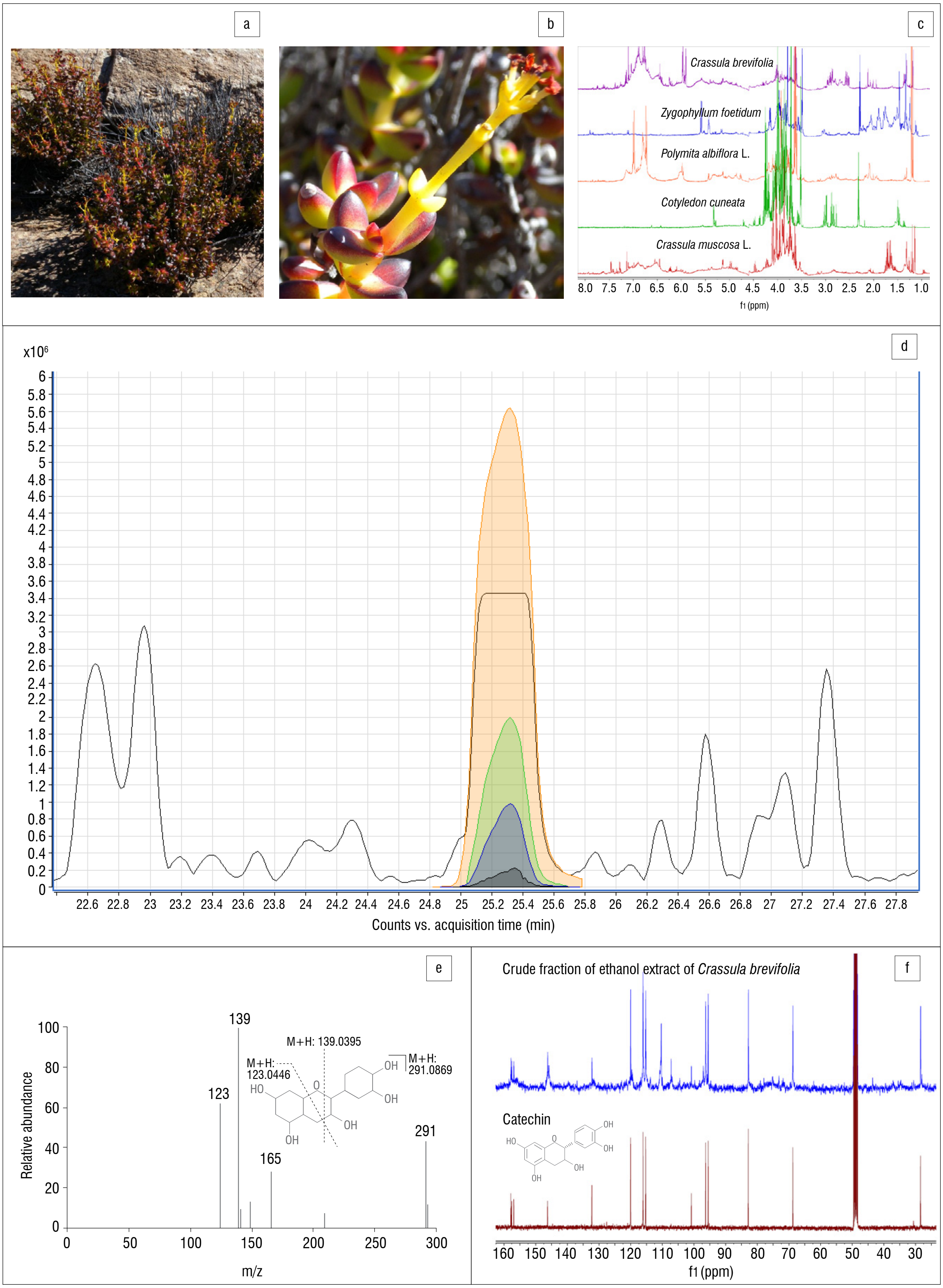


Figure 2: Ways of seeing *Crassula brevifolia*: (a) whole plant, (b) leaves, (c) comparison of proton magnetic resonance spectra of ethanol extract of *C. brevifolia* with similar extracts from other plants, (d) section (22–28 min) of the total ion chromatogram from liquid chromatography–mass spectrometry analysis of the ethanol extract of *C. brevifolia*, (e) tandem mass spectrum of catechin, showing masses of prominent ions, with the structure and fragmentation pattern of the molecule overlaid and (f) carbon magnetic resonance spectrum of pure catechin, and the mixture from which it was recovered.

Hilgart[30] started by monitoring the metabolic, nutrient, physiological and phenological fluctuations across seven common Namaqualand species in a quest for the toxic compound in *G. africana*.[31] *G. africana* is a pioneer species which thrives in the increasingly disturbed soils of southern Africa.[32] Hilgart's work utilised Hoffman's prior interviews with herders in the village of Paulshoek which indicated that there was seasonality to the toxicity of *G. africana*. These approaches and insights offered an opportunity to study toxic compound accumulation and fluctuation on an ecosystem-wide basis, and while considerable progress was made in this regard, the accumulation and statistical analysis of metabolite profiles from liquid chromatography–mass spectrometry data (Figure 3) led to development of molecular barcodes for fine-grained distinguishing of plant species – a kind of molecular 'fingerprint'.

Anthropology: A view of the work of kruiedokters

Around the time that Hilgart arrived at UCT to commence her studies, Gammon's interest in the work of the *kruiedokters* led to a conversation with Lesley Green in the Department of Social Anthropology, who at the time was working on the relationships between sciences and postcolonial knowledge debates, and the questions these raised for universities in South Africa. Green describes her involvement in the project as follows:

My first encounter with Paulshoek was in 2009 via David Gammon who emailed me after searching around at UCT to see who was working on indigenous knowledge. I was struck by the thoughtfulness with which David posed his question: how do we (in his case, scholars of plant chemistry) work with a kruiedokter whose way of working with his patients is serious and considered by locals as effective, yet the resources he is drawing on are in the domain of spirit? It was a really challenging assignment at the nexus of two pressing concerns: how to rethink the simplistic opposition between African knowledge and western science that was being pursued at the time in relation to HIV and the rejection of antiretrovirals, and how to work with very different approaches to knowledge: causality, ontology, epistemology and metaphysics.

Current anthropological debates internationally are focused on problematizing the dividing line between social sciences and life sciences; society

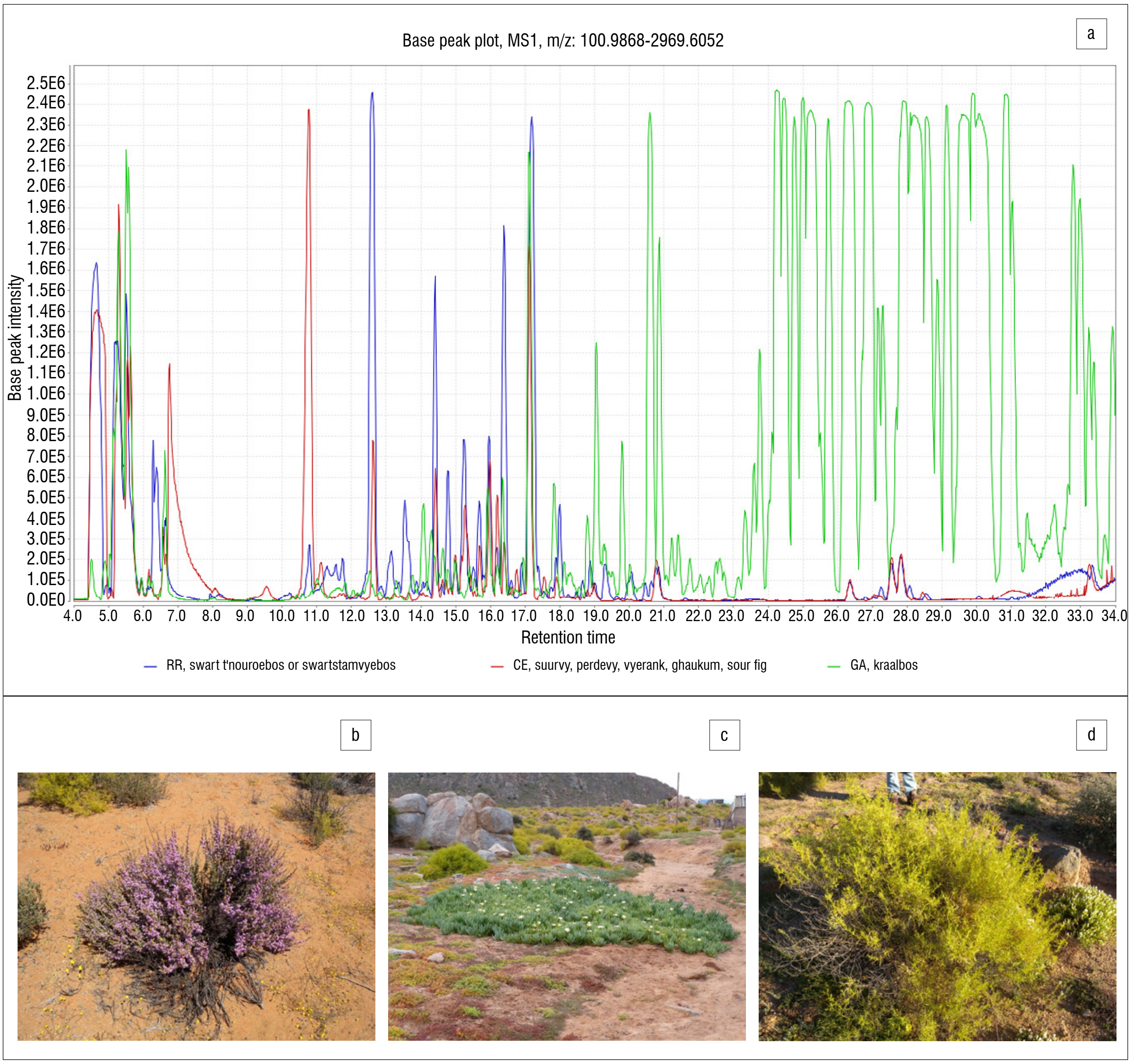


Figure 3: (a) Overlay of total ion chromatograms from liquid chromatography–mass spectrometry analyses of the ethanol extracts of *Tetragonia fruticosa* (red), *Carpobrotus edulis* (blue) and *Galenia africana* (green). Images of (b) *T. fruticosa*, (c) *C. edulis* and (d) *G. africana*.

> and nature. When we rethink this line, it is possible to begin to rethink the ways in which we produce disciplinary knowledges. Doing so requires simultaneous engagement with the criticisms of formal knowledges in the university system from across the south (these debates have various names: 'indigenous knowledge studies', 'decolonial theory' and 'political ontology') and with current philosophers of science such as Isabelle Stengers, whose very broad oeuvre is based on careful and detailed work with a range of knowledge fields. Her collaborators include Nobel Prize winner for chemistry, Ilya Prigogine. Stengers' work is situated in a set of dialogues with a wide range of thinkers who are taking up the challenge to open the frameworks of knowledge to new possibilities, while avoiding the carelessness of solutions to the challenges to knowledge that go under labels such as 'relativism' or 'tolerance' – arguing that 'epistemic charity' (to borrow a word from the critic Nandini Sundar) is not a helpful solution.

In the latter half of 2009, Joshua Cohen was looking for an issue to explore for a doctorate in Anthropology. Through his supervisor, Green, he found himself becoming involved with the Paulshoek plant researchers and began to engage in participative, interdisciplinary research. He describes his initiation into Paulshoek as follows:

> In May 2010, I travelled with Timm Hoffman on one of his monthly research trips to Paulshoek. He introduced me to Gert Julk, one of the region's well-known kruiedokters. Gert left an indelible impression on me from the beginning and instilled an interest in me to work further with kruiedokters. He was quite happy to talk and took us on a whistle-stop tour through his skills and work. I thought his openness and humorous demeanour could help facilitate the kind of collaborative work UCT's molecular biologists were hoping to engage him in.
>
> Sadly, Gert passed away in November 2010. In addition to the tragedy for his friends and family, Gert was the last actively working kruiedokter in Paulshoek. Marianna Lot, however, had already introduced me to Koos, a kruiedokter working in another of the Kamiesberg villages and I shifted my attention to Koos' work. Koos' open, cheerful ways have helped him and I to establish and sustain an excellent relationship based, I believe, on mutual respect and understanding. The relationships I have enjoyed with Koos, his family, friends and patients have been key to my understanding of how Koos and other kruiedokters work with plants.

Cohen's work focused on how *kruiedokters* work with *bossiemedisyne* [bush medicine].[33] In line with Chris Low's work on Khoisan healing,[34-36] two key terms – *krag* and *wind* – emerged. The first of these terms, *krag* (power, vitality, strength), is a commonly used Afrikaans word that can be thought of as a kind of 'body energy' that waxes and wanes with the ups and downs of everyday life. In order to aim their patients toward health, it is important for *kruiedokters* to be able to 'cultivate' their patients' *krag*. To this end, various tools might be used by the *kruiedokter* – jokes, guitar playing, food, and of course *bossiemedisyne*. Many plants are directed at alleviating *krag*-sapping symptoms: high blood pressure, diabetes, colds and flu, swelling in the limbs. The *krag* in the plants themselves, that which enables them to do their healing work, is closely associated with the *krag van die natuur* (the power of nature), this in turn being closely associated with *die Here* (the Lord).

Beyond all other things, *kruiedokters* are skilled at, and known for, their ability to deal with the various kinds of *wind* [wind] that cause so much trouble in people's body-person. There are, generally speaking, two kinds of *wind*: *natuurlike* [natural] and *toor* [magic], both of which retain the propensity for movement and change, the same as the wind that blows in the veld. The first kind of *wind* can be a 'simple' build-up of gas in the intestines and stomach, but can also be 'picked up' in the world and 'lodge' itself in any part of the body – from the head to the muscles of the leg. Health and well-being are often linked internally, to things flowing properly around the body through organs which are not 'stressed'; and externally, to people being able to move along the proper 'path' of life' as laid out for them by God. Therefore, in order to return their patients to these flows, *kruiedokters* use various techniques to unblock these winds, including plants (*kougoed* (*Sceletium tortuosum*) is common) and other substances to aid the expulsion of wind, and massage.

Knowledge of the second kind of *wind*, *toorwind*, is an important area of *kruiedokters'* expertise. These are winds that are 'sent' by malicious others, through a range of media including *gif* [poison] placed in someone's food, in dreams, or through magical 'traps' placed along their victim's path. Such winds can grab on very tightly to their sufferer in the form of a *bose gees* [evil spirit] that 'sits' in the victim's body, growing, sapping their '*liggaam se krag*' [body's energy]. *Kruiedokters* exercise what is seen as a God-given, evil-tackling talent, to cleanse their patients of such winds. They use 'fighting roots', that is the roots of the various 'storm' plants in protective *xaimpies* [medicine bundles] placed in the home or carried around on the person. *Kougoed* encourages the movement of wind out of the body. Existing in what might be called an intersubjective space between a *kwaadaandoener* [evil sender] and their victim, the treatment of these kinds of winds involves the *kruiedokter* effectively placing themselves, as a kind of defender, between the attacker and the attacked.[37-39]

When writing or talking about things like *wind*, or especially *toor*, the question of proof or evidence arises. Rather than getting bogged down in disputes about reality or unreality, Cohen, following the arguments of Viveiros de Castro[40] and Holbraad[41], viewed *wind* and *krag* as effective 'concept phenomena' in their own right. Based in both the experience of phenomena of the world, and their mutually constituting, socially historically generated conceptualisation, such concept phenomena facilitate thinking about humans as 'ecological' beings, that is, as intimately bound both to human and non-human aspects of their environments in ways that do not necessarily conform to the convention of nature or culture.

As conversations progressed, it became clear that one person, Marianna Lot, had been integral to all of the studies. She worked as a field assistant to all of the studies, and had valuable commentary on how they made sense to her personally and to wider debates among people in Paulshoek. Her comments in an interview with Cohen (see below) speak to the ways in which science becomes part of lived experience of the world, and underscore the pleasure and value of being able to fit parts into wholes, and relate scholarship back to publics.

Lot was born on a farm in Bushmanland where her grandfather and father worked tending herds of sheep and goats for a private landowner and her mother worked as a domestic worker. When the children reached school-going age, Lot's family moved to Paulshoek, which was the closest village where they were allowed to settle under apartheid legislation. She completed her primary education in Paulshoek and went to Leliefontein for her high-school years. Thereafter she worked in local towns such as Garies and Vredendal as a shop assistant, and married and had children during this period. She returned to Paulshoek in 1996 soon after her father, Joseph Nero, passed away. Her love for plants and interest in the veld had been kindled by her father who was a well-known *kruiedokter* in the region. People from the village and from as far afield as Springbok and even Cape Town and Namibia were regular visitors to his practice. Soon after returning to Paulshoek, Lot joined the Community Development Forum and started working on a range of community-related projects in the village. When the German-funded BIOTA research programme offered to train eight people from Namaqualand and Namibia as ecological field assistants, Lot applied and was selected for the course. This 'paraecologists' programme, under the leadership of Dr Ute Schmiedel from Hamburg University, provided her with employment for the next 7 years and gave her just the right skill set to provide the necessary high-quality support required by the UCT-based research programme on medicinal plants.

In the paraphrased and edited narrative which follows, Marianna Lot, Paulshoek resident and research assistant to all of the researchers, describes her encounters with different aspects of the project:

> I began with BIOTA in 2004 and I worked for the organisation for seven years. After my work with BIOTA finished, Timm walked with me along the path of learning the processes of working with various students on their projects. I first worked with Nicola on medicinal plants. Her household surveys investigated the number of kruiedokters here in Paulshoek and their roles. Her surveys were also about the difference between a kruiedokter and a healer, and how traditional plant use had changed since 2002 when an earlier survey was undertaken in the village. I went from house to house, sat with people and asked them what kinds of plants they used and for what purposes.
>
> Then came Amelia who worked on the plants that animals do not really eat and like, especially the kraalbos (G. africana) which is very common and used by people as a medicine. It makes other animals sick but people healthy. Amelia collected Galenia to see what precisely is inside this plant. I then met Joshua who came looking for people who are kruiedokters. Unfortunately, in Paulshoek there was only Gert Julk and Joshua wanted to get advice from different doctors. I introduced him to Koos, who is a kruiedokter and asked him if he would like to be part of the project or if he would help Joshua and give him a little advice. Joshua and Koos now have a very good understanding between themselves and work well together. Koos has always been open and shares things with us. He is always ready to help if you need information or if you want to know something.

Holding the material together

Developing a conversation among the disciplines was assisted by academics who were not engaged in Paulshoek research. In 2012 Helen Verran, an Australian philosopher, was a visiting scholar in Cape Town. She enthusiastically entered the discussion and drew on her prizewinning book, *Science and an African Logic*[42], to invite new ways of thinking. She imagined our research as disparate reports of something, which in the beginning, we did not have terms to describe. She pointed to ways of thinking about the different approaches, both as different traditions/histories of knowledge and as a set of questions. These informed dialogue around disciplinary traditions and directed discussion to different parts and wholes. Verran elaborates:

> The project is in part about finding ways to bring in other knowledge traditions: Marianna's phenomenological place-based knowledge, and the kruiedokter's esoteric/arcane doing of plants and particular specified parts of human existence as he understands them. In trying to include their perspectives, we would be trying to reverse in a small way the epistemic erasures that have been part and parcel of the expansion of the academy. Of course, the destruction of precolonial knowledge systems is still a raw wound in both South Africa and Australia.
>
> Whole-parts generalising, which situates rather than abstracts, seems to me to attend to two features of this work that I have been worrying at. The first concerns us as academics. We want to tell our differences, but what could connect our texts? It seems that imagining our projects as each concerned with some emerging part of a vague whole form relating to particularly placed human relations with plants (resisting specificity when it comes to our so-called 'research question'), might be a useful way to begin to connect the projects.
>
> It also seems to be a way to 'bite the bullet' in analytically reading the spoken texts of Marianna and hopefully, at least one *kruiedokter*. We need to do this reading with the same 'disciplining form' we apply to ourselves, so this notion of vague whole and emergent parts is a minimalist standardising form that our collaboration works through. If we envisage our problem through whole-parts generalising, we can imagine it as a version of an old story that Stengers[43] reminds us of:
>
> The famous tale of the three blind men and the elephant, one man recognising a trunk, the second a snake, and the third a fly swatter...The blind men all investigate the elephant...but the diverging ways in which they characterise it appear as an end point [of the story]. The divergence is not a matter of crucial concern to them. If it had been such, the story would not end when the blind men make their first contradictory assessments; they would next move around the elephant to explore the possibility of a coherent account that could turn outright contradictions into very interesting contrasted standpoints. In other words, the blind men would have lent themselves and their respective interpretations to active comparison, giving that which they all address the power to impose 'due attention'.[43]

Following a series of meetings and day-long workshops, and with Verran's guidance, the form of this paper gradually emerged – laying out stories of encountering the plants and landscape as part of the disciplinary-focused interventions in people's own voices. As such, the narrative of encounter with realities that were unexpected in our disciplinary training, could give form to a fresh set of insights about what it was that we were seeing, why that should be so, and how to theorise in different ways the process of knowledge production in which we were all engaged.

Robert Morrell, a social historian and gender sociologist, has been involved in the Paulshoek project since 2010. Based in the UCT Research Office, his brief was to stimulate and support transdisciplinary, Africa-centred research. Morrell hosted and facilitated workshops and meetings with the team. He increasingly took up the role of interlocutor, finding ways to link the researchers to one another and build a common purpose. His contribution was a mix of intellectual and collegial, constantly emphasising the importance of respecting different academic traditions while committing the team collaboratively to a project of knowledge production. New knowledge is not simply something that institutions produce: it is something that people generate in dialogue. The value of nurturing collegiality is not something that, in the age of managerialist approaches to research output, institutions easily see or count. Research funding does not generate new knowledge, people do. Morrell describes his involvement as follows:

> I came to the University of Cape Town after nearly 30 years of lecturing at universities elsewhere in South Africa. During my time as a lecturer, I researched questions of historical inequality and entrenched patterns of violence in South Africa. More recently, I developed my work across disciplines, working with epidemiologists, historians, psychologists, philosophers and sociologists to understand the gendered nature

> of violence. The end of apartheid in 1994 led to calls for 'transformation'. A grant to UCT from the Carnegie Corporation of New York was devoted to transforming the ways in which institutional knowledge was created. The Africa Knowledge Project emerged as a vehicle for bringing together and funding researchers who were interrogating 'Western' knowledge models and assumptions and searching for Africa-centred knowledges.[44] Drawing on diverse theories,[45,46] the project argued for the existence of multiple knowledges with starting points that originated in the lived experiences of the continent's peoples and in its epistemological erasure. This project was part of a move to problematise inequalities in the global knowledge economy and to develop Southern Theory.[47]
>
> The search for Africa-centred knowledges was strengthened by another wave of enquiry to promote interdisciplinary and transdisciplinary work.[48,49] Noting that many problems in the third millennium are complex and cannot be answered with one disciplinary toolkit, Max-Neef commented that transdisciplinarity represented an 'unfinished scientific programme that offers fascinating possibilities for advanced reflection and research'.[50]
>
> To achieve these goals, I sought to create epistemic hospitality. To use Francis Nyamnjoh's expression, my goal was to produce conviviality.[51] The challenge was to allow things to unfold, to allow vague questions to metamorphose into deeper lines of collective investigation.

Gammon framed concerns about transdisciplinarity from a location within current debates in ethnopharmacology. He tabled a complex, difficult and vital set of provocations which follow below. In brief: while natural products chemistry and botany have a great deal to converse about, what does anthropology bring to the table? What relationship does natural products chemistry have to the people who hold the knowledge of plant uses, and can that relationship impact on the way natural products chemists think about their work? The question went to the heart of current debates about the relationship between sciences and indigneous knowledge in South Africa and elsewhere, and it resonated powerfully with current debates in the philosophy and anthropology of science regarding the necessity of being able to think about scientific knowledge as a product of society, without rendering it useless by asserting that it was just a product of vested interests.

The renaissance of natural products chemistry has been associated with technological developments that make possible the analysis of ever smaller quantities of plant (or other) natural products. It has also drawn on other emerging technologies in molecular biology and informatics that have synergistically combined to usher in the 'omics' era and foreshadow an improved capacity to understand organisms in their environment.[52,53] Genomics, proteomics and metabolomics (or metabonomics) are approaches focused on grasping at the totality of the system, by evaluating or mapping the collection of genes, proteins (enzymes) or metabolites, respectively.[54] Recent advances in chromatography and spectroscopic techniques such as high-resolution mass spectrometry and nuclear magnetic resonance, and particularly the combined versions of these such as liquid chromatography–mass spectrometry, allow for quite comprehensive 'fingerprinting' of the array of constituents in an extract, even if a full description of all of these techniques is still beyond reach.[55-57] Coupling of these with sophisticated assay technologies and more specific biological assays,[58] either whole-cell or target-based, suggests new research questions and a realistic situating of natural products research within the more holistic paradigm of systems biology. The new techniques and approaches do not necessarily call into question the more classical approach of painstaking separation and characterisation of constituents, but they have the potential to significantly enhance the process of de-replication of mixtures – a process that follows preliminary screening which searches for new pharmacologically active substances. They also suggest more efficient approaches to searches for active constituents and studies of functions inherent in mixtures. However, both the advances in the science and the warnings and injunctions of various practitioners give pause for thought and raise questions. What is an appropriate and responsible way to proceed with research involving the chemistry of natural products? How does one combine scientific integrity with ecological and cultural sensitivity? Do the recent technological advances draw us further away from, or nearer to, an understanding of different ways of understanding the world and constructing knowledge? In the study of plant natural products, how exactly does one take into account the ecological and community context of the plants and their uses, and does knowledge of and sensitivity towards these go beyond simply providing context? Conversely, what does a detailed understanding of the molecular composition of plants add to ecological or local knowledges and practices of plants and their place in the world? How is knowledge discovered, constructed, crafted? And are these ways mutually exclusive, or are there patterns of thought and practice inherent in our common humanity, independent of educational history and cultural bias?

For anthropologists Lesley Green and Joshua Cohen, what was important in thinking about 'indigenous knowledge research' was the need to resist reducing plants to a pharmacologically active ingredient. Anthropological work, including Cohen's on *krag* and *wind* reflects an effort to understand a different basis for thinking about health.[59] There are thus attempts in many disciplines to move beyond methodological and empirical stricture and the 'valid or invalid' binary that characterises much disciplinary endeavour. Yet the question is how do we have the conversation and begin to pull the threads together? Gammon's concerns compel us to avoid a '*kumbaya*' approach to transdisciplinarity, a cosmopolitan celebration of the wonders of disciplinary diversity in which the 'social science' is a quaint add-on to the science. A social science that simply matches the science would sell short the value of the humanities.

The primary orienting, although often unstated, question in chemical studies of plant medicine concerns pharmacologically active ingredients for antibacterial, antifungal, antiviral and/or anti-inflammatory properties. Yet this orientation depends on equating health and illness with the eradication of a particular taxonomy of pathogens. If the orientation to health includes a wider array of toxins and taxonomies that contribute to the experience of having energy or vitality (*krag*[33] or the different, although not entirely dissimilar concept of 'qi'[59]), – then biochemical research need not necessarily begin with the particular pathway of seeking compounds related to pathogen elimination. Cohen's work suggests the examination of pharma *and* the flow of energy in the body. This suggestion raises questions about plants in an ethnopharmacology of Namaqualand, which would include cleansing, balance, attention to social harms and the toxicities of stress (massage).

Going-on together doing difference

We now discuss the project as a form of whole-parts generalising, which recognises that wholes, or 'bigger pictures', always remain vague. We suggest that such knowledge-making ends up remaking, situating and locating. Any such situating is a call for further efforts, as new puzzles emerge and begin to help clarify what the vague whole at the project's core might become. Verran earlier referred to the elephant and how Stengers mobilised the tale of the difficulties blind men experience in seeking to know an elephant through what their fingers and hands perceive. The work of thinking about how we understand the 'whole' that is our larger area of interest on the basis of the parts that we grasp via our disciplines, does not end with simply presenting divergent discoveries. 'Going on thinking together', we used the divergent findings as provocations for further puzzling. Active comparison arose in allowing divergences to provoke questions of *how* they arose. The Paulshoek medicinal plant research can help attend to the question of 'whether transdisciplinary, Africa-centred knowledge production is worth the trouble it brings'. The question raises two intertwining issues. Does transdisciplinary Africa-centred research offer insight into the problem of epistemic erasures associated with science's past central role in colonising projects? Does it enable us to attend to the challenge which claims scientific knowledge-making as merely the product of vested interests?

These questions can best be approached with the idea of sciences as meaning-making machines, suggesting scientific knowledge is culturally active. This takes us in a different direction than thinking of science's epistemic practices as making truth claims, as revealing 'the true structure of reality'. This is it not a contradiction of the foundationalist way of thinking about knowledge-making. It is possible for a collective to hold simultaneously to both ways of thinking about knowledge – as we do in writing this paper together.

To think of the practices that constitute scientific disciplines as 'a machine' recognises that in the past much careful work has gone into purposefully designing and perfecting that set of practices. Such 'machines' generate objects of knowledge in meeting the requirements and obligations that come with that science.[60,61] Tension in transdisciplinary and cross-cultural research arises precisely because the objects of knowledge generated in different disciplines, by different meaning-making machines, are likely to clash and interrupt. Resisting the impulse to compromise, in good research each discipline (or knowledge tradition) stays explicitly faithful to its objects, thus working in good faith both within and without the discipline. Each team refuses other teams' meanings ('Our meanings of plant are not the same as your meanings of plant.). Thus rather than pretend we are going on together in good will when actually we refuse to compromise over meanings, we acknowledge that we display bad will towards others' meanings. Good transdisciplinary research, like good cross-cultural research, requires explicit good faith and bad will.

When the sciences are considered as 'meaning-making machines', the orthodox, absolutist view that objects are either found or made is left aside. Yet the framing also recognises (and can work with) the actuality that most natural sciences work (and work effectively if often unreflectively) with the assumption that objects are *found*, and most social sciences work with the assumption that their objects are *made*. For example, antibacterial compounds are taken as given in the world, and natural products chemistry *finds* them in ingenious ways. The concept of objects of governance which we introduce below, suggests that in the social sciences, by contrast, various objects, both physical and abstract, become significant when specific actors, recognising the origins of that significance in social activities, apprehend the social roles of these objects.

From a different perspective, one that focuses on processes of knowing, Catherine Elgin argues for 'the ineliminable cognitive contributions of non-literal, non-descriptive symbols':

> *Cognitive advancement is not always a matter of learning something new. We have a vast store of information at our disposal already. Often our problem is what to make of what we've got. This is true even at the level of perception. To a large extent, looking involves overlooking; listening involves discriminating between signal and noise. So a critical epistemological question is: What is worthy of notice? What should be overlooked, marginalised, or ignored? Ordinarily, answers to these questions are simply presupposed. We seldom notice that we notice some things and overlook others. We automatically invoke routine categories to describe or represent phenomena. We adopt familiar orientations and judge by received standards.*[62]

What good transdisciplinary research brings to the fore is the importance of asking about how we know, because different disciplines know in often very different ways.

The objects of knowledge in both phenology and natural products chemistry equally comply with the standards of the so-called scientific method. It is in the particular obligations imposed on scientists within these disciplines by their scientific objects that practices will be experienced as different. Sometimes in meeting the specific obligations that are required to bring these disparate disciplinary objects of knowledge to life and keep them alive, scientists will find themselves needing to disagree with each other, as they struggle to work together. The need to disagree is felt even more strongly when one is struggling to go on together with a practitioner of a disparate knowledge tradition – there we are likely to experience the need to disagree in order to meet the requirements our objects of knowledge impose on our knowledge practices, as much as the specific obligations imposed in a particular situation.

What are the differing objects of knowledge of phenology and natural products chemistry? We follow Stengers[63] and begin with Galileo: he interrogated nature in a mathematical language in his experiments on falling bodies. For him the formalisms of mathematics were the only tool up to that task. The obligation that physicists still feel, to interrogate the world through mathematical formulations, continues to enact this obligation imposed on physicists by the objects known by physics, an obligation that still alienates many beginning science students. Obliged as they are 'to discipline a jungle of diverse molecules…identifying, naming, and classifying on the basis of [elaborate and standardised] tests', chemists' major obligations seem to be caution and scepticism.[61]

By contrast again, plant phenologists observe plants in place, which is quite different to the objects generated in physics' experiments on falling bodies articulated in formal mathematical language, or those generated in chemistry's testing of substances by subjecting them to fire or its equivalent. Phenology requires constancy in observation, and precision in temporal and spatial co-ordination of that observation: going every month to a singular place, walking the same route, observing the state of individual plants. The phenological object of knowledge imposes itself on the life of the scientist, its life partner. 'Phenomena of observable patterns' is a uniquely demanding object of knowledge.

In being both strongly felt and rigorously enacted, these different obligations to disparate objects of knowledge make difference obvious and respected in good transdisciplinary research. The process of writing this paper has led us to focus on the nitty-gritty of separating before considering connecting. In this way the requirements and obligations of the object of knowledge that emerges in a disciplined intersection of phenology and natural products chemistry has been usefully articulated. It might be thought of as an offspring of the two parent objects of knowledge.

The encounter between the three disciplines allowed a fresh appreciation of the differing specific objects of knowledge, and the larger object to which they each contribute. No longer divided into disciplinary-specific components, our transdisciplinary journey brought into the appreciation a new (although still vague) whole. It is as though, in taking care to become familiar with *how* we know, the medicinal plants forced us into a relationship with one another that allowed for this fresh insight of what they are.

Recognising and respecting the alternative objects of knowledge that emerge in transdisciplinary research has a further benefit. It can alert participants to the distinction between objects of knowledge, constituted in epistemic practices, and objects of governance, constituted in practices promoting organisational accountabilities. That distinction is crucial in identifying where and how knowledge-making and governance overlap.

Good transdisciplinary research calls forth a self-consciousness about epistemic practice. This easily extends to a self-consciousness about, and sensitivity towards, the distinctions between objects of knowledge and objects of governance, emergent respectively in epistemic and organisational governance practices. The latter might be thought of as the means of operationalising objects of knowledge. Clearly the objects of knowledge generated in *kruiedokters'* practices have very different requirements and obligations from those of either Western bio-medicine or Chinese traditional medicine. However, just as clearly, the governance practices which operationalise those objects in those differing healing traditions differ; in being operationalised the objects gain different sorts of properties.

In the story that Marianna Lot tells of her involvement in the Paulshoek ABC project, it becomes possible to understand how she has learned to work with people (both scientists and *kruiedokters*) who in their everyday lives must negotiate the requirements and obligations of the objects of knowledge in which they are involved. Lot does not obligate herself directly with those objects, only indirectly through other people. She learned from that indirect involvement to have respect for those objects of knowledge. In her story, she alerts researchers from different disciplines to what is involved in operationalising objects of knowledge, in re-constituting them as objects of governance. In being operationalised, at least one of the practices involves objects of knowledge-accreting stories. Which stories, who can tell them, and when they can be told and to whom, are all expressions of politics. How these are decided addresses the issue of erasure with which this paper began and implications for how postcolonial knowledge is produced.

Learning 'to do our differences' explicitly but together through attending to the issue of *how* we know, we learned to respect medicinal plants as differentially knowable. Having learned to tell each other *how* we know, we now recognise that the next step is to learn to do differing forms of governance together through working with those who know and govern otherwise.

Acknowledgements

This work was supported by the University of Cape Town's Vice Chancellor's Strategic Research Fund and the Programme for the Enhancement of Research Capacity (PERC). Marianna Lot's ongoing work in Paulshoek is gratefully acknowledged by the research team.

Authors' contributions

L.G. oversaw the social anthropology component of the paper, provided theoretical inspiration, knowledge of comparative ontological literature, participated in writing meetings and wrote and edited parts of the text. D.W.G. was the chemistry project leader, received two UCT grants to initiate and facilitate the transdisciplinary project, hosted project and writing meetings and wrote parts of the text. M.T.H. was the botany project leader, hosted project and writing meetings and wrote parts of the text. J.C. drew on his anthropological field research, attended and contributed to writing meetings and wrote part of the text. N.W. and A.H. drew on their field and laboratory research, attended and contributed to writing meetings and wrote parts of the text. R.M. oversaw the transdisciplinary process as a whole, facilitated the meetings, wrote some of the text and edited the entire piece. H.V. attended some of the writing meetings and provided theoretical sophistication to the project and wrote some of the text.

References

1. Verran H. Governance and land management fires effected in engagement between environmental scientists and Aboriginal land owners. Understanding objects of governance as expressing an ethics of dissensus. Learn Communities J. 2014;15:53–64.
2. Hadorn GH, Biber-Klemm S, Grossenbacher-Mansuy W, Hoffmann-Riem H, Joyce D, Wiesmann U, et al. The emergence of transdisciplinarity as a form of research. In: Hadorn GH, Hoffmann-Riem H, Biber-Klemm S, Grossenbacher-Mansuy W, Joye D, Pohl C, et al., editors. Handbook of transdisciplinary research. Dordrecht: Springer; 2008. p. 19–39. http://dx.doi.org/10.1007/978-1-4020-6699-3_2
3. Hubert B, Meuret M, Bonnemaire J. Shepherds, sheep and forest fires: A reconception of grazing land management. In: Hadorn GH, Hoffmann-Riem H, Biber-Klemm S, Grossenbacher-Mansuy W, Joye D, Pohl C, et al., editors. Handbook of transdisciplinary research. Dordrecht: Springer; 2008. p. 103–126.
4. Dubow S. Scientific racism in modern South Africa. Cambridge: University of Cambridge Press; 1995.
5. Fassin D. When bodies remember: Experiences and politics of AIDS in South Africa. Berkeley, CA: University of California Press; 2007.
6. Nattrass N. The moral economy of AIDS in South Africa. Cambridge: Cambridge University Press; 2004.
7. Nattrass N. Mortal combat: AIDS denialism and the struggle for antiretrovirals in South Africa. Pietermaritzburg: University of KwaZulu-Natal Press; 2007.
8. Cullinan K, Thom A, editors. The virus, vitamins and vegetables. The South African HIV/AIDS mystery. Johannesburg: Jacana; 2009.
9. Green L. Beyond South Africa's 'indigenous knowledge – science' wars. S Afr J Sci. 2012;108(7/8):34–43. http://dx.doi.org/10.4102/sajs.v108i7/8.631
10. Verran H, Christie M. Doing difference together: Towards a dialogue with aboriginal knowledge authorities through an Australian comparative philosophical inquiry. Cult Dialogue. 2011;1(2):21–36.
11. Rohde RF, Hoffman MT. One hundred years of separation: The historical ecology of a South African 'coloured reserve'. Africa. 2008;78(2):189–222. http://dx.doi.org/10.3366/E0001972008000132
12. Green LF, Green DR. Knowing the day, knowing the world: Engaging Amerindian thought in public archaeology. Tucson, AZ: University of Arizona Press; 2013.
13. Latour B, Weibel P. Making things public. Atmospheres of democracy. Cambridge, MA: MIT Press; 2005.
14. Latour B. An Inquiry into modes of existence: An anthropology of the moderns. Cambridge, MA: Harvard University Press; 2013.
15. Esler KJ, Rundel PW. Comparative patterns of phenology and growth form diversity in two winter rainfall deserts: The Succulent Karoo and Mojave Desert ecosystems. Plant Ecol. 1999;142:97–104. http://dx.doi.org/10.1023/A:1009830513525
16. Struck M. Flowers and their insect visitors in the arid winter rainfall region of southern Africa: Observations on permanent plots. Insect visitation behaviour. J Arid Environ. 1994;28(1):51–74. http://dx.doi.org/10.1016/S0140-1963(05)80021-7
17. Croteau R, Kutchan TM, Lewis NG. Natural products (secondary metabolites). In: Buchanan B, Gruissem W, Jones R, editors. Biochemistry & molecular biology of plants. Rockville, MD: American Society of Plant Physiologists; 2000.
18. Cordell GA, Colvard MD. Some thoughts on the future of ethnopharmacology. J Ethnopharmacol. 2005;100(1–2):5–14. http://dx.doi.org/10.1016/j.jep.2005.05.027
19. Ngo LT, Okogun JI, Folk WR. 21st Century natural product research and drug development and traditional medicines. Nat Prod Rep. 2014;30(4):584–592. http://dx.doi.org/10.1039/c3np20120a
20. Cordell GA, Colvard MD. Natural products in a world out-of-balance. Arch Org Chem. 2007;7:97–115.
21. Cordell GA. Phytochemistry and traditional medicine – A revolution in process. Phytochem Lett. 2011;4(4):391–398. http://dx.doi.org/10.1016/j.phytol.2011.05.005
22. Cordell GA, Colvard MD. Natural products and traditional medicine: Turning on a paradigm. J Nat Prod. 2012;75(3):514–525. http://dx.doi.org/10.1021/np200803m
23. Etkin NL, Elisabetsky E. Seeking a transdisciplinary and culturally germane science: The future of ethnopharmacology. J Ethnopharmacol. 2005;100(1–2):23–26. http://dx.doi.org/10.1016/j.jep.2005.05.025
24. Andrae-Marobela K, Ntumy AN, Mokobela M, Dube M, Sosome A, Muzila M, et al. 'Now I Heal with Pride'– The application of screens-to-nature technology to indigenous knowledge systems research in Botswana: Implications for drug discovery. In: Chibale K, Davies-Coleman M, Masimirembwa C, editors. Drug discovery in Africa, impacts of genomics, natural products, traditional medicines, insights into medicinal chemistry, and technology platforms in pursuit of new drugs. Heidelberg: Springer; 2012. p. 239–264. http://dx.doi.org/10.1007/978-3-642-28175-4_10
25. Wheat N. An ethnobotanical, phytochemical and metabolomics investigation of plants from the Paulshoek communal area, Namaqualand [PhD thesis]. Cape Town: University of Cape Town; 2014.
26. Verpoorte R, Choi YH, Kim HK. Ethnopharmacology and systems biology: A perfect holistic match. J Ethnopharmacol. 2005;100:53–56. http://dx.doi.org/10.1016/j.jep.2005.05.033
27. Martin GJ. Ethnobotany, conservation and community development. In: Walters M, editor. Ethnobotany: A methods manual. 1st ed. Kew: Chapman and Hall; 1995. p. 223–251. http://dx.doi.org/10.1007/978-1-4615-2496-0_8

28. Gottlieb OR, Borin MRDMB, De Brito NRS. Integration of ethnobotany and phytochemistry: Dream or reality? Phytochemistry. 2002;60(2):145–152. http://dx.doi.org/10.1016/S0031-9422(02)00088-2

29. Sumner LW, Mendes P, Dixon RA. Plant metabolomics: Large-scale phytochemistry in the functional genomics era. Phytochemistry. 2003;62(6):817–836. http://dx.doi.org/10.1016/S0031-9422(02)00708-2

30. Hilgart AA. Metabolomic barcodes from a systems biology-based approach to chemotaxonomy in taxonomically problematic Aizoaceae species and their implications for understanding phytochemical and biological diversity in South Africa [PhD thesis]. Cape Town: University of Cape Town; 2015.

31. Kellerman TS, Coetzer JAW, Naude TW. Plant poisonings and mycotoxicoses of livestock in southern Africa. Cape Town: Oxford University Press; 1988.

32. Allsopp N. Effects of grazing and cultivation on soil patterns and processes in the Paulshoek area of Namaqualand. Plant Ecol. 1999;142(1):179–187. http://dx.doi.org/10.1023/A:1009826412617

33. Cohen J. *Kruiedokters*, plants and molecules: Relations of power, wind, spirit, and atomic mass in Namaqualand [PhD thesis]. Cape Town: University of Cape Town; 2015.

34. Low C. Khoisan healing: Understandings, ideas and practices. Oxford: University of Oxford; 2004.

35. Low C. Khoisan medicine in history and practice. Voßen R, editor. Cologne: Rüdiger Köppe Verlag; 2008.

36. Low C. Birds and Khoesān: Linking spirits and healing with day-to-day life. Africa. 2011;81(2):295–313. http://dx.doi.org/10.1017/S0001972011000027

37. Favret SJ. Deadly words: Witchcraft in the Bocage. Cambridge, MA: Cambridge University Press; 1980.

38. Stoller, P. Fusion of the worlds. An ethnography of possession among the Songhay of Niger. Chicago, IL: University of Chicago Press; 1989. http://dx.doi.org/10.7208/chicago/9780226775494.001.0001

39. Whitehead NL, Wright R. In darkness and secrecy: The anthropology of assault sorcery and witchcraft in Amazonia. J Altern Emergent Relig. 2010;13(4):133–134.

40. Castro EVDE. Cosmological deixis and Amerindian perspectivism. J R Anthropol Inst. 1998;4(3):469–488. http://dx.doi.org/10.2307/3034157

41. Holbraad M. Truth beyond doubt. Ifá oracles in Havana. HAU: Journal of Ethnographic Theory. 2012;2(1):81–109. http://dx.doi.org/10.14318/hau2.1.006

42. Verran H. Science and an African logic. Chicago, IL: University of Chicago Press; 2001.

43. Stengers I. Comparison as a matter of concern. Common Knowl. 2011;17(1):53. http://dx.doi.org/10.1215/0961754X-2010-035

44. Cooper B, Morrell R, editors. Africa-centred knowledges: Crossing fields and worlds. Oxford: James Currey; 2014.

45. Appadurai A. Modernity at large: Cultural dimensions of globalization. St. Paul, MN: University of Minnesota Press; 1996.

46. Rottenburg R. Far-fetched facts: A parable of development aid – Inside technology. Cambridge, MA: MIT Press; 2009.

47. Connell R. Southern theory: The global dynamics of knowledge in social science. Cambridge: Polity; 2007.

48. Klein JT. Interdisciplinarity: History, theory, and practice. Detroit, MI: Wayne State University Press; 1990.

49. Kessel F, Rosenfield P, Anderson N, editors. Expanding the boundaries of health and social science: Case studies of interdisciplinary innovation. New York: Oxford University Press; 2003.

50. Max-Neef MA. Foundations of transdisciplinarity. Ecol Econ. 2005;53(1):5–16. http://dx.doi.org/10.1016/j.ecolecon.2005.01.014

51. Nyamnjoh FB. Perspectives on Africa. Identity Cult Polit. 2004;5:37–59.

52. Sheridan H, Krenn L, Jiang R, Sutherland I, Ignatova S, Marmann A, et al. The potential of metabolic fingerprinting as a tool for the modernisation of TCM preparations. J Ethnopharmacol. 2012;140(3):482–491. http://dx.doi.org/10.1016/j.jep.2012.01.050

53. Hufsky F, Scheubert K, Böcker S. New kids on the block: Novel informatics methods for natural product discovery. Nat Prod Rep. 2014;31(6):807–817. http://dx.doi.org/10.1039/c3np70101h

54. Buriani A, Garcia-bermejo ML, Bosisio E, Xu Q, Li H, Dong X, et al. Omic techniques in systems biology approaches to traditional Chinese medicine research: Present and future. J Ethnopharmacol. 2012;140(3):535–544. http://dx.doi.org/10.1016/j.jep.2012.01.055

55. Ernst M, Silva DB, Silva RR, Vêncio RZN, Lopes NP. Mass spectrometry in plant metabolomics strategies: From analytical platforms to data acquisition and processing. Nat Prod Rep. 2014;31(6):784–806. http://dx.doi.org/10.1039/c3np70086k

56. Halabalaki M, Vougogiannopoulou K, Mikros E, Skaltsounis AL. Recent advances and new strategies in the NMR-based identification of natural products. Curr Opin Biotechnol. 2014;25:1–7. http://dx.doi.org/10.1016/j.copbio.2013.08.005

57. Simmler C, Napolitano JG, McAlpine JB, Chen S-N, Pauli GF. Universal quantitative NMR analysis of complex natural samples. Curr Opin Biotechnol. 2014;25:51–59. http://dx.doi.org/10.1016/j.copbio.2013.08.004

58. Agarwal A, D'Souza P, Johnson TS, Dethe SM, Chandrasekaran C. Use of in vitro bioassays for assessing botanicals. Curr Opin Biotechnol. 2014;25:39–44. http://dx.doi.org/10.1016/j.copbio.2013.08.010

59. Farquhar JB. Knowledge in translation: Global science, local things. In: Green L, Levine S, editors. Medicine and the politics of knowledge. Cape Town: HSRC Press; 2012. p. 153–170.

60. Stengers I. Cosmopolitics I. Minneapolis, MN: University of Minnesota Press; 2010.

61. Bensaude-Vincent B. Philosophy of chemistry. In: Brenner A, Gayon J, editors. French studies in the philosophy of science: Contemporary research in France. Dordrecht: Springer Science & Business Media; 2009. p. 165–186. http://dx.doi.org/10.1007/978-1-4020-9368-5_7

62. Elgin CZ. Creation as reconfiguration: Art in the advancement of science. Int Stud Philos Sci. 2002;16(1):13–25. http://dx.doi.org/10.1080/02698590120118792

63. Stengers I. The invention of modern science. Minneapolis, MN: University of Minnesota Press; 2000.

Determination of a novel size proxy in comparative morphometrics

AUTHOR:
Andrew Gallagher[1]

AFFILIATION:
[1]Centre for Anthropological Research, University of Johannesburg, Johannesburg, South Africa

CORRESPONDENCE TO:
Andrew Gallagher

EMAIL:
agal1815@gmail.com

POSTAL ADDRESS:
Centre for Anthropological Research, Department of Anthropology and Development Studies, University of Johannesburg, Auckland Park, 2006, Johannesburg

DATES:
Received: 02 July 2014
Revised: 30 Oct. 2014
Accepted: 04 Jan. 2015

KEYWORDS:
mammals; size; multivariate scaling; mammalian palaeobiology; vertebrates

HOW TO CITE:
Gallagher A. Determination of a novel size proxy in comparative morphometrics. S Afr J Sci. 2015;111(9/10), Art. 2014-0221, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20140221

Absolute size is a critical determinant of organismal biology, yet there exists no real consensus as to what particular metric of 'size' is empirically valid in assessments of extinct mammalian taxa. The methodological approach of JE Mosimann has found extensive favour in 'size correction' in comparative morphometrics, but not 'size prediction' in palaeontology and palaeobiology. Analyses of five distinct mammalian data sets confirm that a novel size variate (GMSize) derived from $k=8$ dimensions of the postcranial skeleton effectively satisfies all expectations of the Jolicoeur–Mosimann theorem of univariate and multivariate size. On the basis of strong parametric correlations between the $k=8$ variates and between scores derived from the first principal component and geometric mean size (GMSize) in all series, this novel size variable has considerable utility in comparative vertebrate morphometrics and palaeobiology as an appropriate descriptor of individual size in extant and extinct taxa.

Introduction

Absolute size of an organism, typically encapsulated by its body mass or length, is perhaps the most significant variable in comparative biology.[1-4] Body size is intimately intertwined with organismal physiology, ecology, reproductive and, ultimately, evolutionary success.[3-6] Cope's 'rule' of phyletic size increase is a pervasive phenomenon in the vertebrate fossil record,[7-12] and remains a valid prospectus irrespective of any determinant probability governing directional size increases from a lineal founder of comparably diminutive size relative to its terminal members.[7,13]At the population level, there is perhaps no greater testimony to the significance of individual size than the fact that 'length-to-mass' is a critical metric of the intrinsic health of an individual, from ontogeny through to adulthood.[14-18]

In cases where absolute length or mass of an individual organism cannot be reliably determined, as is generally the default in comparative morphometric analyses of specific skeletal elements and in vertebrate palaeontology, a justified linear proxy for body size is required. The pitfalls of directional covariance between $k=2$ or $k=>2$ metrics of a skeletal or dental element in interspecific 'mouse-to-elephant' allometric models has been emphasised,[19-21] and RJ Smith's analysis of dental size and body mass covariance in primates is a constructive exemplar,[19] yet interspecific approaches to individual body mass estimation remains a cornerstone of mammalian palaeontology and palaeobiological inference.[22-27] As a basic objective, we desire a reliable size proxy for either a single case (an individual fossil) or a series of individuals sampling an unknown or indeterminate underlying size distribution. Taxonomically diagnostic dental elements comprise about 90% of the mammalian fossil record and correlate strongly with body mass in broad interspecific contexts, particularly at the class and order levels,[22,23,28-30] yet yield surprisingly poor estimates compared with weight-bearing skeletal elements in narrower taxonomic comparative analyses.[9,31-33]

An alternative approach to mass estimation involves explicit sampling across the skeleton and the application of multiple-dependent least-squares regression procedures (ordinary least squares) or generalised least squares algorithms (ANCOVA) to assess efficacy of a suite of potential variate predictors of vertebrate size.[34-37] While these methods offer considerable improvement over traditional bivariate Model I and Model II regression techniques, their utility is dependent upon access to reasonably complete and associated comparative series in museum repositories and, in the case of interspecific models, effective taxon-specific samples may be little improved over traditional bivariate approaches. There is general acceptance of the size-adjustment approach advocated by JE Mosimann in comparative morphometrics,[38,39] yet there has been hitherto little recognition of the potential primacy of the favoured size variate, the geometric mean (GM), in estimation of 'size'. Following the work of PF Jolicoeur,[40,41] any preferred construct of individual size from a suite of k linear correlates is testable via decomposition of their variance-covariance matrix (VCV) via principal components analysis. Following Jolicoeur's rationale,[38,40,41] if the first principal component (PC) of a VCV matrix of log-transformed k variates accounts for a majority of the total explained variance (>75%), and all k variate loadings on this vector are approximate, then PC1 represents a generalised multivariate size vector and individual variates may be expressed as simple functions of geometric similarity as follows:

$$X_1, \ldots X_k = (1/k)^{1/2}\,(1, \ldots., 1) \qquad \text{Equation 1}$$

From this, our k linear variates are simply re-scaled as components of isometry with the values $\beta=<1$, $\beta=1$, $\beta=>1$ indicating negative allometry, isometry and positive allometry, respectively.[38,41] If this criterion is satisfied, derivation of the *arithmetic mean* (average) of a suite of log-transformed k linear variates for a single individual is the most appropriate metric of its intrinsic size, equivalent to centroid size of a triangle in Euclidean geometry.[42-44]

In geometric morphometric approaches, the natural intrinsic measure of size derived from any constellation of p landmark coordinates in k-dimensional space ($k=2/k=3$) is centroid size, which is simply the sum of all possible squared inter-landmark distances on a single specimen or series of n specimens.[42-44] However, one critical problem with this size metric is that it is entirely dependent upon the number of p landmark points registered on a specimen (or series of specimens) and can differ radically in any given random sequence of restricted landmark points, as in analyses of specific morphological regions of interest.[42-44] The comparative and evolutionary significance of centroid size is further complicated in analyses where Type I (homologous) registration points do not form an overwhelming majority of the registered constellation of *p* x *k* landmarks and where these may be heavily biased towards Type II and even Type III landmarks.[42-44] In extreme cases of the latter, centroid size retains its function as an intrinsic baseline to which all *p* inter-landmark distances are effectively rescaled, which is the goal in statistical shape analysis.[42-44] Nevertheless, as a comparative size variate, any derivation from the *k* x *p* landmark distance space is an inherent intrinsic function of the specific skeletal element under consideration, and cannot be reified as a faithful proxy of size in broader comparative appraisals. From a theoretical perspective, the only available test for allometry in geometric morphometric applications is a simple test of correlation between the first PC on the VCV of the tangent space coordinates,[42-44] in a direct assessment of correlation of size and shape. Assessments of size correspondences across even anatomically proximate structures using centroid size are simply not possible.

In contrast, derivation of the geometric mean of a series of *k* linear dimensions taken on a single element, or across multiple associated elements of the same specimen, offers significant promise as a generalised comparative size variate in normal metric scales of the SI (μm, mm). The geometric mean is simply the *n*th root of the sum of their products (where $n=k$)[45], and the distribution of this size metric in a population of individuals has been demonstrated to conform to expectations of the univariate log-normal and gamma distributions. More critically, the geometric mean of a series of *k* variates is strongly and positively correlated with the PC1 scores derived from a principal components analysis of the VCV of this series.[38,41] A table of parametric correlation coefficients (Pearson's *r*) is an effective assessment of covariance in a series of *k* variates prior to calculation of the GM.

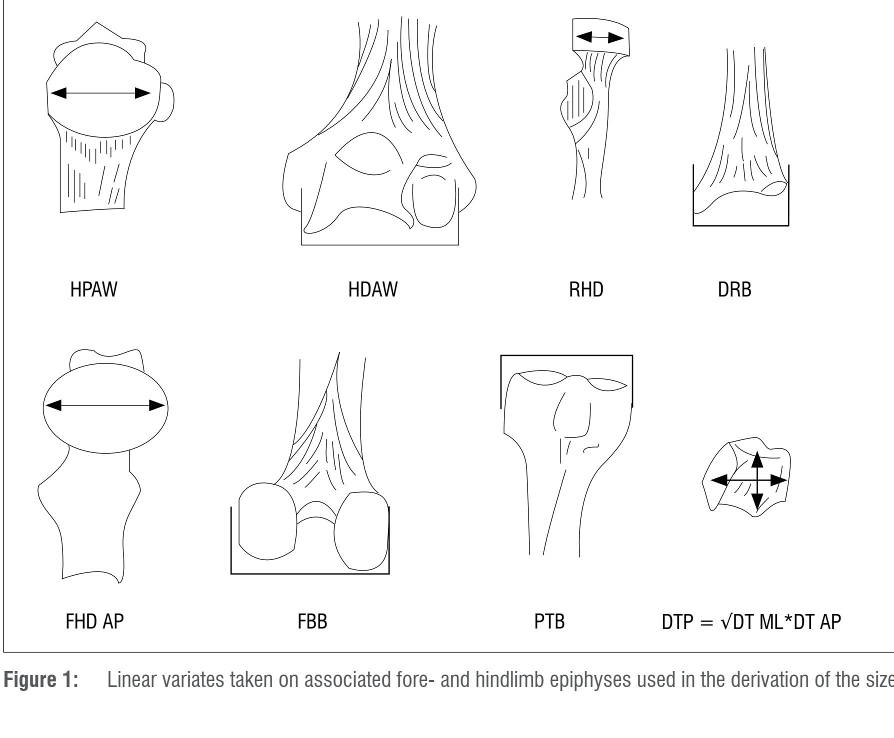
In the event that body length and body mass are unknown, in an individual or a series, an alternative 'proxy' should fulfil the basic prospectus of correspondence with intrinsic organismal size. Weight-bearing epiphyses of the fore- and hindlimb skeleton of living mammals are obvious contenders as individual predictors of body mass in intraspecific and interspecific contexts.[31-34] Nevertheless, such analyses ignore discreet allometric trajectories observed within families, and even between closely related species. The approach favoured here is a global skeletal perspective (Figure 1; Supplementary table 1 online), and follows the size proxy outlined by Reno and colleagues.[46] A series of eight distinct linear dimensions were derived from the proximal and distal epiphyses of the four major long bones in associated individual skeletons. Given that all major weight-bearing epiphyses are sampled, it follows logically that the cumulative proxy of this series, the geometric mean (GMSize), is both intrinsic to an individual and is a faithful approximation of its locus within any hypothetical Guassian normal distribution,[38,40,41,45,46] intraspecifically and at the familial and higher orders of the Linnaean hierarchy. Given a general acceptance of the primacy of postcranial linear variates in the estimation of body mass in extinct mammalian taxa, particularly dimensions of the epiphyses, the GM of $k=8$ linear dimensions of the postcranial epiphyses in associated individual skeletons offers a prospectus for exposition of a generalised size variate in vertebrate morphometrics.[45,46]

Materials and methods

The preferred $k=8$ linear variates of the associated fore- and hindlimb skeletons were taken on a comparative series of extant mammals sampling 247 African hominids (*Gorilla* and *Pan*), 149 Old World monkeys (*Colobus*, *Cercopithecus* and *Papio*) and 62 large-bodied felids (*Panthera* and *Acinonyx*) housed in collections in Africa, Europe and the USA (Supplementary table 2). All data were transformed to natural logarithms (ln), including the GM of the raw series, and parametric correlation matrices (Pearson's *r*) were calculated for these discreet interspecific series. The covariance matrices (VCV) for each of these series were subjected to a principal components analysis and the Eigenvectors, component loadings and PC scores were calculated using PAST version 3.1.[47] In order to assess the efficacy of the proposed size variate at the intraspecific level, pooled-sex series sampling *Pan t. troglodytes* ($n=91$) and *Gorilla g. gorilla* ($n=102$) were assessed. While closely related, these taxa evidence considerable differences in sexual size dimorphism and are sufficiently large to warrant consideration as viable statistical populations.

Results

Correlation coefficients for the $k=8$ fore- and hindlimb dimensions are highly significant across all five data sets and exceed $r=0.92$ in all cases, with the notable exception of *Pn. t. troglodytes* (Supplementary tables 3–7). The poorer correlation coefficients between the linear variate series in common chimpanzees reflects the well-known phenomenon that *centring* any bivariate distribution (*x*,*y*) in a linear regression yields a higher slope in cases in which effective size ranges of *x* and *y* are proportionally large, as in interspecific 'mouse-to-elephant' analyses.[19]



HPAW, mediolateral diameter of the articular surface of the humeral head

DHAW, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum)

RHD, maximum diameter of the radial head

DRB, maximum mediolateral diameter of the distal radial articulation

FHD AP, femoral head diameter (anteroposterior)

FBB, maximum mediolateral diameter of the distal femur

PTB, maximum mediolateral diameter of the tibial articular plateau

DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia

Figure 1: Linear variates taken on associated fore- and hindlimb epiphyses used in the derivation of the size metric (GMSize).

In contrast with *Pn. t. troglodytes*, the considerable linear size range across the variate series observed in *G. g. gorilla* yields coefficients only marginally lower than in the familial Old World monkey and felid data sets (Supplementary tables 3–7). The general consistencies in size correspondences across the $k=8$ fore- and hindlimb joint dimensions in the five data sets is equally supported by the actual proportion of the total variance explained by the first PC across the series (Table 1). Analyses of the pooled-sample African hominids, Old World monkeys, large-bodied felids and *G. g. gorilla* yield a first PC accounting for a staggering 96–98% of the total variance, which is clear confirmation of a dominance of linear size on these axes for these respective series. In contrast, the first PC of the *Pn. t. troglodytes* data set accounts for a considerably depressed percentage of the total variance (76.5%), particularly striking in comparison with *G. g. gorilla*, and is consistent with a scalar decrease in absolute ranges of the $k=8$ variate distributions in this comparably monomorphic taxon. This observed pattern is robust irrespective of whether raw linear data are used in lieu of the log-transformed data (Table 1; Supplementary tables 3–7).

Table 1: Summary statistics for the principal components (PC) analysis

	Natural logarithm			Raw		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
Extant African hominids (*n*=247)						
% Variance	96.35	1.14	0.68	97.47	0.71	0.53
% Cumulative	96.35	97.48	98.76	97.47	98.18	99.14
Extant Old World monkeys (*n*=149)						
% Variance	97.28	0.87	0.64	96.96	1.49	0.69
% Cumulative	97.28	98.14	99.22	96.96	98.45	99.40
Extant Old World felids *(n*=62*)*						
% Variance	97.59	1.35	0.40	98.00	1.01	0.35
% Cumulative	97.59	98.94	99.62	98.00	99.01	99.66
Pan t. troglodytes (n*=91*)						
% Variance	76.48	5.52	4.12	78.94	5.06	4.80
% Cumulative	76.48	82.00	89.76	78.94	84.00	92.31
Gorilla g. gorilla (n*=102*)						
% Variance	94.88	1.71	0.86	95.42	1.44	0.95
% Cumulative	94.88	96.59	98.17	95.42	96.86	98.66

That both Jolicoeur and Mosimann's conditions are met by the size variate preferred here (GMSize) is equally confirmed by data in Figures 2–4. In all five series, there exists a perfect correspondence between the PC1 scores and GMSize ($r=1.00$) for the $k=8$ linear variates of the fore- and hindlimb epiphyses (Figures 2–4). Individual variate loadings on the first PC across the data sets reveals a satisfying consistency within each (Tables 1–6), yet their multivariate isometry coefficients are sufficiently distinct to support family-level and even species-specific allometric scaling trajectories of fore- and hindlimb epiphyseal joints, as revealed in the positive and negative loadings of the various samples on PC2 (Figure 2b). These observed distinctions further caution against the universal efficacy of any 'scaling criterion' derived from interspecific allometric scaling solutions to a single specimen or a series of specimens. Any universal assumption concerning scaling of the proximal femoral articulation in *Pn. t. troglodytes* and *G. g. gorilla* based upon theoretical derivations from pooled-sample analyses of African hominids or Old World monkeys, is not supported by the observation that the proximal femoral articulation scales with negative allometry in these species, as indicated by their pooled-sample multivariate distribution. The proximal femur is actually proportionally smaller in *Gorilla* than in *Pan*. While direct correspondences between multivariate isometry coefficients between the log-transformed and raw linear data series are not possible, it is worth noting that the femoral head loads negatively on the second PC axis of both the log-transformed and raw data series in the pooled African hominid sample, but not in the corresponding tables of the species-specific analyses (Table 1; Supplementary tables 8–12).

Calculation of the Jolicoeur multivariate allometry coefficients in *Pn. t. troglodytes* and *G. g. gorilla* underscores the necessity of sampling all $k=8$ linear variates in the derivation of the preferred size metric, as these taxa also differ in the multivariate scaling of their osseous components of the elbow and knee joints and are not allometrically equivalent (Tables 4 and 5). Observed species-specific or genus-specific allometric scaling constants for any of the $k=8$ variates can be simply tested using conventional post-hoc tests for slopes, y-intercepts and elevations in the bivariate case, yet the observed scalar distinctions in these analyses do not compromise the preferred variate (GMSize) as a valid descriptor of size in comparative contexts. By retaining all $k=8$ linear variates in the analysis, a comparative size proxy is generated which is sufficiently powerful to verify hypotheses of allometric equivalence in the postcranial epiphyses of living and extinct taxa (Figure 2b). On the basis of these data, chimpanzees and gorillas are not allometrically equivalent animals in terms of their relative fore- and hindlimb epiphyseal joint profiles. If we seek to understand the functional and phylogenetic significance of multivariate scaling distinctions in closely allied taxa, then a profitable approach is to assess the significance of shared PCs using common principal components analysis.[48-50]

Darroch and Mosimann[51] have extended the foundations of Jolicoeur's multivariate allometry to canonical component space, subsuming the k-group method of canonical variates analysis.[52,53] Canonical variates analysis is a k-group extension of Fisher's linear discriminant analysis for $k=2$ groups,[52,53] and this extension has both practical and theoretical significance in biological anthropology. Conventional application of a two-sample discriminant function analysis (DFA) in forensic assessment of sex or ancestry[54-58] proceeds from a series of multidimensional ($k=>3$) variates under expectations that the predefined 'sets' sample discreet multivariate universes.[52,53] Nevertheless, substantial overlap exists in observed univariate and multivariate distributions of female and male individuals in all but the most dimorphic mammalian taxa. This observation is confirmed in recent humans by general consistency of classification statistics of about 75–85% in population-specific DFA analyses in sex assessment across the skeletal system and exemplifies the continuous underlying pooled-sample distribution of female and male individuals in multivariate space.[54-58] Following the extension outlined in Darroch and Mosimann[51], if the GM of any suite of k variates is an appropriate descriptor of size, then an equally satisfying correspondence should exist among the total variance explained by PC1, the classification statistics derived using a DFA, and the underlying pooled-sample distributions of the two GM sets.

The *Pn. t. troglodytes* (M=39/F=52) and *G. g. gorilla* (M=56/F=48) series were subjected to DFA based on known sex using an earlier version of PAST (v. 2.3).[47] DFA equations are given in Supplementary table 13 and the correct percentage classifications for *Pan* and *Gorilla* were about 86% and 99%, respectively. The exemplary classification of *Gorilla* is a clear function of the discreet nature of the intraspecific size distribution (bimodal) and is consistent with extreme sexual size dimorphism. Only a single male specimen was incorrectly classified as female. In contrast, the comparably monomorphic *Pan* yields a percentage classification that approximates the upper range of a typical DFA classification in recent humans with 12 specimens incorrectly assigned to their respective sexes. Classification statistics for both raw and log data were equivalent across both samples (Supplementary table 13). Both data sets effectively satisfy expectations based upon canonical components of size and shape.[51] Exploration of pooled-sample distributions of *Pn. t. troglodytes* reveals considerable correspondence between incorrect classifications to the respective sets when individual specimens are expressed as z- and t-scores of sex-specific means and standard deviations, whereas the correspondence in *Gorilla* is perfect (Supplementary tables 14–17). As in conventional DFA of sex assessment in humans, there is a substantially higher incorrect classification of female specimens ($n=8$) than male specimens ($n=4$) in *Pn. t. troglodytes*. The question logically arises as to whether this phenomenon is typical of all monomorphic mammalian taxa, and is certainly worthy of further comparative exploration.

Given that the preferred size variate in this analysis is simply the geometric mean (GMSize) of $k=8$ linear dimensions of the fore- and hindlimb epiphyses, this variate can be reliably constructed from any linear combination of the available series (i.e. $k=<8$). An obvious candidate for redundancy is one of the osseous components of the knee joint (FBB, PTB) (Figure 1; Supplementary table 1), as is one of the elbow joint components (DHAB, RHD), yielding a geometric size variable derived from $k=6$ linear dimensions. As data in Tables 7 and 8 attest, two permutations of GMSize, which reduce the variate series, yield little real improvement to the model (or, alternatively, reduce its efficacy) in terms of the variance explained by the first PC, yet there are subtle distinctions in the loadings of the individual variates on the first and subsequent PCs (Tables 7 and 8). Multivariate allometry coefficients also change subtly, underscoring the observations in Figure 2 and in previous analyses that *Pan* and *Gorilla* are not allometrically equivalent animals. The potential loss of information in more distinct mammalian taxa is graver, as no assumptions of allometric equivalence are made in the entire $k=8$ linear series. Stated simply, the geometric mean of the entire $k=8$ linear dimensions of the fore- and hindlimb epiphyses of the postcranial skeleton retains relevant information pertaining to absolute individual size and equally relevant information about relative joint size, which clearly differs in *Pan* and *Gorilla* and within the large-bodied felids (Figure 2b). On the strength of the correlation coefficients, it is clear that any single variate (such as the proximal femoral articulation) can be used to estimate the preferred size proxy in comparative size appraisals of living and fossil taxa via simple bivariate regression of *x* on *y*.

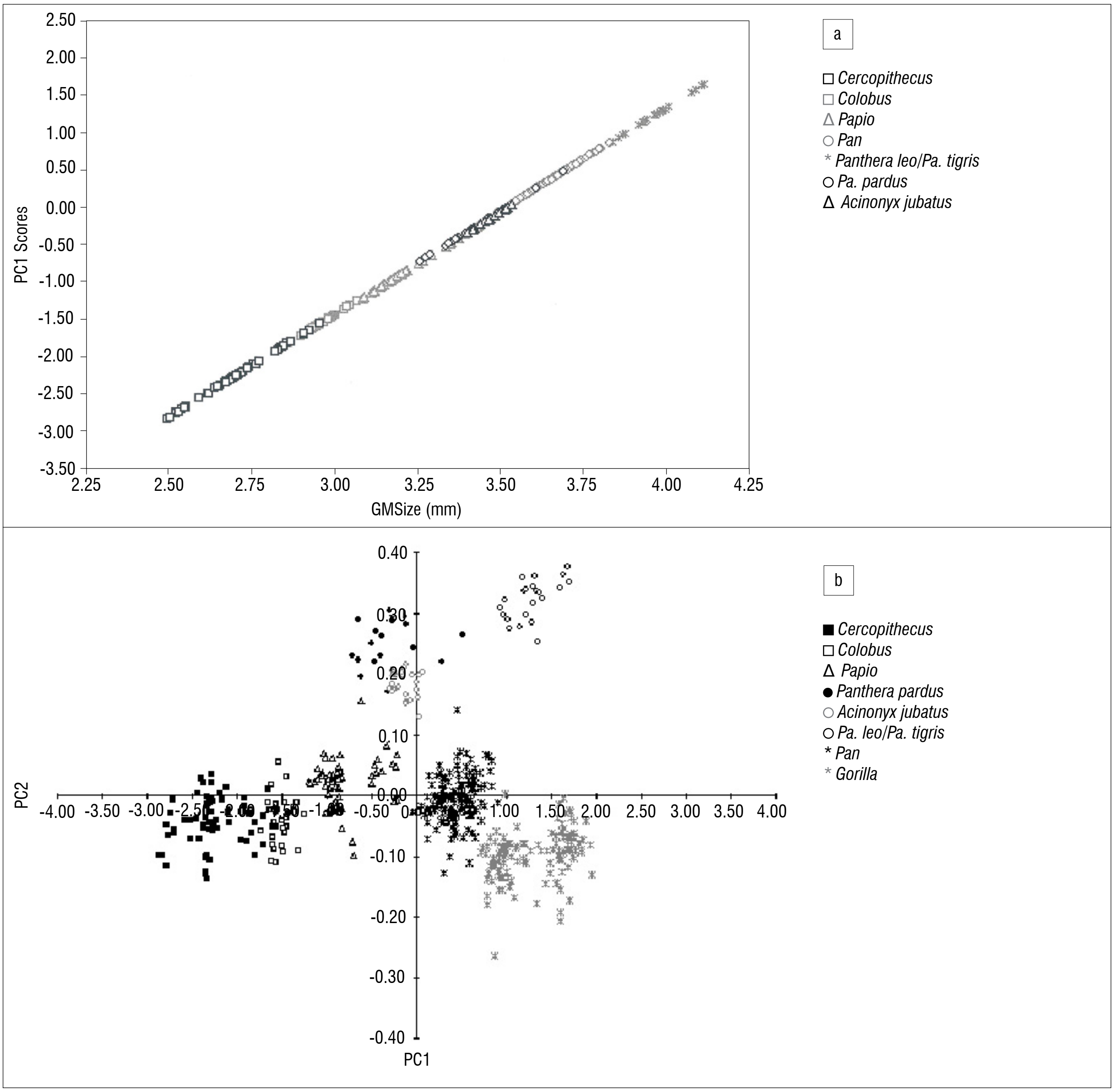


Figure 2: Bivariate scatter plot of (a) PC1 scores (*y*-axis) against GMSize and (b) PC2 to PC1 scores of the $k=8$ postcranial variates in the entire comparative series (*Gorilla* excluded for visual scale).

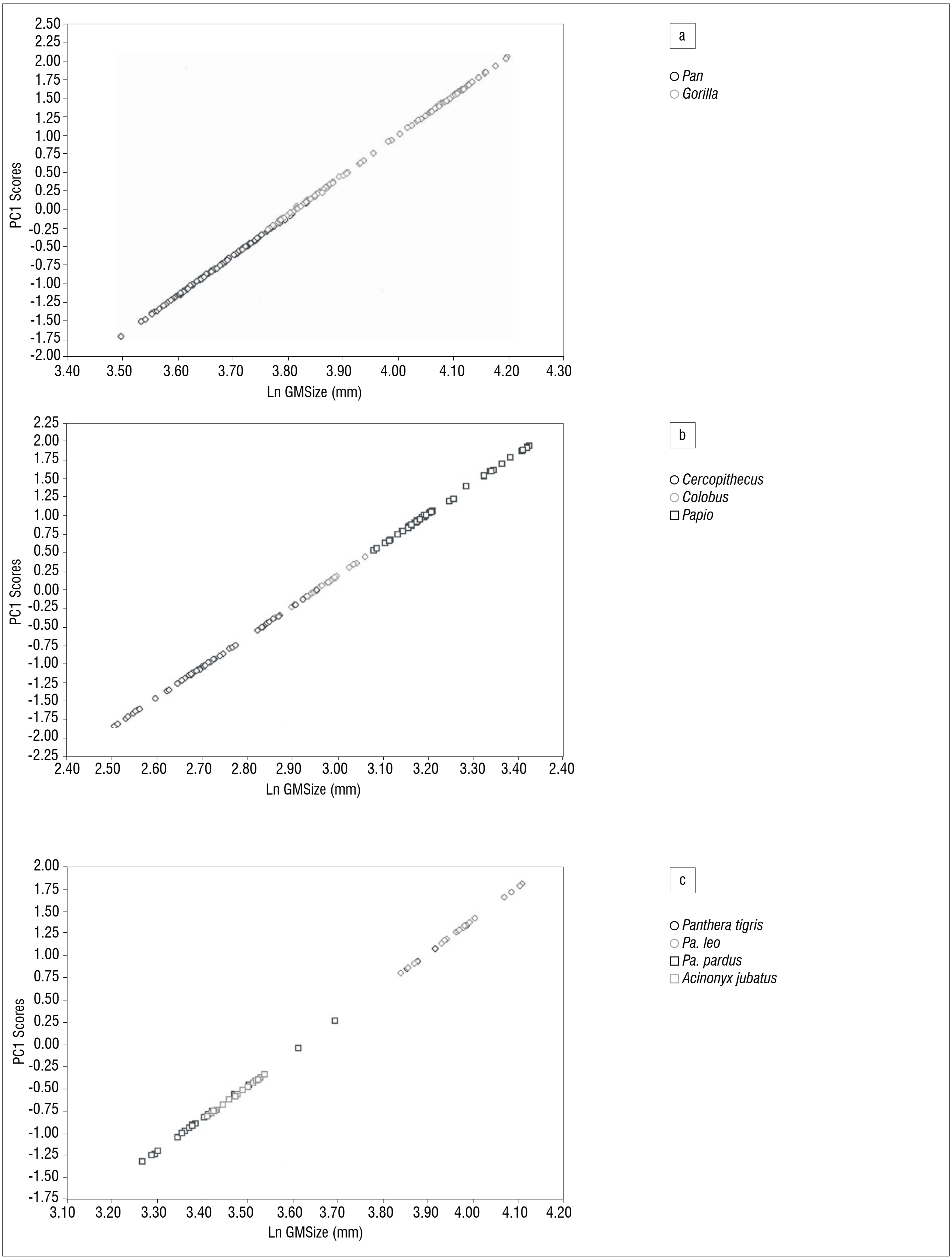


Figure 3: Bivariate scatter plots of PC1 scores (*y*-axis) against GMSize of the $k=8$ postcranial variates in (a) extant African hominids, (b) extant Old World monkeys and (c) extant large-bodied felids. In all cases the parametric correlation (Pearson's) is $r=1.00$.

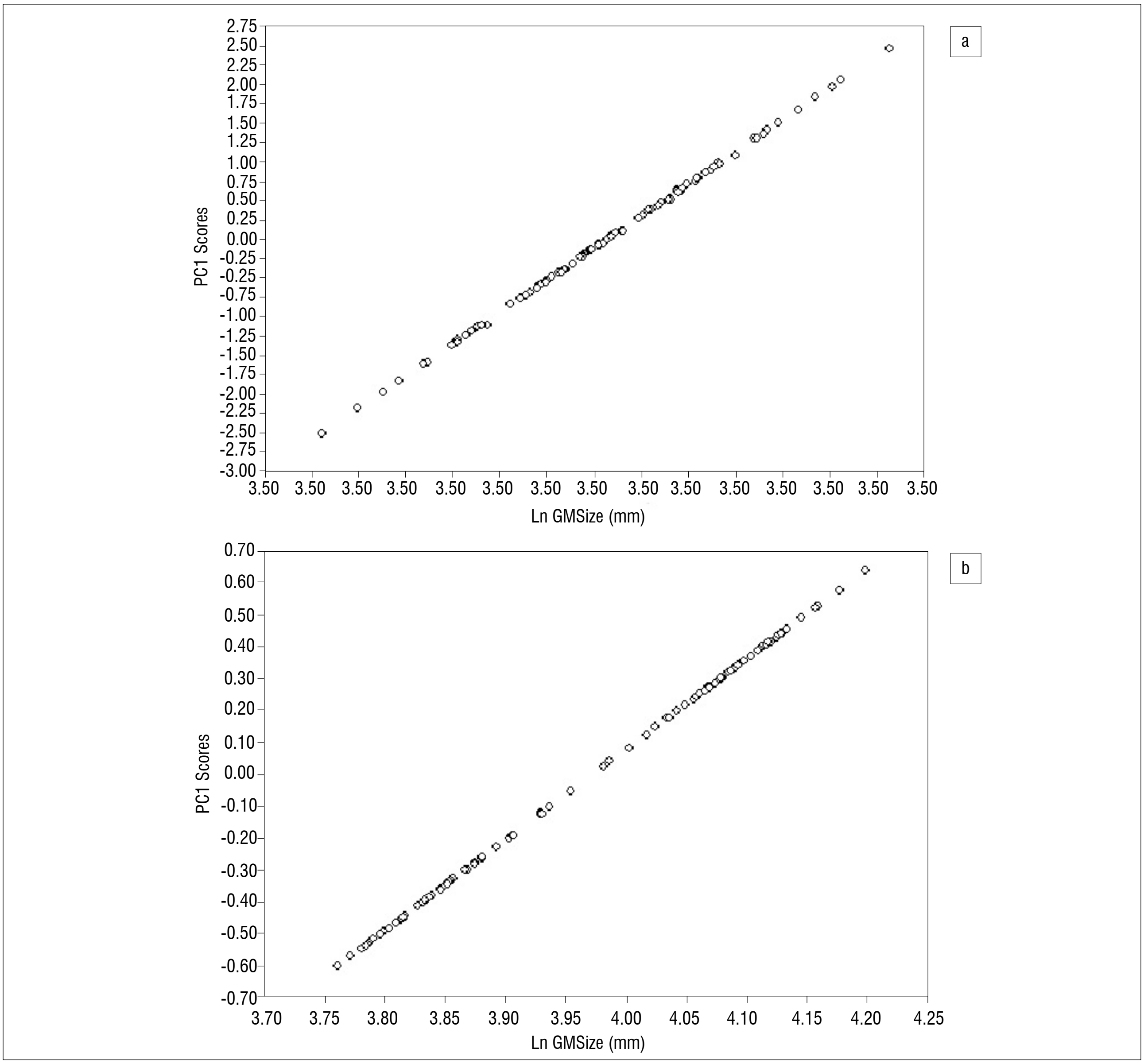


Figure 4: Bivariate scatter plots of PC1 scores (*y*-axis) against GMSize of the *k*=8 postcranial variates in (a) *Pan t. troglodytes* and (b) *Gorilla g. gorilla*. In both cases the parametric correlation (Pearson's) is *r*=1.00.

Table 2: Component loadings for principal component (PC) analysis axes: African hominids

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	JIC
PHAB	0.388	-0.180	-0.022	0.060	0.173	0.160	-0.769	0.407	1.099
DHAB	0.372	-0.050	0.364	0.055	0.298	-0.786	0.030	-0.126	1.052
RHD	0.300	0.460	0.535	0.498	-0.264	0.297	0.071	0.003	0.850
DRB	0.300	0.728	-0.578	-0.076	0.167	-0.114	0.005	-0.002	0.848
FHD	0.364	-0.314	-0.150	0.239	0.572	0.410	0.365	-0.245	1.031
FBB	0.388	-0.243	-0.263	0.032	-0.565	-0.045	-0.192	-0.602	1.097
PTAB	0.365	-0.241	-0.199	0.024	-0.370	-0.125	0.471	0.627	1.033
DTP	0.338	0.087	0.340	-0.825	-0.009	0.263	0.101	-0.053	0.955

JIC, Jolicoeur multivariate allometry coefficients; PHAB, mediolateral diameter of the articular surface of the humeral head; DHAB, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum); RHD, maximum diameter of the radial head; DRB, maximum mediolateral diameter of the distal radial articulation; FHD, femoral head diameter (superoinferior or anteroposterior); FBB, maximum mediolateral diameter of the distal femur; PTAB, maximum mediolateral diameter of the tibial articular plateau; DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia.

Table 3: Component loadings for principal components (PC) analysis axes: Old World monkeys

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	JIC
PHAB	0.355	-0.173	-0.555	0.445	0.524	-0.157	-0.162	-0.107	1.005
DHAB	0.343	-0.366	0.746	0.190	0.345	0.052	0.056	0.175	0.969
RHD	0.418	-0.383	-0.069	-0.601	-0.205	-0.368	-0.366	-0.044	1.182
DRB	0.332	-0.153	-0.014	0.549	-0.748	0.034	0.030	-0.050	0.939
FHD	0.375	-0.165	-0.260	-0.314	0.026	0.538	0.612	-0.035	1.060
FBB	0.331	0.457	0.208	-0.051	0.055	0.344	-0.358	-0.621	0.938
PTAB	0.337	0.440	-0.090	-0.037	-0.019	0.204	-0.282	0.750	0.952
DTP	0.329	0.486	0.103	-0.023	0.024	-0.622	0.504	-0.062	0.930

JIC, Jolicoeur multivariate allometry coefficients; PHAB, mediolateral diameter of the articular surface of the humeral head; DHAB, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum); RHD, maximum diameter of the radial head; DRB, maximum mediolateral diameter of the distal radial articulation; FHD, femoral head diameter (superoinferior or anteroposterior); FBB, maximum mediolateral diameter of the distal femur; PTAB, maximum mediolateral diameter of the tibial articular plateau; DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia.

Table 4: Component loadings for principal components (PC) analysis axes: large-bodied felids

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	JIC
PHAB	0.370	0.149	-0.662	-0.623	0.083	0.020	-0.076	-0.034	1.045
DHAB	0.384	-0.686	-0.018	0.118	0.240	0.514	0.017	0.216	1.086
RHD	0.387	-0.389	-0.118	0.234	-0.400	-0.473	-0.120	-0.481	1.094
DRB	0.363	-0.001	0.727	-0.529	0.150	-0.150	-0.066	-0.104	1.026
FHD	0.327	0.349	0.057	0.288	-0.092	0.117	-0.772	0.260	0.925
FBB	0.316	0.412	0.082	0.180	-0.172	0.554	0.330	-0.496	0.893
PTAB	0.334	0.227	-0.079	0.384	0.688	-0.381	0.253	0.031	0.944
DTP	0.342	0.115	0.046	0.021	-0.492	-0.155	0.455	0.628	0.966

JIC, Jolicoeur multivariate allometry coefficients; PHAB, mediolateral diameter of the articular surface of the humeral head; DHAB, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum); RHD, maximum diameter of the radial head; DRB, maximum mediolateral diameter of the distal radial articulation; FHD, femoral head diameter (superoinferior or anteroposterior); FBB, maximum mediolateral diameter of the distal femur; PTAB, maximum mediolateral diameter of the tibial articular plateau; DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia.

Table 5: Component loadings for principal components (PC) analysis axes: *Pan t. troglodytes*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	JIC
PHAB	0.387	0.123	0.186	0.038	-0.083	-0.475	-0.351	-0.666	1.094
DHAB	0.335	0.132	-0.049	0.278	-0.293	-0.212	0.812	-0.020	0.947
RHD	0.354	-0.088	0.107	0.414	-0.576	0.227	-0.407	0.367	1.002
DRB	0.399	-0.770	-0.413	-0.214	0.048	0.095	0.048	-0.137	1.127
FHD	0.337	0.497	-0.674	0.164	0.336	0.095	-0.168	0.088	0.953
FBB	0.338	0.064	0.159	-0.501	0.087	-0.494	-0.051	0.593	0.956
PTAB	0.349	0.304	0.247	-0.521	-0.119	0.623	0.115	-0.192	0.988
DTP	0.323	-0.153	0.490	0.393	0.662	0.166	0.071	0.081	0.913

JIC, Jolicoeur multivariate allometry coefficients; PHAB, mediolateral diameter of the articular surface of the humeral head; DHAB, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum); RHD, maximum diameter of the radial head; DRB, maximum mediolateral diameter of the distal radial articulation; FHD, femoral head diameter (superoinferior or anteroposterior); FBB, maximum mediolateral diameter of the distal femur; PTAB, maximum mediolateral diameter of the tibial articular plateau; DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia.

Table 6: Component loadings for principal components (PC) analysis axes: *Gorilla g. gorilla*

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	JIC
PHAB	0.363	0.044	0.315	-0.344	-0.045	-0.064	0.215	-0.773	1.026
DHAB	0.372	0.211	0.153	-0.254	-0.558	-0.455	-0.297	0.349	1.053
RHD	0.366	0.119	-0.298	0.323	-0.535	0.592	0.132	-0.068	1.034
DRB	0.364	-0.901	0.080	0.175	-0.018	-0.108	0.006	0.086	1.030
FHD	0.318	0.146	0.531	-0.144	0.299	0.372	0.336	0.484	0.899
FBB	0.354	0.273	-0.249	0.462	0.259	-0.513	0.441	0.025	1.002
PTAB	0.337	0.178	0.160	0.355	0.366	0.119	-0.728	-0.157	0.953
DTP	0.352	-0.036	-0.642	-0.568	0.330	0.109	-0.101	0.098	0.994

JIC, Jolicoeur multivariate allometry coefficients; PHAB, mediolateral diameter of the articular surface of the humeral head; DHAB, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum); RHD, maximum diameter of the radial head; DRB, maximum mediolateral diameter of the distal radial articulation; FHD, femoral head diameter (superoinferior or anteroposterior); FBB, maximum mediolateral diameter of the distal femur; PTAB, maximum mediolateral diameter of the tibial articular plateau; DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia.

Table 7: Summary statistics for the principal components (PC) analysis (trial redundancies)

	Trial 1			Trial 2		
	PC1	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
***Pan t. troglodytes* (*n*=91)**						
% Variance	76.56	7.16	5.28	77.15	6.71	5.27
% Cumulative	76.56	83.72	93.36	77.15	83.86	93.40
***Gorilla g. gorilla* (*n*=102)**						
% Variance	94.86	2.14	1.11	94.68	2.11	1.11
% Cumulative	94.86	97.00	98.85	94.68	96.79	98.80

Discussion and conclusions

The geometric mean of any series of variables is a cumulative dimension inherently dependent on the series of *k* variates employed in its derivation. As a general rule, its only efficacy as a generalised size metric is that it effectively approximates a generalised size vector in multivariate (*n*-dimensional) space which is ultimately testable.[38-41] As Jolicoeur and Mosimann have demonstrated,[38-41,51] both principal and canonical components can be derived and assessed in lieu of any generalised multivariate size distribution (conforming to the Guassian log-normal and gamma distributions) and these effectively approximate the geometric mean. Nevertheless, there has been some recent criticism of the utility of the geometric mean.[59,60] As Auerbach and Sylvester[60] have demonstrated, the Model I slope (least squares regression) of any series of *k* variates regressed upon their respective geometric mean yields a mean slope of β=1.00, irrespective of positive or negative allometry of the independent *k* variates. While this is important, it merely stresses the rationale (theoretical/computational) that bivariate linear regression of any dependent *k* variate upon a geometric mean in a cumulative series of which it is a constituent, is inappropriate.[38,45,51] Irrespective of scalar constraints (i.e. differential size of the *k* dependents), any series of *k* variates is presumed to be highly correlated with its geometric mean and, given the computational mechanics of derivation of the least squares regression slope, the assumption of independence of *x* and *y* is effectively violated. Stated simply, we cannot presume that *x* and *y* are independent nor, for that matter, that error in *y* is independent of error in *x*, when the latter is effectively a cumulative function of unobserved error in a series $y^1 \ldots y^2 \ldots y^3 \ldots y^k$.[38,45] An appropriate solution to this problem is Model II regression.[38]

Table 8: Summary statistics for the principal components (PC) analysis

	k=8		Trial 1		Trial 2	
	PC 1	**JIC**	**PC 1**	**JIC**	**PC 1**	**JIC**
Pan t. troglodytes						
PHAB	0.387	1.094	0.440	1.079	0.430	1.052
DHAB	0.335	0.947	0.383	0.938	NA	NA
RHD	0.354	1.002	NA	NA	0.447	1.096
DRB	0.399	1.127	0.458	1.121	0.421	1.032
FHD AP	0.337	0.953	0.391	0.957	0.381	0.933
FBB	0.338	0.956	NA	NA	0.368	0.901
PTAB	0.349	0.988	0.399	0.977	NA	NA
DTP	0.323	0.913	0.372	0.910	0.397	0.972
Gorilla g. gorilla						
PHAB	0.363	1.026	0.422	1.033	0.440	1.077
DHAB	0.372	1.053	0.432	1.057	NA	NA
RHD	0.366	1.034	NA	NA	0.403	0.987
DRB	0.364	1.030	0.425	1.042	0.463	1.133
FHD AP	0.318	0.899	0.369	0.905	0.382	0.935
FBB	0.354	1.002	NA	NA	0.383	0.939
PTAB	0.337	0.953	0.390	0.955	NA	NA
DTP	0.352	0.994	0.408	0.999	0.371	0.908

JIC, Jolicoeur multivariate allometry coefficients; PHAB, mediolateral diameter of the articular surface of the humeral head; DHAB, mediolateral diameter of the anterior surface of the distal humeral articular surface (trochlea + capitulum); RHD, maximum diameter of the radial head; DRB, maximum mediolateral diameter of the distal radial articulation; FHD, femoral head diameter (superoinferior or anteroposterior); FBB, maximum mediolateral diameter of the distal femur; PTAB, maximum mediolateral diameter of the tibial articular plateau; DTP, the square root of the product of the maximal mediolateral diameter (including the medial malleolus) and the maximum anteroposterior diameter of the distal tibia.

Analysis of the five comparative series included in this study, encompassing the lowest Linnean operational taxonomic unit (i.e. a species) in two cases and in successively higher taxonomic artifices, confirms that the GM of a suite of $k=8$ linear dimensions of the fore- and hindlimb epiphyses of the mammalian postcranial skeleton (GMSize) is both an appropriate and faithful approximate of 'size' in an individual. More crucially, this preferred size variable conforms to *all* logical expectations of the Jolicoeur–Mosimann categorisation of individual organismal size, in both univariate and multivariate space.

Acknowledgements

I thank all the curatorial staff at the various institutions housing the data used in this analysis for their courtesy and hospitality and three anonymous reviewers for their thoughtful critique of the manuscript. I acknowledge the financial support of the Faculty of Health Sciences, University of the Witwatersrand, the Faculty of Humanities, University of Johannesburg, the Ford Foundation, and the Palaeontology Scientific Trust (PAST).

References

1. Bergmann C. Ueber die verhaltnisse der warmeokonomie der thiere zu ihrer grosse. Gottinger Stud. 1847;3:595–708. German.
2. Huxley JS. Problems of relative growth. London: Methuen; 1932.
3. Calder WA. Size, function, and life history. Cambridge, MA: Harvard University Press; 1984.
4. Schmidt-Nielsen K. Scaling: Why is animal size so important? Cambridge, MA: Cambridge University Press; 1984.
5. Clutton-Brock TH, Harvey P. Primate ecology and social organisation. J Zool Lond. 1977;183:1–39. http://dx.doi.org/10.1111/j.1469-7998.1977.tb04171.x
6. Harvey P, Clutton-Brock TH. Life history variation in primates. Evolution. 1985;39:559–581. http://dx.doi.org/10.2307/2408653
7. Stanley SM. An explanation for Cope's rule. Evolution. 1973;27:1–26. http://dx.doi.org/10.2307/2407115
8. Alroy J. Cope's rule and the dynamics of body mass evolution in North American mammals. Science. 1998;280:731–734. http://dx.doi.org/10.1126/science.280.5364.731
9. Van Valkenburgh B, Xiaoming W, Damuth J. Cope's rule, hypercarnivory, and extinction in North American canids. Science. 204;306:101–104.
10. Smith FA, Boyer AG, Brown JH, Costa DP, Dayan T, Morgan Ernest SK, et al. The evolution of maximum body size of terrestrial mammals. Science. 2010;330:1216–1219. http://dx.doi.org/10.1126/science.1194830
11. Sookias RB, Benson RBJ, Butler RJ. Biology, not environment, drives major patterns in maximum tetrapod body size through time. Biol Lett. 2012;8:674–677. http://dx.doi.org/10.1098/rsbl.2012.0060
12. Sookias RB, Butler RJ, Benson RBJ. Rise of dinosaurs reveals major body-size transitions are driven by passive processes of trait evolution. Proc Roy Soc B. 2012;279:2180–2187. http://dx.doi.org/10.1098/rspb.2011.2441
13. Gould SJ. Full house: The spread of excellence from Plato to Darwin. New York: Harmony Rooks; 1996.
14. Cole TJ, Belizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: International survey. BMJ. 2000;320:1–6. http://dx.doi.org/10.1136/bmj.320.7244.1240
15. Cole TJ, Flegal KM, Nicholls DM, Jackson AA. Body mass index cut offs to define thinness in children and adolescents: International survey. BMJ. 2007;335:1–8. http://dx.doi.org/10.1136/bmj.39238.399444.55
16. Ogden CL, Carroll MD, Flegal KM. High body mass index for age among US children and adults. JAMA – J Am Med Assoc. 2004;299:2401–2405. http://dx.doi.org/10.1001/jama.299.20.2401
17. Stevenson RD, Woods WA. Condition indices for conservation: New uses for evolving tools. Integr Comp Biol. 2006;46:1169–1190. http://dx.doi.org/10.1093/icb/icl052
18. Peig J, Green AJ. The paradigm of body condition: A critical reappraisal of current methods based on mass and length. Funct Ecol. 2010;24:1323–1332. http://dx.doi.org/10.1111/j.1365-2435.2010.01751.x
19. Smith RJ. Biology and body size in human evolution: Statistical inference misapplied. Curr Anthropol. 1996;37:451–481. http://dx.doi.org/10.1086/204505
20. Green AJ. Mass/length residuals: Measures of body condition or generators of spurious results. Ecology. 2001;82:1473–1483. http://dx.doi.org/10.1890/0012-9658(2001)082[1473:MLRMOB]2.0.CO;2
21. Peig J, Green AJ. New perspectives for estimating body condition from mass/length data: The scaled mass index as an alternative measure. Oikos. 2009;118:1883–1891. http://dx.doi.org/10.1111/j.1600-0706.2009.17643.x
22. Gingerich PD. Correlation of tooth size and body size in living hominoid primates, with a note on relative brain size in *Aegyptopithecus* and *Proconsul*. Am J Phys Anthropol. 1977;47:395–398. http://dx.doi.org/10.1002/ajpa.1330470308
23. Gingerich PD, Smith BH, Rosenberg KR. Allometric scaling in the dentition of primates and prediction of body weight from tooth size in fossils. Am J Phys Anthropol. 1982;58:81–100. http://dx.doi.org/10.1002/ajpa.1330580110
24. Damuth JM, McFadden BJ. Body size in mammalian paleobiology. Cambridge, MA: Cambridge University Press; 1990.
25. Egi N. Body mass estimates in extinct mammals from limb bone dimensions: The case of the North American hyaenodontids. Palaeontology. 2001;44:497–528. http://dx.doi.org/10.1111/1475-4983.00189
26. De Esteban-Trivigno S, Mendoza M, De Renzi M. Body mass estimation in Xenartha: A predictive equation suitable for all quadrupedal terrestrial placentals? J Morphol. 2008;269:1276–1293. http://dx.doi.org/10.1002/jmor.10659
27. Campione NE, Evans DC. A universal scaling relationship between body mass and proximal limb bone dimensions in quadrupedal terrestrial tetrapods. BMC Biol. 2012;10:60. http://dx.doi.org/10.1186/1741-7007-10-60
28. Rose KD. The beginning of the age of mammals. Baltimore, MA: Johns Hopkins University Press; 2006.
29. Rose KD, Archibald JD. The rise of placental mammals. Origins and relationships of the major extant clades. Baltimore, MA: Johns Hopkins University Press; 2005.
30. Ungar PS. Dental allometry in mammals: A retrospective. Ann Zool Fennici. 2014;51:177–181. http://dx.doi.org/10.5735/086.051.0218
31. McHenry HM. New estimates of body weight in early hominids and their significance to encephalisation and megadontia in "robust" australopithecines. In: Grine FE, editor. Evolutionary history of the robust australopithecines. New York: Aldine de Gruyter; 1988. p. 133–148.
32. McHenry HM. Body size and proportions in early hominids. Am J Phys Anthropol. 1992;87:407–431. http://dx.doi.org/10.1002/ajpa.1330870404
33. McHenry HM. How big were early hominids? Evol Anthropol. 1992;1:15–20. http://dx.doi.org/10.1002/evan.1360010106
34. Mendoza M, Janis CM, Palmqvist P. Estimating the body mass of extinct ungulates: A study on the use of multiple regression. J Zool Lond. 2006;270:90–101. http://dx.doi.org/10.1111/j.1469-7998.2006.00094.x
35. Figueirido B, Pérez-Claros JA, Hunt RM, Palmqvist P. Body mass estimation in amphicynoid carnivoran mammals: A multiple regression approach from the skull and skeleton. Acta Palaeont Polon. 2011;56:225–246. http://dx.doi.org/10.4202/app.2010.0005
36. De Esteban-Trevigno S, Köhler M. New equations for body mass estimation in bovids: Testing some procedures when constructing regression equations. J Mamm Biol. 2011;76:755–761. http://dx.doi.org/10.1016/j.mambio.2011.07.004
37. Field DJ, Lynner C, Brown C, Darroch SAF. Skeletal correlates for body mass estimation in modern and fossil flying birds. PLoS One. 2013;8(11):e82000. http://dx.doi.org/10.1371/journal.pone.0082000
38. Mosimann JE. Size allometry: Size and shape variables with characterisations of the lognormal and gamma distributions. J Am Statist Ass. 1970;65:930–945. http://dx.doi.org/10.1080/01621459.1970.10481136
39. Jungers WL, Falsetti AB, Wall CE. Shape, relative size, and size-adjustments in allometry. Ybk Phys Anthropol. 1995;38:137–161. http://dx.doi.org/10.1002/ajpa.1330380608

40. Jolicoeur PF, Mosimann JE. Size and shape variation in the painted turtle. A principal components analysis. Growth. 1960;24:339–354.

41. Jolicoeur PF. The multivariate generalisation of the allometry equation. Biometrics. 1963;19:497–499. http://dx.doi.org/10.2307/2527939

42. Bookstein FL. Morphometric tools for landmark data: Geometry and biology. Cambridge, MA: Cambridge University Press; 1991.

43. Small CG. The statistical theory of shape. New York: Springer; 1996. http://dx.doi.org/10.1007/978-1-4612-4032-7

44. Dryden IL, Mardia KV. Statistical shape analysis. Chichester: John Wiley and Sons; 1998.

45. Sokal RR, Rohlf FJ. Biometry: The theory and practice of statistics in biological research. 3rd ed. San Francisco, CA: WH Freeman and Co; 1995.

46. Reno PL, McCollum MA, Lovejoy CO, Meindl RS. Adaptationism and the anthropoid postcranium: Selection does not govern the length of the radial neck. J Morphol. 2000;246:59–67. http://dx.doi.org/10.1002/1097-4687(200011)246:2<59::AID-JMOR2>3.0.CO;2-G

47. Hammer Ø, Ryan PD, Harper DAT. PAST: Palaeontological statistics software package for education and data analysis. Palaeont Electr. 2001;4(1), 9 pages. Available from: http://folk.uio.no/ohammer/past

48. Flury BK. Two generalizations of the common principal components model. Biometrika. 1987;74:59–69. http://dx.doi.org/10.1093/biomet/74.1.59

49. Flury BK. Common principal components and related multivariate procedures. New York: Wiley; 1988. Blackith RE, Reyment RA. Multivariate morphometrics. New York: Academic Press; 1970.

50. Neuenschwander B, Flury BK. Common principal components for dependent random vectors. J Multivar Anal. 2000;75:163–183. http://dx.doi.org/10.1006/jmva.2000.1908

51. Darroch JN, Mosimann JE. Canonical and principal components of shape. Biometrika 1985;72:241–252. http://dx.doi.org/10.1093/biomet/72.2.241

52. Blackith RE, Reyment RA. Multivariate morphometrics. New York: Academic Press; 1970.

53. Tabachnick EN, Fidell LS. Using multivariate statistics. 2nd ed. London: Harper and Row; 1991.

54. Bidmos MA, Dayal MR. Further evidence to show population specificity of discriminant function equations using the talus of South African blacks. J Forens Sci. 2004;49:1–6. http://dx.doi.org/10.1520/JFS2003384

55. Dayal MR, Bidmos MA. Discriminating sex in South African blacks using patella dimensions. J Forens Sci. 2005;50:1–4. http://dx.doi.org/10.1520/JFS2004306

56. Dayal MR, Spocter MA, Bidmos MA. An assessment of sex using the skull of black South Africans by discriminant function analysis. HOMO – J Comp Hum Biol. 2008;59:209–221. http://dx.doi.org/10.1016/j.jchb.2007.01.001

57. Kim DI, Kim YS, Lee UY, Han SH. Sex determination from calcaneus in Korean using discriminant analysis. Forens Sci Int. 2013;228(1–3):177.e1–177.e7. http://dx.doi.org/10.1016/j.forsciint.2013.03.012

58. Marinescu M, Panaitescu V, Rosu M, Maru N, Punga A. Sexual dimorphism of crania in a Romanian population: Discriminant function analysis approach for sex estimation. Rom J Leg Med. 2014;22:21–26.

59. Coleman MN. What does geometric mean, mean geometrically? Assessing the utility of geometric mean and other size variables in studies of skull allometry. Am J Phys Anthropol. 2008;135:404–415. http://dx.doi.org/10.1002/ajpa.20761

60. Auerbach BA, Sylvester AD. Allometry and apparent paradoxes in human limb proportions: Implications for scaling factors. Am J Phys Anthropol. 2011;144:382–391. http://dx.doi.org/10.1002/ajpa.21418

Note: This article is supplemented with online only material.

Comparative morphometric analysis of the proximal femur of African hominids and felids

AUTHOR:
Andrew Gallagher[1]

AFFILIATION:
[1]Centre for Anthropological Research (CfAR), University of Johannesburg, Johannesburg, South Africa

CORRESPONDENCE TO:
Andrew Gallagher

EMAIL:
agal1815@gmail.com

POSTAL ADDRESS:
Centre for Anthropological Research, Department of Anthropology and Development Studies, University of Johannesburg, Auckland Park 2006, Johannesburg, South Africa

DATES:
Received: 18 July 2014
Revised: 30 Oct. 2014
Accepted: 04 Jan. 2015

KEYWORDS:
size and shape; proximal femoral morphology; biomechanics

HOW TO CITE:
Gallagher A. Comparative morphometric analysis of the proximal femur of African hominids and large-bodied felids. S Afr J Sci. 2015;111(9/10), Art. 2014-0242, 11 pages. http://dx.doi.org/10.17159/sajs.2015/20140242

Size and shape of the mammalian proximal femur and taxon-specific distinctions in the relative proportions of the proximal articulation, the femoral neck and the proximal femoral diaphysis, are critical determinants in its adaptation to differential biomechanical stresses and observed locomotor habitus in different taxa. The morphometrics of the proximal femur are considered equally critical in the assessment of locomotor habitus of extinct fossil mammals, particularly extinct Miocene anthropoids and Plio-Pleistocene hominins. Analyses of size and shape of $k=10$ dimensions of the proximal femur were undertaken for a large sample series of two extant mammal families – the Felidae and Hominidae – using conventional multivariate statistical procedures, commonly used size-correction methods, and post-hoc tests of significance. While significant differences in form do exist, there are equally striking convergences in the functional morphology of extant hominid and felid taxa. Multivariate and bivariate allometric analyses confirm that the proximal femur of these two mammalian families share a common underlying structure manifest in a shared first common principal component. Nevertheless, while considerable convergences in general form of the proximal femur of African hominids and large-bodied felids are apparent, there exist equally discreet distinctions which are consistent with the differential structural demands imposed by their distinct locomotor and behavioural habitus.

Introduction

The proximal femur of mammals is extremely generalised in its external and internal form[1-3], yet the unique mode of posture and progression in living humans and their extinct Plio-Pleistocene relations is manifest as a suite of consistent structural modifications[4-6]. The principles of osteonal remodelling, and that of its cartilaginous precursors, are universally accepted[7-9] and the adult form of the proximal femur retains considerable potential as a source of quantitative evidence concerning taxonomy, function and phylogeny.[4-6,10]

All quadrupedal vertebrates are essentially hindlimb driven and, the unusual 'diagonal sequence' gait typical of primates notwithstanding, hindlimb drive is employed in recognition of the simple fact that the pelvic and proximal hindlimb musculature is the primary source of initial power input at commencement of the gait cycle from a static position of rest.[11-14] The principal function of the musculoskeletal components of the hip joint are akin to the stroke cycle of a piston rod (or main rod) which powers the main driving wheels in the propulsion of steam-powered locomotives.[15,16] Cyclical rotations and angular excursions of the spherical femoral head within its ball-and-socket joint and the shaft of the femur during contact phases of gait, particularly at optimal and peak velocities, are critical in maintaining momentum and direction of the centre of mass.[14,17,18] From dynamic and static perspectives, the mammalian proximal femur differs from the main rod of a steam locomotive in that it must be 'engineered', for want of a more appropriate descriptor, to withstand potentially critical stresses induced by ground reaction forces which exceed body mass by a considerable magnitude at peak velocities.[9,19]

The femoral neck of mammals varies in its absolute (anatomical) length, its superoinferior angulation, and its degree of anteroposterior anteversion, yet despite observed intraspecific and interspecific functional variance in these parameters, it is generally accepted that the femoral neck performs the function of a cantilever subject to bending in the superioinferior plane.[20-23] These stresses are greatest at the locus of the neck–shaft junction and are structurally mediated by the overall length of the cantilever (anatomical/functional length), its proportions in the primary plane of bending and in the orthogonal neutral axes (x,y) and, by extension, its geometric conformity.[6,21,22,24]

The primary function of the proximal femoral articulation in terrestrial vertebrates lies in the transfer of force, propulsive and reactive, engendered by mass of the organism during posture and ambulation and, to a greater degree, during high-energy ground contact at moderate-to-peak velocities of the gait cycle.[12,19,25,26] The magnitude of these forces ultimately determines the absolute and relative size of the articular surface, which is greater in terrestrial bipeds.[21,27] Absolute length of the neck is a critical determinant of the lever arm of the external and intrinsic hip musculature in quadrupeds and bipeds[5,21] and in terrestrial hominins the neck–shaft junction is subject to considerably elevated compressive stresses compared with those of more arboreal species like *Gorilla* and *Pan*[6,10]. Proportions of the proximal and distal osseous components of the limbs and the digits (manus and pes) attest to considerable modifications to habitual structural demands, whether as levers increasing mechanical efficiency with respect to specific functional demands (i.e. suspension versus support; increased inertia in specialised high-velocity taxa), or in resistance to flexure and ultimate failure during peak loading cycles.[13,28-30]

The extant African hominids (*Gorilla*, *Pan* and *Homo*) and large-bodied felids (*Panthera* and *Acinonyx*) offer a unique prospectus to assess patterns of variance in the general morphology of the proximal femur and its modification to disparate structural demands. Both mammalian families exhibit considerable size variance that was likely exceeded in some extinct Plio-Pleistocene representatives[31,32] and both include one highly specialised extant taxon, namely *Homo* and *Acinonyx*. However, there is good reason to designate the fore- and hindlimb anatomy of *Acinonyx* to being merely functionally derived, rather than unique, whereas the lower limb of *Homo* has no true extant comparator. Living humans were thus excluded from this analysis, which was focused upon the elucidation of distinctions and general underlying convergences in the proximal femoral morphology of other hominids and felids.

This contribution explicitly assesses patterns of underlying convergences and disparities in proximal femoral form of these two phyletically discreet families of extant African mammals. If the proximal femur of mammals is conservative in its general form, it follows logically that extant African hominids and large-bodied African felids will share a common generalised vector of size and shape (H_o 1). However, given the distinct postural repertoires and locomotor modes of extant African hominids and large-bodied felids, whereby the total mass of the individual is supported to a greater extent by the proximal femoral articulation in the former, an equally logical expectation is that any discreet morphometric distinctions of the proximal femur will be consistent with the mechanical demands of their observed disparate posture and locomotor modes.

Materials and methods

A total of 213 specimens of African hominid species and sub-species of the extant genera *Gorilla* (gorillas) and *Pan* (chimpanzees and bonobos) were included in the analysis (Supplementary table 1 of the online supplementary material). The extant large-bodied felid samples included *Panthera* – *Panthera tigris* (tigers), *Panthera leo* (lions) and *Panthera pardus* (leopards) – and *Acinonyx* (*Acinonyx jubatus*; cheetahs) and comprised 69 individuals. Only adult specimens evidencing complete fusion of the postcranial epiphyses were used in this study. A series of $k=10$ linear dimensions of the proximal femur used in conventional appraisals of the taxonomic and functional affinities of extinct Plio-Pleistocene hominin genera (*Australopithecus*, *Paranthropus* and early *Homo*)[33] were taken using a pair of Mitutoyo™ digital calipers (Figure 1). As can be seen, these linear dimensions effectively describe the general form (form=size+shape)[34] of the proximal articulation (superoinferiorly, anteroposteriorly, mediolaterally), the proportions of the femoral neck and proximal femoral diaphysis (mediolaterally and anteroposteriorly).

All linear data were log-transformed (base *e*) in MS Excel and subsequently size corrected using the approach outlined by Mosimann[35], whereby each individual *k* linear dimension for a particular specimen is expressed as a proportion (not a percentage) of the geometric mean of all $k=10$ dimensions. The variates generated using this approach are effectively uncorrelated with size and facilitate both bivariate and multivariate perspectives on relative size distinctions between series of forms.[35] Following the rationale of Jolicoeur, the most appropriate mathematical descriptor of size and shape is the first principal component (PC) of a log-transformed covariance matrix.[35,36] If size is dominant, then the first PC (PC1) should account for an overwhelming proportion of the total observed variance, leaving little residual variance assigned to successive orthogonal (and uncorrelated) PC axes.[37,38] If this prospectus holds, then the loadings of individual *k* variates on this axis can be expressed as multivariate allometry coefficients standardised around a value of 1, the theoretical coefficient of isometry.[34-36] This is simply calculated as $1/\text{sqrt}(k)=0.316$ in the case of our $k=10$ dimensions and all log-transformed raw variates should yield positive scores on the first PC.

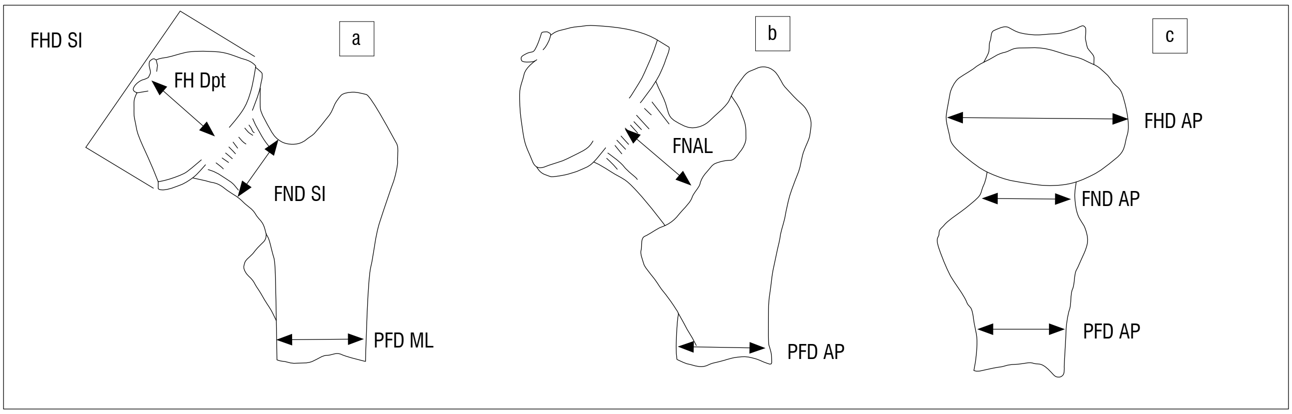
All principal components analyses (PCA) were performed on the variance-covariance matrix (hereafter VCV), as opposed to the correlation matrix, using PAST version 3[39], which extracts a series of components commensurate to the *k*-variate series,[37,38] irrespective of the magnitude of the variance explained by successive PCs, that effectively diminish to a total cumulative 100%. If PC1 does constitute a 'generalised size and shape vector', then successive PCs may be effectively negligible.[34-38] In the first instance, PCAs were performed on the log-transformed raw data of the total sample (pooled hominid and felid) and on the individual family series of hominids and large-bodied felids. Jolicoeur multivariate allometry coefficients were subsequently calculated for the component loadings of the $k=10$ linear variate series on PC1 alone.[35,36]

While the derivation of Jolicoeur multivariate allometry coefficients provides a computationally effective means of assessment of distinctions in $n=2$ (or $n=>2$) samples for a given battery of *k* linear dimensions, explicit tests of underlying structural commonalities in sample-specific VCV matrices were explored via the Flury hierarchy of common principal components analysis which assesses for equality, proportionality and sequential shared common PCs.[40,41] The common principal components hierarchy is an effective test of underlying correspondences and disparities in matrix structure beyond a simple test of equality of *n*-group matrices. These tests were performed using the program written in MS Dos for Windows by Dr Patrick C. Phillips.

A second PCA was performed on the total series data set using the Mosimann shape variates of the $k=10$ dimensions in order to explicitly examine patterns of relative size disparities in the proximal femur of African hominids and large-bodied felids. Model II (reduced major axis) bivariate allometric equations were calculated in post-hoc assessments either as sample-specific (i.e. hominid or felid) or total-sample models using PAST v.3. Post-hoc significance tests of estimates of *y* (the dependent variate) on *x* (the geometric mean) were determined via Fisher's paired randomisation test performed in Rundom Projects 2.[42,43]

Results

Summary data for the PCAs of the total and individual sample series, PC loadings and multivariate allometry coefficients for the total sample series are given in Tables 1 and 2. The PC loadings and multivariate allometry coefficients for the individual series (African hominid, large-bodied felid) are given in the online supplementary material (Supplementary tables 2 and 3). In the three analyses of the log-transformed raw data, the first PC is a generalised size vector accounting for about 88% and 96% of the total variance and is greater in felids than in hominids (Table 1). Pearson's *r* between the PC1 scores and geometric means of the $k=10$ linear variates is perfect in both felids and hominids ($r=1.000$;



FHD SI/AP, superoinferior/anteroposterior diameter of the proximal articulation; FH Dpt, mediolateral diameter of the proximal articulation; FND SI/AP, superoinferior/anteroposterior diameter of the femoral neck; PFD ML/AP, mediolateral/anteroposterior diameter of the proximal diaphysis (distal to the lesser trochanter); FNAL, anatomical length of the femoral neck.

Figure 1: Linear dimensions taken on the proximal femur in (a) anterior, (b) posterior and (c) medial views.

Supplementary figure 1). Nevertheless, subtle distinctions do exist in the sample-specific multivariate allometry coefficients, particularly in dimensions reflecting absolute size of the proximal femur, proximal femoral diaphysis and femoral neck. Yet in both hominids and felids, the articulation is approximately isometric and is greater in the superoinferior plane (Supplementary tables 2 and 3).

As can be seen in Figure 2a, significant contrasts exist in the proximal femur of hominids and large-bodied felids on PC2, which primarily reflects differential loadings of the femoral head on the one hand and dimensions of the neck on the other. An interesting point of note is that the proximal femur of *Pn. paniscus* (bonobos) appears to be intermediate in form between extant felids and the main scatter of *Pn. troglodytes* (chimpanzees) and *Gorilla* on axis 2, reflecting correspondences in the general form of the femoral neck. The anatomical length, the superoinferior height and anteroposterior depth, load positively on this axis and are greater in felids than in hominids. An explicit assessment of commonality of underlying structure in the VCV matrices of log-transformed raw data using the Flury hierarchy confirms that both series share a first, but not a second, common PC. The VCV matrices are neither equivalent nor proportional, and significant contrasts in proximal femoral morphology in hominids and large-bodied felids are confirmed, irrespective of whether the 'jump-up' or 'step-up' approach is preferred (Supplementary tables 4 and 5). [41]

The observed contrasts in the proximal femur of extant African hominids and large-bodied felids on the one hand, and *Pn. paniscus* on the other, are further exemplified in multivariate analyses of the Mosimann

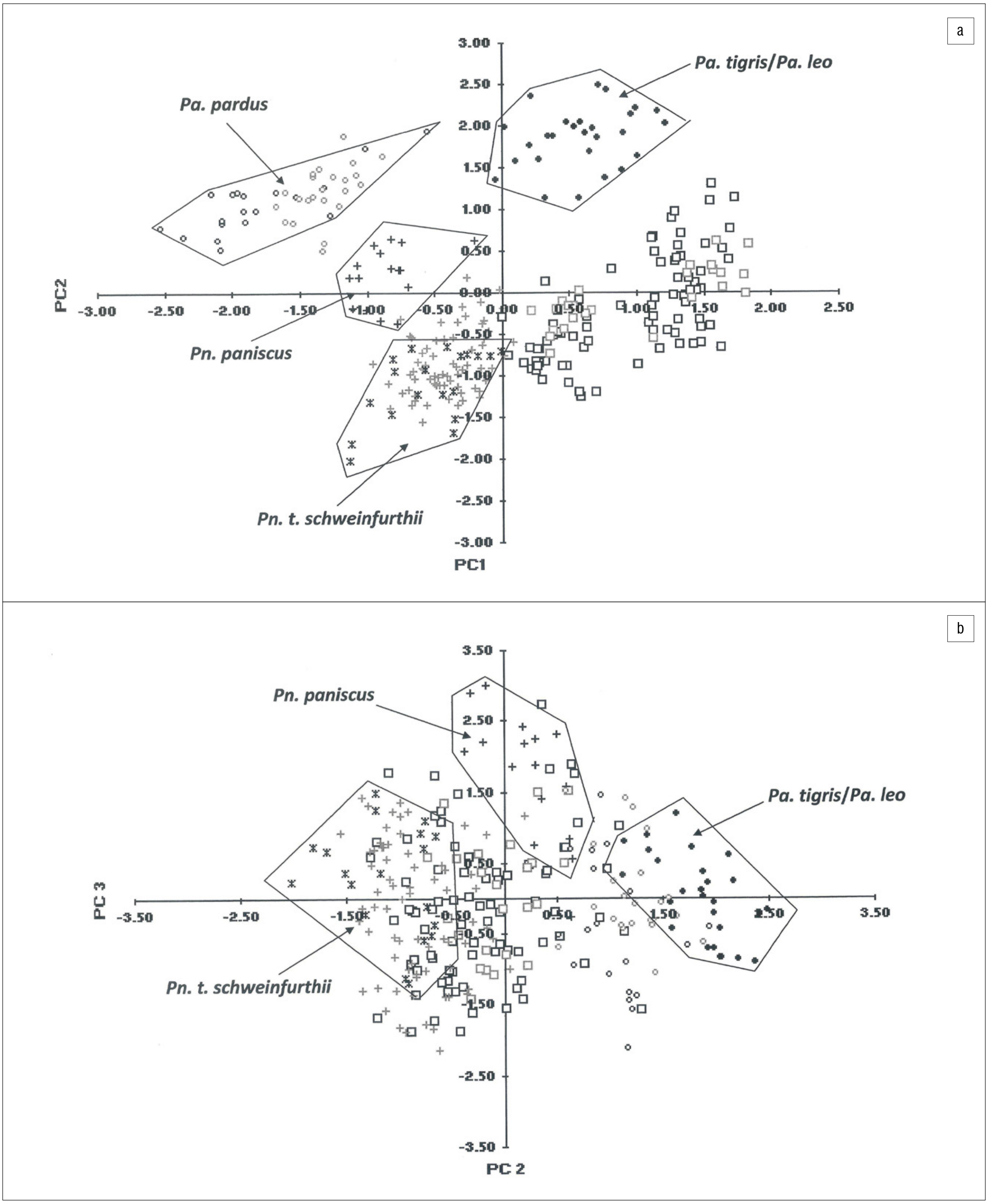
Table 1: Summary statistics for the principal components (PC) analysis: raw data

Total			
PC	**Eigenvalue**	**% Variance**	**% Cumulative**
1	0.5101	87.77	87.77
2	0.0358	6.16	93.93
3	0.0168	2.89	96.82
Hominid			
PC	**Eigenvalue**	**% Variance**	**% Cumulative**
1	0.3684	88.72	88.72
2	0.0199	4.79	93.51
3	0.0119	2.86	96.37
Felid			
PC	**Eigenvalue**	**% Variance**	**% Cumulative**
1	0.6449	95.73	95.73
2	0.011	1.64	97.37
3	0.0051	0.76	98.13

Table 2: Component loadings for the principal components (PC) analysis of the proximal femur: pooled-sample raw data

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**JIC**
FHD SI	0.353	-0.206	0.123	-0.193	0.065	-0.076	-0.082	-0.482	-0.175	-0.708	1.115
FHD AP	0.347	-0.214	0.139	-0.227	0.110	-0.073	-0.147	-0.461	-0.132	0.703	1.096
FH Dpt	0.332	-0.120	0.106	-0.432	0.015	0.008	0.743	0.352	-0.034	0.019	1.049
PFB	0.321	-0.007	0.084	-0.239	-0.095	0.184	-0.322	0.137	0.815	-0.045	1.016
FBNL	0.317	0.028	0.013	-0.140	-0.188	0.303	-0.477	0.494	-0.530	-0.009	1.001
PFD ML	0.331	-0.212	-0.100	0.605	-0.253	0.553	0.263	-0.168	0.036	0.036	1.046
PFD AP	0.352	-0.325	-0.001	0.451	0.045	-0.670	-0.091	0.322	0.057	-0.001	1.114
FNAL	0.299	0.294	-0.875	-0.089	0.201	-0.066	0.017	-0.071	0.006	0.006	0.945
FNM SI	0.257	0.688	0.223	0.077	-0.540	-0.276	0.097	-0.160	-0.029	0.030	0.812
FNM AP	0.232	0.432	0.348	0.255	0.735	0.165	0.013	0.073	-0.008	-0.018	0.732

JIC, Jolicoeur multivariate allometry coefficients; FHD SI/AP, superoinferior/anteroposterior diameter of the proximal articulation; FH Dpt, mediolateral diameter of the proximal articulation; PFB, total mediolateral breadth of the proximal femur; FBNL, biomechanical length of the femoral neck; PFD ML/AP, mediolateral/anteroposterior diameter of the proximal diaphysis (distal to the lesser trochanter); FNAL, anatomical length of the femoral neck; FNM SI/AP, superoinferior/anteroposterior diameter of the femoral neck.



Open black squares, Gorilla g. gorilla; open grey squares, G. b. graueri/G. b. beringei; black crosses, Pan paniscus; grey crosses, Pn. t. troglodytes; black asterisks, Pn. t. schweinfurthii; filled black circles, Panthera tigris/Pa. leo; open black circles, Pa. pardus; open grey circles, A. jubatus.

Figure 2: Principal components (PC) analysis of the pooled sample raw data: (a) PC2 to PC1 and (b) PC3 to PC2.

shape variates (Figure 3). As size has been effectively removed, five PCs are required to account for approximately 93% of the total variance and component loadings of the $k=10$ variates range from negative to positive (Tables 3 and 4; Figure 3). Clear distinctions exist between felids (+ scores) and hominids (- scores) on the first, but not the second, PC axis, which reflects contrasts in the relative proportions of the proximal articulation and proximal diaphysis, which are greater in *Gorilla* and *Pan*, and the femoral neck (Table 4). As in the PCA of the raw data series, the principle morphometric distinctions in the femoral neck of felids and hominids lie in the proportions of the femoral neck, particularly its relative area (superoinferior and anteroposterior diameters). The striking convergence in proximal femoral morphology of *Pn. paniscus* and large-bodied felids is equally apparent in Figure 3a, yet in Figure 3b the equally stark contrasts in absolute size of the head and total and functional length of the proximal femur are demonstrable (Table 4; Figure 3).

In order to further assess the significance of the observed distinctions in relative proportions of the proximal femur in African hominids and large-bodied felids, reduced major axis regression solutions were derived for either pooled-sample or family-specific models, depending upon expectations of a single or independent allometric trend. The geometric mean of the $k=10$ dimensions was used as the independent (x) variate in all comparisons with the exception of the proportions of the neck and proximal diaphysis, in which contrasts in relative anteroposterior/mediolateral proportions were considered (Table 5). With few exceptions, the linear variates of the proximal femur attest to positive allometry in hominids and felids. Bivariate scatter plots of three variates that apparently contribute to effective discrimination of the proximal femur of hominids and large-bodied felids in multivariate space are shown in Figure 4. Only the absolute size of the proximal articulation exhibits a sound bivariate relationship with minimal scatter in the data and further attests consistent proportional distinctions between felids and hominids (Figure 4a). Although there is a general tendency for *Pa. pardus* and *A. jubatus* proximal femoral diaphyses to fall outside the ranges of *Pan* with respect to their femoral neck anteroposterior/mediolateral diaphyseal proportions in their lower size ranges, this tendency is less consistent in their upper ranges and scarcely holds for *Pa. leo* and *Pa. tigris* with *Gorilla* (Figure 4b,c). Despite its functional significance and consistency in the multivariate analyses, relative anatomical length of the neck exhibits inordinate variance. Data in Figure 4c confirm that the major difference in the proximal femur of *Pn. paniscus* compared with *Pn. troglodytes* and *Gorilla* lies in their proportionally shorter, rather than elongate, femoral neck.

A similar overlap is observed in a great majority of other variates regressed against the geometric mean (GM), yet there is some consistency in *Pa. leo* and *Pa. tigris* in the relative proportions of the proximal femoral diaphysis compared with gorillas of equivalent size (Supplementary figures 2–4). These data are generally consistent with post-hoc tests (Table 6) which confirm general distinctions in the relative size of the proximal articulation between hominids and felids and in the proportions of the femoral neck and proximal diaphysis. Nevertheless, comparisons of observed and expected series in felids (medians are presented as a guide) reveal subtle distinctions in the proximal femur of

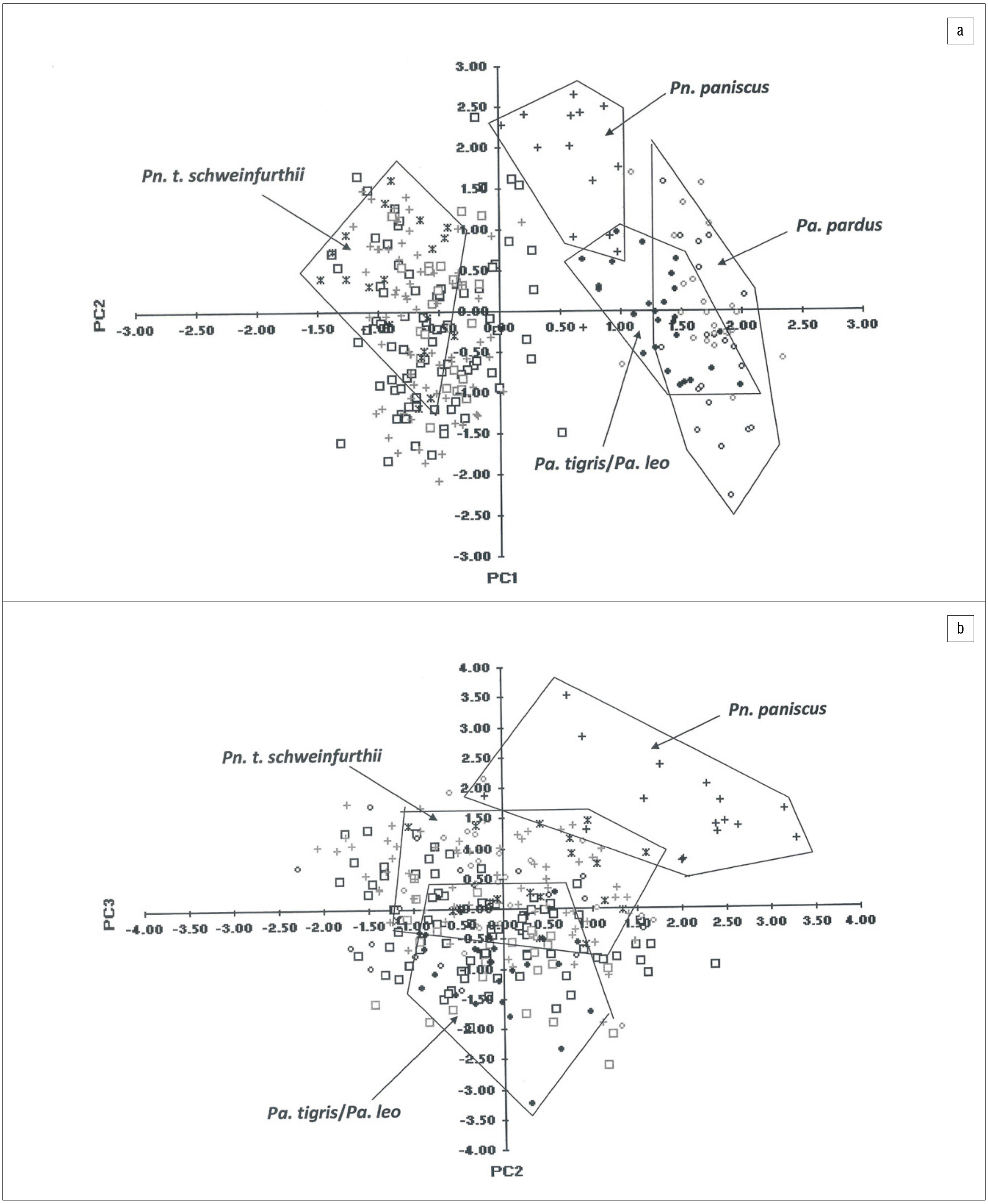
Table 3: Summary statistics for the principal components analysis (PCA): pooled-sample shape data

PCA			
PC	**Eigenvalue**	**% Variance**	**% Cumulative**
1	0.0035	52.541	52.54
2	0.0014	20.823	73.36
3	0.0006	9.378	82.74
4	0.0004	6.320	89.06
5	0.0003	4.376	93.44
6	0.0002	3.428	96.87
7	0.0001	2.045	98.91

Table 4: Component loadings for the principal components (PC) analysis of the proximal femur: pooled-sample shape data

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
FHD SI	-0.258	0.118	0.081	-0.219	-0.016	0.056	0.500	-0.099	-0.704
FHD AP	-0.253	0.140	0.146	-0.199	-0.061	0.113	0.492	-0.064	0.706
FH Dpt	-0.184	0.104	0.160	-0.554	0.079	-0.515	-0.500	0.004	0.025
PFB	0.050	0.112	0.510	0.283	0.075	0.175	-0.110	0.704	-0.047
FBNL	0.062	0.036	0.411	0.372	0.137	0.163	-0.247	-0.693	-0.007
PFD ML	-0.254	-0.118	-0.349	0.571	0.170	-0.566	0.139	0.065	0.042
PFD AP	-0.414	-0.027	-0.503	-0.014	-0.030	0.567	-0.386	0.064	0.002
FNAL	0.238	-0.884	0.057	-0.153	-0.176	0.030	0.058	0.023	0.010
FNM SI	0.613	0.183	-0.322	-0.174	0.582	0.090	0.102	0.023	0.032

FHD SI/AP, superoinferior/anteroposterior diameter of the proximal articulation; FH Dpt, mediolateral diameter of the proximal articulation; PFB, total mediolateral breadth of the proximal femur; FBNL, biomechanical length of the femoral neck; PFD ML/AP, mediolateral/anteroposterior diameter of the proximal diaphysis (distal to the lesser trochanter); FNAL, anatomical length of the femoral neck; FNM SI, superoinferior diameter of the femoral neck.



Open black squares, Gorilla g. gorilla; open grey squares, G. b. graueri/G. b. beringei; black crosses, Pan paniscus; grey crosses, Pn. t. troglodytes; black asterisks, Pn. t. schweinfurthii; filled black circles, Panthera tigris/Pa. leo; open black circles, Pa. pardus; open grey circles, Acinonyx jubatus.

Figure 3: Principal components (PC) analysis of the pooled sample Mosimann shape data: (a) PC2 to PC1 and (b) PC3 to PC2.

these taxa, irrespective of a general convergence in overall form (Tables 5 and 6; Figures 3 and 4; Supplementary figures 2–4).

Discussion and conclusions

Segmental mass distribution in primates differs from quadrupedal mammals primarily in a proportionally greater allocation of the hindlimb musculature.[12,30,44,45] Nonetheless, even among living anthropoids, the terrestrial knuckle-walking gaits of *Gorilla* and *Pan* are unique in that the centre of mass is more distally located towards the hip joints than it is in terrestrial quadrupedal monkeys.[12,13,30,32,45,46] More crucially, irrespective of differences in absolute size, the distribution of mass during forelimb contact in the gait cycle is only marginally higher in *Gorilla* than in *Pan* and does not exceed 20%.[13] If large-bodied felids are accepted as suitable mammalian comparatives, then increased joint-reaction forces generated during normal posture and progression in extant African hominids is sufficient explanative for the observed proportional distinctions in their proximal femoral epiphyses.[13,30,32,45,46] Despite the pooled-sample least-squares regression slope of greater than 1.00, the hip joints in intraspecific comparisons of *Gorilla* and *Pan* likely approximate isometry, and the observed proportional differences between African hominids and felids cannot be simply dismissed as an allometric consequence of extreme size in *Gorilla*.

Given their immense size and exponentially increased hip joint reaction forces, gorillas might be logically expected to exhibit proportionally greater proximal femoral articulations than chimpanzees, particularly in light of the fact that their femora are proportionally shorter relative to the length of the humerus.[13,28,29,32,42,47] Humeral length increases isometrically with body mass ($\alpha=0.333$) within the order Primates,[47] so the elongate humerus of *Gorilla* is entirely expected for an anthropoid of its size. Following this reasoning, the reduced length of the femur, and not the elongation of the humerus, is imperative in their lower humero-femoral indices compared with *Pan*.[13,28,29,32,47] Reduction in the length of the femur in *Gorilla* and their extreme midshaft diaphyseal cross-sectional areas[29,32] compared with *Pan,* reflect necessary structural modifications to their increased bulk. Resultant increases in joint-reaction forces at the hip and knee of the lower limb in *Gorilla* and critical bending stresses at the femoral midshaft may be offset by a reduction in absolute and relative femur length, yet absolute length of the femur may independently mediate absolute size of the proximal and distal epiphyses in adults.[29,32,33]

The observed contrasts in proximal femoral morphology of *Pn. paniscus* and other extant African hominids in this analysis is deserving of consideration. The convergence in proximal femoral form of *Pn. paniscus* with the extant felids reflects subtle distinctions in the relative length and superoinferior/anteroposterior proportions of the femoral neck, which are greater than in *Pn. troglodytes* and *Gorilla*. On the one hand, these observed morphometric contrasts in the proximal femur of *Pn. paniscus* and *Pn. troglodytes* spp./*Gorilla* likely correspond with subtle, yet crucial, distinctions in extrinsic hip musculature architecture (*M. Gluteus medius*, *M. Ischiofemoralis*) and moment arms (*M. Gluteus medius*) in bonobos and their emphasis in increased hip-flexion during climbing activities, which are more influential in the bonobo's lifestyle.[48,49] Nevertheless, no significant contrasts in the general proportions of the neck of *Pn. paniscus* and *Pn. troglodytes* were found in previous geometric morphometric analyses[50,51]; the discrepancies between this analysis and that of Holliday and colleagues[51] may simply reflect sampling constraints (larger samples in this study). However, divergences from the results of Harmon's [50] findings cannot be explained by such factors and require more detailed consideration.

Traditional morphometric methods differ both qualitatively and quantitatively from geometric morphometric approaches in that size standardisation

Table 5: Bivariate allometry equations for the proximal femoral dimensions

Regress	Series	Model	Slope	5% Confidence interval	95% Confidence interval	Intercept	5% Confidence interval	95% Confidence interval	*r*	*p*-value
FHD SI vs GMSize	African hominids	RMA	1.048	1.025	1.070	-0.113	-0.191	-0.031	0.989	0.001
	Large-bodied felids	RMA	1.028	0.999	1.058	-0.161	-0.259	-0.066	0.993	0.001
FHD SI vs GMSize	African hominids	RMA	1.024	1.002	1.017	-0.017	-0.098	0.065	0.989	0.001
	Large-bodied felids	RMA	0.994	0.970	1.018	-0.040	-0.123	0.043	0.995	0.001
FHD Dpth vs GMSize	Total series	RMA	1.089	1.056	1.123	-0.510	-0.634	-0.388	0.967	0.001
PFB vs GMSize	Total series	RMA	1.037	1.017	1.056	0.619	0.548	0.689	0.989	0.001
FBNL vs GMSize	Total series	RMA	1.025	1.004	1.047	0.465	0.392	0.544	0.983	0.001
PFD ML vs GMSize	Total series	RMA	1.105	1.066	1.143	-0.538	-0.672	-0.400	0.948	0.001
PDF AP vs GMSize	African hominids	RMA	1.073	1.014	1.130	-0.573	-0.779	-0.354	0.942	0.001
	Large-bodied felids	RMA	1.036	0.969	1.099	-0.580	-0.793	-0.351	0.973	0.001
FNAL vs GMSize	Total series	RMA	1.108	1.037	1.177	-0.434	-0.657	-0.151	0.864	0.001
FND SI vs GMSize	African hominids	RMA	1.140	1.076	1.202	-0.767	-0.990	-0.537	0.935	0.001
	Large-bodied felids	RMA	1.004	0.955	1.050	-0.001	-0.168	0.159	0.980	0.001
FND AP vs GMSize	African hominids	RMA	0.941	0.889	0.995	-0.283	-0.481	-0.093	0.880	0.001
	Large-bodied felids	RMA	0.947	0.912	0.989	-0.129	-0.266	-0.013	0.980	0.001
PFD AP vs PFD ML	Total series	RMA	1.069	1.028	1.112	-0.413	-0.558	-0.276	0.948	0.001
FND AP vs FND ML	Total series	RMA	0.854	0.806	0.899	0.239	0.094	0.397	0.904	0.001

FHD SI, superoinferior diameter of the proximal articulation; PFB, total mediolateral breadth of the proximal femur; FBNL, biomechanical length of the femoral neck; PFD ML/AP, mediolateral/anteroposterior diameter of the proximal diaphysis (distal to the lesser trochanter); FNAL, anatomical length of the femoral neck; FND SI/AP, superoinferior/anteroposterior diameter of the femoral neck; RMA, reduced major axis.

exemplifies, rather than partitions, linear dimensions that may, or may not, approximate inter-landmark distances in 2D/3D registration space. In contrast, generalised Procrustes rotation (translation, rotation and scaling) does distort linear inter-landmark distances, which may potentially mask real morphometric distinctions in small-scale, localised regions of interest.[52] Further to this, direct post-hoc comparisons of Asian and African apes in Harmon's[50] study did reveal significant contrasts in the relative length of the femoral neck, which is proportionally greater in the former, yet no explicit post-hoc assessment of morphological distinctions in *Pan* was undertaken. At a basic level, these two methodologies are not directly comparable and any potential conflict with previously published analyses does not compromise the consistent contrasts witnessed in this study, nor their significance.

Despite substantial variation in body mass (from <4 kg to nearly 200 kg), extant felids do not exhibit large-scale, size-dependent changes in either limb posture or angular joint excursion during normal progression.[53] Results of these analyses confirm that the proximal femoral morphology of living large-bodied felids exhibits striking consistencies in general form, which is particularly evident in multivariate appraisals of Mosimann shape variates with considerable overlap among taxa. This finding is somewhat unexpected, given the highly specialised gait of living cheetahs (*A. jubatus*) and only the relative mediolateral proportions of the proximal diaphysis attests taxon-specific distinctions when expressed as a function of the GM. Surprisingly, it is *Pa. pardus*, rather than *Acinonyx*, which differs from the morphology of *Pa. leo* and *Pa. tigris* by virtue of their relatively broad proximal shafts in the mediolateral plane. The most logical explanation for this finding is that the elongated anatomical length of the femoral neck in *Pa. pardus* increases joint excursion and yields greater mediolateral bending moments in the proximal femoral diaphysis during normal loading[20-22] – a possible morphological function of increased tree-climbing behaviours observed in this species.[54] Compared with thoroughbred greyhounds, cheetahs evidence considerable hypertrophy and increased moment-arms of their hip-joint flexors and extensors.[55-58]

Table 6: Fisher's paired randomisation comparisons

Variable	Model	Median observed	Median predicted	Sum observed	Mean sum	5% Confidence interval	95% Confidence interval	*p*-value
Hominids								
FHD SI	RMA	3.603	3.605	0.270	0.346	0.028	0.850	0.534
FHD AP	RMA	3.614	3.615	0.262	0.347	0.030	0.851	0.550
FH Dpt	RMA	3.397	3.353	3.160	0.675	0.050	1.650	0.001
PFB	RMA	4.302	4.297	0.877	0.377	0.029	0.927	0.001
FBNL	RMA	4.119	4.101	0.807	0.480	0.041	1.159	0.182
PFD ML	RMA	3.429	3.382	1.953	0.978	0.077	2.427	0.109
PFD AP	RMA	3.252	3.233	0.192	0.820	0.060	1.992	0.852
FNAL	RMA	3.562	3.496	0.380	1.399	0.116	3.543	0.828
FNM SI	RMA	3.281	3.277	0.739	0.912	0.071	2.243	0.514
FNM AP	RMA	3.069	3.055	0.114	1.024	0.078	2.486	0.927
PFD AP	RMA	3.252	3.253	3.112	0.964	0.074	2.354	0.009
FNM AP	RMA	3.069	2.563	105.943	1.875	0.489	14.236	0.001
Felids								
FHD SI	RMA	3.232	3.261	1.983	0.276	0.021	0.679	0.001
FHD AP	RMA	3.249	3.240	0.041	0.159	0.013	0.387	0.838
FH Dpt	RMA	3.046	3.084	3.016	0.506	0.040	1.214	0.001
PFB	RMA	4.040	4.041	1.294	0.258	0.020	0.618	0.001
FBNL	RMA	3.865	3.847	0.310	0.289	0.024	0.706	0.393
PFD ML	RMA	3.094	3.108	2.018	0.392	0.030	0.966	0.001
PFD AP	RMA	2.841	2.839	0.063	0.400	0.031	0.983	0.897
FNAL	RMA	3.410	3.222	8.897	1.103	0.077	2.703	0.001
FNM SI	RMA	3.341	3.312	0.375	0.337	0.027	0.825	0.379
FNM AP	RMA	3.043	3.222	16.684	1.674	0.139	4.113	0.001
PFD AP	RMA	2.841	2.895	3.724	0.620	0.046	1.513	0.001
FNM AP	RMA	3.043	2.615	29.068	2.842	0.220	0.696	0.001

FHD SI/AP, superoinferior/anteroposterior diameter of the proximal articulation; FH Dpt, mediolateral diameter of the proximal articulation; PFB, total mediolateral breadth of the proximal femur; FBNL, biomechanical length of the femoral neck; PFD ML/AP, mediolateral/anteroposterior diameter of the proximal diaphysis (distal to the lesser trochanter); FNAL, anatomical length of the femoral neck; FNM SI/AP/ML, superoinferior/anteroposterior/mediolaterial diameter of the femoral neck; RMA, reduced major axis.

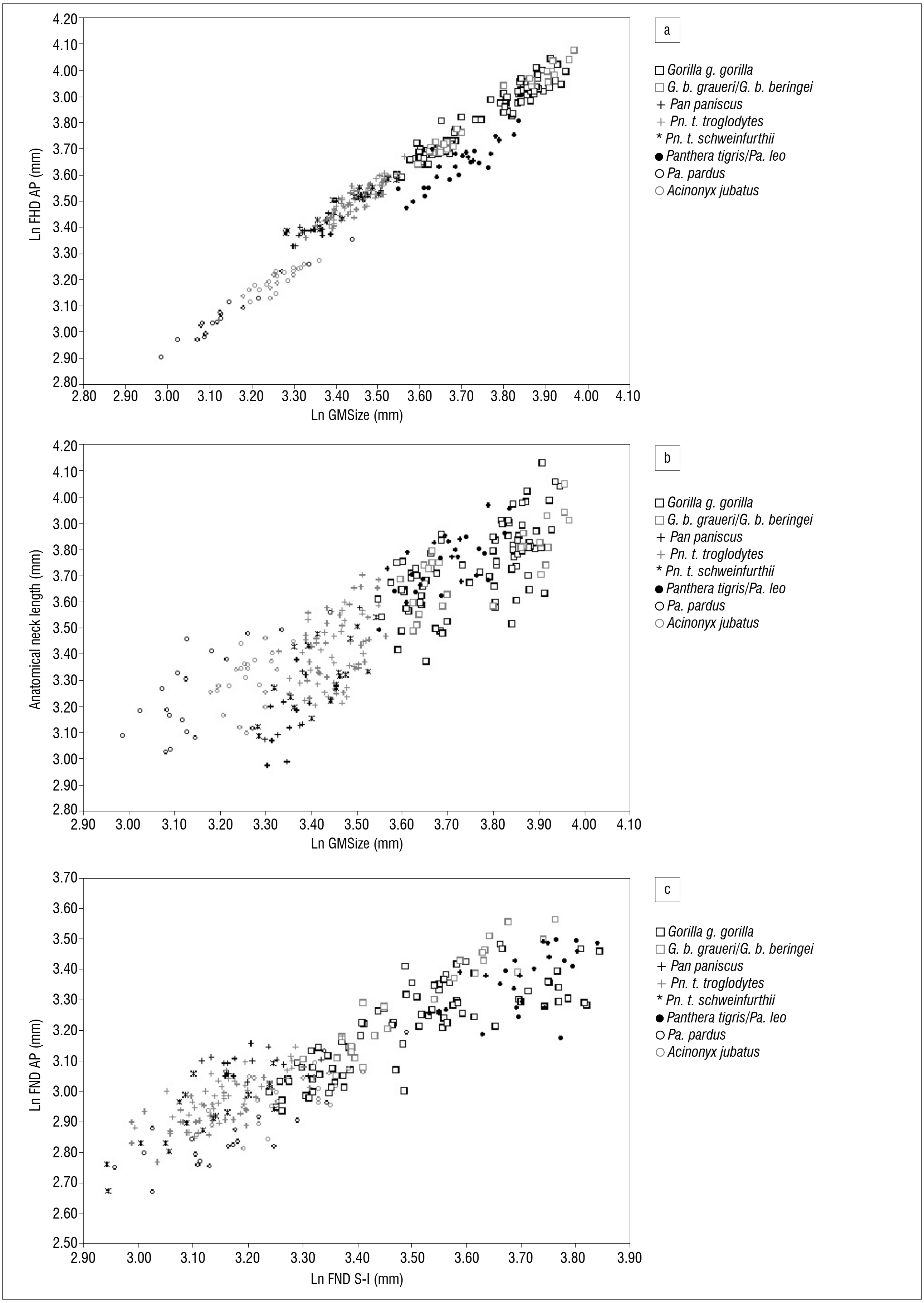


Figure 4: Bivariate scatter plots of selected parameters versus GMSize: (a) femoral head diameter (anterior to posterior); (b) anatomical neck length of the femur and (c) femoral neck (anteroposterior diameter versus superoinferior diameter).

The hypertrophied hip flexors of *Acinonyx*, and presumably of other felid taxa, are critical in the initial power input to the stroke cycle of the proximal segment of the hindlimb during deceleration of the leading limb prior to foot contact, whereas the extensors of the hip and knee ('stifle') joints are equally critical immediately after initial ground contact and the spike in the vertical component of the ground reaction force as the entire limb is extended. The extrinsic muscles of the felid hip, particularly the gluteals, have a broad insertion across the posterior aspect of the proximal femoral diaphysis,[1,2,55,56,58] whereas a principal intrinsic hip flexor, the *M. Psoas* muscle, is extremely well developed in *Acinonyx* compared with greyhounds. The *M. Psoas* muscle and the extrinsic hip-flexors are critical components of the accelerating leading limb during the areal phase of peak velocity galloping. Hypertrophy of the hip flexors and extensors associated with the proximal femoral diaphysis of felids, coupled with an elongate anatomical neck to increase their mechanical advantage,[18,55-58] would effectively reduce, rather than increase, anteroposterior bending moments in the proximal shaft and femoral neck and would account for the distinctive 'pinching' compared with the morphology of *Gorilla* and *Pan*. Functional and structural factors thus mediate the limited morphological distinctions observable in the proximal femur of these two extant mammalian families.

While the biomechanical significance of the proximal femur and its associated musculature cannot be overstated, it is equally critical to remember that osseous structures, particularly the diarthroidal joints, are neither designed nor precision engineered[15,16,21,24,26,27], they emerge via a complex process of axial cellular differentiation at the growth plate[7-10,59]. Recent advances have shed considerable light on the developmental emergence, postnatal growth and underpinning cellular signalling mechanisms that regulate the emerging growth plate and its complexities.[7,9,10,59] Among living mammals, the emerging proximal femoral metaphysis comprises a singular basal growth plate, with discreet secondary ossification centres, either singular or dual, for the articulation and the greater trochanter.[10,60] *Pan* and *Homo* both share separate ossification centres for the articular head and greater trochanter, yet the cellular dynamics of the intervening proximal aspect of the femoral neck are subject to functionally distinct loading patterns which ultimately mediates absolute thickness of the cortex in this region.[6,10] The role of androgens and oestrogens in the cellular dynamics of the proximal femur are also likely critical in the observed sex-specific variance in absolute and relative neck length in humans and other mammals and, by extension, the reduced mechanical integrity and susceptibility to fracture with advanced age.[59,60] Results of this study suggest that the femoral neck of the living African apes is a worthy target for future research.

Results of this study confirm the prospectus that the proximal femur of two distinct families of mammals is extremely generalised in its external form; morphometric consistencies in extant African hominids and large-bodied felids considerably outweigh their observed distinctions. Further to this finding, these results underscore the necessity of a broader comparative context in understanding the nature of form and function in the mammalian postcranium. Nevertheless, where significant differences in proximal femoral form do occur – primarily in the relative size of the proximal articulation and in the anteroposterior/superoinferior and anteroposterior/mediolateral proportions of the neck and proximal diaphysis, respectively – these structural disparities are consistent with the mechanical requirements of their distinct locomotor modes and concomitant habitual stresses upon the proximal hindlimb. Further research into the morphology and functional significance of observed convergences and disparities in proximal femoral forms in other mammalian taxa can only increase our understanding of the underlying mechanisms and adaptive potential of this region of the vertebrate skeleton.

Acknowledgements

I would like to thank the following individuals and Institutions for their assistance and courtesy during the collection of data used in this study: Ms D. Hamerton (Iziko Museum of Natural History, Cape Town); Dr W. van Neer, Dr E. Gilissen and Mr W. Wendelin (Royal Museum of Central Africa, Tervuren); Dr Y. Haile-Selassie, Dr B. Latimer and Mr L. Jellema (Cleveland Museum of Natural History, Cleveland); and Dr S. Badenhorst (Ditsong Museum of Natural History, Tshwane), and the various Collections Managers and the Trustees (Powell-Cotton Museum, Birchington) for access to the specimens in their care and for their generous hospitality. This research was funded by the Natural Environment Research Council (UK), The Ford Foundation (UK), The Palaeontology Scientific Trust (RSA), the Faculty of Health Sciences, University of the Witwatersrand (RSA), and the Faculty of Humanities, University of Johannesburg (RSA). I would also like to acknowledge Prof. Alan G. Morris and three anonymous reviewers for their helpful and thoughtful critique which improved this submission.

References

1. Wagner R. Elements of the comparative anatomy of the vertebrates. New York: JS Redfield; 1845.
2. Mivart G. St. The cat. An introduction to the study of backboned animals, especially mammals. New York: Schribner & Sons; 1881.
3. Reynolds SH. The vertebrate skeleton. Cambridge: Cambridge University Press; 1897.
4. Robinson JT. Early hominid posture and locomotion. Chicago, IL: Chicago University Press; 1972.
5. Lovejoy CO, Heiple KG, Burstein AH. The gait of *Australopithecus*. Am J Phys Anthropol. 1973;38:757-780. http://dx.doi.org/10.1002/ajpa.1330380315
6. Ohman JC, Krochta TJ, Lovejoy CO, Mensforth RB, Latimer B. Cortical distribution of the femoral neck in hominoids: Implications for the locomotion of *Australopithecus afarensis*. Am J Phys Anthropol. 1997;104:117–131. http://dx.doi.org/10.1002/(SICI)1096-8644(199709)104:1<117::AID-AJPA8>3.0.CO;2-O
7. Murray PDF. Bones. Cambridge: Cambridge University Press; 1936.
8. Pearson OM, Lieberman DE. The aging of Wolff's 'law': Ontogeny and responses to mechanical loading in cortical bone. Ybk Phys Anthropol. 2004;47:63–99. http://dx.doi.org/10.1002/ajpa.20155
9. Currey JD. Bones: Structure and mechanics. Princeton NJ: Princeton University Press; 2006.
10. Lovejoy CO, Meindl RS, Ohman JC, Heiple KG, White TD. The Maka femur and its bearing on the antiquity of human walking: Applying contemporary concepts of morphogenesis to the human fossil record. Am J Phys Anthropol. 2002;119:97–133. http://dx.doi.org/10.1002/ajpa.10111
11. Hildebrand M. Symmetrical gaits of Primates. Am J Phys Anthropol. 1967;26:119–130. http://dx.doi.org/10.1002/ajpa.1330260203
12. Demes B, Larson SG, Stern JT Jr, Jungers WL, Biknevicius A, Schmitt D. The kinetics of Primate quadrupedalism: 'Hindlimb drive' reconsidered. J Hum Evol. 1994;26:353–374. http://dx.doi.org/10.1006/jhev.1994.1023
13. Preuschoft H. Mechanisms for the acquisition of habitual bipedality: Are there biomechanical reasons for the acquisition of upright bipedal posture? J Anat. 2004;204:363–384. http://dx.doi.org/10.1111/j.0021-8782.2004.00303.x
14. Schmitt D, Cartmill M, Griffin TM, Hanna JB, Lemelin P. Adaptive value of ambling gaits in Primates and other mammals. J Exp Biol. 2006;209:2042–2049. http://dx.doi.org/10.1242/jeb.02235
15. Edwards E. Modern American locomotive engines: Their design, construction, and management. Philadelphia, PA: Carey Baird & Co; 1883
16. Meyer JGA. Modern locomotive construction. New York: John Wiley & Sons; 1892.
17. Usherwood JR, Wilson AM. No force limit on greyhound sprint speed. Nature. 2005;438:753–754. http://dx.doi.org/10.1038/438753a
18. Hudson PE, Corr SA, Wilson AM. High-speed galloping in the cheetah (*Acinonyx jubatus*) and the racing greyhound (*Canis familiaris*): Spatio-temporal and kinetic characteristics. J Exp Biol. 2012;215:2425–2434. http://dx.doi.org/10.1242/jeb.066720
19. Nigg BM, Herzog W. Biomechanics of the musculo-skeletal system. 2nd ed. New York: John Wiley & Sons; 1999.
20. Pauwels F. Biomechanics of the locomotor apparatus: Contributions to the functional anatomy of the locomotor apparatus. Berlin: Springer; 1980.

21. Ruff CB. Biomechanics of the hip and birth in early *Homo*. Am J Phys Anthropol. 1995;98:527–574. http://dx.doi.org/10.1002/ajpa.1330980412

22. Rafferty K. Structural design of the femoral neck in Primates. J Hum Evol. 1998;34:361–383. http://dx.doi.org/10.1006/jhev.1997.0202

23. Beer FP, Johnston ER, DeWolf JT, Mazurek DF. Mechanics of materials. 6th ed. New York: McGraw Hill; 2012.

24. Lovejoy CO, Burstein AH, Heiple KG. The biomechanical analysis of bone strength: A method and its application to platycnemia. Am J Phys Anthropol. 1976;44:489–506. http://dx.doi.org/10.1002/ajpa.1330440312

25. Paul JP. Forces transmitted by joints in the human body. Proc Inst Mech Engin. 1967;181:8–15.

26. Paul JP. Approaches to design: Force action transmitted by joints in the human body. Proc R Soc Lond B-Bio 1976;192:163–172. http://dx.doi.org/10.1098/rspb.1976.0004

27. Jungers WL. Relative joint size and hominoid locomotor adaptations with implications for the evolution of hominid bipedalism. J Hum Evol. 1988;17:247–265. http://dx.doi.org/10.1016/0047-2484(88)90056-5

28. Schultz AH. The skeleton of the trunk and limbs of higher primates. Hum Biol. 1930;3:303–321.

29. Schultz AH. Proportions, variability, and asymmetries of the long bones of the limbs and the clavicles in man and apes. Hum Biol. 1930;9:281–328.

30. Demes B, Günther MM. Biomechanics and allometric scaling in Primate locomotion and morphology. Folia Primatol. 1989;53:125–141. http://dx.doi.org/10.1159/000156412

31. Eisenberg JF. The mammalian radiations: An analysis of trends in evolution, adaptation, and behaviour. Chicago, IL: Chicago University Press; 1990.

32. Fleagle JG. Primate adaptation and evolution. 2nd ed. London: Academic Press; 1999.

33. McHenry HM, Corruccini RS. The femur in early human evolution. Am J Phys Anthropol. 1978;49;473–488. http://dx.doi.org/10.1002/ajpa.1330490407

34. Gould SJ. Allometry and size in ontogeny and phylogeny. Biol Rev. 1966;41:587–640. http://dx.doi.org/10.1111/j.1469-185X.1966.tb01624.x

35. Mosimann JE. Size allometry: Size and shape variables with characterisations of the lognormal and gamma distributions. J Am Statist Ass. 1970;65:930–945. http://dx.doi.org/10.1080/01621459.1970.10481136

36. Jolicoeur PF. The multivariate generalisation of the allometry equation. Biometrics. 1963;19:497–499. http://dx.doi.org/10.2307/2527939

37. Blackith RE, Reyment RA. Multivariate morphometrics. New York: Academic Press; 1970.

38. Joliffe IT. Principal components analysis. 2nd ed. New York: Springer; 2002.

39. Hammer Ø, Ryan PD, Harper DAT. PAST: Palaeontological statistics software package for education and data analysis. Palaeont Electr. 2001;4(1), 9 pages. Available from: http://folk.uio.no/ohammer/past

40. Flury BK. Common principal components and related multivariate procedures. New York: Wiley; 1988

41. Phillips PC, Arnold SJ. Hierarchical comparison of genetic variance-covariance matrices. I. Using the Flury hierarchy. Evolution. 1999;53:506–1515. http://dx.doi.org/10.2307/2640896

42. Fisher RA. The design of experiments. 8th ed. Edinburgh: Oliver and Boyd; 1966.

43. Jadwisczcak P. Rundom Projects 2; 2003. Available from: http://pjadw.tripod.com

44. Isler K, Payne RC, Günther MM, Thorpe SKS, Yu L, Savage R, et al. Inertial properties of hominoid limb segments. J Anat. 2006;209:201–218. http://dx.doi.org/10.1111/j.1469-7580.2006.00588.x

45. Raichlen DA, Pontzer H, Shapiro LJ, Sockol MD. Understanding hindlimb weight support in chimpanzees with implications for the evolution of Primate locomotion. Am J Phys Anthropol. 2009;138:395–402. http://dx.doi.org/10.1002/ajpa.20952

46. Crompton RH, Vereecke EE, Thorpe SKS. Locomotion and posture from the common hominoid ancestor to fully modern hominins, with special reference to the last common panin/hominin ancestor. J Anat. 2008;212:501–543. http://dx.doi.org/10.1111/j.1469-7580.2008.00870.x

47. Jungers WL. Aspects of size and scaling in Primate biology with special reference to the locomotor skeleton. Am J Phys Anthropol. 1984;27:73–97. http://dx.doi.org/10.1002/ajpa.1330270505

48. Payne RC, Crompton RH, Isler K, Savage R, Vereecke EE, Günther MM, et al. Morphological analysis of the hindlimb in apes and humans. I. Muscle architecture. J Anat. 2006;208:709–724.

49. Payne RC, Crompton RH, Isler K, Savage R, Vereecke EE, Günther MM, et al. Morphological analysis of the hindlimb in apes and humans. II. Moment arms. J Anat. 2006;208:725–742. http://dx.doi.org/10.1111/j.1469-7580.2006.00564.x

50. Harmon EH. The shape of the hominoid proximal femur: A geometric morphometic analysis. J Anat. 2007;210:170–175. http://dx.doi.org/10.1111/j.1469-7580.2006.00688.x

51. Holliday TW, Hutchinson VT, Morrow MB, Livesay GE. Geometric morphometric analyses of hominid proximal femora: Taxonomic and phylogenetic considerations. Homo – J Comp Hum Biol. 2010;61:3–15. http://dx.doi.org/10.1016/j.jchb.2010.01.001

52. Von Cramon-Taubadel N, Frazier BC, Lahr MM. The problem of assessing landmark error in geometric morphometrics: Theory, methods, and modifications. Am J Phys Anthropol. 2007;134:24–35. http://dx.doi.org/10.1002/ajpa.20616

53. Day LM, Jayne BC. Interspecific scaling of the morphology and posture of the limbs during the locomotion of cats (Felidae). J Exp Biol. 2007;210:642–654. http://dx.doi.org/10.1242/jeb.02703

54. Brain CK. The hunters or the hunted? Chicago, IL: Chicago University Press; 1981.

55. Hildebrand M, Goslow GC. Analysis of vertebrate structure. New York: John Wiley and Sons; 2002.

56. Kardong KV. Vertebrates: Comparative anatomy, function, and evolution. 6th ed. New York: McGraw Hill; 2009.

57. Williams SB, Tan H, Usherwood JR, Wilson AM. Pitch then power: Limitations to acceleration in quadrupeds. Biol Lett. 2009;5:610–613. http://dx.doi.org/10.1098/rsbl.2009.0360

58. Hudson PE, Corr SA, Payne-Davis RA, Clancy SN, Lane E, Wilson AM. Functional anatomy of the cheetah hindlimb. J Anat. 2011;218:363–374. http://dx.doi.org/10.1111/j.1469-7580.2010.01310.x

59. Vanderschueren D, Vandenput L, Boonen S, Lindberg SK, Bouillon R, Ohlsson C. Androgens and bone. End Rev. 2004;25:389–425. http://dx.doi.org/10.1210/er.2003-0003

60. Serrat MA, Reno PL, McCollum MA, Meindl RS, Lovejoy CO. Variation in mammalian proximal femoral development: Comparative analysis of two distinct ossification patterns. J Anat. 2007;210:249–258. http://dx.doi.org/10.1111/j.1469-7580.2007.00694.x

Note: This article is supplemented with online only material.

Revealing the social face of innovation

AUTHORS:
Tim G.B. Hart[1,2]
Kgabo H. Ramoroka[3]
Peter T. Jacobs[3]
Brigid A. Letty[4]

AFFILIATIONS:
[1]Economic Performance and Development, Human Sciences Research Council, Pretoria, South Africa

[2]Department of Sociology and Social Anthropology, Stellenbosch University, Stellenbosch, South Africa

[3]Economic Performance and Development, Human Sciences Research Council, Cape Town, South Africa

[4]Institute for Natural Resources, Pietermaritzburg, South Africa

CORRESPONDENCE TO:
Tim Hart

EMAIL:
thart@hsrc.ac.za

POSTAL ADDRESS:
Economic Performance and Development, Human Sciences Research Council, Private Bag X41, Pretoria 0001, South Africa

DATES:
Received: 11 Apr. 2014
Revised: 31 Oct. 2014
Accepted: 16 Jan. 2015

KEYWORDS:
social innovation; inclusive development; rural municipalities; actor-oriented approach

HOW TO CITE:
Hart TGB, Ramoroka KH, Jacobs PT, Letty BA. Revealing the social face of innovation. S Afr J Sci. 2015;111(9/10), Art. #2014-0126, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20140126

Despite the introduction of social innovation in the *1996 White Paper on Science and Technology*, the concept of social innovation has not been actively implemented or even diffused outside of the policy arena in South Africa. Perceptions about what the concept of social innovation should encompass are contested and range from ideas of social welfare outcomes, public goods and a primary focus on the poor. More recently, the emphasis has been on inclusive development that embraces and supports the poor as innovators and which incorporates elements of social and economic development. While contestation in terminology persists, evidence from South Africa's rural areas suggests that although there may be limited state intervention, hampered by structural constraints, and limited understanding of contemporary ideas about innovation and social innovation, local actors practise a variety of forms of social innovation. In most instances, the purpose is to improve social and economic well-being of the poor. Such innovation activities occur almost as widely and as often as strictly commercially oriented innovation activities. However, it is unclear from observed social innovation practices *who* should benefit from these practices (the poor or everyone), *how* (directly or indirectly) and *when* (immediately or gradually). It is suggested that extensive use of the actor-oriented sociological approach to understanding social dynamics in both science and development can provide a means of understanding the subtleties involved in innovation practices and its use should be adopted to address structural challenges within the National System of Innovation that mediate against the contribution of innovations to the poor for inclusive development.

Introduction

Two years after South Africa's first democratic elections, the National Department of Arts, Culture, Science and Technology introduced the *1996 White Paper on Science and Technology*[1] – a necessary and progressive document that aimed to change the thinking about innovation in South Africa and restructure the country's National System of Innovation (NSI). The White Paper was guided by the experiences of innovation systems in other countries, particularly those in North America and Western Europe, and in emerging economies in Asia and Latin America. It emphasised the changes required to shift South Africa's relatively dysfunctional NSI away from its historical focus on labour-intensive commodity production and the military-industrial complex of the latter apartheid years. Unfortunately, policy implementation and related strategies have struggled:

- to effectively change the structure of the NSI
- to make the NSI more representative of key innovation actors by accommodating the marginalised, the private sector and civil society
- to broaden the understanding around innovation activities and reduce the bias of the traditional focus on technical- and business-oriented outputs
- to expand the ideas about and practices related to social innovation as a means of making the outputs of innovation more relevant to society to encourage inclusive development.

While many of the structural constraints of the system undoubtedly remain, the notion of social innovation is probably the least examined area, being largely overlooked beyond any 'trickle-down' social and economic benefits that arise from technical and business innovations. The 2012 *Final Report of the Ministerial Review of South Africa's Science, Technology and Innovation Landscape*[2] also drew attention to the need to focus on the poor and ensure they benefit from NSI activities, but the report struggles to portray what the idea of social innovation should encompass, its purpose beyond notions of poverty reduction and how to ensure its place in the NSI so that it has far-reaching, direct and positive social outcomes for marginalised members of society.

Despite the trend in recent years to acknowledge the desirability for the use of innovations to improve society at large, as well as specific vulnerable groups within society,[3,4] the meaning and use of social innovation is contested and numerous broad and narrow definitions exist. Their existence makes it challenging to implement such an idea effectively and efficiently into South Africa's science, technology and innovation (STI) landscape. In trying to address this challenge, we ask some pertinent questions and attempt to answer them pragmatically:

- Do we need a special category of innovation such as social innovation?
- Does social innovation require its own specific place in the NSI?
- Are current definitions and understandings of this term useful to policymaking and intervention?
- How could we improve our understanding of innovation generally, and social innovation particularly?

We start by briefly reviewing key innovation policy documents since 1996 to understand how social innovation is interpreted within South Africa's policy environment. To contextualise this understanding more broadly, our focus then shifts to the contemporary global understandings of innovation and social innovation in particular. Using evidence obtained from the Rural Innovation Assessment Toolbox (RIAT) pilot study conducted during 2012 and 2013, we show that understandings of innovation and social innovation are extremely blurred in South

Africa's rural areas because STI policy ideas have not been effectively disseminated to innovators in these areas. Why does this situation arise? The absence of a clear understanding and reasons for differences in innovation perceptions and practices suggests that social innovation cannot be simply included into STI policies with the expectation that doing so will ensure broader societal benefits. Drawing on work from the sociology of science and rural development, we argue that in order to bring about change that will reduce structural obstacles within the NSI, particularly those that mediate against social and economic innovation for and by the poor and vulnerable sectors of society, one needs to explore the roles of actors throughout the innovation system and consider their perceptions and world views and what motivates them. These factors are fundamentally different. The most pragmatic way to do this is to adopt an actor-oriented sociology of innovation approach to understanding the social subtleties in the innovation process. Subtleties include what models are dominant and why; who gets what resources for innovation purposes; and what ideas and products are diffused and encouraged. Greater understanding of existing challenges will enable us to effectively address challenges and encourage the STI landscape to be more inclusive, not only in what it does, but in how it does it.

Innovation policy and strategy in South Africa

The 1996 White Paper recognised the importance of both formal – government, higher education and research institutions, private sector and civil society and informal – households and individuals – actors in the NSI.[1] Highlighting the experiences of developing economies in Asia and Latin America, the 1996 White Paper acknowledged that an exclusive focus on technical- and business-oriented innovation was insufficient and that social innovation should be included in the national innovation strategy. Unfortunately, social innovation was never clearly articulated and was only vaguely considered to relate to innovations that produced or improved social or welfare benefits. This was qualified with a need for a primary focus on the vulnerable sectors of society, particularly the poor and unemployed. Only the first and the last of the six goals indicated in the White Paper made any mention of social innovation, but failed to define this clearly. Goal One emphasised the need to 'establish an efficient, well-coordinated and integrated system of technological and social innovation' while Goal Six acknowledged a need for improved 'support to all types of innovation fundamental to sustainable economic growth, employment creation, equity through redress and social development'[1]. Despite the lack of clear articulation of social innovation, the impression conveyed is that social innovation and development are intertwined in some way.

Succeeding policy documents, such as the *Innovation towards a Knowledge-based Economy: Ten-year Plan for South Africa 2008–2018,*[5] generally underplayed the idea of social innovation in terms of improved social or welfare benefits in favour of highly technical innovations in line with the Global Grand Challenges, through which any benefits to the poor would occur slowly and indirectly at best. As part of the Human and Social Dynamics Global Grand Challenge, the Department of Science and Technology (DST) recognised the importance of human behaviour in relation to science and technology and proposed areas of multidisciplinary research to better understand such behaviour as a means of using and informing technological innovations and contributing to greater global understanding of human behaviour. It is proposed that the reduction of persistent and chronic poverty in the country will transpire through technological intervention in the provision of better and affordable services, such as health, energy, water and sanitation. Information and communication technologies (ICTs) are seen as crucial for information dissemination and education. However, these ideas are not clearly linked to the development of the poor and marginalised and aim instead at macro or national contributions to specific sectors such as agriculture, space technology, energy and the green economy, and security. This 10-year plan is partly a quick reaction to some of the key criticisms of the 2007 Organisation for Economic Cooperation and Development (OECD) *Review of the Innovation Policy of South Africa*[6] and focuses on developing high-level innovation skills in ways that could increase business and research collaboration with OECD countries. The OECD Review noted that 'trickle-down' benefits are insufficient to immediately address unemployment and poverty. The reviewers highlighted the importance for improved prioritising of the NSI's contribution to technology and innovation for poverty reduction strategies and improved coordination in terms of resource provision, development and implementation. However, there is no indication in these two documents of the acknowledgement or relevance of less technologically centred innovations such as social organisation and innovative networks that mitigate risk and increase broader participation in society.

Acknowledging the absence of clarity about social innovation and the poor in previous strategies and policy instruments, the 2012 Ministerial Review specifically aimed to illustrate what the idea of social innovation could encompass.[2] Firstly, the Review Committee equated social innovation with innovation for development (a term itself shrouded in multiple understandings, meanings and practices[7]) and stated that social innovation should address priorities arising from unemployment and poverty. For the Committee, this means that social innovation must have social purposes and involve the full range of societal actors, including the public sector, private sector, civil society and the poor. In some way it must be inclusive. However, this focus appears to be strongly related to national development priorities and fails to emphasise the importance of social innovation to local development.[8,9] Secondly, following observations by rural researchers,[10] the Review Committee also acknowledged that development for the poor should consider not only the poor as consumers of innovation, but also their immense potential for creative and active agency,[2] while noting that structural conditions limit their ability to exercise their agency to the fullest. The recognition of the poor as both innovators and consumers of innovations is in itself an important shift away from the narrow view that innovations benefit the poor by means of technological trickle-down effects.

Rather than adopting a broad definition of social innovation, following Petersen[11], the Review Committee considered the primary focus of social innovation in the South African context to be 'on any appropriate technologies or interventions that can address the challenges of poor communities'[2]. Of course such challenges faced by poor communities and the possible solutions to these are not simply social but include economic, historical, political and spatial considerations. Although not clearly stated, the adoption of Petersen's perspective suggests that the Committee is in fact more concerned with promoting innovations that are inclusive and have a developmental focus, rather than simply focusing on innovations that have welfare and social benefits. In other words, the Committee seems to be promoting innovations for development, which include the poor as actors and beneficiaries, and involve technical and social innovations with poverty mitigating economic and social outcomes. The remainder of the Review Committee's discussion on social innovation concentrates on listing current and potential flagship projects, strategies and organisations across civil society, the private and the public sectors who could participate in such projects.[2] Perhaps the most salient features of the Review Committee's discussion is the acknowledgement that the poor are themselves creative and innovative actors and agents in their livelihood and social improvement strategies, but that these activities are often constrained. Unfortunately, ways to overcome fundamental constraints, such as access to resources as a result of prevailing structural conditions within the NSI and society, are not addressed by the Review Committee, possibly because there is limited awareness of the multiple constraints, and especially how these are manifested amongst marginal groups and actors. Also, the Committee seems to confuse social innovation with processes and outcomes that are perhaps better-termed pro-poor innovation given that the focus is not simply on welfare and well-being but more on the role technological products can play with regard to creating economic and social improvements in the lives of the poor and marginalised. This perspective is only one perspective on social innovation.

What do we understand innovation to be?

Increased interest in and research about innovation has acknowledged the importance of innovation in service and low-technology manufacturing industries, i.e. those outside the mainstream research and development intensive industries, such as high-technology manufacturing.[3,12-14] Such acknowledgement has led to broadening the definition of innovation to one that moves beyond traditional industries and the traditional approach of concentrating on technological product and process innovation. The result is the consideration of social arrangements or organisational structures and social outcomes or products that are equally and fundamentally important for an innovative economy and society as a whole.

According to the third edition of the *OSLO Manual*[12] and a review of innovation by Gault[13], the contemporary understanding of innovation is that it involves both processes and the outputs of these processes. The innovation processes are generally accepted to consist of four activities: adoption (the use of innovations), adaption (the improvement of innovations), diffusion (the sharing or transfer of innovations) and invention (the creation of new innovations).[12] These activities need not be linear although they can be. The outputs of the innovation processes are now generally agreed to include four main types: product (goods and services) innovations, process innovations, marketing strategies and organisational arrangements.[12,13] In order to be considered an innovation, the product, process, marketing strategy or organisational arrangement must at least be novel to the user, and must be valorised or wanted.[13] Novelty or newness need not extend beyond the first-time user to society[15], while traditionally, value has been couched in commercial terms in the sense that the innovation exclusively improves profits or improves processes that in turn improve profits. Value can also be simply the improvement of knowledge, such as in the case of research undertaken by different disciplines for immediately contributing to the body of knowledge, improving methods to do so, or having the expectation of contributing to social and economic needs over the long term. It is probably only since the turn of this century that social value (such as improving well-being and welfare) has been acknowledged as a means to valorise innovation.[16] Very simply, value implies usefulness to the user – the more useful, the greater the value.

Despite the multiple and contrasting ideas about innovation and how to achieve greater impacts from innovation, some neoliberal economists still argue that the benefits of innovation ultimately trickle-down indirectly to the most needy and that therein lies the social and economic value of all types of innovations, irrespective of purpose. For example, innovative firms can employ a small number of extra people to develop products that generate revenue as well as improve other aspects of society, such as lowering the carbon footprint or providing improved medicines, transportation systems or communication systems – activities that can be considered socially innovative. However, the increasing awareness of the need for more direct benefits for the large number of marginalised and vulnerable members of society, and their role as innovators, has resulted in the call for more direct socially focused innovations.

Contemporary understandings of social innovation

A review of recent international literature on social innovation[17] suggests that although the term is contested, there are three primary definitions of social innovation used globally. However, each definition has its own dissenters who propose further qualifications to emphasise what should be categorised as a social innovation.

Firstly, social innovations are largely considered to be products (goods and services) with human welfare or social benefit outcomes, including better health, education, improved water access, cost-efficient energy devices and products that improve communication and transportation. The output is socially oriented and for our purposes, we term this the social product definition. However, some emphasise that to be considered as social innovations, such products must be social and public goods.[18] In this regard, innovations in the private sector, such as vaccines, are excluded, as they are not public goods although they have social benefit. Others argue that Internet search engines developed in the private sector and owned by private enterprises are social innovations because the value to society outweighs profits to private sector.[19]

The second definition of social innovation considers the organisation or arrangement of people and things within enterprises or social settings (informal or formal organisations and arrangements).[17] There is a social process. Examples include trade unions, bargaining councils, worker forums, job sharing, *stokvels*, neighbourhood watch committees, rural neighbourhood or kin-based work parties, grazing and land-management committees, and even various product distribution and sales methods. We term this the social collaboration definition.

The third definition of social innovation is a combination of the first two. Social innovations are those new products, services, models and practices that concurrently meet social requirements and involve new social collaborations. To be considered social innovations under this definition, innovations must have both a social means and end, in that potential recipients must decide what has to be done and do it (but can draw on external resources and advice). We term this the social means and ends definition. Furthermore, such innovations must achieve broad systemic transformation,[14] in the sense that the prevailing structure of the innovation system (global, national, regional and local) is altered and improved. However, like the first definition, innovations occurring in commercial enterprises are excluded. A social process is used to bring about a social outcome and it is the presence of new social processes that catalyse systemic change.

However, is it correct, or even relevant and helpful, to distinguish innovations with social purposes and means from other innovations, purely based on the processes involved and the intended outcomes, benefits and beneficiaries? Clearly, given the numerous examples, the social collaboration definition focuses on the social arrangement of people irrespective of their location in formal or informal commercial or social enterprises and the formal or informal nature of the social arrangement. Consequently, if we are to talk about social innovation, it seems rather naïve to ignore social innovations occurring in commercially oriented enterprises[20], irrespective of the scale of such enterprises[21]. The sustained presence of social arrangements suggests that they must have benefits for all actors involved. Evidence from the RIAT pilot study suggests that perhaps we should not be too quick to narrow the parameters of social innovation.

Evidence from innovating enterprises in South Africa

Using a purposive snowball sampling technique as part of the RIAT pilot study in four South African rural district municipalities (RDMs), we formally interviewed representatives from 482 formal and informal enterprises using a structured questionnaire which comprised qualitative and quantitative questions. The methodology was approved by the Research Ethics Committee of the Human Sciences Research Council (protocol no. REC5/24/04/13). Identified enterprises were screened by interviewees and had to have been engaged in at least one of the four innovation activities during 2011 or 2012. Enterprises from the primary sector (agriculture, forestry, mining and minerals) accounted for 30% of the sample. Those from the secondary sector (manufacturing and energy) accounted for 16% of the sample and those from the tertiary sector (providers of tertiary services including ICT, health, education, finance and community services) accounted for 54%. While not conclusive, this evidence suggests that the tertiary sector enterprises are important innovation actors in the sampled RDMs. Although the public enterprises had the largest share of actors (71%) in the tertiary sector, almost half of the private (49%) and non- profit (48%) enterprises were also involved in this sector. This evidence suggests that social innovation, at least in terms of providing basic services and welfare benefits, is undertaken by a significant share of enterprises based in the RDMs.

Main innovation activities

Most respondents perceive that innovation is something new but should involve technology and improve the revenue of the user. Such ideas reaffirm traditional perceptions of innovation being linked to technology and business. Broad ideas of innovation activities were not initially acknowledged by respondents but when directly asked about innovation activities, we see in Table 1 that of those respondents involved in innovation activities during 2012, many were actively engaged in activities such as adoption (53%), adaption (29%) and diffusion (24%) of existing innovations. Only a handful of rural-based enterprises were engaged in the invention of new innovations (7%). Public enterprises and non-profit enterprises, many of which are involved in the community services sector, tended to be more active with regard to adopting and diffusing innovations. On the other hand, the private sector enterprises, which are mainly profit driven, tended to be more involved in adoption and subsequent adaption.

Awareness of science, technology and innovation policy and support

Only 28% of the enterprises acknowledged an awareness and understanding of South Africa's STI policies. Often this knowledge was based on specific sector-related polices, such as those emanating from specific line departments other than DST, rather than the national strategy as a whole. Almost two-thirds (63%) of the enterprises investigated were aware of government support for innovation activities in the private sector. However, the most aware were the non-profit enterprises (73%). Slightly more than a third (38%) of all enterprises had applied for government support, with the private sector enterprises having the lowest share of applicants (22%) and the non-profit enterprises having the highest share, at 59%.

Social innovation awareness

The awareness of social innovation amongst these rural-based enterprises was even lower than their understanding of STI policies. In Table 2 we see that only 22% of the respondent rural enterprises were aware of the concept of social innovation. The greatest share of awareness (36%) was found among public enterprises and less than 20% of both private and non-profit organisations had any awareness of the concept. These figures are fairly dismal, but reinforce the fact that social innovation has not been actively articulated or disseminated since its initial inclusion into South African innovation policy in 1996.

Reasons for innovating

Table 3 illustrates the responses of the enterprises with regard to their main reasons for innovating. Commercial purposes include increasing profits of enterprise and market share, meeting subsistence and survival needs, and innovation to improve the body of knowledge (activities undertaken by research institutes). Social improvement includes products, services and arrangements that directly improve society and the poorer members in particular, and is involved in research with this focus. While 2% of the respondents were uncertain why their enterprises were innovating, more than half (56%) reported that this was for commercial purposes. The majority of public sector enterprises (76%) were innovating for social improvement purposes and the majority of private sector enterprises (86%) were innovating for commercial purposes. Interestingly, the gap between these purposes is not so great for the non-profit enterprises although the greatest share (56%) focused on social improvement. This gap might be because of the high level of competition for resources to provide services in this sector and the resultant need to supplement grant income in creative ways – ways that require innovations that increase income. In addition, it might reflect the fact that some non-profit enterprises, especially those linked to government projects and

Table 1: Share (%) of enterprises engaged in innovation activities by enterprise type during 2012

Innovation activity		Valid observations†	Share (%) of public enterprises (*n*=96)	Share (%) of private enterprises (*n*=201)	Share (%) of non-profit enterprises (*n*=179)	Share (%) of all enterprises innovating during 2012 (*n*=476)
Invent	Yes	34	7	7	7	7
	No	441	93	93	93	93
Adopt	Yes	252	45	58	52	53
	No	224	55	42	48	47
Adapt	Yes	139	20	37	26	29
	No	338	80	63	74	71
Diffuse	Yes	116	29	19	28	24
	No	359	71	81	72	76

†Valid observations refers to the number of non-missing values

Table 2: Share (%) of enterprises aware of the term social innovation by enterprise type

Aware of social innovation	Share (%) of public enterprises (*n*=97)	Share (%) of private enterprises (*n*=202)	Share (%) of non-profit enterprises (*n*=179)	Share (%) of all enterprises (*n*=478)
Yes	37	19	17	22
No	63	81	83	78

Table 3: Share (%) of enterprises engaging in innovation activities for commercial or social welfare purposes by enterprise type

Main purpose of innovation activities	Share (%) of public enterprises (*n*=98)	Share (%) of private enterprises (*n*=202)	Share (%) of non-profit enterprises (*n*=182)	Share (%) of all enterprises (*n*=482)
Commercial purposes	22	86	41	56
Social improvement purposes	76	13	57	42
Uncertain	2	1	2	2

community groups, aim to generate an income for members as part of poverty-reduction strategies.

Local understandings of social innovation

Local understandings of social innovation are often contested and do not always coincide with those used by policymakers and researchers. Responses to in-depth qualitative questions indicate that a fair number of respondents representing rural enterprises believe they are involved in innovation activities that have social or welfare purposes, in the sense that there is a very direct link to providing or improving social services, addressing community needs and helping others in their immediate proximity. This belief was particularly so for respondents from public and non-profit enterprises who mentioned targeting the poor, less fortunate and marginalised. However, some of the respondents representing private enterprises perceived their profit-making products, processes and strategies as being social innovations in that these ultimately have a social benefit, even if only indirectly.

Some farmers indicated that the use of improved seeds, plant material and inputs that improved food quality and availability improved national food security, even if the use of these technological innovations ultimately increased food prices. Similarly, some credit providers argued that the use of innovative microfinance arrangements that ensure the repayments of loans, rather than their affordability, were also social innovations in that by ensuring repayment they were ultimately able to provide credit to more people. Others suggested that creating a few jobs for other people, while significantly increasing the income of the innovating entrepreneur, were also social innovations because these activities enabled others to earn an income, thereby reducing unemployment, even if this income was far below that of the entrepreneur.

Undoubtedly, the above are innovations – but are they social innovations? While the social impacts of these examples are very indirect and at best have a limited and gradual effect on the well-being of the less fortunate, they also have potentially far-reaching negative effects. These negative effects include raising the price of foodstuffs and decreasing farm employment opportunities, increasing the debt of those whom can ill afford it and increasing the number of low-paid jobs. Although there are some elements of social benefit, the primary aim of such 'social innovations' still appears to be the immediate improvement of enterprise turnover and revenue, and as a result there is no tangible benefit for the poor.

In these four RDMs, the perceptions of innovators about the purpose and beneficiaries of social innovation are blurred, although there is an indication that the poor or less fortunate should benefit more directly – a view mainly expressed by public and non-profit sector organisations. Given this context, to simply introduce a new and rather broad concept, such as social innovation, into South Africa's innovation system is likely to create more problems than it solves. Already there are numerous terms used in the development discourse that are simply buzzwords adopted by actors and agencies to promote and generate support for their ideals. These terms include empowerment, pro-poor, participation and poverty reduction. Unfortunately, they have multiple meanings and can be used and translated by actors and agencies as they desire and for very different purposes.[22] The concept of social innovation, with its multiple foci and qualifications faces similar challenges.[23]

Towards an actor-oriented sociology of innovation

At present, South African innovation policy tends to lump social innovation with pro-poor development, focusing on products that better serve the poor, and closely linking the definition of social innovation to a social product. We see from the international literature that this definition is insufficient.[20,21,23] It ignores innovation as social collaboration – i.e. innovative social processes. It also excludes the broader product and means definition, with its structural altering intentions. However, our research indicates that clear definitions of social innovation are not possible, as they do not exist neatly in South African rural areas. Furthermore, innovation of a social nature is not simply confined to producing social goods and services exclusively for the poor. Alarmingly, while innovation occurs in rural areas, very little seems to be immediately linked to or acknowledged by the broader system of innovation unless it is highly formalised. This is especially the case for non-traditional innovations and those undertaken by the poor in marginal areas. These factors make it necessary to consider the prevalence of structural challenges that possibly permeate the NSI and mediate against increased innovation in rural areas. A simple focus on social innovation in any of the three definitions is inadequate. The focus must be broader in our view, so that it includes both social and economic purposes of innovation, as well as technological and social processes and products. Currently any 'labelling' of something as a social innovation runs the risk of confining it too simply to the social product definition.

Similar to Neumeier[23], we propose that a first step out of the existing policy and real-world impasse would be to adopt a sociological actor-oriented approach to understanding the social life of innovation generally, irrespective of the social, technical or economic outputs, purposes and class of beneficiaries. An actor-oriented sociological approach to innovation would enable a better understanding of the social, cultural, political, economic, historical and structural factors that influence innovation activities and the capability to innovate. The actor-oriented works of Latour[24] on the sociology of science, Long[25] on the sociology of rural development and Mosse and Lewis[26] on local and international development, provide relevant insights into the 'social life' or 'real world' of science and development projects. By focusing on the actors we are able to understand what is taking place and why, along with the relationships between the various actors involved at different levels of the global, national, regional and local innovation system and innovation activity.[24] A similar sociological perspective of innovation would show that actors are often in conflict with one another over resources and have different perceptions on needs and contrasting world views, despite outward appearances of collaboration.[23,25] Inconsistencies in processes and support are revealed when often there first appear to be none, deepening our understanding of the dynamic and changing relationships between people and people, and people and things.[24]

Any discussion about the social qualities of innovation must also look at the social process or life of innovation. This is a necessary first step to understanding innovation processes, challenges and contributions to society. To ignore this would be to overlook a crucial element of innovation, i.e., it is not simply a technocratic process, but one that is governed by political, social, economic and historical factors. Most innovations have some degree of social benefit, either for immediate users or for others who are indirect beneficiaries. This characteristic enables people to suggest that almost all innovations have a social benefit. As noted in the current study, some innovations may have negative long-term effects that are not immediately acknowledged or realised. It is important to understand why some innovations are acknowledged, encouraged and supported, rewarded and diffused while others are simply ignored or unrecognised as innovations. This necessitates a deep understanding of the actors involved at different tiers of the innovation system.

The RIAT pilot study shows that, in rural areas, awareness of STI policy by innovating enterprises is low, especially innovations with social objectives, and that while awareness of the availability of support is relatively high, the support seeking activities are low. The study also suggests that a significant share of innovation takes place in the tertiary or service sector. However, innovation in rural areas is generally equated with agricultural innovation and less so with mining (the two key actors in the primary sector). Why does innovation in the tertiary sector remain unacknowledged? Invention is low in rural areas but there is no clear explanation for this unless it has to do with how resources are currently distributed within the NSI. Innovations with social foci are not as prevalent as those with a commercial focus but comprising 42% of the sample they are significant and deserve to be acknowledged. Even if the focus of such innovations is not in line with the multiple current definitions, their contribution to society needs to be a focus of policy research, and better understood.

Conclusions

The evidence from the RIAT pilot study used in this paper shows that there are gaps in policy implementation since the introduction of the *White Paper on Science and Technology* in 1996, while the review of recent policy papers shows that these papers do not sufficiently address the gaps. As Neumeier[23] points out, there is a need to go beyond the simple assessment of rural innovation projects and activities to explicitly investigate and interrogate factors that promote or constrain innovation in the area between top-down and bottom-up rural development approaches. A sociological approach presents a refreshing research focus to understand the social face of innovation activities and outcomes generally, without the obscurity created by multiple definitions of social innovation. It is an approach that is more encompassing than those currently suggested in policy documents and is drawn from evidence of innovating enterprises in marginal rural district municipalities.

An actor-oriented sociological approach to innovation provides a point of departure for a deeper understanding of capabilities and access to resources, decision-making and the reasons for these. Perhaps more importantly, when structural transformation is necessary, as in South Africa, this approach could highlight the existing power relations within the NSI, across the broader STI and development policy-making frameworks, and the effects of these relationships at different levels of these frameworks. Inevitably, this approach will deepen our awareness of who benefits (and who does not) from innovation policies and interventions and why and how this occurs. It would also identify more clearly the social and economic factors that underpin the innovation process, purposes and outcomes more clearly and would put these in a context in which it would be easier to understand the developmental impacts of innovation and innovations broadly without being confined simply to process and outcomes.

In order to determine the broader developmental contributions of innovation, the adoption of an actor-oriented sociology of innovation appears to have much to offer to our understanding of innovation actors, agency, processes and their outcomes along with the nature of the innovation system. This can be considered a necessary step to deepening our understanding of the national, sub-national, regional and sectoral innovation systems in South Africa and lead towards identifying the structural and systemic changes, beyond repetitive quantitative monitoring and the simple creation of new entities and new foci, as proposed in the 2012 Ministerial Review document. Innovation, including social innovation in its broadest sense, cannot be truly understood without a sociological perspective.

Acknowledgements

In this article, we draw on evidence obtained during the Rural Innovation Assessment Toolbox (RIAT) Pilot study undertaken by Economic Performance and Development (EPD) researchers of the Human Sciences Research Council during 2012 and 2013. We acknowledge the financial support of the Department of Science and Technology as well as the collegial inputs and discussions with fellow researchers Brandon Bodenstein, Hlokoma Mangqalaza, Alex Mhula and Makale Ngwenya. The comments and suggestions of the Associate Editor, Pearl Sithole, and two unknown reviewers are deeply appreciated. The views expressed are those of the authors and do not necessarily reflect those of any other party.

Authors' contributions

T.G.B.H. was the lead author who conceptualised the study and conducted the literature review and qualitative analyses. K.H.R. and P.T.J. conducted the quantitative analysis and contributed to rewriting various drafts. B.A.L. assisted in the qualitative and thematic analysis and the rewriting of various drafts.

References

1. DACST (Department of Arts, Culture, Science and Technology). White paper on science & technology: Preparing for the 21st century. Pretoria: Department of Arts, Culture, Science and Technology; 1996.
2. DST (Department of Science and Technology). Department of Science and Technology Ministerial Review Committee on the science, technology and innovation landscape in South Africa: Final report 2012 March. Pretoria: Department of Science and Technology; 2012.
3. Marcelle G. Editorial. Int J Technol Learn Innov Dev. 2012;5(1/2):1–11.
4. Gupta AK. Innovations for the poor by the poor. Int J Technol Learn Innov Dev. 2012;5(1/2):28–39. http://dx.doi.org/10.1504/IJTLID.2012.044875
5. DST (Department of Science and Technology). Innovation towards a knowledge-based economy: Ten-year plan for South Africa (2008–2018). Pretoria: Department of Science and Technology; 2007.
6. Organisation for Economic Cooperation and Development (OECD). OECD reviews of innovation policy: South Africa. Brussels: OECD; 2007.
7. Willis K. Theories and practices of development. New York: Routledge; 2009.
8. Committee for Scientific and Technological Policy, Organisation for Economic Cooperation and Development (OECD). Fostering innovation to address social challenges. Paris: OECD; 2011.
9. Gerometta J, Haussermann H, Longo G. Social innovation and civil society in urban governance: Strategies for an inclusive city. Urban Stud. 2005;42(11):2007–2021. http://dx.doi.org/10.1080/00420980500279851
10. Cousins B. Rural innovation systems. Paper prepared for the Ministerial Review Committee on the science, technology and innovation landscape in South Africa. Pretoria: DST; 2011 [unpublished report].
11. Petersen F. 2011. Specialist report on social innovation and sustainability. Paper prepared for the Ministerial Review Committee on the science, technology and innovation landscape in South Africa. Pretoria: DST; 2011 [unpublished report].
12. Organisation for Economic Cooperation and Development/Statistical Office of the European Communities (OECD/Eurostat). The measurement of scientific and technological activities – Oslo manual: Guidelines for collecting and interpreting innovation data. 3rd ed. Paris: OECD/Eurostat; 2005.
13. Gault F. Innovation strategies for a global economy: Development, implementation, measurement and management. Ottawa: International Development Research Centre; 2010.
14. Young Foundation / Social Innovation eXchange (YF/SIX). Study on social innovation. A paper prepared by the Social Innovation Exchange (SIX) and the Young Foundation for the Bureau of European Policy Advisors [document on the Internet]. c2010 [cited 2012 Dec 06]. Available from: http://youngfoundation.org/wp-content/uploads/2012/10/Study-on-Social-Innovation-for-the-Bureau-of-European-Policy-Advisors-March-2010.pdf
15. Rogers EM. Diffusion of innovations. New York: Free Press; 1995.
16. Moulaert F, Martinelli F, Swyngedouw E, Gonzalez S. Towards alternative model(s) of local innovation. Urban Stud. 2005;42(11):1969–1990. http://dx.doi.org/10.1080/00420980500279893
17. Hart T, Jacobs P, Mangqalaza H. Key concepts in innovation studies – Towards working definitions. RIAT Concept Paper Series – Concept Paper 2. Pretoria: Human Sciences Research Council; 2012.
18. Harris M, Albury D. The innovation imperative. The LAB Discussion Paper, 2009 March. London: NESTA; 2009.
19. Phillis JA, Deiglmeier K, Miller DT. 2008. Rediscovering social innovation. Stanford Soc Innov Rev. 2008;Fall:33–43.
20. Datta PB. Exploring the evolution of a social innovation: A case study from India. Int J Technol Manag Sustain Dev 2011;10(1):55–75. http://dx.doi.org/10.1386/tmsd.10.1.55_1
21. Godoi-deSousa E, Valadao Junior VM. Social enterprises in Brazil: Socially produced knowledge versus social innovation. J Technol Manag Innov 2013;8:166–176.
22. Cornwall A, Brock C. What do buzzwords do for development policy? A critical look at 'participation', 'empowerment' and 'poverty reduction'. Third World Q. 2005;26(7):1043–1060. http://dx.doi.org/10.1080/01436590500235603
23. Neumeier S. Why do social innovations in rural development matter and should they be considered more seriously in rural development research? – Proposal for a stronger focus on social innovations in rural development research. Sociol Ruralis. 2011;52(1):48–69. http://dx.doi.org/10.1111/j.1467-9523.2011.00553.x
24. Latour B. Aramis or the love of technology [translated by Catherine Porter]. Cambridge, MA: Harvard University Press; 1996.
25. Long N. Development sociology: Actor perspectives. London: Routledge; 2001. http://dx.doi.org/10.4324/9780203398531
26. Mosse D, Lewis D. Theoretical approaches to brokerage and translation in development. In: Lewis D, Mosse D, editors. Development brokers and translators: The ethnography of aid and agencies. Bloomfield, CT: Kumarian; 2006. p. 1–26.

The qualification of coal degradation with the aid of micro-focus computed tomography

AUTHORS:
Jacob Viljoen[1]
Quentin P. Campbell[1]
Marco le Roux[1]
Jakobus Hoffman[2]

AFFILIATIONS:
[1]School of Chemical and Minerals Engineering, North-West University, Potchefstroom, South Africa
[2]Necsa – Radiation Science, Pretoria, South Africa

CORRESPONDENCE TO:
Jacob Viljoen

EMAIL:
13037242@nwu.ac.za

POSTAL ADDRESS:
School of Chemical and Minerals Engineering, North-West University, Private Bag X6001, Potchefstroom 2522, South Africa

DATES:
Received: 21 Jan. 2014
Revised: 21 Oct. 2014
Accepted: 17 Jan. 2015

KEYWORDS:
coal microstructure; compressive breakage; impact breakage; thermal breakage; X-ray

HOW TO CITE:
Viljoen J, Campbell QP, Le Roux M, Hoffman J. The qualification of coal degradation with the aid of micro-focus computed tomography. S Afr J Sci. 2015;111(9/10), Art. #2014-0025, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20140025

The production of unwanted coal fines during the handling and utilisation of coal is a serious problem in processes that rely on large or closely sized particles. Coal degradation occurs at many different steps within the beneficiation or utilisation processes and through many different mechanisms, none of which are understood thoroughly. In an effort to describe the degradation mechanisms, the changes within a number of coal particles were tracked using micro-focus X-ray computed tomography (μ-CT). The observed changes were caused by impact loading, compressive loading and thermal shock. The resolution of the μ-CT tomograms enabled the identification and tracking of changes in the coal microstructure. A comparison of the tomograms taken before, during and after breakage and fracture showed that the microstructure of coal had an influence on the breakage characteristics. For impact- and compressive loading as well as during thermal treatment, the biggest structural contributor was shown to be the network of pre-existing cracks and cleats within a particle. Lower density macerals contributed more to breakage than the higher density macerals and any structure (pre-existing cracks, lithotypes boundaries and mineral boundaries) present within the particles had the potential to either act as a crack initiation site, change the direction of a propagating crack or arrest crack propagation. The direction of the applied loads during compressive- and impact loading was the biggest contributor to the directionality of newly formed cracks. For thermal treatment, the vitrinite rich microlithotypes showed more new crack formation compared to the other microlithotypes present. The particles also showed no evidence of devolatilisation (an increase in the porosity of the particle) but did show evidence of thermal drying (new cracks formed perpendicular to existing cracks).

Introduction

One of the biggest challenges in coal beneficiation as well as in coal utilisation is the unwanted production of coal fines, known as degradation.[1,2] The presence of coal fines in processes that rely on a closely sized feed can reduce the process efficiency.[3,4] The presence of fines in a beneficiated product can also have financial implications for the suppliers as many contracts specify penalties for product that does not meet specifications.[1,2,5-7] Fine coal is also hard to clean, relying on wet processes like flotation, spirals and dense medium cyclones to wash the coal.[5,8,9] The wet, fine coal is then significantly harder to dewater,[9-11] and due to the increased moisture content, the transport cost of fine coal is greatly increased.[8,9] The high moisture content can also incur financial penalties for out-of-spec product.[8,11]

Fines generation can occur at any step during the beneficiation process where the coal is mechanically stressed[2,6-8,12,13]: during comminution, screening, conveying, loading, stockpiling and reclaiming.[2,7,11] Fines are also generated when coal particles enter a reactor (be it a pyrolysis or combustion process) due to thermal shock and devolatilisation.[3,14-17] The presence of fines in a large particle reactor can reduce bed permeability[3,13,14,16], process efficiency[15,17-19] and throughput[13,18].

In both the case of physical degradation and thermal fracture, the breakage- and fracture-mechanisms are not well understood.[20,21] Clarification of these mechanisms can help to reduce the production of fines as well as aid the design of better comminution machines.[20,22,23] Over the years, various studies have tried to shed some light on the effect that physical properties have on the breakage characteristics of coal.[1,3,4,7,8,22-25] The effect of comminution machine properties on the breakage characteristics of coal has also been studied.[2,6,7,13,22,24,25]

For the impact breakage of coal it has been found that the drop height (impact energy) increases coal degradation.[6,7,13,22,24,25] The impact surface also influences the degradation – the harder the surface the higher the degradation.[2,6,7] The shape and orientation of the impacted particle has an influence on the breakage of coal[8,22]: rounded particles show less degradation than slab-like particles (particles where one of the three orthogonal dimensions is significantly smaller than the other two). The composition of the coal also plays a role. With an increased amount of vitrinite an increase in degradation will occur.[25] For single particles, bigger particles have lower specific breakage strengths and produce more degradation products.[1,7,22,23,25,26] This is due to bigger particles having more inherent weaknesses, and is referred to as the particle size effect.[4,23,25-27] For samples containing multiple particles and subjected to multiple drops, the removal of fines after each drop increased the total fines generation when compared to the same amount of drops without removal of the fines. This is referred to as cushioning.[2,6,13]

The size effect is also documented for the compressive loading of coal.[27] During uniaxial strength tests, coal samples undergo a reduction in strength as the sample size increases, until a characteristic value is reached in samples with a diameter larger than 1.5 m.[27] This is again due to an increased amount of defects present in the larger samples.[21,27] These defects include micro-cracks, cleats and inclusions. Fatigue is another mechanism that has an effect on the compressive breakage strength of coal.[7,21,23] Fatigue is the repeated application of a force, which although not strong enough to damage a particle, eventually destroys the particle due to damage accumulation within the particle.

For thermal fracture, it was found that an increase in furnace temperature causes the degradation to increase.[17,28] An increase in particle size increases both the size of the progeny as well as the particle count of the progeny.[17,19,28] In addition, the longer the particle stayed in the furnace, the lower the particle count became due to the progeny being burnt away.[17,28] It was also found that an increase in the volatile content of the particles increased the degradation,[15,19] although some authors found no connection between volatile content and fragmentation.[28] Taking this into account, Dacombe et al.[28] postulated that there are two mechanisms responsible for thermal degradation: exfoliation and fragmentation. Exfoliation is the formation of many small progeny from the outside of the particle[15,28] and occurs due to thermal stresses that occur at higher heating rates.[15,17,28] Fragmentation is when the particle centre remaining after exfoliation breaks into larger (when compared to exfoliation) progeny and happens due to the build-up of internal pressure as the volatile matter in the coal particle is released.[15,17,28]

A common theme in all of the degradation scenarios is the presence of micro-defects within the particles. These micro-defects can act as sites where crack propagation can either start or arrest.[29] Micro defects that can occur within coal are mineral inclusions[23,25], pre-existing cracks and cleats[23,25,27,30,31] as well as microlithotype boundaries[23,25]. The macerals that occur in coal can be grouped into three main classes: vitrinite, inertinite and liptinite.[32,33] For coal degradation, vitrinite is the most important one, as it is known to be a very brittle, weak maceral.[25,27] It also has the most flaws present[25,27], as it experiences the highest degree of shrinkage during coalification[32]. Cleats are cracks that are inherent to coal and are formed during the coalification process by the shrinkage and expansion of the coal matrix.[27,30,32,34] The mineral inclusions present in coal are usually in the form of either discrete mineral inclusions[35] or cleat filling minerals[32,35]. The minerals can become part of the coal matrix through detrital deposition[35] and through biogenic-, syngenetic- or epigenetic precipitation.[32,35] Various studies have applied micro-focus X-ray computed tomography to identify these microstructures in coal.[36-42]

X-ray computed tomography (μ-CT) is a technique that was originally developed in the medical sciences where it is commonly used during diagnostic radiology.[31,36,37,43] Since the 1980s, the technique has also been widely used in the geosciences for non-destructive characterisation.[31,36,37,39,43] During the application of the technique, an X-ray cone beam is projected at and through the sample to be analysed. As the beam travels through the sample, the intensity of the beam diminishes due to absorption and scatter from the direct beam path. A detector records the intensity of the beam after it passes through the sample as a radiograph.[30,33,43-45] A number of radiographs, made by rotating the sample along its axis, is combined to form a tomogram, or a three-dimensional representation of the scanned sample, using filtered backprojection.[33,36,40,41,43,44] A detailed mathematical treatment of the technique and reconstruction is beyond the scope of this study. There is however a large body of literature available on the subject.[33,36,38,39,43,46]

The attenuation coefficient or tomo-density determines the absorption and scattering of the X-ray energy from the direct beam path and is a combination of the physical density, mean atomic number of the sample elements and the energy of the X-ray beam.[38,39,45]

Micro-focus X-ray computed tomography (μ-CT) is an improvement over traditional medical CT. These improvements are mostly because there is no limit to the radiation dose that can be applied to samples typically scanned using μ-CT.[36] Higher energies, longer exposure times and higher resolutions are therefore possible.[36] The resolution of a medical CT is in the range of 250 X 250 X 250 μm[39] to 0.6 X 0.6 X 1 mm[16,31,37], while the resolution of a μ-CT tomogram can be as low as 10 X 10 X 10 μm[31,33,37,46], depending on the size of the sample that is to be analysed[45,46]. This reduction in resolution is due to the minimisation of the X-ray spot size by focusing the electron beam prior to hitting the target[39,45] as well as by lowering the X-ray current[39,43]. The focussing of the electron beam also has another advantage, namely the reduction of the geometrical unsharpness.[45]

The aim of this study is to explore the applicability of μ-CT as an analytical tool in the study of coal breakage and degradation, and forms part of a larger study on coal degradation. To this end, a small number of samples were loaded, mechanically and thermally, and μ-CT was used to track the breakage that took place during loading. Changes to the experimental set-ups that may improve the experimental accuracy were also recommended.

Experimental methods

All of the tomograms for the experiments below were generated on the Nikon XTH 225 ST system, manufactured in England and housed at the South African Nuclear Energy Corporation (Necsa) in their Micro-focus X-ray Radiography and Tomography (MIXRAD) department. The system is operational between 30 kV and 225 kV, and between 0 mA and 1 mA with a point source of between 1 μm and 3 μm. The detector is a 400 X 400 mm flat Perkin Elmer panel detector with 200 X 200 μm pixels. All tomograms were generated with 1000 projections in 360º and an exposure time of 1 s per projection. The resolutions obtained during these studies varied from 17 μm to 50 μm depending on the size of the samples. Hoffman and De Beer[40] give a detailed description of the MIXRAD facility. All of the reconstructions were performed using Metris CT-Pro and the analyses were done using the VGStudio Max 2.1 visualisation software package.

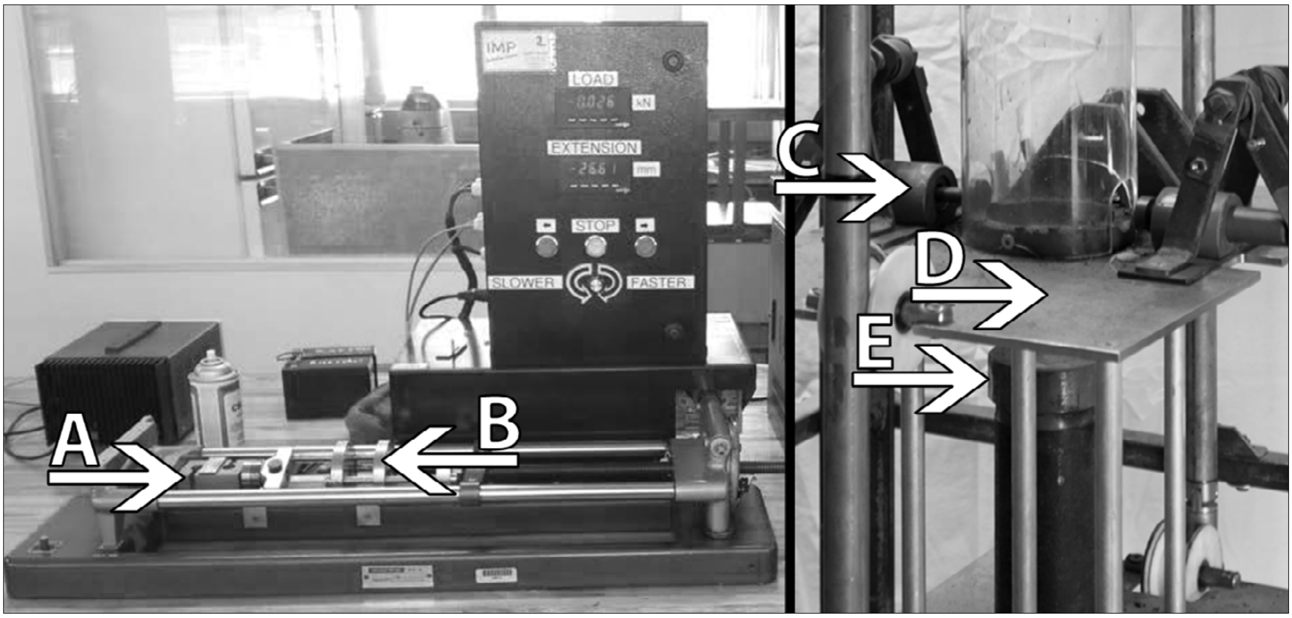
All of the mechanically loaded particles were wrapped in GLAD® cling film to minimise the loss of material during breakage and to simplify the comparison of the before and after tomograms. It is assumed that the radio translucent cling film did not deteriorate the quality of the tomograms, as the cling film has a low physical density (0.92 g/cm^3)[47] and the sample was wrapped in a very thin layer of the film.

The experimental procedures for all three types of degradation experiments (compression loading, impact loading and thermal treatment) are given below.

Compressive loading

To determine the effect of the internal structures on the compression breakage characteristics, two experiments were conducted. During the first experiment, the samples were prepared from a single large (approximately 1 x 0.5 x 0.4 m) block of run of mine (ROM) coal sourced from the Waterberg coalfield in South Africa. The samples, prepared from the heterogeneous, bright-dull banded lithotypes, were cut into 30 X 30 X 30 mm square particles. Two surfaces, on opposite sides of the particle, were machined flat and uniaxial compression was applied to these surfaces. Six samples were sent for data acquisition at the MIXRAD facility. After data acquisition, the particles were wrapped in cling film, loaded into a Monsanto tensometer, supplied by IMP, and compressed while the displacement was measured. Figure 1 shows the Monsanto tensometer (left) that was used during the first round of compression tests and the experimental set-up used during the impact breakage tests (right). The labels on the left of Figure 1 indicate the load cell (A) that was used to measure the applied load during the compression tests and two surfaces between which the particles were compressed (B). The load was applied either perpendicular or parallel to the bedding plane until the first crack that split the parent particle into two or more progeny was detected and the force required to generate the first fatal crack (load at first fracture) was recorded. The load at first fracture was measured as the applied force where the displacement showed a sudden increase. After loading, the particles were again scanned and the tomograms generated before and after fracture compared. From the comparison, new crack formation, probable crack initiation sites and propagation routes were identified. The tomograms were generated at 100 kV and 100 μA and an average spatial resolution of 24 μm was achieved.

During the first set of experiments, it was found that it is very hard to come to a solid conclusion as to the probable propagation routes of the cracks if the only information available is at the start and at the end of the loading cycle. A second set of experiments were undertaken in which two additional samples were subjected to compressive loading that was incrementally increased. The particles were placed into a specifically designed, radio translucent test cartridge manufactured (at the North-West University) from poly(methyl methacrylate) in such a fashion that



A, load cell; B, the compression surfaces of the tensometer; C, sample clamps; D, impact carriage; E, impact anvil.

Figure 1: Experimental set-ups for the compressive breakage tests (left) and the impact breakage tests (right).

the uniaxial compressive force applied to the particle could be increased incrementally. The particle placed in the cartridge was scanned, the applied load increased and the particle scanned again. This process of tomogram generation and load increase was repeated until the particle failed. All of the tomograms generated were compared to determine new crack formation, crack initiation sites and crack propagation routes. Tomograms were generated at 130 kV and 100 μA with a spatial resolution of 29 μm.

Impact loading

A number of coal particles were prepared from a single block of ROM coal sourced from the Waterberg coalfield. The samples were cut to 50 X 50 X 50 mm square particles and three sides machined flat – one side where the impact was to take place and two sides, opposite one another, where the sample clamps gripped the particles. In order to ensure that there were as many maceral boundaries and pre-existing cracks as possible to study, the samples were prepared from the bright-dull banded lithotypes of the block and impacted parallel and perpendicular to the bedding planes. The crack network for each particle was characterised using μ-CT, wrapped in cling film and dropped onto a steel anvil at 4.5 m/s. The impact velocity is equivalent to a drop from 1 m. The rig was developed at the North-West University and manufactured in such a way that the bedding plane orientations of the particles could be accurately controlled. Half the particles were dropped with the bedding planes parallel to the impact surface and the other half with the bedding planes perpendicular to the impact surface. After the particles were dropped, they were again scanned using μ-CT. Figure 1 shows the Monsanto tensometer (left) that was used during the first round of compression tests and the experimental set-up used during the impact breakage tests (right). The labels on the right side of Figure 1 indicate the sample clamps (C), impact carriage (D) and impact anvil (E). The tomograms generated both before impact and after were generated at 160 kV and 70 μA and an average spatial resolution of 48 μm was achieved. A comparison of the before and after tomograms were made and new crack formation, crack initiation sites and probable propagation routes were identified.

Fracture from thermal treatment

In order to determine the effect of maceral boundaries and crack network on the thermal breakage properties of coal, a number of cylindrical samples, with a height of 30 mm and a diameter of 19 mm, were prepared from large particles of a Witbank coal. Witbank coal was used to prevent any damage to the graphite crucibles as the Witbank coal is non-swelling while the Waterberg coal is a medium-swelling coal. The prepared samples were placed in graphite crucibles that could be rapidly heated in a radio frequency induction furnace built by Necsa. The mechanical sample preparation was undertaken to ensure that the samples fitted tightly into the radio translucent graphite crucibles. This was to ensure that a clear tomogram was generated. If the samples did not tightly fit in the crucibles, there would be the possibility of the samples moving while being scanned, introducing errors and reducing the sharpness of the tomogram. The crucibles containing the samples were scanned before and after being heated and comparisons of the tomograms made to determine new crack formation, crack initiation sites and probable crack propagation routes. The heating rate within the graphite crucible was controlled by changing the potential difference through the radio frequency furnace coil and the maximum temperature controlled by the residence time within the furnace. The heating rates ranged from 27 ºC/s to 76 ºC/s with final temperatures reaching between 700 ºC and 800 ºC. The residence times varied between 9 s and 29 s. The heating rates, maximum temperatures and residence times were varied for all samples. All experimental runs were conducted in an air atmosphere. The temperatures stated are the outside crucible temperatures, and were measured using a pyrometer that determines the crucible temperature by the infrared radiation it emits. Owing to the exploratory nature of the study and the difficulty of measuring the internal temperature without introducing errors, the internal temperatures were not measured or estimated. The particle temperature was, however, high enough to cause thermal drying but not so high as to cause any detectable devolatilisation. The tomograms were generated at 130 kV and 100 μA with voxel sizes of 17 μm. A comparison to determine new crack formation, crack initiation sites and probable propagation routes was made.

Results and discussion

Figures 2 to 9 show comparisons of slices from the tomograms generated. The slices on the left in all figures are from the tomograms generated before loading (or thermal treatment) while the slices on the right are from the tomograms generated after loading. Where there are three slices in a figure, the centre slice is from a tomogram generated during loading.

In all of the tomogram slices presented in this study, the varying shades of grey represent the variation in the linear attenuation of the particles. Black areas are areas where no sample material is present; dark grey areas are areas of lower attenuation; light grey areas are areas of higher attenuation and white areas represent mineral matter. The light grey macerals are assumed to be either inertinite- or carbominerite-rich microlithotypes, while the darker grey areas are assumed to be vitrinite rich microlithotypes. This is due to the higher density of the inertinite- and carbominerite-rich macerals compared to the vitrinite rich macerals.[48,49]

Compressive loading

Figure 2 shows a comparison of the before and after tomogram of a particle that was compressed perpendicular to the bedding plane in the Monsanto tensometer, while Figure 3 shows a particle loaded parallel to the bedding plane. The load application perpendicular to the bedding plane indicates that the particle was compressed from the top and bottom of the slice given in Figure 2, while the parallel load application indicates that the particle in Figure 3 was compressed from the left and right of the slice.

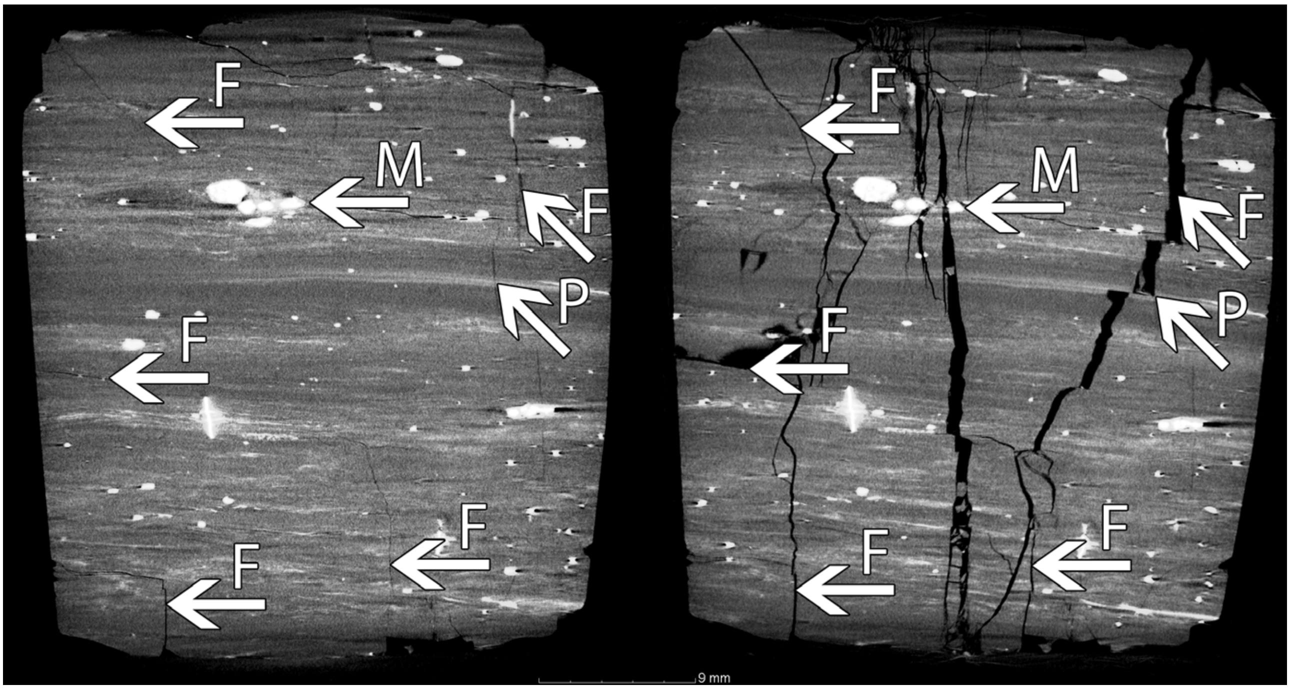
Both particles shown were prepared from the bright-dull banded lithotype of a single block of ROM coal from the Waterberg coalfield. All the arrows marked (F) in both Figure 2 and 3 show cleats or cracks that existed in the pre-loaded particles that enlarged during compression. The arrows marked (M) show mineral inclusions that may have had an influence on either crack initiation or crack propagation. Lastly, arrows marked (P) show either maceral-maceral boundaries or maceral-mineral boundaries that may have had an influence on crack initiation or propagation.

Figure 2 shows a number of cracks present in the original crack network that enlarged during compression. The largest new crack formed during compression, however, is the crack that runs down the middle of Figure 2 (through M), in the post compression slice (right-hand slice). This newly formed fracture did not form from any crack present and, although it did propagate through the mineral inclusions at M, shows no influence of any microstructure in the particle. It is speculated that the newly formed crack is due to load concentration where the tensometer contacted the particle, thus forming a crack in the direction of the applied load. Because of the manual sample preparation, the sides of the particle could not be made perfectly smooth and parallel. This created some features where stress concentration could take place.

In Figure 3 it can again be seen that the major, newly formed cracks (through P_1; M_1 and M_2) formed parallel to the applied load, i.e. from the left and right. The left-hand slice shows a significant number of cracks (marked F) that contributed to the new crack network; the majority of the cracks that are perpendicular to the load direction are due to the enlargement of these pre-existing cracks. P_1 shows where more new cracks formed in the vitrinite rich microlithotypes compared to the inertinite rich microlithotypes. P_2 shows where a boundary between two macerals caused the propagating crack to change direction. M_1 shows a probable crack initiation site at a mineral inclusion, while M_2 shows a mineral inclusion that was in the path of a propagating crack. This inclusion may have helped with the crack propagation.

From the comparisons above, it seems that the direction of the applied load has a great influence, with the majority of the newly formed cracks propagating in the direction of load application. The cracks that did not form in the same direction as the load application are due to pre-existing cracks. In addition, of the microstructures present, the existing crack or cleat network has the greatest influence on crack initiation and propagation, with very few of the pre-existing cracks showing no change. It is also clear that the lower density macerals show more new crack formation than the higher density macerals and in some cases, the boundaries between two macerals influenced the crack propagation.

Although it is possible to speculate which of the microstructures influenced the propagation and the chronology of the newly formed cracks, it is impossible to say so with certainty. To clarify this, another set of experiments was performed where the load application to the samples was increased incrementally until the particle failed, with the internal structure of the sample investigated after every increase. This was done to try to determine, with a higher degree of certainty, which of the microstructures would have a higher probability of influencing the crack propagation. To this end Figure 4 and Figure 5 show three steps in the compression process of two particles that were compressed perpendicular (Figure 4, loaded top to bottom) and parallel (Figure 5, loaded top to bottom) to the bedding planes. Although there were many tomograms generated during the incremental load increase, Figures 4



F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; M indicates mineral inclusions that may have had an influence on either crack initiation or crack propagation; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

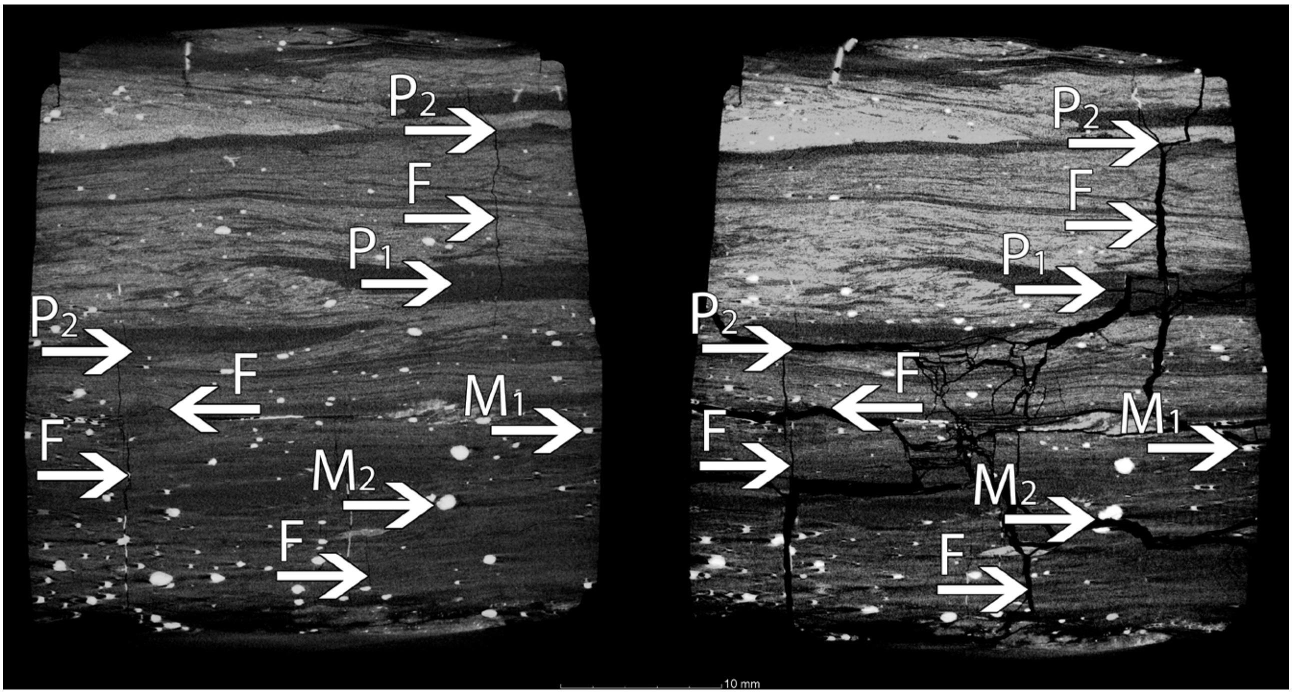
Figure 2: Comparison of the before (left) and after (right) tomogram of a particle compressed perpendicular to the bedding plane in the Monsanto tensometer.

and 5 show only the tomograms generated before loading, after loading and a single tomogram generated during compression. The samples were again prepared from the bright-dull banded lithotypes of the large block of Waterberg ROM coal. The arrows marked F, M and P have the same meaning as in Figures 2 and 3.

The labels marked F_1 in Figure 4 show a number of cracks that are thought to contribute to the formation in the large crack in the far right slice of Figure 4. The label P shows a maceral-maceral boundary that influenced the crack propagation while M shows a mineral inclusion that did the same. F_2 shows a number of cracks that were present in the particle, but due to the pressure applied, closed in the middle slice. Another crack that closed while being compressed is shown at F_3. Even though the closed crack is no longer apparent in the middle slice, this does not mean that the closed crack no longer contributes to crack propagation. In the final slice it is clear that the closed crack influenced the propagation of the crack through F_1 and F_3.

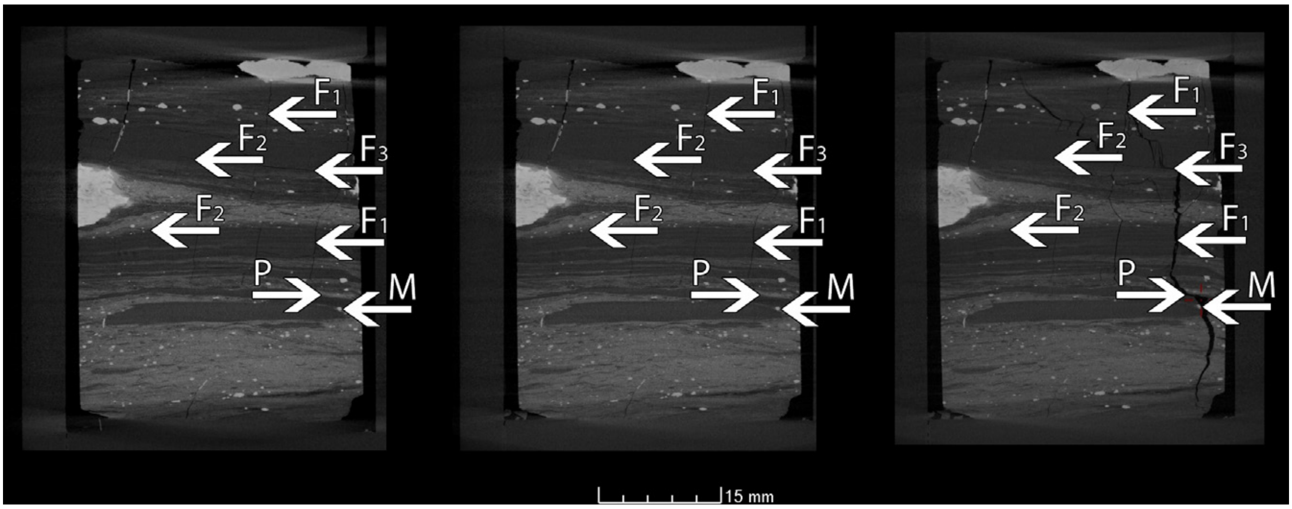
Figure 5 again shows some pre-existing cracks (F) that contributed to the final crack network, as well as a lower density maceral (P_1) where a higher amount of newly formed cracks can be observed. P_2 shows a maceral-maceral boundary that also contributed to the propagation of a crack. Figure 5 shows no cracks that closed due to the applied load, in fact, the opposite is observed: there are some cracks that, despite the load applied perpendicular to the cracks, enlarged.

From the incremental loading, the same conclusions can be reached regarding the effect of the directionality of the applied load, as well as to the influence of the original microstructure on the crack evolution. It was also observed that in the particle where the compression was



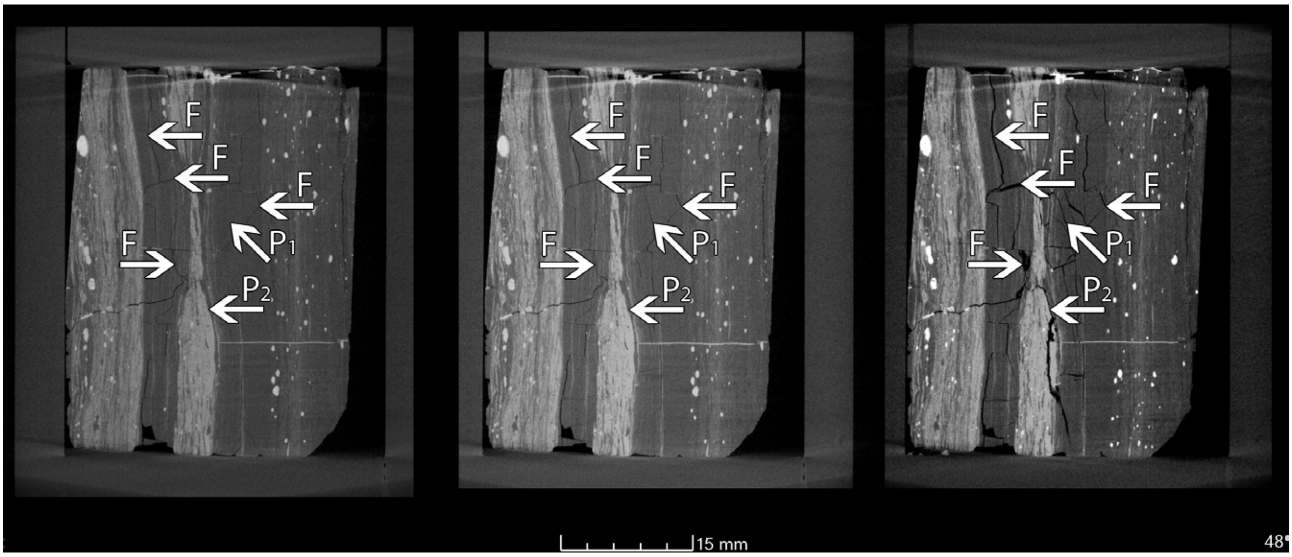
F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; M indicates mineral inclusions that may have had an influence on either crack initiation or crack propagation; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

Figure 3: Comparison of the before (left) and after (right) tomogram of a particle compressed parallel to the bedding plane in the Monsanto tensometer.



F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; M indicates mineral inclusions that may have had an influence on either crack initiation or crack propagation; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

Figure 4: Comparison of a particle before (left) and after (right) compression incrementally perpendicular to the bedding plane.



F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

Figure 5: Comparison of a particle before (left) and after (right) compression incrementally parallel to the bedding plane.

perpendicular to the bedding plane, some of the cracks that were perpendicular to the applied load closed due to the applied load. This was only observed in the particle that was compressed perpendicular to the bedding plane.

From the comparison it is clear that the incremental load increase is still too large to deduce with any higher degree of certainty the propagation routes and hierarchy. Owing to the time required to generate a tomogram, the crack propagation cannot be observed, only the after effects i.e. the newly formed cracks. If the load increase with each increment is reduced and more tomograms generated during compression (6–8 tomograms generated instead of 3–4), it may be possible to observe the order in which new cracks form and the propagation routes deduced.

Impact loading

Figures 6 and 7 show comparisons of the tomograms generated before and after impact of two samples prepared from a large block of ROM coal sourced from the Waterberg coalfield. The samples in both Figure 6 and Figure 7 were impacted onto a steel anvil at a velocity of 4.5 m/s. The impact velocity was equivalent to a drop from 1 m, a low impact velocity selected specifically to study the degradation at low impact energy. Figure 6 shows a particle that was impacted with the bedding plane parallel to the anvil while Figure 7 shows a particle that was impacted with the bedding plane perpendicular to the anvil. In both Figure 6 and Figure 7, the surface of the particle that was in contact with the impact anvil (impacted surface) is the bottom edge of the slice or the edge closest to the scale bar.

In Figure 6, the majority of the cracks that formed, formed from pre-existing cracks (F) within the particle. There were however some new cracks that formed in the lower density, vitrinite rich, macerals (P).

Figure 7 also shows a number of existing cracks that enlarged during impact. Some of the other trends observed in the compression loaded particles can also be seen in Figure 7. Here P_1 shows where a newly formed crack propagated along a maceral boundary and P_2 indicates cracks formed in a vitrinite rich microlithotype. Several cracks end at the mineral inclusion indicated by M in Figure 7.

From the comparisons made in both Figure 6 and Figure 7, it is clear that the main contribution to crack initiation and propagation is the existing crack network of the coal. The final crack network did not show a high dependence on the direction of the applied impact load when compared to the compressive loading in Figures 2 to 5. It is hypothesised that this is due to the low impact velocity, equivalent to a drop of approximately 1 m onto a steel plate. This is supported by various studies where it was found that that a higher drop height increased the breakage and degradation.[6,7,13,22,24,25]

Fracture from temperature increase

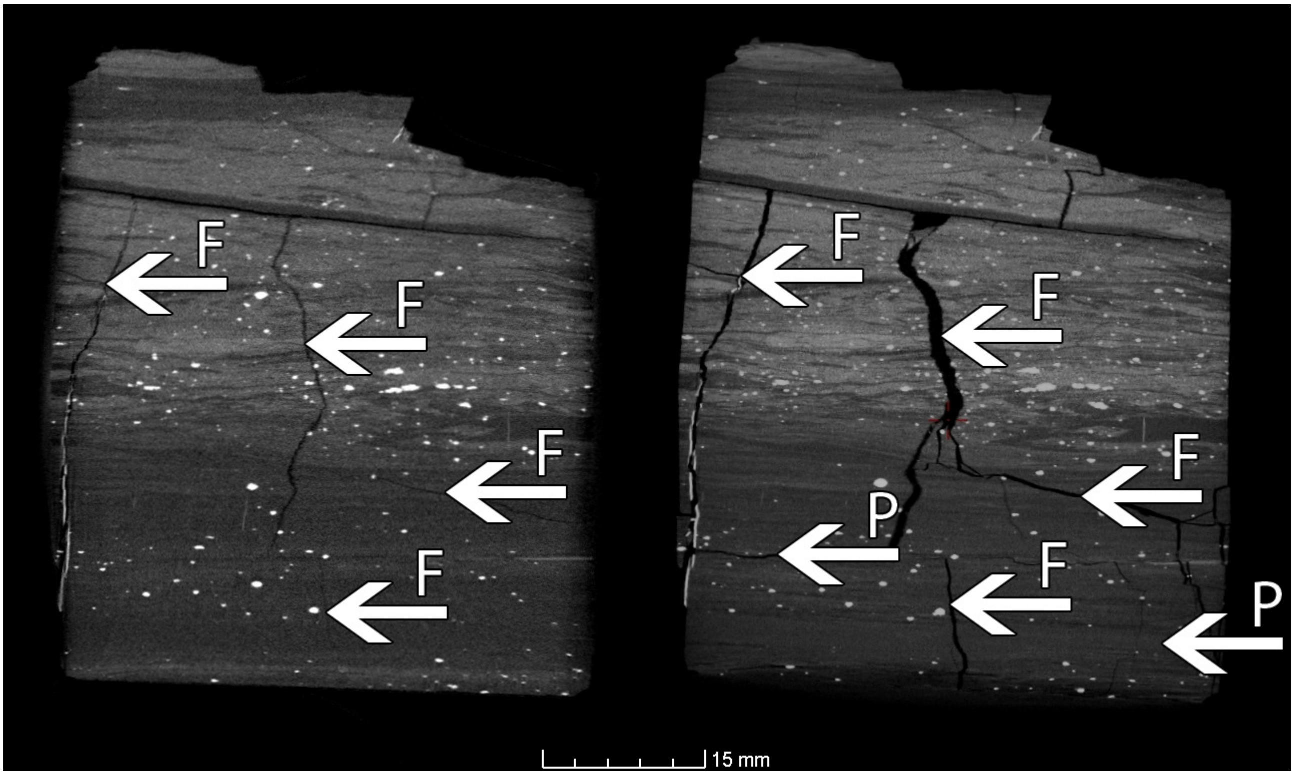
Figures 8 and 9 show comparisons of the tomograms generated before and after thermal treatment. A Witbank coal was used rather than a Waterberg coal to prevent damage to the graphite crucibles. Figure 8 shows the comparison of a particle heated to 800 °C at a rate of 31 °C/s. Figure 9 shows the comparison of a particle heated to 700 °C at a rate of 51 °C/s.

Figure 8 shows two fractures (F) present in the coal before heating, both of which contributed to crack propagation. The crack that propagated along the line indicated by the arrows marked M shows the crack did not propagate through the mineral inclusions but along the outside of the tiny mineral inclusions. It is assumed that the two cracks that formed from the initial cracks marked F were the first to form. Note that many of the cracks that radiate from these initially formed cracks do so almost perpendicular to the main cracks. According to Mathews et al.[50], this occurs during thermal drying due to particle shrinkage. Figure 8 also shows no structural changes indicative of devolatilisation, indicating a discrepancy between the outside crucible temperature and the particle temperature.

Figure 9 shows that during the thermal treatment of coal, pre-existing cracks (F1) present in coal enlarge. Again it can be seen in Figure 9 that new cracks form perpendicular to pre-existing cracks (F2) indicating that thermal drying took place. The enlarged sections in Figure 9 show that the vitrinite rich layers, indicated by P, show an increase in new crack formation when compared to the rest of the particle. The enlargements in Figure 9 also show how the vertical, mineral filled cleats act as discontinuities, with cracks terminating at the mineral filled cleats.

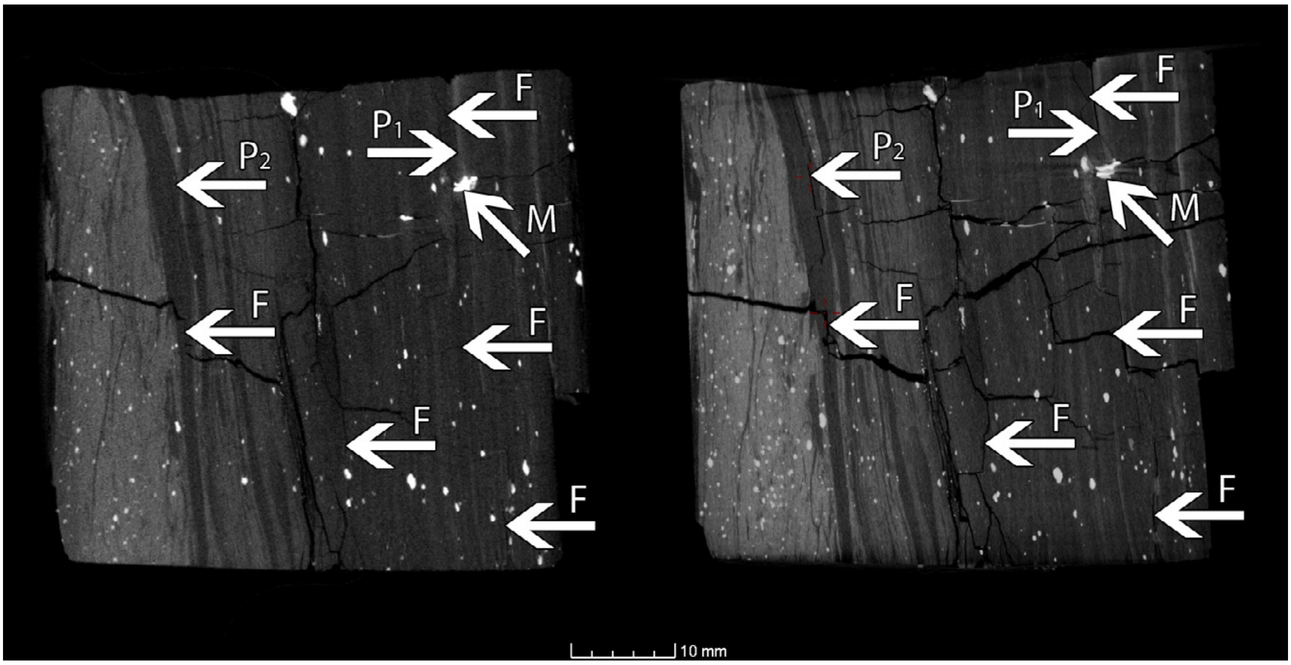
In both Figures 8 and 9, it was observed that pre-existing cracks expanded during the thermal treatment of coal particles. It is also assumed that any new cracks that form will do so either from an initial crack or in the lower density macerals, and crack propagation will occur along lines of weakness such as maceral-mineral boundaries.

Even though the heating rates (27 °C/s, 51 °C/s, 31 °C/s and 76 °C/s), final temperatures (700 °C, 750 °C and 800 °C) and residence times (9 s, 13 s 25 s and 29 s) were varied, no influence of these changes



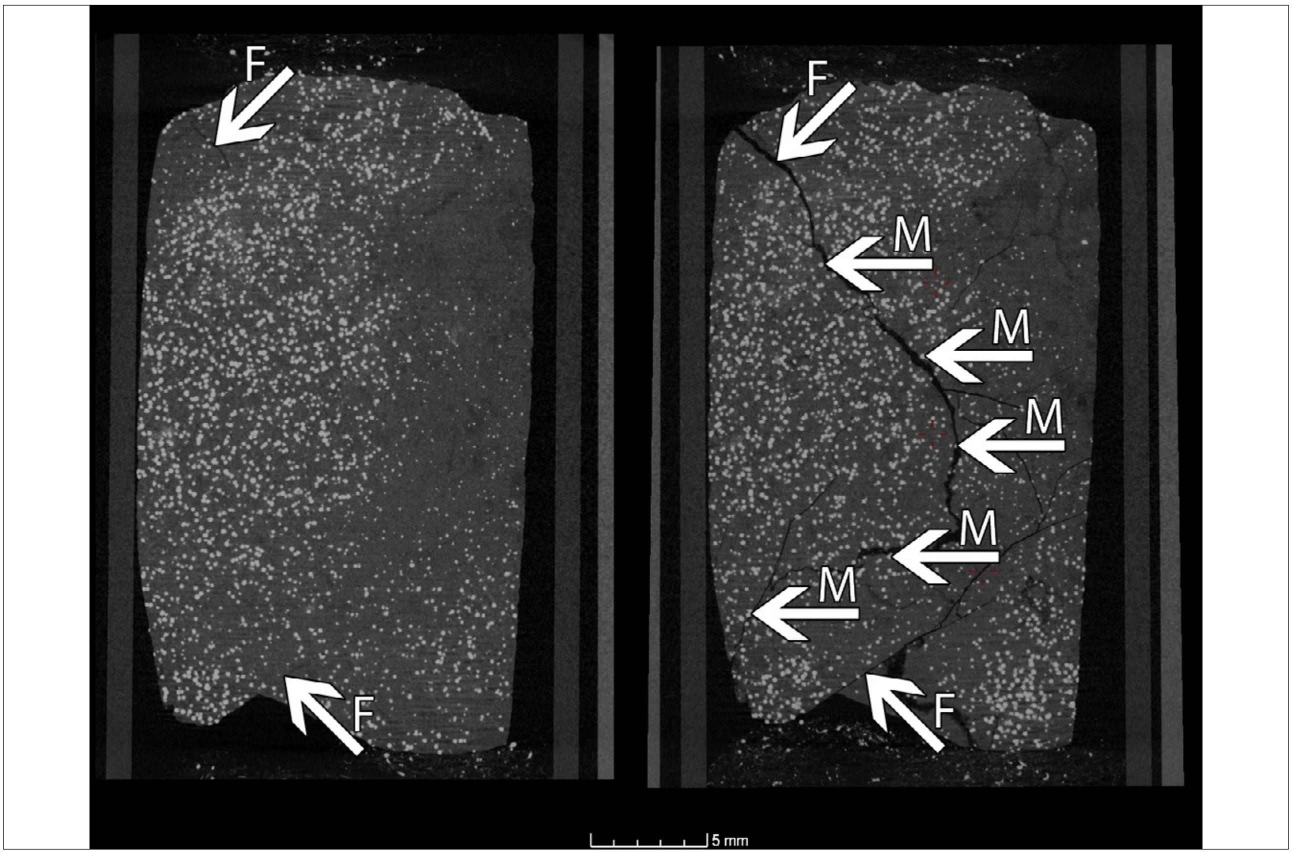
F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

Figure 6: Comparison of a particle before (left) and after (right) impaction with the bedding plane parallel to the anvil.



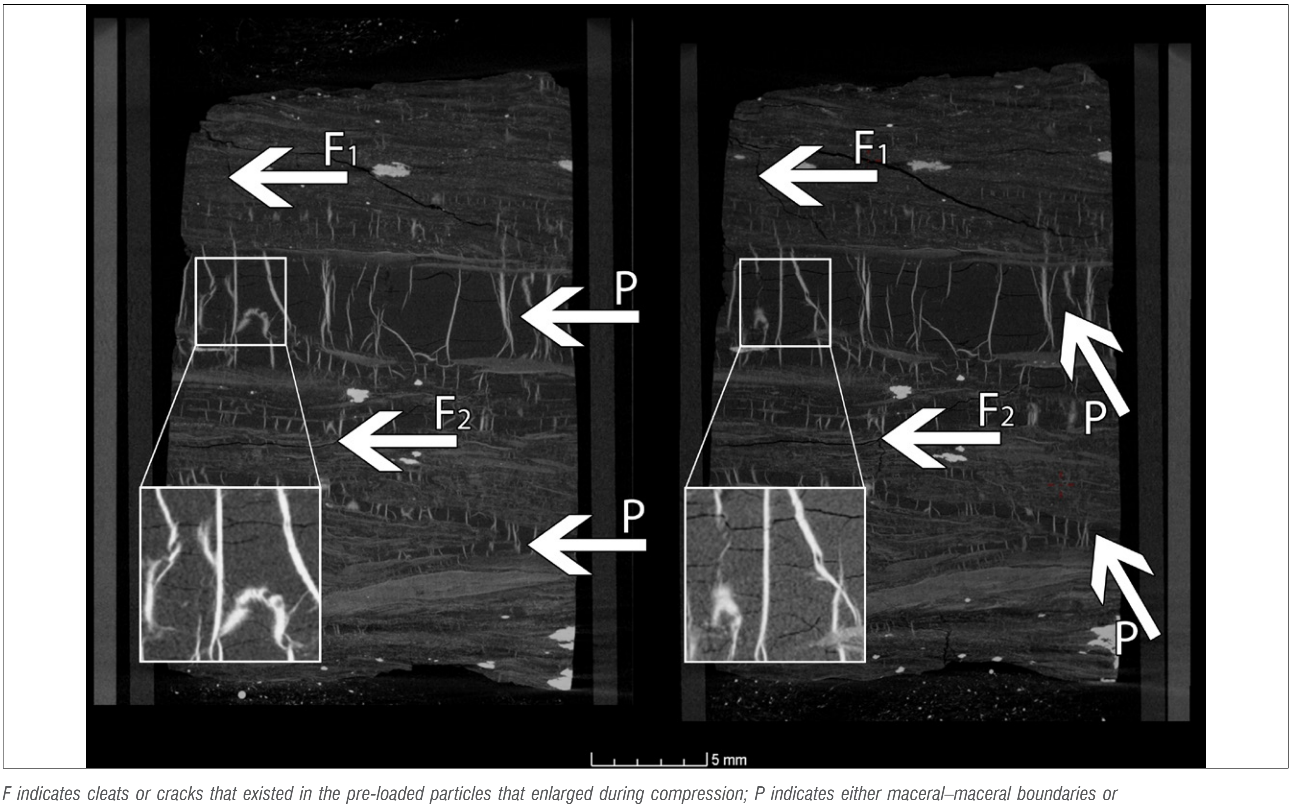
F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

Figure 7: Comparison of a particle before (left) and after (right) impaction with the bedding plane perpendicular to the anvil.



F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; M indicates mineral inclusions that may have had an influence on either crack initiation or crack propagation.

Figure 8: Comparison of a particle before (left) and after (right) it was heated to 800 ºC at a rate of 31 ºC/s.



F indicates cleats or cracks that existed in the pre-loaded particles that enlarged during compression; P indicates either maceral–maceral boundaries or maceral–mineral boundaries that may have had an influence on crack initiation.

Figure 9: Comparison of a particle before (left) and after (right) it was heated to 700 ºC at a rate of 51 ºC/s.

could be detected. It is assumed that this is due to the residence times being far too short for the particle temperature to reach the maximum crucible temperature. Although there is no evidence of devolatilisation taking place, the particle did show evidence of thermal drying.

Conclusion

The microstructure of coal was investigated using μ-CT to determine if the influence of the microstructure on the coal breakage characteristics can be ascertained. To this end, a number of degradation processes were simulated using single particles: slow compression breakage, impact breakage and primary (thermal) fracture. During the mechanical loading, a number of particles were prepared from a ROM sample of Waterberg coal and loaded in such a way that the influence of the microstructures could be identified. The following conclusions were drawn from the comparisons:

- During the compressive loading of coal, one of the biggest influences on the enlargement of existing cracks and the propagation of newly formed cracks is the direction of the applied load.
- It was found that if a particle is compressed with the bedding plane perpendicular to the load direction, some of the cracks along the bedding plane will close. The cracks that do close will still influence the development of the crack network around them, as other cracks can still terminate and initiate within them. This was not observed with the particles loaded parallel to the bedding plane.
- In both the compressive and impact loading scenarios, the pre-existing cracks within a particle have the biggest influence of the microstructures present.
- The macerals and maceral boundaries also influence the formation of new cracks and the crack propagation; the lower density, vitrinite rich, macerals showing a tendency towards increased crack formation and some of the maceral boundaries can affect the propagation of a crack.

For the thermal loading of coal, samples were hand selected and prepared from a large sample of washed and sized Witbank coal and exposed to various temperatures and heating rates. The conclusions drawn are as follows:

- The pre-existing crack network influences the final network by increasing both in length and aperture; the network also acts as an initiation site for new cracks.
- The lower density macerals show a propensity to form more new cracks than do the denser macerals.
- Cracks propagate along the maceral-mineral boundaries.

Thus, it is clear that the microstructure of coal can be identified using μ-CT as an analytical technique. The change in the microstructure can also be tracked using μ-CT, but some modifications to the experimental procedures and set-ups are required to tell with greater certainty how the cracks formed and propagated. For the compressive loading of coal, the incremental load increase should be reduced to try to isolate the decisive moments in the very fast crack initiation and propagation process. For the impact loading of coal, the impact energy range should be expanded (i.e. higher impact energies and lower impact energies) to enable the study of the effect of impact energy on crack development. This will allow for a better understanding of the mechanisms by which coal breaks during impact. Finally, according to literature, temperature, heating rate and residence time affects the thermal degradation of coal but none of these effects were seen in this study. Should the ranges for the temperature, heating rate and residence times be increased, and the internal temperature gradients determined, a better understanding of crack development during heating can be gained.

Acknowledgements

We thank Juan Nelson, Wynand Breytenbach, Charlotte Badenhorst and Carla Hatting for their help with the experimental work; the personnel at the MIXRAD facility, Necsa for their assistance, patience and enthusiasm and two anonymous reviewers for their invaluable input. This project was funded by the Southern African Coal Processing Society. The work presented in this paper 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 (Coal Research Chair grant no. 86880). Any opinion, finding, conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard.

Authors' contributions

J.W.H. is a tomographic instrument expert and was responsible for the generation of the tomograms used during this study and significant editing of the manuscript. M.L.R. and Q.P.C. were the project leaders and initiators; they were also responsible for the experimental concept and significant editing of the manuscript. J.V., M.L.R. and Q.P.C. were responsible for the experimental design. J.V. was responsible for some of the experimental work, the analysis of the generated tomograms and writing the manuscript.

References

1. Teo CS, Waters AG, Nicol SK. Quantification of the breakage of lump materials during handling operations. Int J Miner Process. 1990;30(3/4):159–184. http://dx.doi.org/10.1016/0301-7516(90)90013-O
2. Sahoo RK, Roach D. Effect of different types of impact surface on coal degradation. Chem Eng Process. 2005;44(2):253–261. http://dx.doi.org/10.1016/j.cep.2004.02.019
3. Bunt JR, Waanders FB. An understanding of lump coal physical property behaviour (density and particle size effects) impacting on a commercial-scale Sasol-Lurgi FBDB gasifier. Fuel. 2008;87(13):2856–2865. http://dx.doi.org/10.1016/j.fuel.2008.03.022
4. Eswaraiah C, Gupta A, Nagarajan R, Rajavel M, Nandakumar K. Minimization of fines generation in size reduction of coals by impact crusher. Fuel Process Technol. 2008;89(7):704–714. http://dx.doi.org/10.1016/j.fuproc.2008.01.001
5. England T, Hand PE, Micheal DC, Falcon LM, Yell AD, editors. Coal preparation in South Africa. 4th ed. Pietermaritzburg: The South African Coal Processing Society; 2002.
6. Sahoo R. Degradation characteristics of steel making materials during handling. Powder Technol. 2007;176(2):77–87. http://dx.doi.org/10.1016/j.powtec.2007.02.013
7. Tavares LM, De Carvalho RM. Modeling ore degradation during handling using continuum damage mechanics. Int J Miner Process. 2011;101(1):21–27. http://dx.doi.org/10.1016/j.minpro.2010.07.008
8. Le Roux M. The effect of thermal drying on the mechanical strength of South African coals. J S Afr I Min Metall. 2008;108:783–787.
9. Le Roux M, Campbell QP. An investigation into an improved method of fine coal dewatering. Minerals Eng. 2003;16(10):999–1003. http://dx.doi.org/10.1016/j.mineng.2003.08.004
10. Le Roux M, Campbell QP, Watermeyer MS, De Oliveira S. The optimization of an improved method of fine coal dewatering. Minerals Eng. 2005;18(9):931–934. http://dx.doi.org/10.1016/j.mineng.2005.01.033
11. Oberholzer V, Van der Walt J. Investigation of factors influencing the attrition breakage of coal. J S Afr I Min Metall. 2009;109(4):211–216.
12. Broadbent SR, Callcott TG. Coal Breakage Processes: III. The Analysis of a Coal Transport System. J I Fuel. 1957;30:13–25.
13. Sahoo R, Roach D. Quantification of the lump coal breakage during handling operation at the gladstone port. Chem Eng Process. 2005;44(7):797–804. http://dx.doi.org/10.1016/j.cep.2004.09.004
14. Paprika MJ, Komatina MS, Dakic DV, Nemoda SD. Prediction of coal primary fragmentation and char particle size distribution in fluidized bed. Energ Fuel. 2013;27(9):5488–5494. http://dx.doi.org/10.1021/ef400875q
15. Senneca O, Urciuolo M, Chirone R. A semidetailed model of primary fragmentation of coal. Fuel. 2013;104:253–261. http://dx.doi.org/10.1016/j.fuel.2012.09.026
16. Van Dyk JC. Development of an alternative laboratory method to determine thermal fragmentation of coal sources during pyrolysis in the gasification process. Fuel. 2001;80(2):245–249. http://dx.doi.org/10.1016/S0016-2361(00)00089-2

17. Zhang H, Cen K, Yan J, Ni M. The fragmentation of coal particles during the coal combustion in a fluidized bed. Fuel. 2002;81(14):1835–1840. http://dx.doi.org/10.1016/S0016-2361(02)00111-4

18. Bunt JR, Wagner NJ, Waanders FB. Carbon particle type characterization of the carbon behaviour impacting on a commercial-scale Sasol-Lurgi FBDB gasifier. Fuel. 2009;88(5):771–779. http://dx.doi.org/10.1016/j.fuel.2008.11.021

19. Senneca O, Russo S, Chirone R. Primary fragmentation of coal particles at high heating rate. Chem Eng Trans. 2009;18:569–574. http://dx.doi.org/10.3303/CET0918092

20. Powell MS, Morrison RD. The future of comminution modelling. Int J Miner Process. 2007;84(1):228–239. http://dx.doi.org/10.1016/j.minpro.2006.08.003

21. Han T, Kalman H, Levy A. Theoretical and experimental study of multi-compression particle breakage. Adv Powder Technol. 2003;14(5):605–620. http://dx.doi.org/10.1163/156855203322448372

22. Chandramohan R, Holtham P, Powell M. The influence of particle shape in rock fracture. Paper presented at: XXV International Mineral Processing Congress; 2010 Sept 6–10; Brisbane, Australia.

23. Tavares LM, King RP. Single-particle fracture under impact loading. Int J Miner Process. 1998;54(1):1–28. http://dx.doi.org/10.1016/S0301-7516(98)00005-2

24. Sahoo R. Review: An investigation of single particle breakage tests for coal handling system of the gladstone port. Powder Technol. 2006;161(2):158–167. http://dx.doi.org/10.1016/j.powtec.2005.09.001

25. Esterle JS, Kolatschek Y, O'Brien G. Relationship between in situ coal stratigraphy and particle size and composition after breakage in bituminous coals. Int J Coal Geol. 2002;49(2):195–214. http://dx.doi.org/10.1016/S0166-5162(01)00077-5

26. Shi F, Kojovic T. Validation of a model for impact breakage incorporating particle size effect. Int J Miner Process. 2007;82(3):156–163. http://dx.doi.org/10.1016/j.minpro.2006.09.006

27. Poulsen BA, Adhikary DP. A numerical study of the scale effect in coal strength. Int J Rock Mech Min Sci. 2013;63:62–71. http://dx.doi.org/10.1016/j.ijrmms.2013.06.006

28. Dacombe P, Pourkashanian M, Williams A, Yap L. Combustion-induced fragmentation behavior of isolated coal particles. Fuel. 1999;78(15):1847–1857. http://dx.doi.org/10.1016/S0016-2361(99)00076-9

29. Tavares LM, Das Neves PB. Microstructure of quarry rocks and relationships to particle breakage and crushing. Int J Miner Process. 2008;87(1–2):28–41. http://dx.doi.org/10.1016/j.minpro.2008.01.007

30. Mazumder S, Wolf KAA, Elewaut K, Ephraim R. Application of X-ray computed tomography for analyzing cleat spacing and cleat aperture in coal samples. Int J Coal Geol. 2006;68(3–4):205–222. http://dx.doi.org/10.1016/j.coal.2006.02.005

31. Van Geet M, Swennen R. Quantitative 3D-fracture analysis by means of microfocus X-ray computer tomography: An example from coal. Geophys Res Lett. 2001;28(17):3333–3336. http://dx.doi.org/10.1029/2001GL013247

32. Falcon RMS, Snyman CP. An introduction to coal petrography: Atlas of petrographic constituents in the bituminous coals of southern Africa. Johannesburg: The Geological Society of South Africa; 1986.

33. Van Geet M, Swennen R, Wevers M. Towards 3-D petrography: Application of microfocus computer tomography in geological science. Comput Geosci. 2001;27(9):1091–1099. http://dx.doi.org/10.1016/S0098-3004(00)00154-0

34. Laubach SE, Marret RA, Olson JE, Scott AR. Characteristics and origins of coal cleat: A review. Int J Coal Geol. 1998;35:175–207. http://dx.doi.org/10.1016/S0166-5162(97)00012-8

35. Ward CR. Analysis and significance of mineral matter in coal seams. Int J Coal Geol; 2002;50(1):135–168. http://dx.doi.org/10.1016/S0166-5162(02)00117-9

36. Ketcham RA, Carlson WD. Acquisition, optimization and interpretation of X-ray computed tomographic imagery: Applications to the geosciences. Comput Geosci. 2001;27(4):381–400. http://dx.doi.org/10.1016/S0098-3004(00)00116-3

37. Mees F, Swennen R, Van Geet M, Jacobs P. Applications of X-ray computed tomography in the geosciences. In: Mees F, Swennen R, Van Geet M, Jacobs P, editors. Applications of X-ray computed tomography in the geosciences. London: Geological Society of London; 2003. p. 1–6. http://dx.doi.org/10.1144/gsl.sp.2003.215.01.01

38. Duliu OG. Computer axial tomography in geosciences: An overview. Earth-Sci Rev. 1999;48(4):265–281. http://dx.doi.org/10.1016/S0012-8252(99)00056-2

39. Cnudde V, Boone MN. High-resolution X-ray computed tomography in geosciences: A review of the current technology and applications. Earth-Sci Rev. 2013;123:1–17. http://dx.doi.org/10.1016/j.earscirev.2013.04.003

40. Hoffman JW and De Beer FC. Characteristics of the Micro-Focus X-Ray Tomography Facility (MIXRAD) at Necsa in South Africa. Paper presented at: 18th World Conference on Nondestructive Testing; 2012 Apr 16–20; Durban, South Africa.

41. Ketcham RA, Slottke DT, Sharp JM. Three-dimensional measurement of fractures in heterogeneous materials using high-resolution X-ray computed tomography. Geosphere. 2010;6(5):499–514. http://dx.doi.org/10.1130/GES00552.1

42. Keller A. High resolution, non-destructive measurement and characterization of fracture apertures. Int J Rock Mech Min Sci. 1998;35(8):1037–1050. http://dx.doi.org/10.1016/S0148-9062(98)00164-8

43. Cnudde V, Masschaele B, Dierick M, Vlassenbroeck J, Hoorebeke LV, Jacobs P. Recent progress in X-ray CT as a geosciences tool. Appl Geochem. 2006;21(5):826–832. http://dx.doi.org/10.1016/j.apgeochem.2006.02.010

44. Krimmel S, Stephan J, Baumann J. 3D computed tomography using a microfocus X-ray source: Analysis of artifact formation in the reconstructed images using simulated as well as experimental projection data. Nucl Instrum Meth A. 2005;542(1–3):399–407. http://dx.doi.org/10.1016/j.nima.2005.01.171

45. Simons FJ, Verhelst F, Swennen R. Quantitative characterization of coal by means of microfocal X-ray computed microtomography (CMT) and color image analysis (CIA). Int J Coal Geol 1997;34(1):69–88. http://dx.doi.org/10.1016/S0166-5162(97)00011-6

46. Van Geet M, Swennen R, David P. Quantitative coal characterisation by means of microfocus X-ray computer tomography, colour image analysis and back-scattered scanning electron microscopy. Int J Coal Geol. 2001;46(1):11–25. http://dx.doi.org/10.1016/S0166-5162(01)00006-4

47. Dhillon RK, Singh P, Gupta SK, Singh S, Kumar R. Study of high energy (MeV) N6+ ion and gamma radiation induced modifications in low density polyethylene (LDPE) polymer. Nucl Instrum Meth B. 2013;301:12–16. http://dx.doi.org/10.1016/j.nimb.2013.02.014

48. Naudé G, Hoffman J, Theron SJ, Coetzer G. The use of X-ray computed tomography in the characterisation of coal and associated char reductants. Minerals Eng. 2013;52:143–154. http://dx.doi.org/10.1016/j.mineng.2013.05.012

49. Van Geet M, David P, Swennen R. Three dimensional coal characterisation (maceral, mineral and cleats) by means of X-ray microfocus computer tomography (μCT). Proceedings of the European Coal Conference IV; 2000 Sept 26–28; Poland, Uströn. Warsaw: Polish Geological Institute; 2002. p. 263–270.

50. Mathews JP, Pone JDN, Mitchell GD, Halleck P. High-resolution X-ray computed tomography observations of the thermal drying of lump-sized subbituminous coal. Fuel Process Technol. 2011;92(1):58–64. http://dx.doi.org/10.1016/j.fuproc.2010.08.020

Nematode pests threatening soybean production in South Africa, with reference to *Meloidogyne*

AUTHORS:
Hendrika Fourie[1]
Dirk de Waele[1,2]
†Alexander H. Mc Donald[1]
Charlotte Mienie[1]
Mariette Marais[3]
Annelie de Beer[4]

AFFILIATIONS:
[1]Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa
[2]Afdeling Plantenbiotechniek, University of *Leuven,* Leuven, Belgium
[3]Plant Protection Research Institute, Agricultural Research Council, Pretoria, South Africa
[4]Grain Crops Institute, Agricultural Research Council, Potchefstroom, South Africa

CORRESPONDENCE TO:
Hendrika Fourie

EMAIL:
driekie.fourie@nwu.ac.za

POSTAL ADDRESS:
[1]Unit for Environmental Sciences and Management, North-West University, Private Bag X6001, Potchefstroom 2520, South Africa

DATES:
Received: 19 June 2014
Revised: 21 Oct. 2014
Accepted: 19 Jan. 2015

KEYWORDS:
host plant resistance; legume; molecular markers; plant-parasitic nematodes; root-knot nematodes

HOW TO CITE:
Fourie H, De Waele D, Mc Donald AH, Mienie C, Marais M, De Beer A. Nematode pests threatening soybean production in South Africa, with reference to *Meloidogyne*. S Afr J Sci. 2015;111(9/10), Art. #2014-0212, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140212

†Deceased

The area planted to soybean in South Africa has increased by 54% since the 2009 growing season, mainly as a result of the increasing demand for protein-rich food and fodder sources. Moreover, the introduction of advanced technology, namely the availability of genetically modified herbicide tolerant soybean cultivars also contributed towards increased soybean production. The omnipresence of plant-parasitic nematodes in local agricultural soils, however, poses a threat to the sustainable expansion and production of soybean and other rotation crops. *Meloidogyne incognita* and *M. javanica* are the predominant nematode pests in local soybean production areas and those where other grain-, legume- and/or vegetable crops are grown. The lack of registered nematicides for soybean locally, crop production systems that are conducive to nematode pest build-ups as well as the limited availability of genetic host plant resistance to root-knot nematode pests, complicate their management. Research aimed at various aspects related to soybean-nematode research, namely, audits of nematode assemblages associated with the crop, identification of genetic host plant resistance in soybean germplasm to *M. incognita* and *M. javanica*, the use of molecular markers that are linked to such genetic resistance traits as well as agronomic performance of pre-released cultivars that can be valuable to producers and the industry are accentuated in this review. Evaluation of synthetically-derived as well as biological-control agents are also discussed as complementary management tactics. It is important that lessons learned through extensive research on soybean-nematode interactions in South Africa be shared with researchers and industries in other countries as they might experience or expect similar problems and/or challenges.

Introduction

Soybean (*Glycine max* (L.) Merr) is a major source of protein and oil, both for local human and animal consumption.[1,2] During the 2012/2013 growing season, sunflower ranked first in terms of its production (860 000 t)[1], followed by soybean (710 000 t) and dry bean (60 200 t)[3]. These three are the main oilseed- and protein crops being produced in South Africa. Locally, the hectares planted to soybean have increased by 54% from 2008/2009 to the 2013/2014 growing seasons.[1] Furthermore, the introduction of advanced technology in the form of genetically-modified, herbicide-tolerant, Roundup® Ready (RR) soybean material, was experienced in 2004 when such cultivars were released for commercial production in South Africa.[4] These trends reflect the increasing and urgent need for oil and protein sources to feed a fast growing nation as well as its cattle industry.[2]

Although soybean was traditionally cultivated in the Free State, KwaZulu-Natal, Mpumalanga and Gauteng Provinces,[1] its production was and still is extended to areas where predominantly maize and crops such as groundnut, sunflower, potato and others were traditionally grown. The initiative to expand and stimulate local soybean production resulted in exposure of the crop to new pests and diseases that have the potential to seriously reduce local soybean production.[2] For example, the soybean leaf miner, *Aproaerema modicella* (Deventer) that was introduced into South Africa and reported as a pest of groundnut during the early 2000s also attacks soybean in certain areas of the country.[5] A similar scenario was experienced in 2001 when the economically important soybean rust disease, caused by the newly-introduced fungus *Phakopsora pachyrhizi* Sydow, was first recorded as a major pathogen of soybean crops in traditional local production areas.[6] Not only should soybean cultivars be adapted to local environmental conditions to optimise crop performance, it should also exhibit resistance to diseases and pests such as plant-parasitic nematodes, bacteria, fungi and insect pests.[2]

Although not always perceived as pests of soybean and other crops, plant-parasitic nematodes are economically one of the most important production constraints in crop production areas of sub-Saharan Africa.[7] The latter include current and potential soybean production areas in South Africa.[8-10] Towards the end of the 1980s, the estimated annual soybean yield losses resulting from plant-parasitic nematode parasitism amounted to approximately 9%.[11] The 9%, however, referred to damage caused by various plant-parasitic nematode communities and did not distinguish between the contribution by particular nematode species. More recent assessments of the pest status of root-knot nematodes on soybean revealed yield losses that ranged from 25–70%.[8,12,13] In addition, two of the national soybean cultivar trials that are annually conducted by the Agricultural Research Council's Grain Crops Institute (ARC-GCI) were terminated during 1999 as a result of high root-knot nematode infections causing total crop failure.[14] Distinct root-galling (Figure 1a and b) represents below-ground symptoms and is caused by feeding of female root-knot nematodes (Figure 2). On the other hand, above-ground symptoms in fields where high population levels of these pests occur can include stunted plants with yellowing leaves (Figure 3).

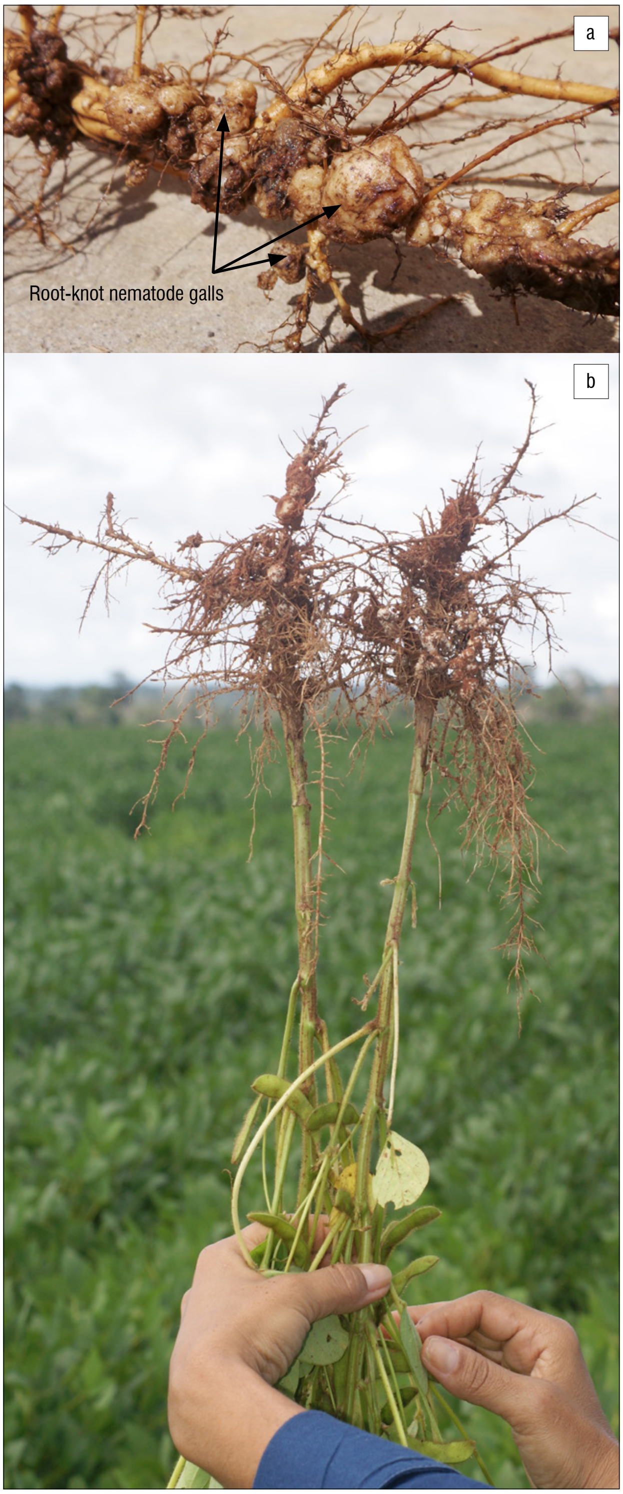


Photo: (a) Philip Holtzhuizen; (b) Driekie Fourie

Figure 1: (a) Root-knot nematode galls, resulting from parasitism by *Meloidogyne* sp. females, on the roots of a soybean plant that was sampled during the 2013/2014 growing season in the Sasolburg area (Free State Province of South Africa). (b) Heavily galled root systems of two root-knot nematode infected soybean plants that grew in the Viljoenskroon area (Free State Province of South Africa) during the 2012/2013 growing season.

The increased awareness and adverse impact of plant-parasitic nematodes on soybean crops and the expansion of the crop resulted in the initiation of several research projects. Subsequently, the significant body of knowledge regarding soybean-nematode interactions that was accumulated since the middle 1950s is discussed. As soybean production on the African continent increases, South African knowledge of this crop could be applied in the rest of Africa where soybean production is also increasing[15] and where similar environmental conditions occur as those in South Africa. For these reasons, aspects are highlighted regarding the most important nematode pests of soybean, expected problems encountered with the introduction of soybean into production areas where it was not grown before as well as tactics that can be used to manage these pests.

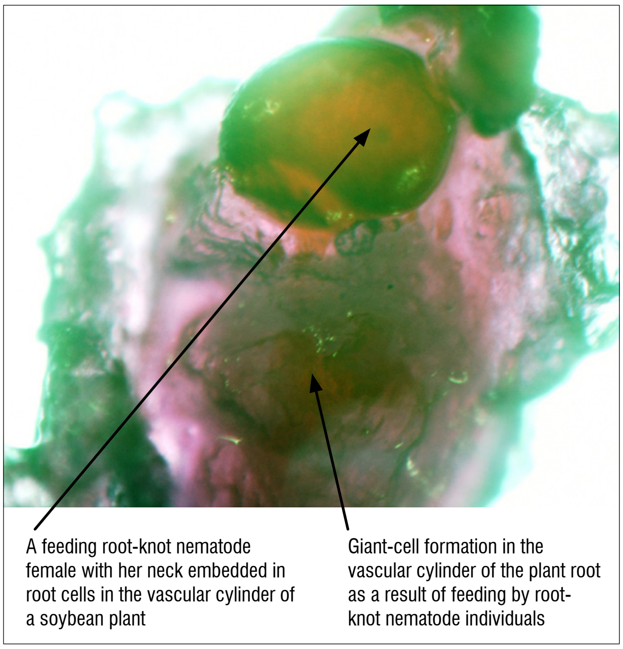

A feeding root-knot nematode female with her neck embedded in root cells in the vascular cylinder of a soybean plant
Giant-cell formation in the vascular cylinder of the plant root as a result of feeding by root-knot nematode individuals

Photo: Driekie Fourie

Figure 2: The swollen, roundish body of a red-stained female root-knot nematode positioned with her neck embedded in root cells in the vascular cylinder of the soybean plant in which she is feeding.

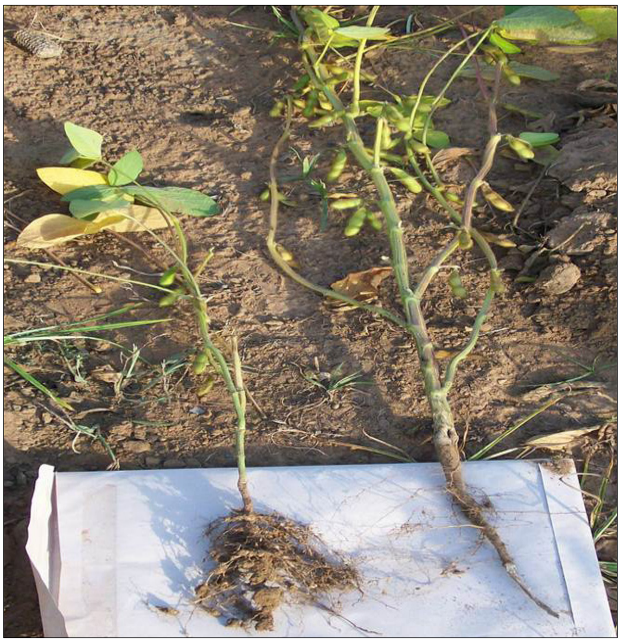


Photo: Driekie Fourie

Figure 3: A root-knot nematode infected and galled soybean root system (left) and an uninfected one (right). The infected soybean plant (left) has yellow leaves and is stunted.

Plant-parasitic nematodes associated with soybean in South Africa

To date, 18 plant-parasitic nematode genera and 48 species have been associated with soybean in South Africa (Table 1). In 1959, individuals of the endoparasitic root-knot nematode species *M. arenaria* were reported to parasitise soybean[16], followed by listings of *M. hapla*, *M. incognita* and *M. javanica* being associated with the crop in 1968.[17] During the early 1980's, the species list was extended when numerous plant-parasitic nematode species identified from rhisosphere soil and roots of soybean plants were added.[18] Since then, more nematodes associated with soybean were reported from material deposited in the National Collection of Nematodes (Nematology Unit of the Agricultural Research Councils' Plant Protection Research Institution) and samples collected during surveys that formed part of the South African Plant-Parasitic Nematode Survey.[9,19]

The first extensive nematode survey was conducted during 1995/1996 at 17 localities situated within the local soybean production areas.[20] As a result, two nematode genera (*Longidorus* and *Tylenchorhynchus*) and 11 species were listed as new records for soybean in South Africa. The latter species were *Criconemoides sphaerocephalus, Helicotylenchus digonicus, H. microcephalus, Longidorus pisi, Meloidogyne ethiopica Pratylenchus crenatus, P. teres, P. thornei, Scutellonema truncatum, Tylenchorhynchus goffarti,* and *Xiphinema elongatum.* The predominant endoparasitic nematode pests identified from soybean roots during the survey were *Meloidogyne* spp. (*M. ethiopica, M. hapla, M. incognita* and *M. javanica*) and *Pratylenchus* spp. (*P. brachyurus* and *P. zeae*). Moreover, root-knot nematode second-stage juveniles (J2) were present in 91% of all root samples. It was also evident that the occurrence of the predominant endoparasitic nematodes was not restricted to sandy soil, but that they also occurred at localities containing soils with clay content as high as 35%. Although present in low population density levels, another economically important endoparasitic nematode species *Ditylenchus africanus* (peanut-pod nematode) was also identified from soybean roots*.* This nematode represents a definite production constraint for groundnut crops throughout local production areas.[21] The soybean cyst nematode, *Heterodera glycines*, that poses a significant threat to soybean production in other parts of the world[22,23] has, however, not been reported locally during the survey or to date (Marais M 2014, oral communication, June 6).

That root-knot nematodes are generally the predominant plant-parasitic nematodes associated with local soybean crops[8,20] corresponds with reports that these pests are also considered as a serious constraint to production of the crop worldwide[22,23]. Diagnostic nematode analyses revealed exceptional high root-knot nematode population density levels of 11 401 eggs and J2/50 g roots from RR plants that grew in the Bothaville area (Free State Province) during the 2011 season.[24] During April 2013, 161 213 *Meloidogyne* sp. eggs and J2/50 g roots were extracted from roots of a conventional soybean cultivar that was cultivated in the Edenville area (Free State Province).[25] The latter areas include those to where soybean production has been expanded recently. Increased infection of soybean and rotation crops included in soybean-based cropping systems by single or mixed populations of *M. incognita* and *M. javanica* is thus imminent because of the damage potential of such pests. The latter species commonly occur in areas where soybean was traditionally cultivated in South Africa as well as in those areas where maize is grown[8,10] and where soybean is now being introduced. This scenario particularly applies where soybean is included in conservation agriculture systems in which the use of herbicide tolerant cultivars is often preferred. To date, all genetically modified herbicide tolerant, RR soybean cultivars evaluated for their host suitability to *M. incognita* have been reported as susceptible.[26,27] This scenario emphasises and complicates the challenge faced by producers and the industry to manage these pests in future. The emphasis on soybean-nematode research has been on the use and exploitation of genetic resistance as a viable and environmentally safe tactic to reduce population levels of particularly root-knot nematodes. Initiatives in this regard will be discussed below, followed by knowledge gained in terms of other management tactics that may add value to soybean producers and the industry.

Table 1: Plant-parasitic nematodes associated with soybean in South Africa since 1959[9,16,17,18,20,64,65]

Nematodes	
Criconema corbetti[64]	*Pratylenchus brachyurus*[18]
C. mutabile[65]	*P. crenatus*[20]
C. pauciannulatum[9]	*P. neglectus*[9]
Criconema sp.[64]	*P. penetrans*[64]
Criconemoides parvus[9]	*Pratylenchus* sp.[65]
C. sphaerocephalus[20]	*P. teres*[20]
Ditylenchus africanus[20]	*P. thornei*[20]
Geocenamus brevidens[18]	*P. zeae*[9]
Helicotylenchus digonicus[20]	*Rotylenchulus parvus*[9]
H. dihystera[9]	*Rotylenchulus* sp.[64]
H. martini[64]	*Rotylenchus incultus*[64]
H. microcephalus[20]	*Rotylenchus* sp.[64]
H. paraplatyurus[65]	*R. unisexus*[9]
H. pseudorobustus[9]	*Scutellonema brachyurus*[18]
Helicotylenchus sp.[64]	*S. commune*[9]
Hemicriconemoides strictathecatus[64]	*Scutellonema* sp.[64]
Hemicycliophora sp.[64]	*S. truncatum*[20]
H. typica[64]	*S. unum*[64]
Longidorus pisi[20]	*Subanguina* sp.[64]
Longidorus sp.[64]	*Tylenchorhynchus brevilineatus*[64]
Meloidogyne arenaria[16]	*T. goffarti*[20]
Meloidogyne ethiopica[20]	*T. mashhoodi*[64]
Meloidogyne hapla[17]	*Tylenchorhynchus* sp.[65]
Meloidogyne incognita[17]	*T. ventralis*[64]
Meloidogyne javanica[17]	*Xiphinema clavatum*[65]
Meloidogyne sp.[18]	*X.elongatum*[20]
Nanidorus minor[9,64]	*X. mampara*[65]
Paratrichodorus lobatus[64]	*X. ornatizulu*[65]
P. porosus[64]	*Xiphinema* sp.[64]
Paratrichodorus sp.[64]	*X. vanderlindei*[9]
	X. zulu[64]

Genetic resistance to root-knot nematodes

The use of root-knot nematode resistant soybean cultivars is one of the most economically justified strategies for controlling root-knot nematode pests.[28] The rest of this review will thus focus on this strategy as the most popular, cost-effective and efficient strategy for sustainable production of soybean, while only a concise summary on other potential management strategies will be given.

The use of cultivars that exhibit resistance to root-knot nematodes generally results in substantial reductions in population levels of these pests.[12,22,28] Despite the phenomenon that J2s penetrate roots of resistant cultivars to the same extent as that of susceptible cultivars, sub-optimal development of J2s within the roots of resistant host plants follow with subsequent retarded development of all J2 life-cycle stages.[29-32] Significantly lower numbers of eggs are thus produced by mature root-knot nematode females that feed in roots of resistant cultivars opposed to those that parasitise roots of susceptible cultivars. Although a wide range of genotypes with varying levels of resistance to root-knot nematode species and races is available,[23,23] such material is not necessarily adapted to local environmental conditions. They are also not necessarily resistant to local root-knot nematode species and races as will be illustrated below.

Studies on the host status of local soybean cultivars to root-knot nematodes were first reported in the 1990s when 19 commercially available cultivars were screened for their host suitability to *M. javanica* and *M. incognita* race 4, respectively.[33] The cultivars differed with regard to their host suitability to the two respective root-knot nematode species, with relatively low to moderate levels of resistance being identified. During the end of the 1990s, further screening of local cultivars using various nematode life history parameters as criteria, namely egg-laying female indices, egg and J2 numbers/root system and reproduction factor (Rf) values, followed.[34] The latter parameters varied substantially for the 38 soybean cultivars that were screened against *M. incognita* race 1, 2 and 4 as well as *M. javanica*. According to (Rf)values, none of the cultivars exhibited resistance to *M. incognita* race 2 (Table 2). However, several were considered to have some level of resistance to *M. incognita* races 1 and 4 as well as *M. javanica* (Table 2). Sources of resistance in local soybean cultivars against *M. incognita* races 2 and 4 and *M. javanica* (Table 2) were also reported during the mid 2000s.[26,35] Of the 85 local and foreign soybean genotypes that were evaluated for host suitability to *M. incognita* race 2,[26] LS5995 exhibited the highest level of resistance (Rf=0.01) followed by PI96354, PAN780, Egret, PAN660, LS688, Potties, PAN564, G93-9106, G93-9009, G93-9201 and LS666 (Table 2). Interestingly, Forrest that was recorded with partial resistance to USA populations of *M. incognita,*[36] proved to be susceptible to some local *M. incognita* race 2 populations.[26,27]

Although gall ratings and egg mass indices were commonly used as criteria for determining root-knot nematode resistance in soybean,[28,37] egg production is generally regarded as a more reliable criterion.[28,38] In some cases, soybean genotypes exhibited low gall ratings but high egg-laying female indices and high numbers of eggs/plant. This unexpected crop reaction to root-knot nematode infection was also reported for exotic soybean cultivars.[36] It implies that using gall ratings alone can lead to inaccurate interpretation of data regarding cultivar resistance to root-knot nematodes. This phenomenon is further illustrated as plant resistance could be affected through one or several different mechanisms of resistance.[28,37] Several criteria describing the possible resistance mechanism involved should thus be applied during the identification of resistance in crop cultivars. Results obtained during screenings, however, ultimately resulted in valuable knowledge being available for use in the planning of crop rotation systems as well as the exploitation of sources of resistance for breeding purposes. Undoubtedly, the continuous screening of cultivars that enter the market is crucial because producers should be updated annually on poor-host cultivars that could be used in their rotation systems. In this way, root-knot nematode populations can be reduced on a continuous basis to allow for the sustainable production of crops.

Table 2: Soybean genotypes resistant to various South African root-knot nematode species and races[26,27,34,35,66]

Genotype	*Mi* 1	*Mi* 2	*Mi* 4	Mj	Genotype	*Mi* 1	*Mi* 2	*Mi* 4	Mj
A5308[34]	√				GCI-PRF 7[27]		√		
A5409[34,35]			√	√	Highveld Top[35]			√	
A7119[34,35]	√		√	√	Hutton[34]	√		√	
Bakgat[35]		√	√	√	LS5995[26,27]		√		
Bamboes[34,35]	√	√	√		LS5995[66]				√
Crawford[35]		√	√	√	LS688[26]		√		
Columbus[35]			√	√	LS666[26]		√		
CRN2233[34]	√		√		Nyala[34]			√	√
D82-3298[26]		√			PAN494[34]				√
Egret[26]		√			PAN494[35]		√	√	
Forrest[35]		√	√		PAN581[34]				√
Forrest[34]	√			√	PAN723[34]	√		√	
Gazelle[34]				√	PAN780[26]		√		
G93-9201[26]		√			PAN790[35]		√	√	
G93-9009[26]		√			PAN812[34]	√			√
G93-9106[26]		√			PAN812[35]		√	√	
GCI-PRF 1[27]		√			PI96354[26]		√		
GCI-PRF 2[27]		√			SCSI[34]				√
GCI-PRF 3[27]		√			SNK60[34]	√			√
GCI-PRF 4[27]		√			Talana[34]	√			√
GCI-PRF 5[27]		√			Zebra[34]				√
GCI-PRF 6[27]		√							

Molecular markers linked to root-knot nematode resistance

Molecular methods, i.e. marker-assisted selection (MAS) during breeding, have been applied widely to improve the success rate and levels of root-knot nematode resistance selection[28] and to accelerate the development of soybean cultivars that exhibit this trait.[39-44] Genetic markers associated with resistance to *M. incognita*[41-43], *M. javanica*[39,44] and *M. arenaria*[40] have been identified using amplified fragment length polymorphism (AFLP), restriction fragment length polymorphism (RFLP), sequence characterised amplified regions (SCAR) and/or micro-satellite or simple sequence repeat (SSR) markers. The use of such markers linked to root-knot nematode resistance traits in soybean cultivars and subsequent application of MAS is a quick and effective way to expedite nematode resistance breeding processes[28,39-44], which has also been exploited in local resistance breeding programmes[45,46]. The latter include identification and verification of *M. javanica*[45] and *M. incognita* resistance[46], being additional milestones for soybean-nematode research in South Africa.

Meloidogyne javanica

AFLP markers linked to *M. javanica* resistance in the local soybean cultivar Gazelle and subsequent conversion thereof to SCARS[45], was the first successful attempt for such research on soybean in South Africa. A close linkage of RFLP marker B212 was reported for the resistance trait, accounting for 62% of the variation in *M. javanica* gall index measurements. Marker data obtained in this regard corresponded with those for a marker located in the same region on LG-F for the exotic *M. javanica*-resistant soybean line PI230977.[39] However, the other marker, A725-2 situated on LG-D1, that accounts for only 13% of gall index variation in the latter exotic line, was not polymorphic for the two parents used in the local mapping population and thus not detected in Gazelle.

The AFLP fragments identified in Gazelle were then used to develop a marker system that is easily and economically applicable in MAS in local breeding programmes.[45] Marker E-AAC/M-CAT1 (LG-F) that linked in the repulsion phase accounted for the greatest variation in gall indices (42%), while marker E-ACC/M-CTC2 (LG-F) that linked in the coupling phase explained 25% of the variation for the same nematode parameter. Both markers associated with the gall index parameter were linked to LG-F. The quantitative trait locus (QTL) for gall-index resistance mapped between markers B212 and E-AAC/M-CAT1 (SOJA7), which according to MAPMAKER-EXP analysis are only 2.4 cM apart. E-ACC/M-CTC2 (SOJA6), mapped near B212, with the QTL being recorded as 3.8 cM from marker B212. The latter indicates that the combined use of these two markers in MAS could be very effective. Subsequently the two AFLP markers mapping closely to and bracketing the *M. javanica* resistance trait in Gazelle were successfully converted to SCARs (SOJA6 and SOJA7, respectively) and employed for MAS in a breeding population. SOJA6 distinguished between homozygotic and heterozygotic progeny and SOJA7 against homozygous resistant plants only. Successful conversion of AFLP markers to either RFLP or polymerase chain reaction (PCR) markers has been reported by only a few authors.[47-50] Additional QTLs, if present, would likely be of minor importance in terms of a contribution to explaining variation in gall index for *M. javanica* resistance in local cultivar Gazelle.

Meloidogyne incognita

For identification of molecular markers associated with *M. incognita* resistance, the soybean cultivar LS5995 was used as the resistant parent in crosses to obtain a segregating F_2 mapping population.[46] A lack of polymorphism for SSR markers between the parents, however, complicated the identification and location of QTLs associated with the *M. incognita* resistance in LS5995. As a result of the lack of polymorphism, accurate mapping of a major QTL associated with resistance to *M. incognita* in the resistant USA soybean line PI96354[41] could not be achieved for the local resistant LS5995. The F_2 population was screened with a number of SSR evenly distributed throughout the soybean genome, with Satt201, Satt358, Satt487 and Satt590 being the SSR markers identified as those linked to the resistance trait.[46] QTL data obtained for LS5995 differed from those published by other authors[41,42] because no SSR markers could be identified in association with resistance to *M. incognita* in LS5995 on LG-G. In contrast, a minor QTL (Satt012) on LG-G that explained 18% of the variation in *M. incognita* gall indices[41] as well as three SNPs in Satt199 source-sequences were linked near a major QTL in the exotic line PI96354.[42] In the local F_2 mapping population, two QTLs were identified using variation in gall indices and egg and J2 numbers/root system. One of these was located on LG-O (close to the one identified for the exotic line PI96354),[42] while the other major QTL was located on LG-M. The QTL located near Satt358 explained 56% of the variation in gall indices on LG-O for PI96354, while in local cultivar LS5995, it only explained 32% of the phenotypic variation for this parameter. On the other hand, the major QTL on LG-M identified in this study explained 80% of the variation in *M. incognita* eggs and J2 numbers/root system. The latter QTL was not identified in the exotic line PI96354,[42] suggesting that different resistance mechanisms are involved in the two genotypes. The frequency distribution of the F_2 progeny for both *M. incognita* gall indices and eggs and J2 numbers/root system suggested that resistance in LS5995 was quantitatively inherited and is thus controlled by a number of genes and not by partially dominant inheritance of one major gene.[51] Validation of molecular markers associated with resistance to *M. incognita* present in local and foreign cultivars and an F_6 progeny of a LS5995 x Prima2000 cross[52] further emphasises and confirms the value of MAS in soybean breeding programmes. Important to note, however, is that the presence of only one of the several markers identified in a particular root-knot nematode-resistant cultivar does not necessarily guarantee the introgression of such resistance. For example, the *M. incognita*-resistant soybean cultivar Forrest contains three markers (Satt201, Satt487 and Satt358)[46] but proved to be susceptible when screened to a local population of this species in several glasshouse experiments[46,34]. This scenario also emphasises the use of MAS in combination with traditional screening procedures to ensure that resistance is successfully introgressed into a genotype. Ultimately, interventions to pyramid both minor and major genes linked to *M. incognita* and *M. javanica* resistance should enhance the level of these traits in local soybean cultivars. This way, superior and polyspecific levels of root-knot nematode resistance can be selected for in germplasm.

Resistance mechanism(s)

The resistance mechanism(s) exhibited by the *M. incognita*-resistant cultivar LS5995 was determined by means of penetration, development, reproduction and histopathology studies.[29] These studies showed that J2s initially penetrated roots of the resistant LS5995 and susceptible Prima2000 in equal numbers. However, the J2 penetration rate was significantly lower in roots of LS5995 10 days after inoculation, which corresponds with results for exotic *M. incognita*-resistant genotypes reported by other authors.[30-32] Furthermore, numbers of J2 developmental stages were 4.6-fold higher in the roots of the susceptible Prima2000 compared to that in LS5995. Ultimately, the number of eggs/egg mass and numbers of eggs/root system, which are important indicators of antibiosis resistance, were also significantly lower in LS5995. These studies therefore indicated that the major mechanism of resistance in the resistant cultivar represented typical post-infectional antibiosis.

Histopathology investigations on the other hand illustrated that *M. incognita* J2s penetrated roots of both the resistant and susceptible cultivars and migrated intercellularly to the parenchyma cells in the vascular cylinder 2 days after inoculation.[29] Pronounced cellular changes were also observed in the roots of both cultivars between 10 and 30 days after inoculation and generally represented those reported for other exotic resistant and susceptible cultivars.[53,54] The presence of sub-optimal giant cells, some with distinctly thicker cell walls that was recorded for the resistant LS5995 had not been reported for other resistant soybean cultivars or resistant cultivars of other crops before.[53,54]

Whether the presence of such atypical giant cells with thicker giant cell walls in roots of LS5995 can be ascribed to differences in genetic markers associated with *M. incognita* resistance in LS5995 compared to those reported in exotic cultivars[42] is unknown and warrants further investigation. Giant cells in the roots of the resistant LS5995 were also smaller and fewer (Figure 4) compared to those in roots of its susceptible counterpart (Figure 5). The association of *M. incognita* individuals that showed retarded development with sub-optimal giant cell formation (including necrosis around the giant cells; Figure 6) and ultimately reduced reproduction and fecundity in LS5995, further illustrated the presence of multiple defence strategies at genetic level to withstand parasitism by *M. incognita*. These findings also complemented the quantitative nature of the resistance identified in LS5995. Finally, studies at cellular level substantiated and gave insight into host plant defence mechanisms employed in LS5995. This can contribute to enhancing the development of strategies to better understand and engineer resistance against this species.

Verification of resistance

Damage-threshold levels

The damage-threshold level that was determined for *M. incognita* in the resistant cultivar LS5995 in semi-field studies was 10 times higher compared to that of its susceptible counterpart.[12,55] This study illustrated that a relatively small economic loss will be sustained when a resistant cultivar is planted compared to a susceptible one. However, to extrapolate and apply such information to other areas is complicated because numerous factors have an effect on nematode threshold levels, such as environmental conditions, the cultivar planted, the nematode species/race present as well as other geographic and edaphic factors.[56] Therefore, it is suggested that damage-threshold values for root-knot nematodes in agricultural crops, soybean in particular, should be considered circumspectively and at most, be used as guidelines for implementation of control action. This implies that successful management of root-knot nematodes in soybean cannot be done by using a single strategy such as host plant resistance. Additional nematode control strategies such as chemical control (if available), cultural control and others should for

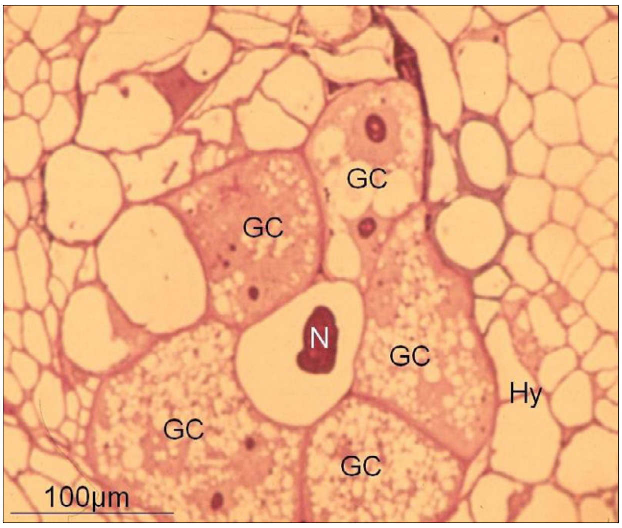


Image: Driekie Fourie; 100x magnification

GC, giant cell; Hy, hyperplasia; N, nematode

Figure 4: A light-microscope micrograph of a transverse section showing five distinct and optimal giant cells visible in the mature provascular cylinder in the root of a susceptible soybean cultivar 10 days after inoculation with *Meloidoyne incognita* second-stage juveniles (J2s).

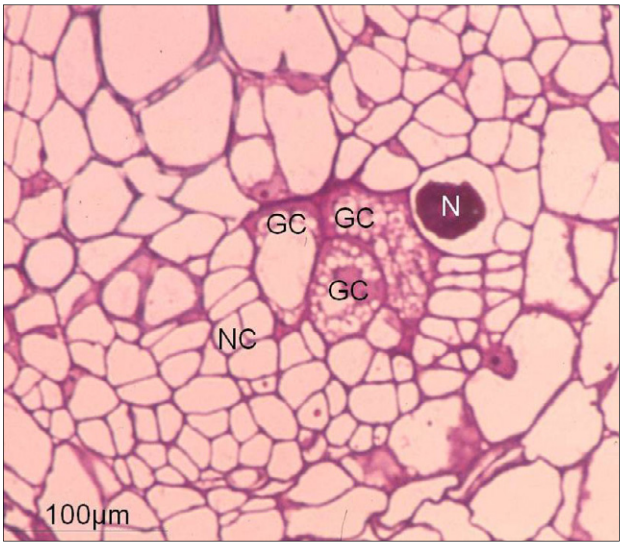


Image: Driekie Fourie; 100x magnification

GC, giant cell; N, nematode; NC, normal cell

Figure 5: A light-microscope micrograph of a transverse section showing three distinct, non-optimal giant cells visible in the mature provascular cylinder in the root of a resistant cultivar, LS5995, 10 days after inoculation with *Meloidoyne incognita* second-stage juveniles (J2s).

example be used on an integrated basis to ensure effective long term suppression of nematode populations.

Glasshouse and semi-field studies

Resistance against *M. incognita* was verified by determining the effect of increasing initial population density levels (Pi) on population dynamics and yield of the resistant LS5995 as well as a susceptible counterpart in semi-field studies.[12] Strong, non-linear relationships existed between Pi for all the nematode variables used, namely, number of egg masses, egg-laying females indices, Rf-values and percentage yield loss. Non-linear models for Pi against percentage yield loss indicated that yield loss in LS5995 was at least six times lower than that of the susceptible cultivar, which demonstrated the monetary benefit provided by the resistant cultivar. Similar results and benefits of differential yield loss in exotic resistant soybean cultivars grown in *M. incognita*-infested soil have been reported.[57]

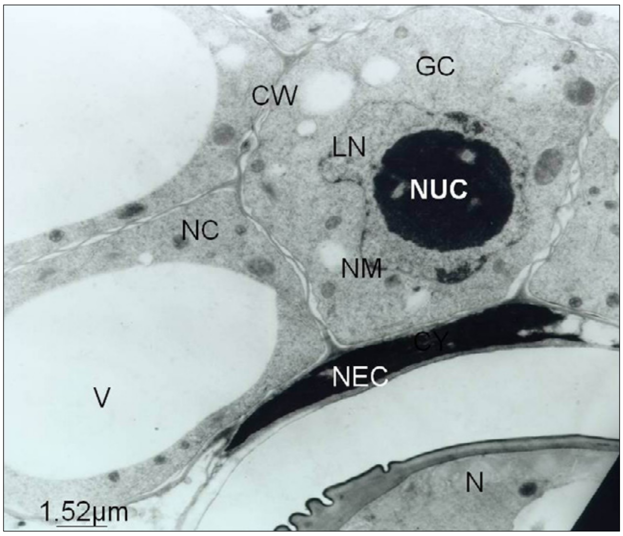


Image: Wilna Pretorius

CW, cell-wall; GC, giant cell; N, nematode; NC, normal cell; NEC, necrotic cell; NM, nucleus membrane; LN, lobed nucleus; NUC, nucleolus; V, vacuole

Figure 6: A transmission electron micrograph of a transverse section through an undifferentiated provascular cylinder showing necrotic cell tissue visible in the root of a resistant cultivar, LS5995, near the head of a feeding *Meloidoyne incognita* second-stage juvenile (J2) 2 days after inoculation.

Distinct differences in the response of *M. incognita* individuals in terms of their life history parameters, furthermore confirmed the resistance trait in LS5995.[29] The importance of considering as many as possible nematode parameters when investigating aspects such as host plant resistance was highlighted as a result of this research. It is also suggested that the use of host plant resistance in only one crop (soybean in this case) in local rotation cycles with other susceptible crops such as maize[8,10], sunflower[58], potato[59] or dry bean[9,11] may not be sufficient for keeping *M. incognita* population levels below damage-threshold levels in the medium to long term. Also, the use of host-plant resistance as the only management strategy may not provide sufficient protection against nematode pests. Therefore, careful consideration of as many factors as possible during the planning and development of production systems where nematode-susceptible crops will be cultivated in root-knot nematode infested soils needs to be considered.

Field studies

Host and yield responses of *M. incognita*-resistant genotypes identified in initial screenings, including LS5995, together with two susceptible soybean cultivars were furthermore verified under natural environmental conditions.[55] Root-knot nematode numbers in both soil and root samples were significantly higher for all genotypes inoculated with *M. incognita* eggs and J2 compared to the uninoculated control plants. Furthermore, the number of eggs and J2 in the roots of nematode-infected plants was significantly higher in the susceptible cultivars compared to the resistant genotypes, except for the resistant cultivar Potties in one of the trial sites. In contrast with the high reproduction of root-knot nematodes in roots of the susceptible Prima/Prima2000, LS5995 in particular, consistently maintained significantly lower *M. incognita* population levels in all field experiments. Also, in the majority of the experiments, yield of the resistant genotypes did not differ significantly between the uninoculated and the nematode-infected plants. Yield response was, however, generally dependent on environmental effects, as was also indicated by authors,[56,57] and thus limited further qualification of resistant or susceptible soybean genotypes as tolerant, intolerant and/or hypersensitive.

Eight F_8 lines with superior levels of resistance to *M. incognita*[52], resulting from breeding efforts to identify molecular markers in the local resistant cultivar LS5995 were additionally evaluated for their agronomic performance during two seasons at eight different localities[60]. Yield data of two of these lines were similar to that of Egret, which is the only cultivar with resistance to *M. incognita* that is currently available commercially. These genotypes do not contain the RR gene and no such root-knot nematode-resistant soybean cultivars are registered at this stage in South Africa, therefore, exploitation of such material and conversion thereof to RR material will add value to the local soybean industry as an investment to complement sustainable production of the crop.

Other management strategies

Chemical and biological control

Although several nematicides are registered for use on soybean worldwide[22,23], but not in South Africa[61], application of such products is seldom economically justifiable[22,23]. Locally, evaluation of nematicides on soybean that was cultivated in *M. javanica*-infested soil included ethylene-dibromide (EDB®), aldicarb and chlorpirifos as well as two biological products. The latter contained *Paecelomyces lilacinus* and *Bacillus* spp., respectively.[62] Plots where the soil was fumigated with EDB® consistently resulted in the lowest *M. javanica* population levels and highest yields, followed by those treated with aldicarb and terbufos. The bionematicide treatments did, however, not always differ significantly from the untreated control or other treatments in terms of their efficacy.

Nematicide efficacy was also conducted under field conditions where high infestation levels of *M. incognita* were present.[63] Fourteen nematicide treatments were included and represented various dosages of aldicarb, abamectin, cadusafos, oxagran, oxamyl and terbufos. Significant differences in efficacy of nematicides existed with regard to the number of egg and J2 counts/50 g roots at all localities. Only oxamyl SL, terbufos GR and abamectin/seed treatments resulted in a significant reduction of *M. incognita* numbers in 50g roots and generally showed a higher income/ha compared to the untreated control. Although data from these trials indicated that synthetically-derived nematicides may provide relief to producers where root-knot nematodes attain high pest status, cost-efficacy analyses did not allow registration of these products for use in local soybean production systems at the time. However, the evaluation of newly-developed products with nematicidal properties should proceed. The use of sustainable strategies where such products can contribute to reduce plant-parasitic nematodes in an integrated management approach can be advantageous to producers.

Conclusion

The magnitude of plant-parasitic nematode problems, focussing on root-knot nematodes, in soybean has been illustrated in this article. Sustainable production of soybean is likely to be jeopardised as a result of the build-up of various root-knot nematode species, in particular, in soybean-based cropping systems. This necessitates research aimed at quantifying the impact of these and other nematode pests in producing areas where cultivation of the crop is planned. In addition, it is crucial that funding be secured on a continuous basis to develop improved management systems for these pests in soybean-based cropping systems. Current approaches towards environmentally-friendly strategies to combat plant-parasitic nematodes increase the pressure on researchers and decision-makers in the soybean industry to coordinate research initiatives and seek sustainable solutions. Future research should particularly address the identification of alternative sources of resistance to economically important plant-parasitic nematodes, such as *Meloidogyne* spp. In addition, the development of integrated strategies to combat nematode pests of soybean should be addressed.

Acknowledgements

The research was partially financed by the Protein Research Foundation, while the North-West University and Agricultural Research Council-Grain Crops Institute supplied the infrastructure. The following research assistants provided valuable support and inputs during the execution of research activities related to this manuscript: Samuel Kwena, Rita Jantjies, Lizette Bronkhorst, Nicolene de Klerk and Heila Vermeulen from the Agricultural Research Council-Grain Crops Institute as well as Edith du Randt, Emily Phetoe and Helena Strydom from the North-West University. The financial assistance of the National Research Foundation towards this research is hereby also acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not to be attributed to the NRF.

Authors' contributions

H.F. initiated the writing of this review article and compiled the initial manuscript. As the promotor and co-promoter of H.F.'s MSc and PhD studies, respectively, which represents a major part of nematology research done on soybean locally, D.d.W. and A.H.M. edited and provided valuable inputs to the manuscript. C.M. is the molecular specialist that was involved in the execution of the molecular identification of genetic markers associated with the resistance in soybean cultivars to two *Meloidogyne* species. M.M is a nematode taxonomist who identified plant-parasitic nematodes associated with soybean to species level and supplied the information from the South African Plant-Parasitic Nematode Survey database about plant-parasitic nematodes reported from soybean in South Africa. A.D. has been involved in executing field trials to assess yield and other data on the agronomic performance of S8 progeny of soybean lines in which *Meloidogyne incognita* and *M. javanica* resistance has been introgressed by H.F. and a soybean breeder (Pieter Herbst from the former LinkSeed). M.M. and A.D. also contributed towards inclusion of data and information and the final editing of the manuscript.

References

1. Protein Research Foundation (PRF). Statistics and estimates [homepage on the Internet]. No date [cited 2014 Mar 27]. Available from: http://www.proteinresearch.net/index.php?dirname=html_docs_030statistics_and_estimates

2. Liebenberg A. Soybean production manual: Your guide to successful soybean production. Potchefstroom: Agricultural Research Council; 2012.

3. South African Grain Information Service (SAGIS). CEC crop estimates [homepage on the Internet]. No date [cited 2014 Mar 27]. Available from: http://www.sagis.org.za / /Flatpages/Oesskatting.htm; 2013.

4. James C. International service for the acquisition of Agri-biotech applications (ISAAA). Report on global status of biotech/GM crops [homepage on the internet]. c2009 [cited 2013 Nov 23]. Available from: http://www.isaaa.org/.

5. Du Plessis H. First report of groundnut leafminer *Aproaerema modicella* (Deventer) (Lepidoptera: Gelechiidae) on groundnut, soybean and lucerne in South Africa. S Afr J Plant Soil. 2001;20(1):48. http://dx.doi.org/10.1080/02571862.2003.10634906

6. Jarvey JA. A review on soybean rust from a South African perspective. SA J Soil Sci. 2009;105:103–108.

7. Coyne DL, Fourie HH, Moens M. Current and future management strategies in resource-poor farming. In: Perry R, Moens M, Starr JL, editors. Root-knot nematodes. Wallingford: CABI; 2009. p. 444–475. http://dx.doi.org/10.1079/9781845934927.0444

8. Riekert HF, Henshaw GE. Effect of soybean, cowpea and groundnut rotations on root-knot nematode build-up and infestation in dryland maize. Afr Crop Sci J. 1998;6:377–383. http://dx.doi.org/10.4314/acsj.v6i4.27789

9. Kleynhans KPN, Van den Berg E, Swart A, Marais M, Buckley NH. Plant nematodes in South Africa. Pretoria: Business Print; 1996.

10. Riekert HF. Economic feasibility of nematode control in dryland maize in South Africa. Afr Crop Sci J. 1996;4:477–481.

11. Keetch DP. A perspective of plant nematology in South Africa. S Afr J Sci.1989;85:506–508.

12. Fourie H, Mc Donald AH, De Waele D. Relationships between initial population densities of *Meloidogyne incognita* race 2 and nematode population development in terms of variable soybean resistance. J Nematol. 2010;42:55–61.

13. Fourie H, Mc Donald AH. Control of root-knot nematodes on soybean. Presented at the Fifteenth Symposium of the Nematological Society of Southern Africa. Afr Plant Prot. 2002;8:79–80.

14. Smit MA, De Beer GP. Report of the national soybean cultivar trials 1998/99. Potchefstroom: Agricultural Research Council; 1998.

15. Food Agricultural Organization (FAO). Food and Agriculture Organization of the United Nations [homepage on the Internet]. No date [cited 2014 Mar 27]. Available from: www.faostat.fao.org/

16. Van der Linde WJ, Clemitson JG, Crous ME. Host-parasite relationships of South African root-knot eelworm (*Meloidogyne* spp.). Dep Agric Techn Serv Repub S Afr Ent Ser. 1959;44:3–16.

17. Coetzee v. The distribution of the family Heteroderidae (Filipjev, 1934) in South Africa and some host records of Meloidogyne species. S Afr J Agr Sci. 1968;11:775–788.

18. Keetch DP, Buckley NH. A check-list of the plant-parasitic nematodes of southern Africa: Technical Communication No. 195. Pretoria: Department of Agriculture, Government Printer; 1984.

19. Marais M. South African plant-parasitic nematode survey (SAPPNS). Plant Prot News. 2006;67:6.

20. Fourie H, Mc Donald AH, Loots GC. Plant-parasitic nematodes in field crops in South Africa 6: Soybean. Nematology. 2001;3:447–454. http://dx.doi.org/10.1163/156854101753250773

21. De Waele D, Jones BL, Bolton C, Van Den Berg E. *Ditylenchus destructor* in hulls and seeds of peanut. J Nematol. 1989;21:10–15.

22. Bridge J, Starr JL. Plant nematodes of agricultural importance. Boston, MA: Academic Press; 2007. http://dx.doi.org/10.1201/b15142

23. Sikora RA, Greco N, Silva JFV. Nematode parasites of food legumes. In: Luc M, Sikora RA, Bridge J, editors. Plant parasitic nematodes in subtropical and tropical agriculture. Wallingford: CABI; 2005. p. 259–318. http://dx.doi.org/10.1079/9780851997278.0319

24. Fourie H, Bekker S, Mc Donald AH, Engelbrecht E. Prospective extended benefits of plant-nematode diagnostic and advisory services should closer cooperation between laboratories be attainable. Presented at the Twentieth NSSA Symposium. S Afr J Plant Soil. 2011;28(4):265.

25. Fourie H. Research Report No 3: Nematode survey (2012/2013) in local soybean production areas. Potchefstroom: Nematology Unit, Plant Protection, North-West University; 2014.

26. Fourie H, Mc Donald AH, De Waele D. Host suitability of South African and foreign soybean cultivars to *Meloidogyne incognit*a race 2. S Afr J Plant Soil. 2006;23:132–137. http://dx.doi.org/10.1080/02571862.2006.10634743

27. Venter C. Exploitation and characterisation of resistance to the root-knot nematode *Meloidogyne incognita* in soybean [MSc dissertation]. Potchefstroom: North-West University; 2014.

28. Starr JL, Mercer CF. Development of resistant varieties. In: Perry R, Moens M, Starr JL, editors. Root-knot nematodes. Wallingford: CABI; 2009. p. 326–337. http://dx.doi.org/10.1079/9781845934927.0326

29. Fourie H, Mc Donald AH, De Waele D, Jordaan A. Comparative cellular responses in resistant and susceptible soybean cultivars infected with *Meloidogyne incognita*. Nematology. 2013;15:695–708. http://dx.doi.org/10.1163/15685411-00002712

30. Moura RM, Davis EL, Luzzi BM, Boerma HR, Hussey RS. Post-infectional development of *Meloidogyne incognita* on susceptible and resistant soybean genotypes. Nematropica. 1993;23:7–13.

31. Hadisoeganda WW, Sasser JN. Resistance of tomato, bean, southern pea, and garden pea cultivars to root-knot nematodes based on host suitability. Plant Dis. 1982;66:145–150. http://dx.doi.org/10.1094/PD-66-145

32. Kaplan DT, Davis EL. Mechanisms of plant incompatibility with nematodes. In: Veech JA, Dickson W, editor. Vistas on nematology. Hyattsville, MD: Society of Nematologists Inc.; 1987. p. 267–276.

33. Van den Berg E, Mc Donald AH. Screening of soybean cultivars for resistance against *M. javanica* and *M. incognit*a race 4. Presented at the Tenth Symposium of the Nematological Society of Southern Africa; 1991 Apr 7–10; Wilderness, South Africa. Available from: http://www.sanematodes.com/Documents/Abstracts1991.pdf.34.

34. Fourie H, Mc Donald AH, Loots GC. Host suitability of South African commercial soybean cultivars to two root-knot nematode species. Afr Plant Prot. 1999;2:119–124.

35. Van Biljon ER. Crop rotation as part of an integrated pest management programme for the control of plant parasitic nematodes. Presented at: The Fourteenth Soilborne Plant Diseases Interest Group Symposium; 2004 Sept 15–16; Stellenbosch, South Africa. Available from: http://www.saspp.co.za.

36. Luzzi BM, Boerma HR, Hussey RS. Resistance to three species of root-knot nematode in soybean. Crop Sci. 1987;27:258–262. http://dx.doi.org/10.2135/cropsci1987.0011183X002700020027x

37. Cook R, Starr JL. Resistant cultivars. In: Perry R, Moens M. Plant nematology. Wallingford: CABI; 2006. p. 370–391. http://dx.doi.org/10.1079/9781845930561.0370

38. Windham GL, Williams WP. Reproduction of *Meloidogyne javanica* on corn hybrids and inbreds. Ann Appl Nematol. 1988;2:25–28.

39. Tamulonis JP, Luzzi BM, Hussey RS, Parrott WA, Boerma HR. DNA markers associated with resistance to Javanese root-knot nematode in soybean. Crop Sci. 1997;37:783–788. http://dx.doi.org/10.2135/cropsci1997.0011183X003700060039x

40. Tamulonis JP, Luzzi BM, Hussey RS, Parrott WA, Boerma HR. DNA marker analysis of loci conferring resistance to peanut root-knot nematode in soybean. Theor Appl Genet. 1997;95:664–670. http://dx.doi.org/10.1007/s001220050610

41. Tamulonis JP, Luzzi BM, Hussey RS, Parrott WA, Boerma HR. RFLP mapping of resistance to southern root-knot nematode in soybean. Crop Sci. 1997;37:1903–1909. http://dx.doi.org/10.2135/cropsci1997.0011183X003700060039x

42. Li Z, Jakkula L, Hussey RS, Tamulonis JP. SSR mapping and confirmation of QTL from PI96354 conditioning soybean resistance to southern root-knot nematode. Theor Appl Genet. 2001;103:1167–1173. http://dx.doi.org/10.1007/s001220100672

43. Ha BK, Hussey RS, Boerma HR. Development of SNP assays for marker-assisted selection of two southern root-knot nematode resistance QTL in soybean. Crop Sci. 2007;47:73–82. http://dx.doi.org/10.2135/cropsci2006.10.0660tpg

44. Fuganti R, Beneventi MA, Silva JFV, Arias CAA, Silvana RR, Binneck ME, et al. Identification of microsatellite markers for the selection of soybean genotypes resistant to *Meloidogyne javanica*. Nematol Bras. 2004;28:125–130.

45. Mienie CMS, Fourie H, Smit MA, Van Staden J, Botha FC. Identification of AFLP markers in soybean linked to resistance to *Meloidogyne javanic*a and conversion to sequence characterized amplified regions (SCARS). Plant Growth Regul. 2002;37:157–166. http://dx.doi.org/10.1023/A:1020585023976

46. Fourie H, Mienie CMS, Mc Donald AH, De Waele D. Identification of genetic markers associated with *Meloidogyne incognita* race 2 resistance in soybean (Glycine max L. Merr). Nematology. 2008;10:651–661. http://dx.doi.org/10.1163/156854108785787235

47. Cho YG, Blair WM, Panaud O, McCouch SR. Cloning and mapping of variety-specific rice genomic DNA sequences: Amplified fragment length polymorphisms (AFLP) from silver-stained polyacrylamide gels. Genome. 1996;39:373–378. http://dx.doi.org/10.1139/g96-048

48. Meksem K, Leister D, Peleman J, Zabeau M, Salamini F, Gebhardt C. A high-resolution map of the vicinity of the R1 locus on chromosome V of potato based on RFLP and AFLP markers. Mol Gen Genet. 1995;249:74–81. http://dx.doi.org/10.1007/BF00290238

49. Qu LJ, Foote TN, Roberts MA, Aragon-Alcaide L, Snape JW, Moore G. A simple PCR-based method for scoring the ph1b deletion in wheat. Theor Appl Genet. 1998;96:371–375. http://dx.doi.org/10.1007/s001220050751

50. Shan X, Blake TK, Talbert LE. Conversion of AFLP markers to sequence-specific PCR markers in barley and wheat. Theor Appl Genet. 1999;98:1072–1078. http://dx.doi.org/10.1007/s001220051169

51. Fehr WR. Principles of cultivar development: Vol. 1: Theory and technique. New York: Macmillan; 1987.

52. Fourie H, Mc Donald AH. Report on project O14/01: Introduction of root-knot nematode resistance into local, popular soybean genotypes using marker-assisted selection (MAS) and enhanced generation advancement. Potchefstroom: Agricultural Research Council – Grain Crops Institute; 2010. p. 8–9.

53. Barker KR, Hussey RS. Histopathology of nodular tissues of legumes infected with certain nematodes. Phytopathology. 1976;66:851–855. http://dx.doi.org/10.1094/Phyto-66-851

54. Pedrosa EMR, Hussey RS, Boerma HR. Penetration and post-infectional development and reproduction of *Meloidogyne arenaria* races 1 and 2 on susceptible and resistant soybean genotypes. J Nematol. 1996;28:343–351.

55. Fourie H, Mc Donald AH, De Waele D. In vivo host and yield responses of *Meloidogyne incognit*a-resistant and susceptible soybean genotypes. Int J Pest Manag. 2013;59(2):111–121. http://dx.doi.org/10.1080/09670874.2013.772261

56. Barker KR, Olthof THA. Relationships between nematode population densities and crop response. Annu Rev Phytopathol. 1976;14:327–353. http://dx.doi.org/10.1146/annurev.py.14.090176.001551

57. Herman M, Hussey RS, Boerma HR. Response of resistant soybean plant introductions to *Meloidogyne incognit*a in field microplots. J Nematol. 1990;22:237–241.

58. Bolton C, De Waele D, Loots GC. Plant-parasitic nematodes on field crops in South Africa 3: Sunflower. Rev Nématol. 1989;12:69–76.

59. Onkendi EM, Moleleki LN. Distribution and genetic diversity of root-knot nematodes (*Meloidogyne* spp.) in potatoes from South Africa. Plant Pathol. 2013;62(5):1184–1192. http://dx.doi.org/10.1111/ppa.12035

60. De Beer A, Fourie H, Venter C, De Klerk N. Evaluation of root-knot nematode (*Meloidogyne incognita*) resistant soybean genotypes for their host status and yield potential. Presented at: The Combined Congress; 2014 Jan 20–23; Grahamstown, South Africa. p. 39.

61. Van Zyl K. A guide to crop pest management in South Africa. A compendium of acaracides, insecticides, nematicides, molluscicides, avicides and rodenticides. A CropLife South African Compendium. 1st ed. Pinetown: VR Print; 2013.

62. Fourie H, Mc Donald AH. Report for Project M151/60: Chemical control options for plant-parasitic nematodes associated with soybean in South Africa. Potchefstroom: Agricultural Research Council – Grain Crops Institute; 2001.

63. Fourie H, Mc Donald AH. Chemical control options for plant-parasitic nematodes associated with soybean in South Africa: Report for Project M151/60. Potchefstroom: Agricultural Research Council – Grain Crops Institute; 2007.

64. Marais M. Identification job sheet N3092: Dataset from South African Plant-Parasitic Nematode Survey database. Pretoria: Nematology Unit, Biosystematics Division, Plant Protection Research Institute, Agricultural Research Council; 2012.

65. Marais M, Swart A. Plant nematodes in South Africa 8: Bizana, Lusikisiki and Port St Johns area, Eastern Cape Province. Afr Plant Prot. 2007;13:16–27.

66. De Beer, G Pretorius, AJ Fourie H. Soybean cultivar recommendations for 2001/2002. Potchefstroom: Agricultural Research Council – Grain Crops Institute; 2002.

The time to degree or dropout amongst full-time master's students at University of KwaZulu-Natal

AUTHORS:
Temesgen Zewotir[1]
Delia North[1]
Mike Murray[1]

AFFILIATION:
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Temesgen Zewotir

EMAIL:
zewotir@ukzn.ac.za

POSTAL ADDRESS:
School of Mathematics,
Statistics and Computer Science,
University of KwaZulu-Natal,
Private Bag X54001, Durban
4000, South Africa

DATES:
Received: 26 Aug. 2014
Revised: 06 Nov. 2014
Accepted: 22 Jan. 2015

KEYWORDS:
competing risk; hazard; model fitting; survival function; throughput

HOW TO CITE:
Zewotir T, North D, Murray M. The time to degree or dropout amongst full-time master's students at University of KwaZulu-Natal. S Afr J Sci. 2015;111(9/10), Art. #2014-0298, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20140298

Universities around the world are grappling with strategies to increase throughput and minimise dropout rates of postgraduate students. This study focuses on students at the University of KwaZulu-Natal and we attempt to estimate the time that it takes for these students to successfully complete or drop out from a master's programme. We used survival analysis to identify the factors which affect this. The results of this analysis showed that having some form of financial aid and/or being a student in the Faculties of Humanities or Management, all significantly shortened the length of time that it took to eventually drop out from a master's programme. For students who successfully completed a master's degree, having some form of financial aid, being of international origin and/or being registered in the Faculties of Health, Humanities, Law or Management, all helped to significantly shorten the length of time it took to successfully complete a master's programme. Students in the Faculty of Medicine, however, took longer to successfully complete their studies. Black Africans took less time to complete their master's degrees when compared with otherwise identical students from the other race groups.

Introduction

Over the past few years, the traditional concept of what constitutes a 'knowledge base' for a given country has evolved to include knowledge production and the development of innovation at the highest levels so that the country can stay ahead in a globally competitive world. Because PhD graduates have the necessary skills to make greater contributions to the knowledge base of a country, universities, which are the primary producers of this high-end knowledge, have become tasked with increasing this pool of graduates.[1] In particular, the South African Department of Science and Technology has set in place a Ten-Year Plan that seeks to promote innovation in research[2] by funding South African universities (through the Department of Higher Education and Training) according to a formula based on the drivers that promote research and teaching outputs.[3] The funding formula is heavily weighted towards rewarding institutions that graduate PhD candidates, with the result that the registration and throughput of PhD students have now become a very important area of focus for all higher education institutions in this country. Likewise, master's programmes are geared to promote contribution to research. The higher education institution funding formula is weighted in favour of research-based master's programme throughput rather than structured master's programmes based on structured course modules taught through lectures, seminars, laboratory work, or distance learning. A research-based master's programme requires the student to undertake his or her own research project in a specialised field of study.

A study conducted by the Academy of Science of South Africa[4] on all the universities in the country found that South Africa is lagging behind other countries with regard to the production of PhD students. South African universities only produce 26 doctoral graduates per million citizens which is far below other countries (Brazil: 52 per million, Korea: 187 per million, Australia 264 per million, and Sweden: 427 per million).[5] South Africa's inability to produce enough doctoral graduates who can at best, help to build the 'knowledge base' of our economy or at worst, simply replace the existing cohort of academics in our higher education system, has therefore become a huge challenge that needs to be addressed by all the universities in South Africa.[6]

A study conducted by the Department of Education in 2000 based on a cohort of first-time undergraduate students, found that only 30% of these students had graduated 5 years after entering the higher education institution, 56% had dropped out of university, and the remaining 14% were still pursuing their studies 5 years after having enrolled.[7] Focusing on first-time enrolments for a master's degree, a Council on Higher Education study[8] in 2008 found that nationally, enrolment increased at an average annual rate of 4.4% between 2000 and 2005. This increase, however, was tempered by a steady decline in completion rates for the degree (from 67% in 2001 to 52% in 2005).[8] Furthermore, focusing only on students who had successfully completed their master's degrees, the study found that they were taking on average 3 years to complete their degree and were graduating at a relatively late age (34 years). This means that many master's students typically interrupt their studies after completing their bachelor's and honours degrees to enter the job market, only to take up their master's studies later on. The interruption in studies is probably due to a lack of financial resources impacting on their preparedness for advanced study, and could be a contributor to these students taking longer to graduate.

These conclusions are drawn from a study prior to 2005[8] at national level; however, nothing is known about University of KwaZulu-Natal (UKZN) students in particular. One of the aims of this paper is therefore to focus only on students enrolled for a master's programme at UKZN and to establish whether they exhibit a graduation pattern that is different from that of other universities in the country. Although a detailed study has been conducted on the attrition rate of undergraduate students,[9] very little is known about the throughput rate of postgraduate students. It is therefore important to initiate research that will identify factors that will help to improve the throughput rate among UKZN master's degree students.

Data

This study focused on a cohort of master's students who registered at UKZN between the years 2004 and 2011. It followed each student's progress until completion of their degrees or until they dropped out from their studies.

A single record for each full-time registered master's student was created along with a set of variables indicating: their year of entry into the programme; whether or not they received some form of financial aid; the faculty in which they were registered; their race, gender and age (when first registered); and whether or not they were international students. Two response variables were also created: the first recording the total number of years for which the student was registered and the second whether the student had graduated, dropped out, or was still studying at the end of our study period.

Race was categorised as: African, white or Asian. Faculty was categorised as: Education, Engineering, Health Sciences, Law, Humanities, Development and Social Sciences, Science and Agriculture, Medicine, or Management Studies. Admission into the master's programme was based on an excellent undergraduate academic record, a letter of recommendation, and/or some form of previous professional experience. Typically, one would want to include the undergraduate record of each student as a predictor for success in the programme but most of the students who enrolled for a master's programme at UKZN came from universities and countries that had their own grading systems. For the purposes of this study, we were not able to gain access to these records. Instead, we have used the variable labelled 'international' to adjust for this; recording the variable as 'yes' if the student had no South African residency and 'no' if otherwise. Financial aid was given a yes/no response.

Focusing on the number of students who enrolled for a master's programme at UKZN, Figure 1 indicates that this number steadily increased over the period of our study. An exponential fit, $y_t = 221.6 * exp(0.0782 \times t)$ provided the best fit ($R^2 = 0.9981$), indicating that the year-on-year admission rate of master's students at UKZN between 2004 and 2007 increased on average by 8.13% each year (this follows because the multiplicative effect in the quoted exponential fit of a single unit increase in the academic year takes on the value $exp(0.0782) = 1.08134$). Figure 1 needs to be contrasted with the national average of 4.4% growth experienced by all other universities in South Africa.[8]

Descriptive statistics relating to the age of master's students at the time of first admission into a particular faculty at UKZN are given in Table 1. It is clear that the average age for admission of students into the Faculties of Science and Agriculture and Engineering is lower than in other faculties. A possible reason for this anomaly may be that students in other faculties are being forced to complete some form of mandatory internship programme before enrolling for a master's degree. Across all faculties, the average age on admission into a master's programme was found to be 27.3 years for the period of our study.

Figure 2 shows student status at the end of the registration period. The modal number of study years for successful completion of a master's degree is two. The first year is the modal dropout year. After one year of registration, 11.6% were able to graduate, 35.4% dropped out and 53% registered for the second year to continue their studies. After 2 years of registration, 40.6% of these students successfully completed their master's studies, 21.5% dropped out and 37.9% continued their studies into a third year. Figure 2 shows that the majority of students who completed their master's degrees did so in year two. In year three and four, the probability of degree attainment was much higher than dropping out or extending the study for another year. After 4 years, only a few students remained in the

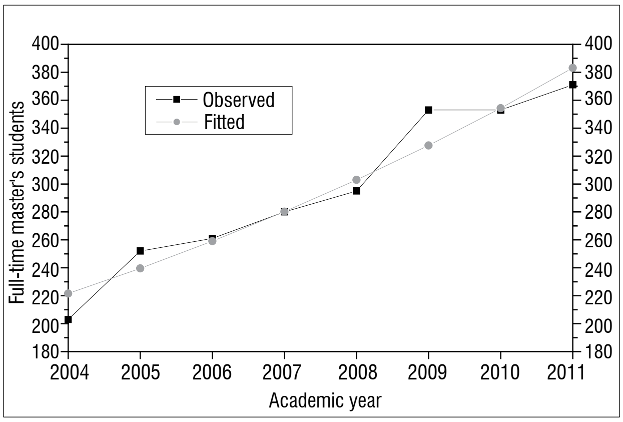


Figure 1: Admission pattern of full-time master's students enrolled at University of KwaZulu-Natal.

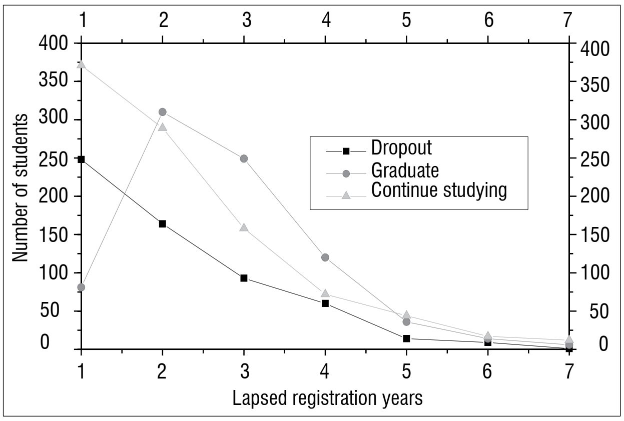


Figure 2: Number of master's students with lapsed registration enrolled at University of KwaZulu-Natal between 2004 and 2011.

Table 1: Age of master's students on admission to the University of KwaZulu-Natal according to faculty descriptive statistics

Faculty	Statistics					
	n	Mean	SD	First quartile	Median	Third quartile
Education	68	34.9	8.71	26.5	34.5	43
Engineering	268	24.6	3.81	23	23	25
Health Sciences	125	27.3	6.84	23	25	28
Humanities and Development Studies	568	29.5	8.74	23	25	33
Law	16	28.9	7.80	23	25.5	33.5
Management Studies	219	30.3	7.86	24	28	35
NRM School of Medicine	107	26.9	6.28	23	24	27
Science and Agriculture	997	25.5	5.03	23	24	26
Total	2368	27.3	6.99	23	24	28

master's programme. After year four, the dropout rate remained small, with an equal likelihood of graduating or continuing with the study.

Methodology

In this study, it was important to make a clear distinction between the actual occurrence of a particular event and the time it took for the event to occur. Focusing on two possible sources of exit, namely graduation and dropout, we started by attempting to link the probability of exit (from either cause) to some of the other variables recorded in the data set (e.g. gender, age, race, financial aid, international student). We then made use of survival analysis techniques[10] to help identify how the above-mentioned factors affected the length of time (in years) that it took for graduation or dropout to occur.

While some students, called censored individuals, were still pursuing their studies when the data collection period expired, survival analysis provided a methodology for dealing with the problem, provided that the censoring took place in a non-informative manner.[10] Given our context, one would expect students to carefully weigh up the benefits of completing a master's qualification with the costs and possible benefits of entering the job market at an earlier stage. When viewed from this perspective, the decision to complete or drop out of university becomes a competing risk problem.

Scott[11] developed a model that makes use of this competing risks methodology. We included this model in our modelling framework, where we were interested in determining how some of the demographic factors (associated with each student) affected the length of time (in years) that it took for graduation to occur (where dropout is being treated as a competing risk) and the length of time (in years) it took for dropout to occur (where graduation is treated as a competing risk).

Focusing on the actual length of time it took for a student to graduate or drop out from a master's programme at UKZN, the hazard rates associated with dropout and degree completion for each year after start of registration for master's study were computed. As the data cover the period 2004 until 2011, the last year for which these hazard rates could be modelled was $t=7$ (7 years after the start of master's study).

At this point, it is useful to outline two essential concepts in survival analysis. The risk set in a particular year t ($t = 1, 2,\ldots,7$) represents the group of students who we know can experience one of the above-mentioned outcomes during year t. The risk set does not include individuals who have already experienced one of the above events. That is, the hazard rate associated with outcome k ($k=1,2$) in year t ($t=1,\ldots,7$) [$h(k,t)$] represents the probability that a randomly selected individual experiences outcome k (dropout or degree completion) in year t, given that he/she has experienced no such outcome before year t. While the hazard rate functions can be used to determine the probability of experiencing a particular outcome in a given year, they can also be used to compute the probability associated with having experienced a particular event *by* a certain year *t.*

Results

Figure 3 displays a set of cause specific cumulative hazard and survival functions for graduation and dropout in our data set. Before year three, the cumulative hazard associated with dropout is higher than the cumulative hazard associated with master's completion. Notably, the median time associated with a successful master's completion is slightly more than 2 years with the median dropout time being slightly longer than 2 years. It is interesting to note that within 2 years of registration, at least 50% of the master's students in our study period had experienced one or another of the above events.

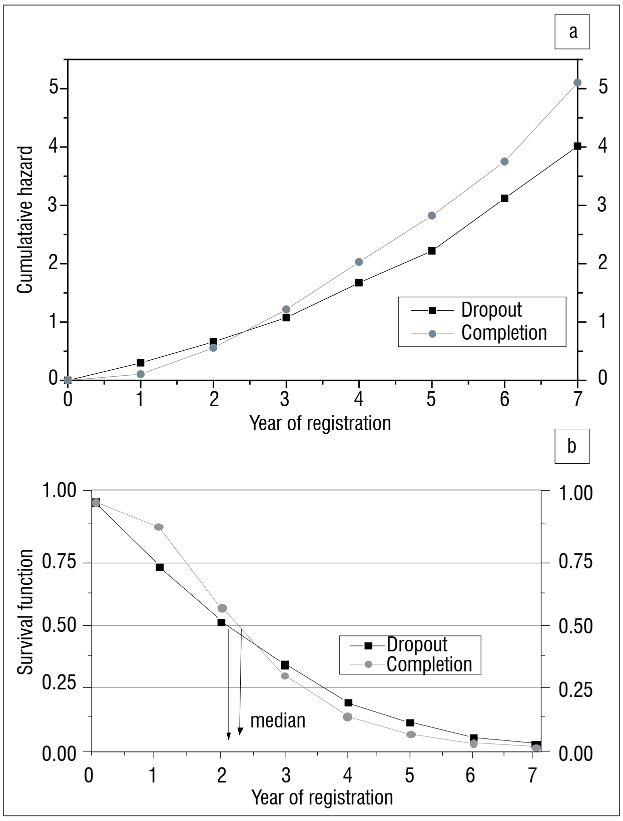


Figure 3: (a) Cumulative hazard and (b) survival functions for master's students who graduated or dropped out of study at University of KwaZulu-Natal.

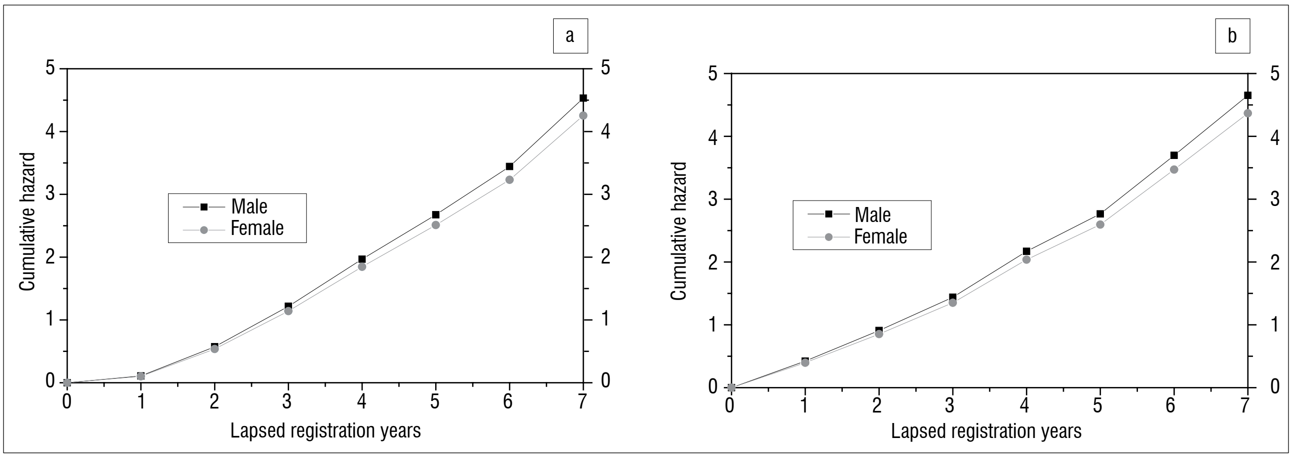


Figure 4: Cumulative hazard functions by gender for master's students who (a) graduated or (b) dropped out of study at the University of KwaZulu-Natal.

In order to gain a better understanding of the effect of certain explanatory variables on the time to completion or dropout after registration for a master's degree, the cumulative hazard functions for dropping out or successfully completing the degree are presented separately for men and women in Figure 4. The hazard rates for dropout or eventual graduation show that gender has no influence on dropout or on completion.

The cumulative hazard functions for race are presented in Figure 5. No significant difference appears to exist between the race groups.

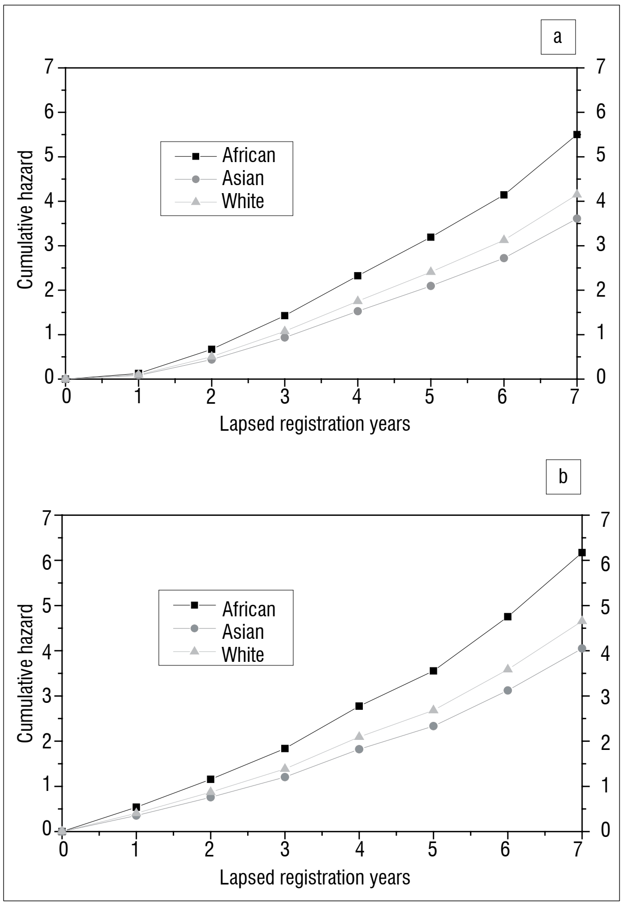


Figure 5: Cumulative hazard functions by race for master's students who (a) graduated or (b) dropped out of study at the University of KwaZulu-Natal.

The hazard functions for students who received some form of financial aid are given in Figure 6. Table 2 shows that financial aid appears to have a significant effect on dropout and graduation, particularly from year three onwards.

The exploratory hazard functions presented in Figures 3–7 control for a single variable. An extension of the Cox[12] proportional hazards regression model to a multinomial logistic regression model allows for modelling the effect of several explanatory variables in a competing risks model based setting.[3,11,13] Under this competing risks model, a cause *k* specific hazard rate, $h_t(k, i)$, for subject *i* that takes into account the competing risks assumption ($k=1$ for dropout, $k=2$ for master's completion) at time *t* ($t=1, 2,...,7$) can be given by

$$h_t(k,i) = \frac{\exp(X_{1ti}\beta_{k1} + X_{2ti}\beta_{k2} + ... + X_{pti}\beta_{kp})}{1 + \sum^{2}_{k=1} \exp(X_{1ti}\beta_{k1} + X_{2ti}\beta_{k2} + ... + X_{pti}\beta_{kp})}$$

where X_1, X_2,...,X_p denote a set of p explanatory variables that help to form this hazard rate at time *t*.

The coefficients, $\beta_{k1},\beta_{k2},...,\beta_{kp}$ for $k=1,2,...$ can be estimated using a maximum likelihood methodology. The results obtained can be interpreted in much the same way as the results for a Cox proportional hazards model. More specifically, a positive valued estimate for β_{kj} would suggest that an increase in the value of the variable *j* increases the hazard rate associated with the occurrence of the cause specific outcome *k* (i.e. shortens the length of time until the event occurs). Similarly, a negative valued estimate for β_{kj} would suggest that an increase in the value of variable *j* decreases the hazard rate associated with the occurrence of the cause specific outcome *k* (i.e. lengthens the time until occurrence of the event). The fitted model results along with their *p*-values are presented in Table 2. For all analyses, SAS software was used; $p \leq 0.05$ was considered statistically significant.

The hazard ratios given in Table 2 suggest that race, gender, nationality and age have no significant effect on the hazard rate associated with students dropping out from the master's programme. Financial aid, however, exerts a significant effect on the hazard rate associated with dropout. If we exponentiate the estimated coefficient of 0.140 obtained for the financial aid variable, we obtain a percentage change in the hazard rate associated with eventually dropping out that can be associated with a single unit increase in the covariate variable that we have called financial aid. Thus, a student with financial aid has a 15.0% higher hazard rate associated with dropping out from their studies when compared with a student who has not received financial aid. Using students from the Faculty of Science and Agriculture as a baseline category, the statistically significant positive valued estimates obtained for students in the Faculties of Management and/or Humanities suggest that they drop out more quickly from their

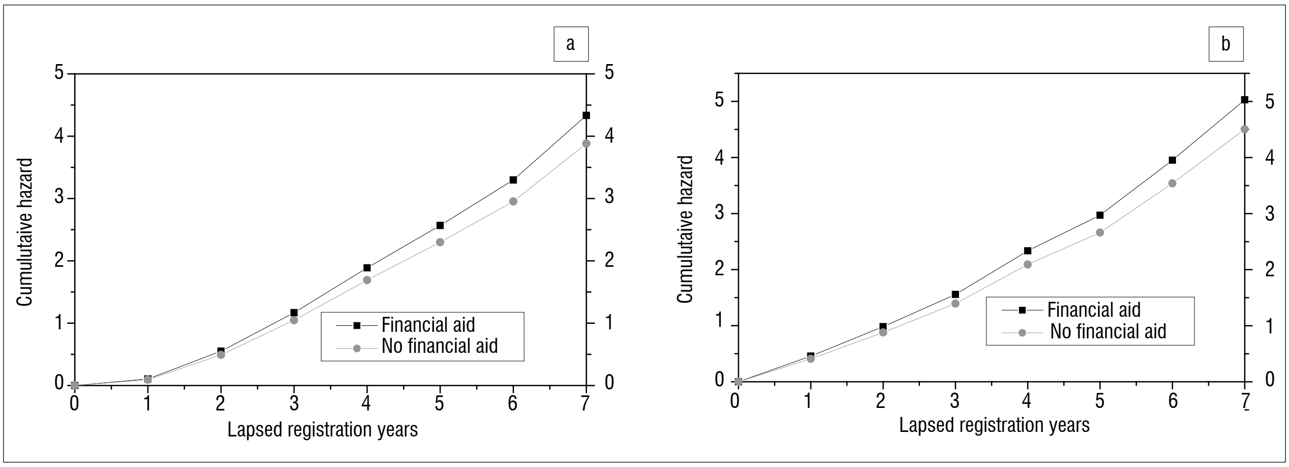


Figure 6: Cumulative hazard functions by financial aid for master's students who (a) graduated or (b) dropped out of study at the University of KwaZulu-Natal.

studies compared with otherwise identical students (in terms of the factors selected) in the Faculty of Science and Agriculture.

Focusing on those students who will eventually successfully complete their master's programme, the results in Table 2 suggest that gender has no effect on the successful completion of a master's degree. Chosen field of study does, however, seem to exert an influence on completion amongst students in the Faculties of Health, Humanities, Law and Management, all taking a significantly shorter period of time to complete their studies compared with otherwise identical students in the Faculty of Science and Agriculture. However, students in the Faculty of Medicine appear to take a significantly longer period to successfully complete their studies compared with otherwise identical students from the Faculty of Science and Agriculture.

The positive valued estimates obtained in Table 2 for international students and students with financial aid suggest that these students are successfully completing their studies more quickly than otherwise identical (in terms of gender, age, faculty and race) students who have no financial aid or who are of local origin. Using the white race group as a baseline category, the African race group shows a significantly higher completion hazard, implying that they are taking a shorter period of time to successfully complete their master's degree compared with otherwise identical white race group counterparts.

The negative valued estimate obtained for the age based covariate indicates that older students are taking longer to successfully complete their master's degrees. Figure 7 shows the cumulative hazard functions plotted for the effect of age on dropout and successful completion.

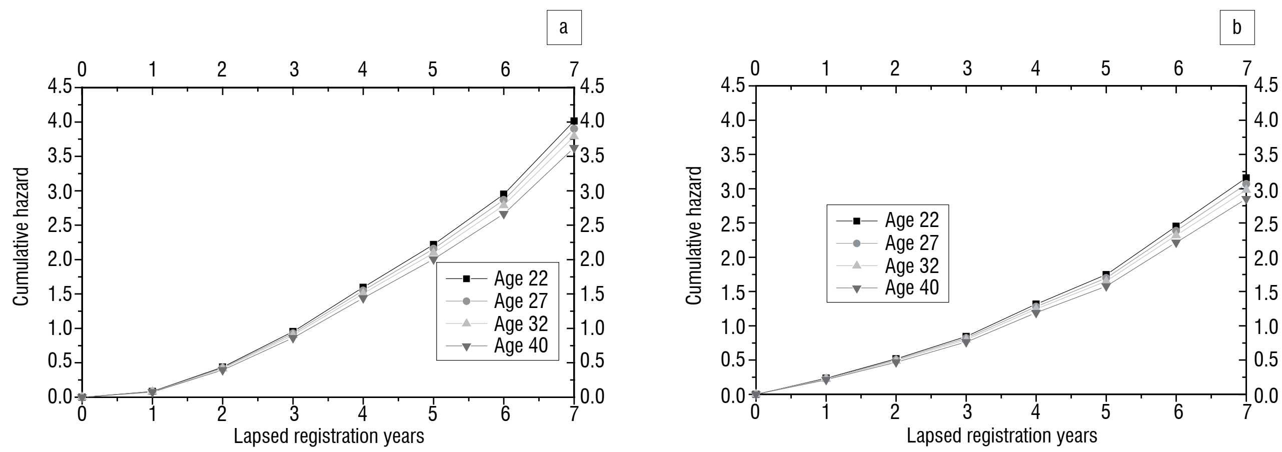


Figure 7: Age-specific cumulative hazard functions at the reference values for master's students who (a) graduated or (b) dropped out of study at the University of KwaZulu-Natal.

Table 2: Results of the competing risk survival model fit for master's students who received some form of financial aid

Risk factor	Dropout			Master's completion		
	Estimate	*p*-value	Hazard ratio	Estimate	*p*-value	Hazard ratio
Race (Reference=White)						
African	0.083	0.265	1.087	0.144	0.025*	1.155
Asian	-0.027	0.732	0.973	-0.040	0.548	0.961
Gender (Reference=Female)						
Male	0.083	0.181	1.087	-0.064	0.243	0.938
International (Reference=No)						
Yes	0.036	0.669	1.037	0.143	0.033*	1.154
Financial aid (Reference=No)						
Yes	0.140	0.036*	1.150	0.175	0.001*	1.191
Faculty (Reference=Science and Agriculture)						
Education	-0.141	0.464	0.868	0.169	0.321	1.184
Engineering	0.222	0.054	1.249	-0.093	0.219	0.911
Health	0.188	0.146	1.207	0.688	0.000*	1.990
Humanities	0.407	0.000*	1.502	0.170	0.009*	1.185
Law	0.393	0.050	1.482	1.174	0.001*	3.235
Management	0.477	0.000*	1.612	0.364	0.029*	1.439
Medicine	-0.244	0.174	0.783	-0.574	0.000*	0.563
Age (Continuous variable)	**0.001**	**0.723**	**1.00**	**-0.57**	**0.000***	**0.563**

**Significant at 5% level of significance*

Interestingly enough, the results in Table 2 indicate that age does not appear to have a significant effect on the time to dropout of students in the master's programme.

Summary and conclusion

The study shows that gender does not have a significant effect on the throughput rate of students when one considers the time that it takes to successfully complete or drop out from a master's programme. This may indicate equal competiveness in both genders for success, bearing evidence of the success of our policies which promote equal academic opportunities for both genders.

Age on admission has a negative effect on the time it takes to graduate. Older students seem to take longer to successfully complete a master's degree compared with otherwise identical students who are younger. One of the reasons for this variation might be that older learners have more family or personal commitments.

Receiving some form of financial aid appears to reduce the length of time that it takes for a student to drop out from a master's programme. Likewise, receiving financial aid reduces the length of time that it takes to successfully complete a master's programme. This might be due to the contractual conditions set by funders in the form of repayment or limited period of support. The other possible explanations for the paradox that funding helps a student to graduate more quickly and also to drop out more quickly, could be that students who successfully complete the master's programme use this source of funding to purchase extra books and other academic support materials, and attend conferences which helps them to graduate more quickly. Students who eventually drop out 'because of this funding' may indicate that the funding is no longer sufficient for their needs. Similarly, according to Lewin[14], more than a third of American college dropouts reported that even if they were given grants to pay for their books and tuition, it would be hard to go back to school given their work and family commitments.

Race had no effect on those who eventually dropped out from their studies. For students who eventually graduated, there was, however, evidence that the time it took to graduate was shorter for African students.

International students (i.e. students from beyond South African borders) all seemed to perform very favourably considering their throughput rates. In particular, amongst the students who eventually graduated, being of international origin seemed to shorten the length of time it took to complete their degree.

Different faculties seemed to have varying levels of success with regard to dropout and throughput rates. Focusing on the students who eventually dropped out, the ones in the Faculty of Medicine appeared to take longer to drop out or graduate from their studies compared with students from other faculties.

Considering the limited resources available for research funding, it is important to understand the influences on dropout or completion rates in the master's programme. We identified some of the demographic factors that may be affecting the throughput rates at UKZN. The list of factors selected were by no means exhaustive as the study was limited to data available through the UKZN data management and information archives. As such, this study should be viewed as a starting point for further reflection on what drives throughput rates for master's study at UKZN.

Acknowledgement

We are grateful for funding support received through a teaching-learning competitive grant from the University of KwaZulu-Natal.

Authors' contributions

T.Z. was responsible for the data analysis, design of the study and writing the manuscript. D.N. was responsible for the data organisation and for critically revising the manuscript. M.M. made editorial and conceptual contributions.

References

1. Barnacle R, Usher R. Assessing the quality of research training: The case of part-time candidates in full-time professional work. High Educ Res Dev. 2003;22(3):345–358. http://dx.doi.org/10.1080/0729436032000145185

2. Kahn MJ, Vlotman N, Steyn C, Van der Schyff M. Innovation policy and higher education in South Africa: Addressing the challenge. S Afr Rev Soc. 2007;38(2):176–190. http://dx.doi.org/10.1080/21528586.2007.10419174

3. Essack SY, Barnes G, Jackson L, Majozi M, McInerney P, Mtshali N, et al. Maximizing income via the higher education funding framework in health sciences. S Afr J High Educ. 2009;23(2):275–292.

4. Academy of Science of South Africa (ASSAf). The PhD study: An evidence-based study on how to meet the demands for high-level skills in an emerging economy. Pretoria: ASSAf; 2010. Available from: http://www.assaf.org.za/wp-content/uploads/2010/11/40696-Boldesign-PHD-small.pdf

5. Samuel M, Vithal R. Emergent framework of research teaching and learning in a cohort-based doctoral programme. Perspect Educ. 2011;29(3):76–87.

6. Dell S. South Africa: Decline in PhD numbers problem. University World News. 2010 August 22; Issue 60. Available from: http://www.universityworldnews.com/article.php?story=20100820150736361.

7. Scott I, Yeld N, Hendry J. Higher Education Monitor no. 6: A case for improving teaching and learning in South African higher education. Pretoria: Council on Higher Education; 2007. Available from: http://www.che.ac.za/sites/default/files/publications/HE_Monitor_6_ITLS_Oct2007_0.pdf

8. Council on Higher Education (CHE). Higher Education Monitor no. 7: Postgraduate studies in South Africa: A statistical profile. Pretoria: Council on Higher Education; 2009. Available from: http://www.che.ac.za/sites/default/files/publications/CHE_MonitorProjectV7.pdf

9. Zewotir T, North D, Murray M. Student success in entry level modules at the University of KwaZulu-Natal. S Afr J High Educ. 2011;25(6):1233–1244.

10. Klein PJ, Moeschberger ML. Survival analysis: Techniques for censored and truncated data. New York: Springer-Verlag; 1997. http://dx.doi.org/10.1007/978-1-4757-2728-9

11. Scott MA, Kennedy BB. Pitfalls in pathways: Some perspectives on competing risks event history analysis in education research. J Educ Behav Stat. 2005;30(4):413–442. http://dx.doi.org/10.3102/10769986030004413

12. Cox DR. Regression models and life tables. J Roy Statist Ser B. 1972;34(2):187–220.

13. Therneau TG. Modeling survival data: Extending the Cox model (statistics for biology and health). New York: Springer-Verlag; 2000. http://dx.doi.org/10.1007/978-1-4757-3294-8

14. Lewin T. College dropouts cite low money and high stress. The New York Times. 2009 Dec 9; page A27. Available from: http://www.nytimes.com/2009/12/10/education/10graduate.html?_r=0

Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe

AUTHORS:
Delson Chikobvu[1]
Retius Chifurira[2]

AFFILIATIONS:
[1]Department of Mathematical Statistics and Actuarial Sciences, University of the Free State, Bloemfontein, South Africa
[2]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Delson Chikobvu

EMAIL:
chikobvu@ufs.ac.za

POSTAL ADDRESS:
Department of Mathematical Statistics and Actuarial Sciences, University of the Free State, PO Box 339, Bloemfontein 9300, South Africa

DATES:
Received: 06 Aug. 2014
Revised: 04 Nov. 2014
Accepted: 01 Feb. 2015

KEYWORDS:
minima; return level; mean annual rainfall; Bayesian approach; severe meteorological drought

HOW TO CITE:
Chikobvu D, Chifurira R. Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. S Afr J Sci. 2015;111(9/10), Art. #2014-0271, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140271

We modelled the mean annual rainfall for data recorded in Zimbabwe from 1901 to 2009. Extreme value theory was used to estimate the probabilities of meteorological droughts. Droughts can be viewed as extreme events which go beyond and/or below normal rainfall occurrences, such as exceptionally low mean annual rainfall. The duality between the distribution of the minima and maxima was exploited and used to fit the generalised extreme value distribution (GEVD) to the data and hence find probabilities of extreme low levels of mean annual rainfall. The augmented Dickey Fuller test confirmed that rainfall data were stationary, while the normal quantile-quantile plot indicated that rainfall data deviated from the normality assumption at both ends of the tails of the distribution. The maximum likelihood estimation method and the Bayesian approach were used to find the parameters of the GEVD. The Kolmogorov–Smirnov and Anderson–Darling goodness-of-fit tests showed that the Weibull class of distributions was a good fit to the minima mean annual rainfall using the maximum likelihood estimation method. The mean return period estimate of a meteorological drought using the threshold value of mean annual rainfall of 473 mm was 8 years. This implies that if in the year there is a meteorological drought then another drought of the same intensity or greater is expected after 8 years. It is expected that the use of Bayesian inference may better quantify the level of uncertainty associated with the GEVD parameter estimates than with the maximum likelihood estimation method. The Markov chain Monte Carlo algorithm for the GEVD was applied to construct the model parameter estimates using the Bayesian approach. These findings are significant because results based on non-informative priors (Bayesian method) and the maximum likelihood method approach are expected to be similar.

Introduction

Relatively extreme low rainfall attributed to global warming, although rare, is a natural phenomenon that affects people's socio-economic activities worldwide. Extreme droughts occur from time to time in Zimbabwe, and impact negatively on the country's economic performance. The drought of rainfall season year 1991/1992 was one of the worst in the recorded history of Zimbabwe. Its impact was felt even in the insurance industry which received high claims for crop failure.[1] Droughts can be viewed as extreme events outside of the normal rainfall occurrences, such as exceptionally lower amounts of mean annual rainfall.[2] In Zimbabwe, at least 50% of the gross domestic product is derived from rain-fed agriculture.[3] With more low technology indigenous farmers entering commercial agriculture through the accelerated land-reform programme, modelling and prediction of extreme low annual rainfall and the associated probabilities of drought become more relevant.

Developing methods that can give a suitable prediction of meteorological events is always interesting for both meteorologists and statisticians. The use of standard statistical techniques in modelling, forecasting and prediction of extremes in average rainfall and rare events is less prudent because of gross under-estimation.[4] Extreme value theory is an alternative and superior approach to quantify the stochastic behaviour of a process at unusually large or small levels.[4] Extreme value theory provides the statistical framework to make inferences about the probability of very rare and extreme events. It is based on the analysis of the maximum (or minimum) value in a selected time period.

Recently there has been growing interest in modelling extreme events, especially in situations in which scientists underestimated the probabilities of extreme events that subsequently occurred and caused catastrophic damage.[5] Work has been done which provides evidence of the importance of modelling rainfall from different regions of the world: Nadarajah and Choi[6] used extreme value theory for rainfall data from South Korea; Koutsoyiannis[7] applied extreme value theory to rainfall data from Europe and the USA; Koutsoyiannis and Baloutsos[8] applied extreme value theory to Greece's rainfall data; and Crisci et al.[9] applied extreme value distributions to rainfall data from Italy. The use of extreme value distributions is not restricted to meteorology events; examples appear in energy[10]; insurance[11]; fish management[12] and ecology[13]. There is no work known to us on rainfall extremes in Zimbabwe. In this paper, we provide the first application of extreme value distributions to model minimum annual rainfall in Zimbabwe.

Rainfall in Zimbabwe is associated with the behaviour of the inter-tropical convergence zones whose oscillations are influenced by changing pressure patterns to the north and south of the country.[10] Zimbabwe lies in the Southwest Indian Ocean zone, which is often affected by tropical cyclones. Tropical cyclones are low pressure systems that have well-defined clockwise (in the southern hemisphere) wind circulations which spiral toward the centre where the winds are strongest and rains are heaviest. Cyclones that develop over the western side of the Indian Ocean occasionally affect the rainy season. The amount and intensity of rainfall during a given wet spell is enhanced by the passage of upper westerly wind waves of mid-latitude origin.[14,15]

Studies of extreme low rainfall are beneficial to decision-makers in government, non-governmental organisations involved in early warning systems and food security, poverty alleviation and disaster management and risk management. This study will also inform climatologists about the behaviour of extreme low rainfall. Appropriate decisions and plans can be made based on the results of this study to prepare the general public for changes brought on by extremely low rainfall. The objective of this study was to quantify and describe the behaviour of extremely low rainfall in Zimbabwe. In particular, the aim was to model the extreme low rainfall using the generalised extreme value distribution (GEVD) by using the maximum likelihood estimation method and the Bayesian statistics approach. The mean return period – that is, the number of years on average before another drought of equal or greater intensity – was also calculated.

Research methodology

Normal distribution

A normal distribution is symmetrical and has a bell-shaped density curve with a single peak. The normal density function, which gives the height of the density at any value *x* is given by:

$$g_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{ -\frac{1}{2\sigma^2}\,(x_p - \mu)^2 \right\} \text{ with } \sigma > 0 \qquad \text{Equation 1}$$

where μ is the mean (where the peak of the density occurs) and σ is the standard deviation (which indicates the spread or girth of the bell curve).

Generalised extreme value distribution

In climatology, meteorology and hydrology, maxima of temperatures, precipitation and river discharges have been recorded for many decades.[16] The extreme value theorem provides a theoretical framework to model the distribution of extreme events and the three-parameter GEVD was recommended for meteorology frequency analysis.[17] The three parameters are: location, scale and shape. The GEVD is a family of continuous probability distributions developed within the extreme value theorem. The GEVD unites the Gumbel, Fréchet and Weibull family of distributions into a single family to allow a continuous range of possible shapes. Based on the extreme value theorem, the GEVD is the limiting distribution of properly normalised maxima of a sequence of independent and identically distributed random variables.[18] Thus, the GEVD is used to model the maxima of a long (finite) sequence of random variables. The unified GEVD for modelling maxima is given by:

$$G_{\xi,\mu,\sigma}(x) = exp\left\{ -\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} \right\} \text{ with } \xi \neq 0 \qquad \text{Equation 2}$$

with $\mu \in R$, $\sigma > 0$ and $1 + \xi\,(\frac{x-\mu}{\sigma}) > 0$,

where μ, σ and ξ are the location, scale and shape parameters, respectively. The probability density function is sometimes called the Fisher–Tippett distribution and is obtained as the derivative of the distribution function:

$$g_{\xi,\mu,\sigma}(x) = \frac{1}{\sigma}\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1\frac{1}{\xi}} exp\left\{ -\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} \right\},\ \xi \neq 0$$

Equation 3

The shape parameter ξ is also known as the extreme value index. The parameter ξ^{-1} is the rate of tail decay of the GEVD. If $\xi > 0$, *G* belongs to the heavy-tailed Fréchet class of distributions such as Pareto, Cauchy, student-*t* and mixture distributions. If $\xi < 0$, *G* belongs to the short-tailed with finite lower bounds Weibull class of distributions which includes distributions such as uniform and beta distributions. If $\xi = 0$ then *G* belongs to the light-tailed Gumbell class of distributions which includes distributions such as normal, exponential, gamma and log-normal distributions.[19]

Modelling minima random variables

The classical GEVD for extremes is based on asymptotic approximations to the sampling behaviour of block maxima. The block maxima size (hourly, daily, weekly, monthly or yearly) varies according to instrument constraints, seasonality and the application at hand. The only possible limiting form of a normalised maximum of a random sample (when a non-degenerate limit exists) is captured by the GEVD. The data set is partitioned into blocks of equal length and distribution and GEVD is fitted to the set of block maxima. In this study, minima rainfall was modelled using GEVD. In order to model minima random variables we use the duality between the distributions for maxima and minima. If $M_N = \min\{X_1, X_2, ..., X_N\}$ where $X_1, X_2, ..., X_N$ is a sequence of independent random variables having a common distribution function and $Y_i = -X_i$ for $i = 1,..,N$ the change of sign means that small values of X_i correspond to large values of Y_i. So if $M_N = \min\{X_1, X_2, ..., X_N\}$ and $\widetilde{M}_N = \max\{Y_1, Y_2, ..., Y_N\}$, then $M_N = -\widetilde{M}_N$. The minima becomes:

$$\widetilde{M}_N = -\max\{-X_1, -X_2, ..., -X_T\} \qquad \text{Equation 4}$$

where X_i for $i = 1,2,3,...,T$ represents mean annual rainfall in period *i*. Extreme maxima theory and methods are then used to model extreme minima.[5,20] Based on the extreme value theorem that derives the GEVD, we can fit a sample of extremes to the GEVD to obtain the parameters that best explain the probability distribution of the extremes.

Parameter estimation

There is a wide variety of methods to estimate the GEV parameters in the independent and identically distributed settings.[21] The three parameters are estimated by method of moments, maximum likelihood method, method of textiles[22] and probability weighted moments or equivalent L-moments.[17] Hosking[23] showed that the probability weighted moments quantile estimators for the GEVD are better than the maximum likelihood method for small samples ($n < 50$). Madsen et al.[24] also showed that method of moments quantile estimators perform well when the sample size is modest. In this study, the maximum likelihood method was exploited because $n > 50$.

Maximum likelihood method

Under the assumption that $X_1, ..., X_m$ are independent random samples having a GEVD, the log-likelihood for the GEVD parameters when $\xi \neq 0$ is:

$$l(\mu,\sigma,\xi) = -mlog\sigma - (1+\frac{1}{\xi})\sum_{i=1}^{m} log\,[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)] - \sum_{i=1}^{m}[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)]^{-\frac{1}{\xi}}$$

Equation 5

provided that $1 + \xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ for $i = 1,...,m$.[5] We differentiated the log-likelihood of GEVD to find a set of equations which we solved using numerical optimisation algorithms (see Appendix 1 in the online supplementary material). For computational details, we refer to previous studies.[23,25-27] Because the support of *G* depends on the unknown parameter values, the usual regularity conditions underlying the asymptotic properties of maximum likelihood estimators are not satisfied. This problem is studied in depth by Stephens[28]. In the case $\xi > -0.5$, the usual properties of consistency, asymptotic efficiency and asymptotic normality hold.

Test for stationarity

The augmented Dickey Fuller (ADF) stationarity test is performed on the data to test for stationarity. The null hypothesis of the ADF test is that there is no trend while the alternative hypothesis is that there is a trend in the data.

Goodness of fit

To access the quality of convergence of the GEVD, the Kolmogorov–Smirnov (K-S) and Anderson–Darling goodness-of-fit tests were used. The K-S test, based on the empirical cumulative distribution function, is used to decide if a sample comes from the hypothesised continuous distribution. The K-S test is less sensitive at the tails than at the centre of the distribution. The Anderson–Darling test, which is an improvement of the K-S test, compares the fit of an observed cumulative distribution function to an expected cumulative distribution function; this test gives more weight to the tails of a distribution than does the K-S test.[29]

Return period or level estimates

We can estimate how often the extreme quantiles occur with a certain return level. The return level is defined as a level that is expected to be equalled or exceeded on average once every interval of time (*T*) with a probability of *p*. For the normal distribution we set:

$$G_{\mu,\sigma} = \int_{-\infty}^{xp} \frac{1}{\sqrt{2\pi\sigma^2}} exp \left\{ -\frac{1}{2\sigma^2} \ (x_p - \mu)^2 \right\} dx = 1 - p \qquad \text{Equation 6}$$

where *T* is the return period and x_p the return level. By setting the return period and solving the equation (see Appendix 1 in the online supplementary material), the return level, x_p, can be calculated:

$$x_p = \sigma\phi^{-1}(1\text{-}p) + \mu \qquad \text{Equation 7}$$

which can be re-written as:

$$x_p = \sigma Z_{1\text{-p}} + \mu \qquad \text{Equation 8}$$

Similarly, for the GEVD we set:

$$G_{\xi,\mu,\sigma}(\mathrm{x}) = exp \left\{ - \left(1 + \xi \left(\frac{x_p - \mu}{\sigma} \right) \right)^{-\frac{1}{\xi}} \right\} = 1 - p \qquad \text{Equation 9}$$

The return level (see Appendix 1 in the online supplementary material) is given by:

$$x_p = \mu + \tfrac{\sigma}{\xi} \left(\left(— \ln \left(1 - \tfrac{1}{\mathrm{T}} \right) \right)^{-\xi} - 1 \right), \xi \neq 0 \qquad \text{Equation 10}$$

Return levels are important for prediction purposes and can be estimated from stationary models. The mean return period is the number of years we expect to wait on average before we observe another drought of equal or greater intensity. If the exceedance probability of observing a drought of a given severity in any given year is *p* then the mean return period *T* is such that $T = \frac{1}{\mathrm{p}}$.

Bayesian analysis of extreme values for GEVD

Inference on the extremes of environmental processes is important to meteorologists, civil engineers, agriculturalists and statisticians. Naturally, data at extreme levels are scarce. Bayesian inference allows any additional information about the processes to be incorporated as prior information. The basic theory of Bayesian analysis of extreme values is well documented (see Coles[5], Coles and Tawn[30] and Gamerman[31] for more information). The Markov chain Monte Carlo techniques are applied in this paper to give Bayesian analyses of the annual minima rainfall data for Zimbabwe. [The annual maximum rainfall data are given in Appendix 2 of the online supplementary material.] Markov chain Monte Carlo techniques provide a way of simulating from complex distributions by simulating from Markov chains, which have the target distributions as their stationary distributions.[32] In this paper, the prior is constructed by assuming there is no information available about the process (rainfall) apart from the data. The annual rainfall data have a GEVD, i.e. $X_i \sim GEVD(\mu, \sigma, \xi)$ and the parameters μ, σ and ξ are treated as random variables for which we specify prior distributions. For specification of the prior, the parameterisation $\phi = \log \sigma$ is easier to work with because σ is constrained to be positive. The specification of priors enables us to supplement the information provided by the data. The prior density is:

$$\pi(\mu,\phi,\xi) = \pi_\mu(\mu)\pi_\phi(\phi)\pi_\xi(\xi), \qquad \text{Equation 11}$$

where each marginal prior is normally distributed with large variances. The variances are chosen to be large enough to make the distributions almost flat, corresponding to prior ignorance. The joint posterior density is the product of the prior and the likelihood and is given as:

$$\pi(\mu,\sigma,\xi\,|\,x) \propto \pi(\mu,\phi,\xi)L(\mu,\phi,\xi\,|\,x) \qquad \text{Equation 12}$$

where

$$L(\mu,\sigma,\xi\,|\,x) = \frac{1}{\sigma^{\mathrm{m}}} exp \left\{ - \left(1 + \xi \left(\frac{x_i - \mu}{\sigma} \right) \right)^{-\frac{1}{\xi}} \right\} \prod_{i=1}^{\mathrm{m}} \left[1 + \xi \left(\frac{x_i - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}}$$

is the likelihood with σ replaced by e^ϕ. The Gibbs sampler is used to simulate from each of the full conditionals. The posterior is of the form:

$$\pi(\mu,\sigma,\xi\,|\,x)L(\mu,\phi,\xi\,|\,x) \propto \pi_\mu(\mu)\pi_\phi(\phi)\pi_\xi(\xi)L(\mu,\phi,\xi\,|\,x)$$

so the full conditionals are of the form:

$$\pi(\mu\,|\,\phi,\xi) = \pi_\mu(\mu)L(\mu,\phi,\xi\,|\,x)$$

$$\pi(\phi\,|\,\mu,\xi) = \pi_\phi(\phi)L(\mu,\phi,\xi\,|\,x)$$

$$\pi(\xi\,|\,\mu,\phi) = \pi_\xi(\xi)L(\mu,\phi,\xi\,|\,x)$$

For details of the Markov chain Monte Carlo algorithm refer to R package (evdbayes version1.1-1).

Results

Description of data

The analysis was based on the historical mean annual rainfall data recorded from all 62 weather stations in Zimbabwe, dating from as far back as year 1901 to year 2009. A mean annual rainfall figure for the country was calculated. The mean data were obtained from the Zimbabwe Department of Meteorological Services. From Figure 1, it seems reasonable to assume that the pattern variation has stayed stationary over the observation period, and we can model the rainfall data as independent observations from the GEVD. In fitting a 109-year data set to a GEVD, a block size had to be chosen so that individual block minima had a common distribution; yearly blocks were therefore used in this study. Figure 1 shows the graph of x_i, $i = 1,\ldots, n$ the annual rainfall for Zimbabwe.

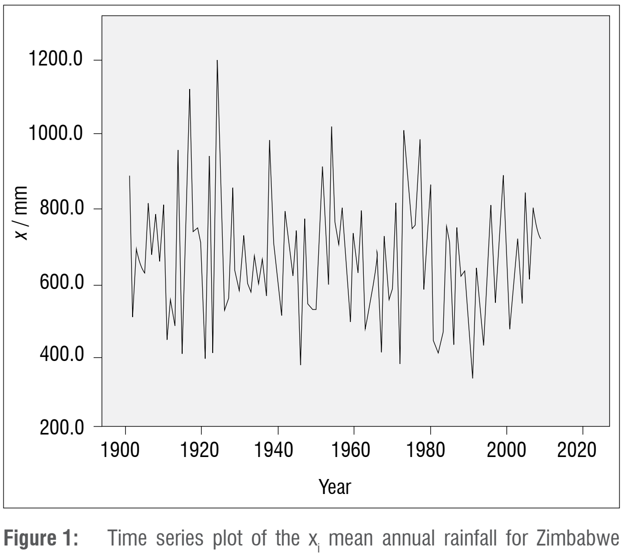


Figure 1: Time series plot of the x_i mean annual rainfall for Zimbabwe from 1901 to 2009.

The duality principle between the distribution of minima and maxima to fit the distribution of minimal rainfall for Zimbabwe was employed. Maximum likelihood estimates of parameters were estimated (see Appendix 1 in the online supplementary material). Figure 2 shows the graph of $-x_i$ annual rainfall for Zimbabwe.

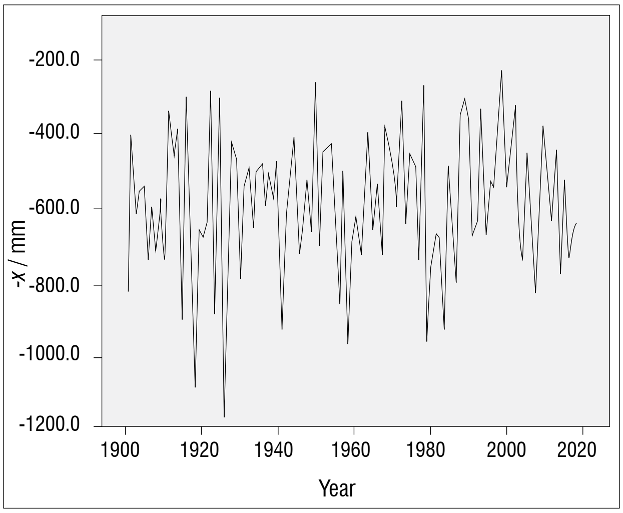


Figure 2: Time series plot of the $-x_i$ annual rainfall for Zimbabwe from 1901 to 2009.

Unit root test for stationarity

The augmented Dickey Fuller (ADF) test was used to check for stationarity of $-x_i$ annual rainfall for Zimbabwe. Table 1 shows the results of the ADF test.

The *p*-values of the ADF test statistics are significant; we therefore reject the null hypothesis of no stationary at 1%, 5% and 10% levels of significance and conclude that the rainfall data are stationary. The ADF test also indicates that the data do not follow any trend. Therefore, we can determine the return levels of minima annual rainfall.

Table 1: Unit root test to determine stationarity of minimum annual rainfall data for Zimbabwe for the period 1901–2009

Test	Test's critical values			Test statistic
	1%	5%	10%	
Augmented Dickey Fuller	-4.05	-3.45	-3.15	-4.26 (0.0000)

Descriptive statistics

Table 2 shows the descriptive statistics – specifically the coefficient of skewness and Jarque-Bera normality test – of the 109 years of annual rainfall data. The coefficient of skewness of minima annual rainfall ($-x_i$) is negative. This observation suggests that the rainfall data fit a distribution which is relatively long left tailed.

Table 2: Summary statistics of normality tests of annual rainfall data from 1901 to 2009

N	Minimum	Maximum	Mean	Standard deviation	Coefficient of skewness	Jarque–Bera statistic
109	-1192.60	-335.30	-659.93	169.25	-0.45	3.85 (0.15)

Fitting distributions to minimum mean annual rainfall

Normal distribution

Figure 3 shows the normal probability density function of minima mean annual rainfall data from 1901 to 2009.

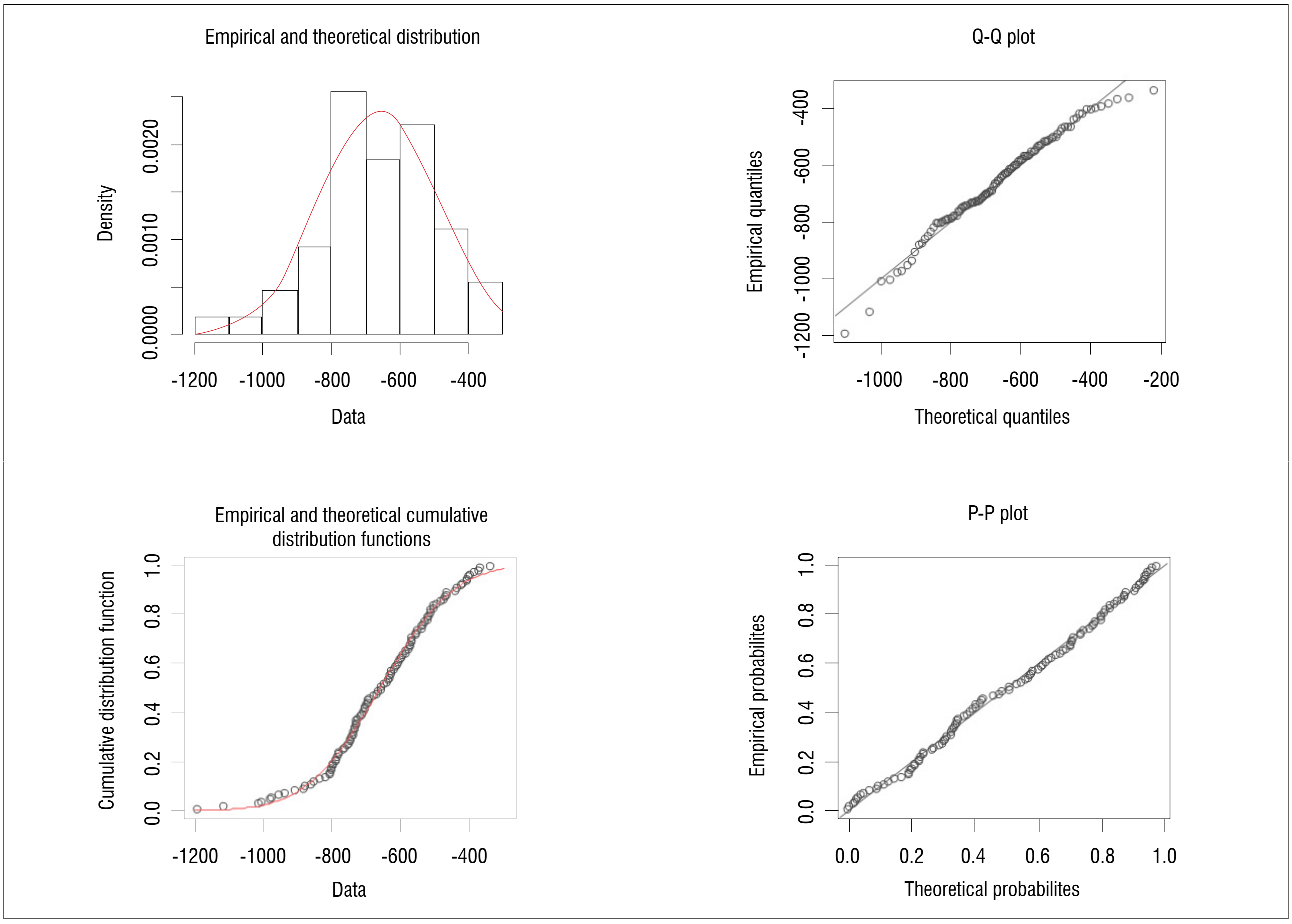


Figure 3: The normal probability density function of minimum annual rainfall for Zimbabwe for the period 1901–2009.

The parameter estimates and their corresponding standard errors in brackets are:

$\hat{\mu}$ = -659.9312 (16.13623)
$\hat{\sigma}$ = 168.4675 (11.41010)

Figure 3 shows the minima mean annual rainfall data normal quantile–quantile (Q-Q) plot of minima annual rainfall for Zimbabwe. The normal Q-Q plot of minima annual rainfall shows deviation from a normal distribution at both lower and upper tails of the data. However, based on the *p* value of the Jarque–Bera test, we fail to reject the null hypothesis of normality. The question is: If annual rainfall is normally distributed, then how do we account for extremely low rainfall (severe droughts) or extremely high rainfall (severe floods) events that have been recorded? The normal distribution approximates these events as negligible or close to zero. If the distribution of minima annual rainfall is heavy-tailed or skewed, the normal distribution may be misleading. Thus, the normal distribution is not a good fit for these rainfall data. The further one gets into the tails of the distribution, the rarer the event, but the event will be catastrophic if it happens. It is important to fit a distribution that is able to capture the probability of extreme minimum annual rainfall.

Generalised extreme value distribution

Figure 4 shows the diagnostic plots for the goodness of fit of the minima annual rainfall for Zimbabwe from 1901 to 2009.

Table 3 shows the maximum likelihood estimates of the GEVD model with their corresponding standard errors in brackets.

These results show that the data can be modelled using a Weibull class of distribution because $\hat{\xi} < 0$ (bounded tail). Combining estimates and standard errors, the 95% confidence intervals for ξ, σ and μ are [-0.5561; -0.3259], [154.8222; 203.7196] and [-744.4090; -670.9278], respectively. ξ is significantly different from zero because zero is not contained in the interval.

Table 3: Maximum likelihood estimates (standard errors) of the generalised extreme value distribution parameters

Shape $\hat{\xi}$	Scale $\hat{\sigma}$	Location $\hat{\mu}$
-0.44 (0.06)	179.27 (13.66)	-707.67 (18.75)

Model diagnostic

It is important to confirm that the data adequately fit the GEVD. Figure 3 shows the Q-Q plot and the P-P plot of the data. The quantiles of minima rainfall regressed against the quantiles of GEVD shows a straight line. This finding suggests that the data do not deviate from the assumption that they follow a GEVD. Table 4 shows the K-S and Anderson–Darling statistics.

The Anderson–Darling statistic is less than its 5% critical value and the K-S statistic test leads to a decision of non-rejection of the null hypothesis. We conclude that the minimum annual rainfall for Zimbabwe follows the specified GEVD.

The maximum likelihood estimate for ξ is negative, corresponding to a bounded distribution, in which the 95% confidence interval does not contain zero. Greater accuracy of the confidence interval is achieved by the use of the profile likelihood. Figure 5 shows the profile likelihood of the generalised extreme value parameter ξ, from which a 95% confidence interval for ξ is obtained as approximately [-0.55; -0.45], which is almost the same as the calculated 95% confidence interval.

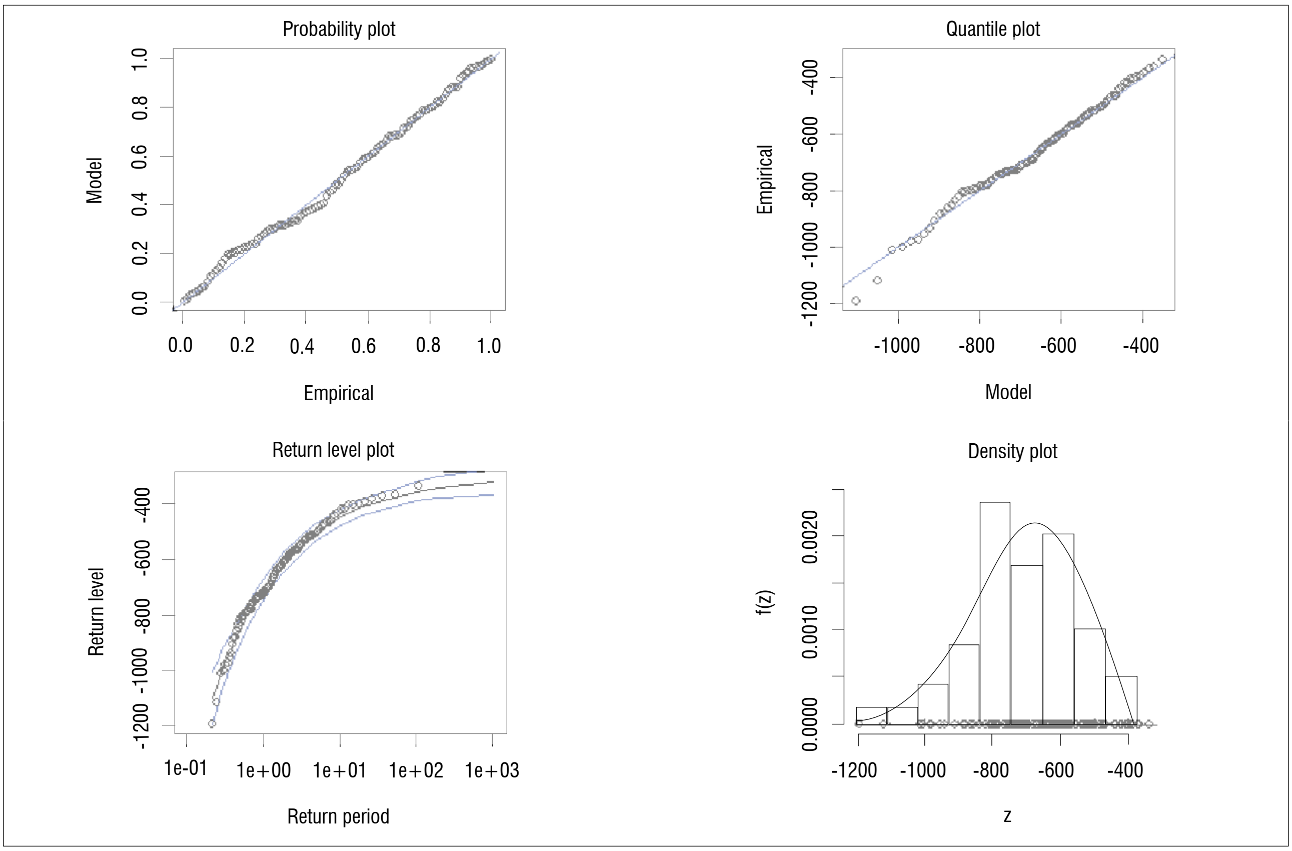


Figure 4: The generalised extreme value probability density function of minima annual rainfall for Zimbabwe for the period 1901–2009.

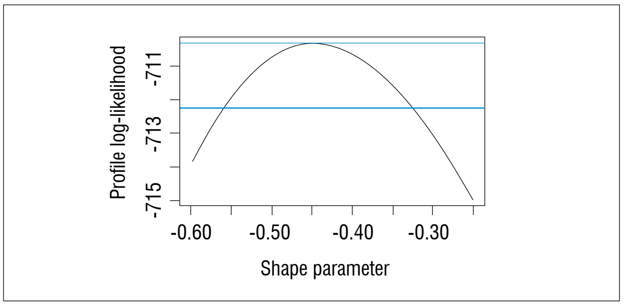


Figure 5: Profile likelihood for the generalised extreme value parameter shape, for minimum annual rainfall for Zimbabwe for the period 1901–2009.

Table 4: Kolmogorov–Smirnov and Anderson–Darling tests to determine whether annual rainfall data for Zimbabwe for 1901–2009 follow a generalised extreme value distribution

Kolmogorov–Smirnov test		Anderson–Darling test	
Statistic	**Critical value**	**Statistic**	**Critical value**
0.058 (0.84)	0.13	0.24	2.50

Return level estimate

The return levels or periods are estimated using the GEVD. Rainfall less than 473 mm per annum is categorised by the Department of Meteorological Services in Zimbabwe as a meteorological drought. Table 5 shows the return level estimates at selected return intervals *T* using the GEVD. Mean annual rainfall is expected to be below the drought threshold value of 473 mm in a return period of *T*=8 years.

Table 5: Return level estimates (mm) at selected return intervals (*T*) determined using the generalised extreme value distribution

T = 5	*T* = 10	*T* = 15	*T* = 20	*T* = 25	*T* = 30	*T* = 35
510.95	451.84	426.18	410.86	400.35	392.55	386.45

The minimum mean annual rainfall for Zimbabwe was 335.3 mm, recorded in the 1991/1992 rainfall season. This is the worst drought in the recorded history of the country. The return level estimate of 335.2 mm is associated with a mean return period of about 90 years, that is, we expect a drought of similar or worse magnitude in 90 years.

Bayesian analysis of minima annual rainfall data

The Markov chain Monte Carlo method was applied to the annual minimum rainfall data. The GEVD scale parameter was re-parameterised as $\phi = \log \sigma$ to retain the positivity of this parameter. The prior density was chosen to be

$$\pi(\mu,\phi,\xi) = \pi_\mu(\mu)\pi_\phi(\phi)\pi_\xi(\xi), \qquad \text{Equation 13}$$

where the marginal priors, $\pi_\mu(.)$, $\pi_\phi(.)$ and, $\pi_\xi(.)$ are

$\mu \sim N(0,400000)$

$\phi \sim N(0,400000)$

$\xi \sim N(0,10000)$

for the three parameters of the GEVD, where, for example, *N(0,400000)* denotes a Gaussian distribution with mean 0 and variance 400 000. These are independent normal priors with large variances. The variances were chosen to be large enough to make the distributions almost flat, corresponding to prior ignorance. In this paper, 30 000 iterations of the algorithm were carried out. Figure 6 shows the Markov chain Monte Carlo trace plots.

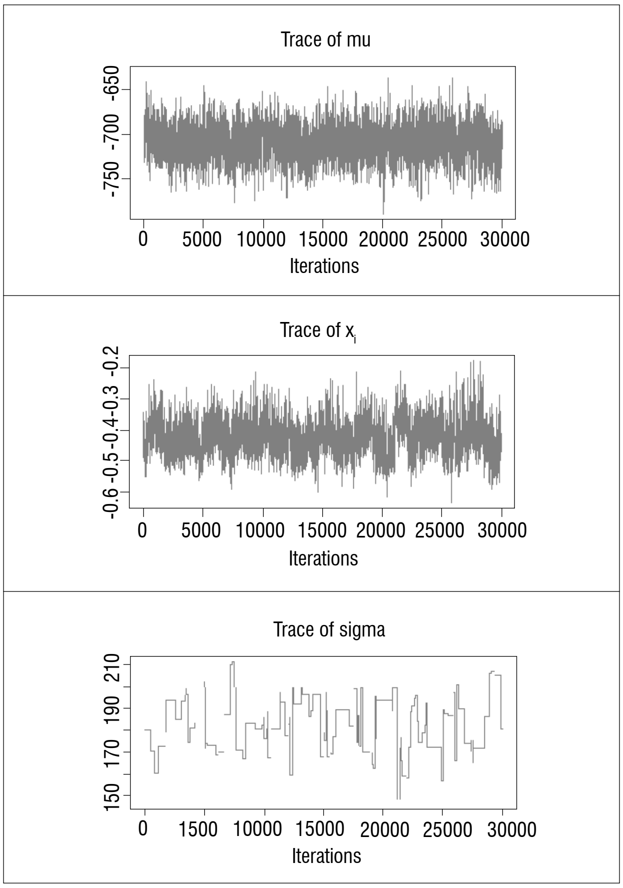


Figure 6: Trace plots of the generalised extreme value distribution parameters using non-informative priors for minimum annual rainfall for Zimbabwe for the period 1901–2009.

To check that the chains had converged to the correct place, different starting points were used. All the chains converged. The estimated posterior densities for the GEVD parameters for Zimbabwe are given in Figure 7.

The posterior means and standard deviations for the GEVD parameters are given in Table 6. Using non-informative priors, which are almost flat and add very little information to the likelihood, the posterior means are close to the maximum likelihood estimates of the GEVD parameters given in Table 3. The frequentist properties are preserved by using non-informative priors in the Bayesian statistics approach.

Conclusion

We modelled extreme minimum annual rainfall in Zimbabwe using the GEVD. Exploring the duality of maxima and minima, annual rainfall data from 1901 to 2009 were fitted to the GEVD. The maximum likelihood estimation method was used to obtain the estimates of the parameters. Model diagnostics, which included the Q-Q plot and the K-S and Anderson–Darling tests, showed that the minimum annual rainfall follows a Weibull class of distribution. The ADF test showed that the minima annual rainfall data were stationary and had no trend. Return level estimates, which are the return levels expected to be exceeded in a certain period, were calculated for Zimbabwe.

The 1992 record drought is likely to return in a mean return period of *T* = 90 years. The Department of Meteorological Services in Zimbabwe categorises a year with mean annual rainfall below 473 mm as a meteorological drought year. The mean annual rainfall is expected to be

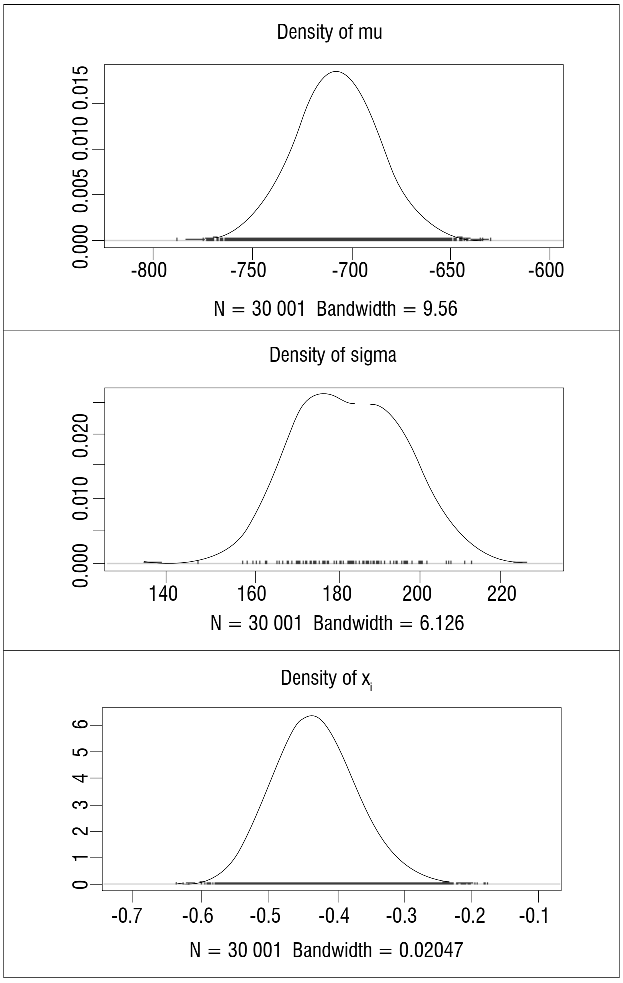


Figure 7: Posterior densities of the generalised extreme value distribution parameters using non-informative priors for minimum annual rainfall for Zimbabwe for the period 1901–2009.

Table 6: Posterior means (standard deviations) of the generalised extreme value distribution parameters

Shape $\hat{\xi}$	Scale $\hat{\sigma}$	Location $\hat{\mu}$
-0.43 (0.15)	182.26 (38.00)	-705.65 (55.63)

Mean annual rainfall is expected to be below the drought threshold value of 473 mm in a return period of 8 years.

less than the drought threshold value of 473 mm in a mean return period of $T = 8$ years.

The GEVD parameter estimates using the Bayesian approach were close to the maximum likelihood estimates with smaller standard deviations. Using non-informative priors, the frequentist properties of the model are preserved. Using classical statistics is therefore akin to using Bayesian statistics with non-informative priors. Expert opinion, when available in future, can be used to further improve the model. An area of further research is modelling and predicting minimum annual rainfall in Zimbabwe using the Bayesian approach using informative priors. The use of expert priors may improve the precision of the parameters over the maximum likelihood estimates.

Authors' contributions

R.C. performed the data analysis and wrote the paper; D.C. supervised R.C., came up with the concept of statistics of extremes to the problem and introduced it to R.C. together with the preliminary analysis and draft on outcomes, provided guidance, and proofread and corrected the paper.

References

1. Makarau A, Jury MR. Predictability of Zimbabwe summer rainfall. Int J Climatol. 1997;17:1421–1432. http://dx.doi.org/10.1002/(SICI)1097-0088(19971115)17:13<1421::AID-JOC202>3.0.CO;2-Z
2. Panu US, Sharma TC. Challenges in drought research: Some perspectives and future directions. Hydro Sci J. 2002;47(S):S19–S30.
3. Jury MR. Regional tele-connection pattern associated with summer rainfall over South Africa, Namibia and Zimbabwe. Int J Climatol. 1996;16:135–153. http://dx.doi.org/10.1002/(SICI)1097-0088(199602)16:2<135::AID-JOC4>3.0.CO;2-7
4. Hasan HB, Ahmad NF, Suraiya BK. Modeling of extreme temperature using generalized extreme value (GEV) distribution: A case study of Penang. In: Ao SI, Gelman L, Hukins DWL, Hunter A, Korsunsky AM, editors. Proceedings of the World Congress on Engineering volume 1; 2012 July 4–6; London, UK. Hong Kong: Newswood Limited; 2012. p. 1–6.
5. Coles S. An introduction to statistical modeling of extreme values. London: Springer; 2001. http://dx.doi.org/10.1007/978-1-4471-3675-0
6. Nadarajah S, Choi D. Maximum daily rainfall in South Korea. J Earth Sys Sci. 2007;116(4):311–320. http://dx.doi.org/10.1007/s12040-007-0028-0
7. Koutsoyiannis D. Statistics of extremes and estimation of extreme rainfall: 11 Empirical investigations of long rainfall records. Hydrol Sci J. 2004;49(4):591–610. http://dx.doi.org/10.1623/hysj.49.4.591.54424
8. Koutsoyiannis D, Baloutsos G. Analysis of a long record of annual maximum rainfall in Athens, Greece and design rainfall inferences. Nat Hazards. 2000;22:29–48. http://dx.doi.org/10.1023/A:1008001312219
9. Crisci A, Gozzini B, Meneguzzo F, Pagliara S, Maracchi G. Extreme rainfall in a changing climate: Regional analysis and hydrological implications in Tuscany. Hydrol Process. 2002;6:1261–1274. http://dx.doi.org/10.1002/hyp.1061
10. Chikobvu D, Sigauke C. Modeling influence of temperature on daily peak electricity demand in South Africa. J Energy South Afr. 2013;24(4):63–70.
11. Smith RL, Goodman DL. Bayesian risk analysis. Technical report. Chapel Hill, NC: Department of Statistics, University of North Carolina; 2000.
12. Hilborn R, Mangel M. The ecological detective: Confronting models with data. Monographs on population biology. Princeton, NJ: Princeton University Press; 1997.
13. Ludwig D. Uncertainty and the assessment of extinction probabilities. Ecol Appl. 1996;6(4):1067–1076. http://dx.doi.org/10.2307/2269591
14. Buckle C. Weather and climate in Africa. Harlow: Longman; 1996.
15. Smith SV. Studies of the effects of cold fronts during the rainy season in Zimbabwe. Weather. 1985;(40):198–203. http://dx.doi.org/10.1002/j.1477-8696.1985.tb06869.x
16. Hosking JRM, Wallis J. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics. 1987;29:339–349. http://dx.doi.org/10.1080/00401706.1987.10488243
17. Bunya PK, Jain S, Ohjha C, Agarwal A. Simple parameter estimation technique for three-parameter generalized extreme value distribution. J Hydrol Eng. 2007;12(6):682–689. http://dx.doi.org/10.1061/(ASCE)1084-0699(2007)12:6(682)
18. Beirlant J, Goegebeur Y, Segers J, Teugels J. Statistics of extremes: Theory and applications. Chichester: John Wiley & Sons; 2004. http://dx.doi.org/10.1002/0470012382
19. Bali TG. The generalized extreme value distribution. Econ Lett. 2003;79:423–427. http://dx.doi.org/10.1016/S0165-1765(03)00035-1
20. Hosking JRM. Testing whether the shape parameter is zero in the generalized extreme value distribution. Biometrika. 1984;71:367–374.
21. Diebolt L, Guillou A, Rached I. Application of the distribution of excesses through a generalized probability-weighted moments method. J Stat Plan Infer. 2007;137:841–857. http://dx.doi.org/10.1016/j.jspi.2006.06.012
22. National Environmental Research Council (NERC). Flood studies report. Wallingford: Institute of Hydrology; 1975.

23. Hosking JRM. Algorithm AS 215: Maximum likelihood estimation of the parameters of the generalized extreme value distribution. Appl Stat. 1985;34:301–310. http://dx.doi.org/10.2307/2347483

24. Madsen H, Pearson CP, Rosbjerg D. Comparison of annual maximum series and partial duration series methods for modelling extreme hydrological events II: Regional modelling. Water Resour Res. 1997;17:1421–1432.

25. Prescott P, Walden AT. Maximum likelihood estimation of the parameters of the generalized extreme value distribution. Biometrika. 1980;67:723–724. http://dx.doi.org/10.1093/biomet/67.3.723

26. Prescott P, Walden AT. Maximum likelihood estimation of the parameters of the three-parameter generalized extreme value distribution from censored samples. J Stat Comput Sim. 1983;16:241–250. http://dx.doi.org/10.1080/00949658308810625

27. Macleod AJ. AS R76 – A remark on the algorithm AS 215: Maximum likelihood estimation of the parameters of the generalized extreme value distribution. Appl Stat. 1989;38:198–199. http://dx.doi.org/10.2307/2347695

28. Smith RL. Maximum likelihood estimation in a class of non-regular cases. Biometrika. 1985;72:67–90. http://dx.doi.org/10.1093/biomet/72.1.67

29. Stephens MA. EDF statistics for goodness of fit and some comparisons. J Amer Stat Assoc. 1974;69:730–737. http://dx.doi.org/10.1080/01621459.1974.10480196

30. Coles S, Tawn J. Bayesian modelling of extreme surges on the UK east coast. Philos T Roy Soc A. 2005;363:1387–1406. http://dx.doi.org/10.1098/rsta.2005.1574

31. Gamerman D. Markov chain Monte Carlo: Stochastic Bayesian inference. London: Chapman and Hall; 1997.

Note: This article is supplemented with online only material.

Digital terrain model height estimation using support vector machine regression

AUTHORS:
Onuwa Okwuashi[1]
Christopher Ndehedehe[2]

AFFILIATIONS:
[1]Department of Geoinformatics and Surveying, University of Uyo, Uyo, Nigeria
[2]Department of Spatial Science, Curtin University, Perth, Australia

CORRESPONDENCE TO:
Onuwa Okwuashi

EMAIL:
onuwaokwuashi@gmail.com

POSTAL ADDRESS:
Department of Geoinformatics and Surveying, University of Uyo, PMB 1017, Uyo, Aks, Nigeria

DATES:
Received: 03 May 2014
Revised: 01 Nov. 2014
Accepted: 03 Feb. 2015

KEYWORDS:
height estimation; artificial neural network; nearest neighbour; algorithm; machine learning

HOW TO CITE:
Okwuashi O, Ndehedehe C. Digital terrain model height estimation using support vector machine regression. S Afr J Sci. 2015;111(9/10), Art. 2014-0153, 5 pages. http://dx.doi.org/10.17159/sajs.2015/20140153

Digital terrain model interpolation is intrinsically a surface fitting problem, in which unknown heights H are estimated from known X-Y coordinates. Notable methods of digital terrain model interpolation include inverse distance to power, local polynomial, minimum curvature, modified Shepard's method, nearest neighbour and polynomial regression. We investigated the support vector machine regression (SVMR) as a new alternative method to these models. SVMR is a contemporary machine learning algorithm that has been applied to several real-world problems aside from digital terrain modelling. The SVMR results were compared with those from notable parametric (the nearest neighbour) and non-parametric (the artificial neural network) techniques. Four categories of error analysis were used to assess the accuracy of the modelling: minimum error, maximum error, means error and standard error. The results indicate that SVMR furnished the lowest error, followed by the artificial neural network model. The SVMR also produced the smoothest surface followed by the artificial neural network model. The high accuracy furnished by SVMR in this experiment attests that SVMR is a promising model for digital terrain model interpolation.

Introduction

Engineers and other related scientists are often charged with the responsibility of producing digital maps that represent the three-dimensional visualisation of the earth's surface. These maps usually serve as auxiliary data for engineering designs of roads, bridges, drainage systems and general landscaping. These digital three-dimensional maps are referred to generically as digital terrain models (DTMs). A DTM is referred to as a 'form of computer surface modelling which deals with the specific problems of numerically representing the surface of the earth'[1].

A DTM is created by using one of two methods: triangulation or gridding. In a gridding method, the corners of regular rectangles or squares are calculated from the scattered control points. In triangulation, triangles are created based on the scattered control points. These triangles do not intersect each other and represent the terrain surface as a linear or non-linear function. In both methods, the heights of grid points and triangular points with unknown heights in the modelled area are estimated by interpolation using their control points.[2] Notable methods of DTM interpolation include inverse distance to power, local polynomial, minimum curvature, modified Shepard's method, polynomial regression, radial basis function, Kriging and nearest neighbour.[3,4]

The aim of this study was to investigate the support vector machine regression (SVMR) model,[5-7] as a new method of DTM interpolation. The goal of the SVMR model is to construct a hyperplane that lies close to as many data points as possible, by choosing a hyperplane that has a small norm that simultaneously minimises the sum of the distances from the data points to the hyperplane. SVMR attempts to minimise the generalisation error bound so as to achieve generalised performance, instead of minimising the observed training error. The idea of SVMR is based on the computation of a linear regression function in a high-dimensional feature space in which the input data are mapped via a non-linear function.[7] In this experiment, the SVMR results were compared with those from a notable parametric technique (nearest neighbour, NN) and a notable non-parametric technique (artificial neural network, ANN).

Support vector machine regression algorithm

For the linear case, given the set of data,

$$(y_i,x_i),\dots,(y_l,x_l),x\in R^n,y\in R$$

with a linear function

$$f(x)=(w\cdot x)+b,^{8} \qquad \text{Equation 1}$$

hence the regression function is given by the minimum of the functional

$$\Phi(w,\xi^*,\xi)=\frac{1}{2}|w|^2+C\left[\sum_{i=1}^{l}\xi_i+\sum_{i=1}^{l}\xi_i^*\right] \qquad \text{Equation 2}$$

where C is a pre-specified penalty value, and ξ, ξ^* are slack variables representing upper and lower constraints, respectively.[9] Using an ε-insensitive loss function

$$L_\varepsilon(y)=\begin{cases}0 \text{ for } |f(x)-y|<\varepsilon \\ |f(x)-y|-\varepsilon \text{ otherwise}\end{cases} \qquad \text{Equation 3}$$

the solution is given by:

$$\max_{\alpha,\alpha^*}W(\alpha,\alpha^*)=\max_{\alpha,\alpha^*}\begin{Bmatrix}-\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(x_i\cdot x_j) \\ +\sum_{i=1}^{l}\alpha_i(y_i-\varepsilon)-\alpha_i^*(y_i+\varepsilon)\end{Bmatrix} \qquad \text{Equation 4}$$

with constraints

$$0\leq\alpha_i\leq C, \qquad i=1,...,l \qquad \text{Equation 5}$$

$$0\leq\alpha_\mathrm{i}^*\leq C, \qquad i=1,...,l \qquad \text{Equation 6}$$

$$\sum_{i=1}^{l}(\alpha-\alpha_i^*)=0 \qquad \text{Equation 7}$$

$$\overline{w}= \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i \qquad \text{Equation 8}$$

$$\overline{b}=-\frac{1}{2}\,\overline{w}\cdot[x_r\cdot x_s] \qquad \text{Equation 9}$$ [6]

The Karush–Kunn–Tucker conditions that are satisfied by the solution are:

$$\overline{\alpha_i\alpha_i}^*=0, \qquad i=1,...,l \qquad \text{Equation 10}$$

$$\min_{\beta}\frac{1}{2}\sum_{i=1}^{l}\ \sum_{j=1}^{l}\beta_i\beta_j(x_i\cdot x_j)-\sum_{i=1}^{l}\beta_i y_i \qquad \text{Equation 11}$$

with constraints

$$-C\leq\beta_i\leq C, \qquad \text{where } i=1,...,l \qquad \text{Equation 12}$$

$$\sum_{i=1}^{l}\beta_i=0 \qquad \text{Equation 13}$$

and the regression function is given by Equation 1, where:

$$\overline{w}=\sum_{i=1}^{l}\overline{\beta}_i \mathrm{x}_i \qquad \text{Equation 14}$$

and

$$\overline{b}=-\frac{1}{2}\,\overline{w}\cdot[x_r\cdot x_s] \qquad \text{Equation 15}$$ [10]

The non-linear SVMR solution using an ε-sensitive function is given by:

$$\max_{\alpha,\alpha^*}W(\alpha,\alpha^*)=\max_{\alpha,\alpha^*}\left\{\begin{matrix}\sum_{i=1}^{l}\alpha_i^*(\mathrm{y}_i-\varepsilon)-\alpha_i(\mathrm{y}_i+\varepsilon)\\ -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)\end{matrix}\right\} \qquad \text{Equation 16}$$

with constraints

$$0\leq\alpha_i\leq C, \qquad i=1,...,l \qquad \text{Equation 17}$$

$$0\leq\alpha_i^*\leq C, \qquad i=1,...,l \qquad \text{Equation 18}$$

$$\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0 \qquad \text{Equation 19}$$

Solving Equation 16, with constraints Equations 17–19, determines the Lagrange multipliers α_i , α_i^* and the regression function is given by:

$$f(x)=\sum_{SVs}(\overline{\alpha}_i-\overline{\alpha}_i^*)K(x_i,x)+\overline{b} \qquad \text{Equation 20}$$

where

$$\overline{w}\cdot x=\sum_{SVs}(\overline{\alpha}_i-\overline{\alpha}_i^*)K(x_i,x) \qquad \text{Equation 21}$$

$$\overline{b}=-\frac{1}{2}\sum_{SVs}(\overline{\alpha}_i-\overline{\alpha}_i^*)[K(x_r,x_i)+K(x_s,x_i)] \qquad \text{Equation 22}$$

The kernel $K(x_i,x_j)$can be any of the following common kernel functions: the linear kernel $x\cdot x_i$, polynomial kernel $(x\cdot x_i+1)^d$ and radial basis function kernel

$$K(x_i,x_j)=\exp\left(-\frac{|x_i-x_j|^2}{2\gamma^2}\right).$$ [7]

Methodology

The eastings (E), northings (N) and orthometric heights (H) of 601 points were sampled from a 3.068-ha portion located in southern Nigeria (Figure 1). The sampled area was bounded by UTM coordinates 381810E–382060E and 559700N–559980N (Figure 2). Out of the 601 sampled points, 200 points were selected for training the SVMR algorithm.

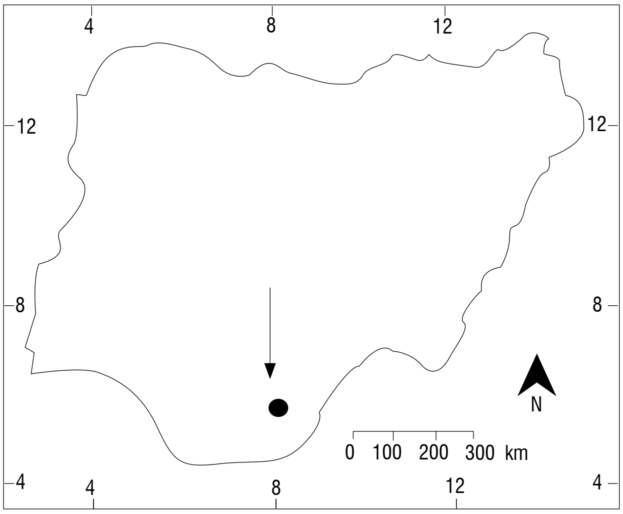


Figure 1: A map showing the location of the 3.068-ha area sampled in southern Nigeria.

The experiment was done in MATLAB. The training data contained E, N and H values. E and N were the explanatory variables while H was the target variable. The test data contained just E and N in order to predict the values of H. The SVMR model was trained with known E, N and H values in order to estimate H values of points not used as training data.

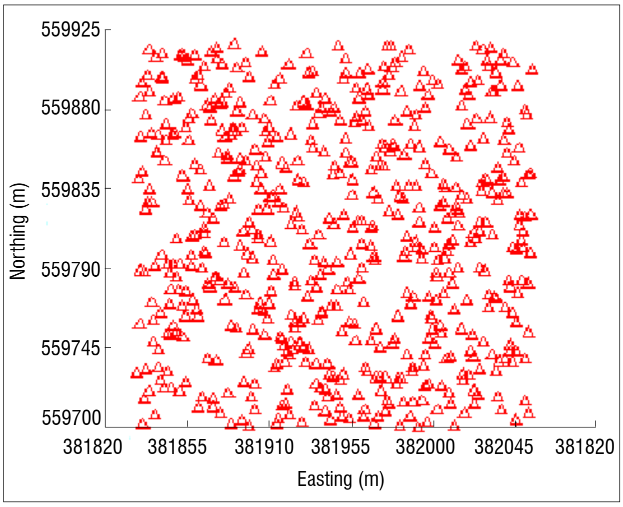


Figure 2: UTM coordinates of the sampled points in the study area.

The stratified random sampling was used to select the training data. The spatial dependency of the observed data was examined by plotting its semivariances against the lag distance to produce a semivariogram (Figure 3). The calculated nugget was 0.6. The range was 70, while the sill was 29.6. Beyond the range value the data are spatially independent, whereas the data within the range area are spatially dependent. The nugget value of 0.6 indicated that the error in observation was minimal. The sill value of 29.6 indicated the maximum value of the semivariance that corresponded to the range value.

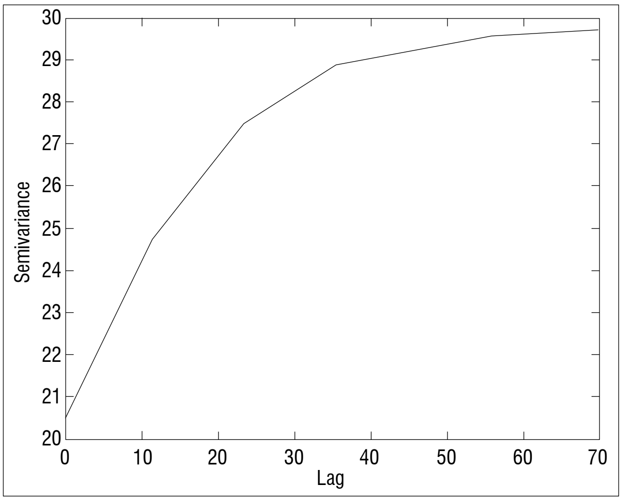
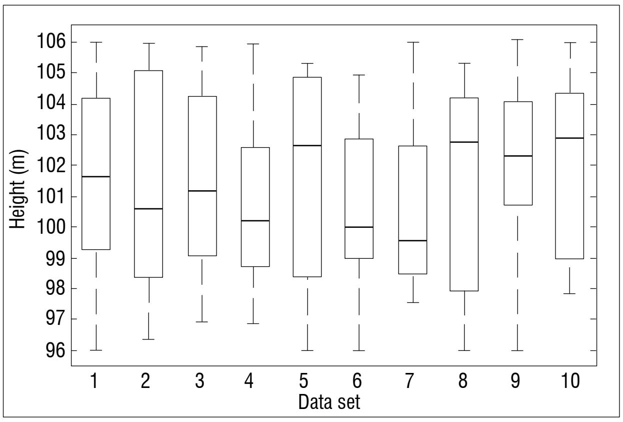


Figure 3: Semivariogram of the observed data.

The 200 points selected for training and testing were split into classes of 10 (Figure 4). The box plot in Figure 4 shows the statistical distribution of the data. The inner line of the box indicates the median of the data. The top of the upper tail indicates the highest value of the data, while the bottom of the lower tail indicates the lowest value of the data. The experiment was implemented with an optimal ε-sensitive function value $\varepsilon=0.1$. The polynomial, radial basis function and the linear kernels were investigated to select the best kernel function for the experiment, through the method of cross-validation.



The inner lines of the box indicate the median of the data. The top of the upper tail indicates the highest value of the data, while the bottom of the lower tail indicates the lowest value of the data.

Figure 4: Box plots of the 10 data sets of 20 sample points each.

The selection of the optimum kernel parameters' values of degree *d*, gamma and penalty value *C* was done using the k-fold cross-validation process where k=10.[11] In each experiment, nine data sets (k-1 data sets) were put together to train the SVMR while the remaining one data set was held to test the accuracy of the experiment. The experiment was repeated in 10 folds until all 10 data sets were used for both training and testing (Figure 5).

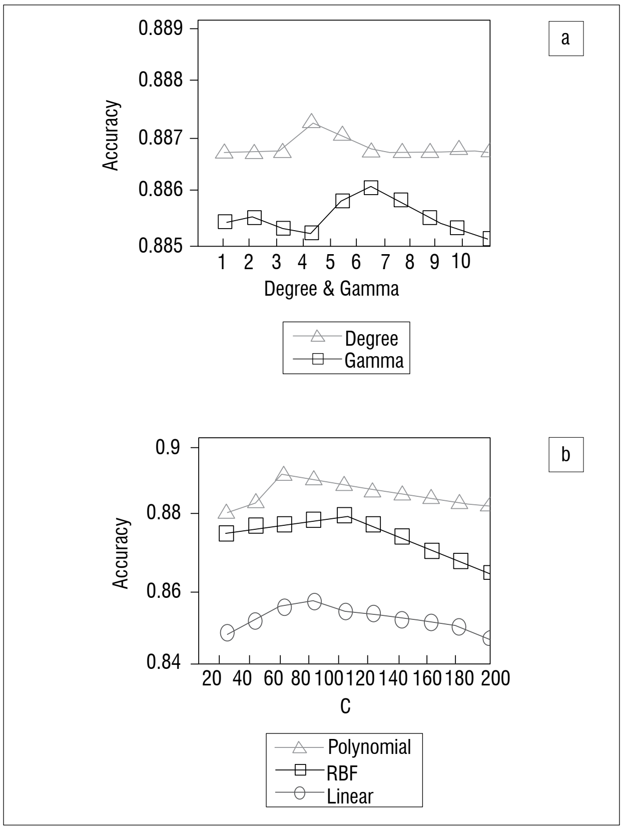


Figure 5: Cross-validation results for (a) degree and gamma and (b) penalty value, *C*.

The cross-validation result from Figure 5a shows the accuracy for degree and gamma using values 1 to 10. The value of 4 was found to be the value for degree with the highest accuracy, while 6 was found to be the highest value of gamma with the highest accuracy. From Figure 5b, the values of *C* with the highest accuracy were 60, 100 and 80 for polynomial, radial basis function and linear kernels, respectively. The polynomial kernel yielded the highest accuracy, and the best value of *d* was 4 (Figure 5).

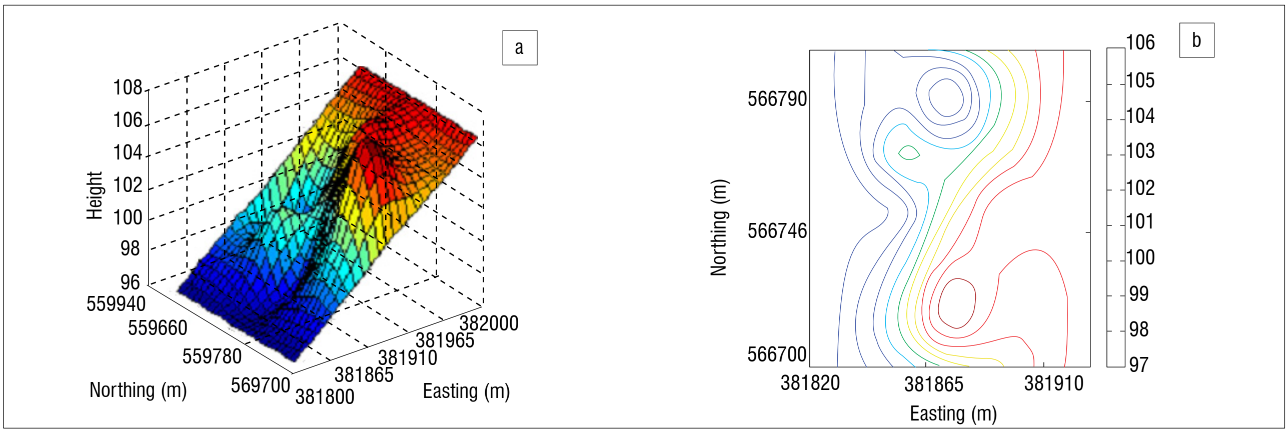


Figure 6: Support vector machine regression method (a) digital terrain model plot and (b) contour plot.

Results

The resulting DTM and contour plots from interpolation using SVMR are depicted in Figure 6. After interpolation of the heights of the unsampled points using SVMR, the same was implemented with the ANN and NN models.

The NN technique predicts the value of an attribute at an unsampled point based on the value of the nearest sample by drawing perpendicular bisectors between sampled points (n), thereby forming polygons (V_i, $i=1,2,...,n$).[4] One polygon is produced per sample and the sample is located in the centre of the polygon, such that in each polygon all points are nearer to its enclosed sample point than to other sample points.[12-14] The estimations of the attribute at unsampled points within polygon V_i are the measured value at the nearest single sampled data point x_i, that is $\hat{z}(x_0)=z(x_i)$. The weights are:

$$\lambda_i = \begin{cases} 1 \text{ if } x_i \in V_i \\ 0 \text{ otherwise} \end{cases} \qquad \text{Equation 23}$$

All points within each polygon are assigned the same value.[12,14] The NN DTM and contour plots are presented in Figure 7.

The ANN was programmed using multilayer perceptron with a sigmoidal hidden-layer transfer function and linear output neurons. The multilayer perceptron neural network was trained with a back-propagation algorithm, using a two-layer feed-forward neural network. The network had an input layer, an output layer and one or two hidden layers; however, there is no limit to the number of hidden layers.[15] Basically a signal from neuron i of the first input layer of a cell x at time t received by a neuron j of the hidden layer can be expressed as:

$$net_j(x,t)=\sum_i W_{i,j} S'_i(x,t) \qquad \text{Equation 24}$$

where $S'_i\,(x,t)$ denotes the site attributes given by variable (neuron) i; $W_{i,j}$ is the weight of the input from neuron i to neuron j; $net_j(x,t)$ is the signal received for neuron j of cell x at time t.[15] Based on the method of cross-validation, the random seed number was set and the required *number of neurons* in the hidden layer was set between 1 and 50. The ANN was initialised with initial weights; hence different results were obtained every time the ANN model was run. To ensure the results remained the same at every run of the neural network the random seed number was kept constant. The random seed number is an arbitrary constant chosen by trial and error.

After the random seed number had been set, the number of hidden neurons was the single parameter that was adjusted to obtain simulation results of the ANN. The training of the neural network was done by simply adjusting the number of neurons in the hidden layer in order to minimise the training error. The training error is the discrepancy between the predicted and the actual value. The adjustment of the number of neurons was sustained until the training error fell below a pre-determined threshold.[16-19] The ANN DTM and contour plots are presented in Figure 8.

In Figure 9, the sampled heights were plotted against the 200 points used for the validation of the models. The discrepancies between the observed points and the predicted points using SVMR, ANN and NN are depicted in Figure 9. The plots in Figure 9 show that SVMR yielded the best fit, followed by ANN.

The experimental errors were calculated by comparing the estimated values using these models with their actual values from the sample. Out of the 601 sampled points, 200 points were selected for training while 50 points were used to test the accuracy of the modelling. The final result was selected from several results obtained by repeating the process, and by re-selecting the training and test data (Figure 10).

The box plot in Figure 10a shows the statistical error distribution of the data. The inner lines of the box indicate the median of the data. The top of the upper tail indicates the highest value of the data and the bottom of the lower tail indicates the lowest value of the data. From Figure 10a, the calculated minimum error was -1.39, the maximum error was 2.05 and the mean error was -0.07, using the SVMR model. For the ANN model, the calculated minimum error was -1.44, the maximum error was 2.08 and the mean

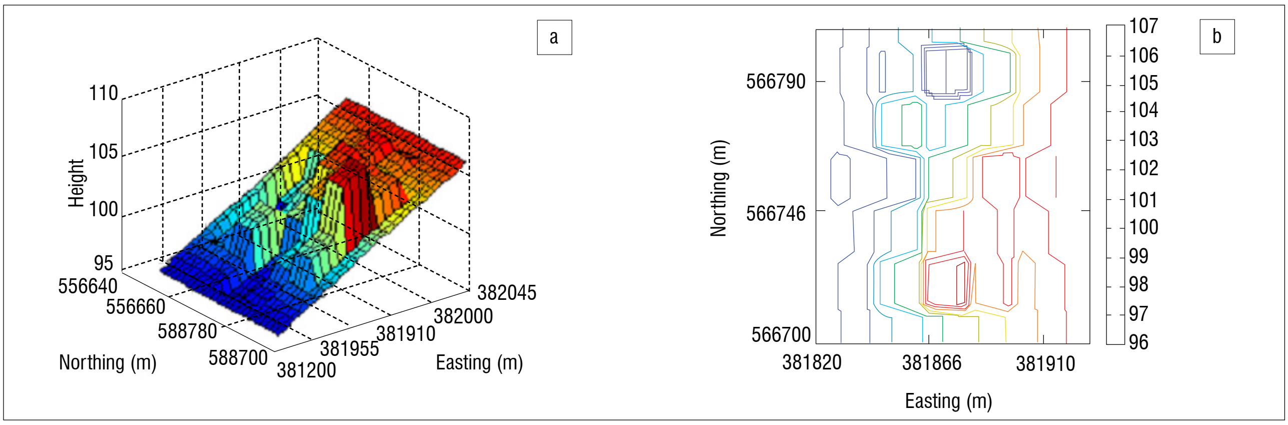


Figure 7: Nearest neighbour method (a) digital terrain model plot and (b) contour plot.

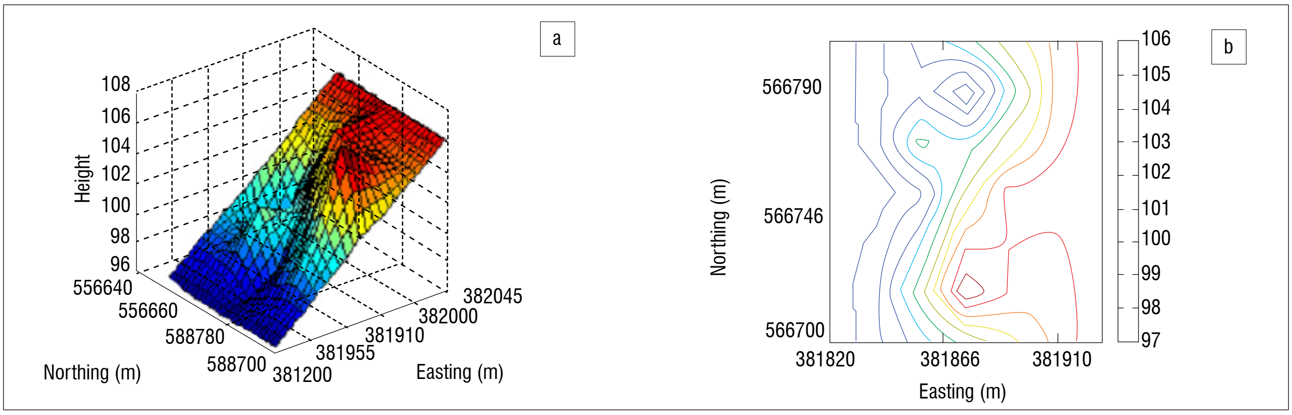


Figure 8: Artificial neural network method (a) digital terrain model plot and (b) contour plot.

error was -0.09. For the NN model, the calculated minimum error was -2.97, the maximum error was 3.86 and the mean error was 0.35. From Figure 10b, the calculated standard errors for SVMR, ANN and NN were 0.60, 0.65 and 0.81, respectively.

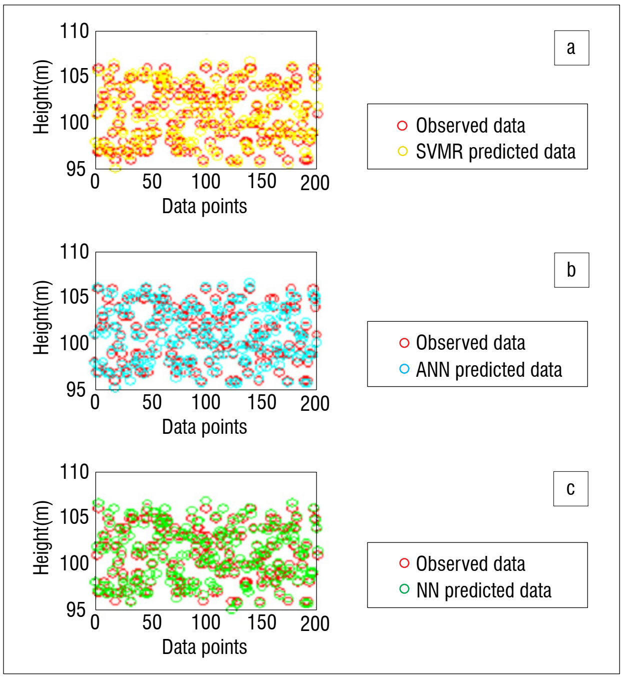


Figure 9: Plots showing discrepancies between the observed heights and the heights predicted using (a) support vector machine regression (SVMR), (b) artificial neural network (ANN) and (c) nearest neighbour (NN) methods.

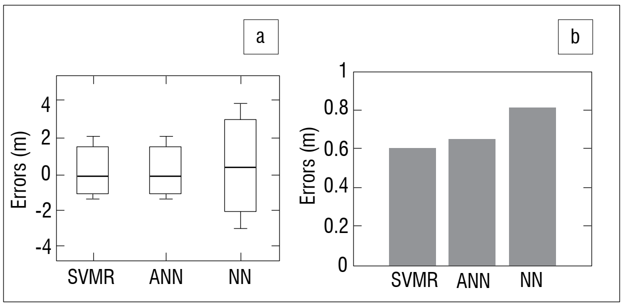


Figure 10: (a) Box plot showing statistical error distribution and (b) bar plot showing the calculated standard errors for support vector machine regression (SVMR), artificial neural network (ANN) and nearest neighbour (NN) models.

Conclusion

The SVMR results were compared with those of the NN and ANN techniques. The NN and ANN are common parametric and non-parametric models, respectively, that have been used in previous studies.[3] From Figure 10, we can see that SVMR produced the best results of all the models, followed by the ANN model. The results from the NN model were the least accurate. SVMR also produced the smoothest surface, followed by ANN, while NN produced the roughest surface (Figures 6–8). Even though the SVMR is not a common method of DTM interpolation, our results show that it is a robust technique and can be considered for spatial surface interpolation. However, the differences between the SVMR and ANN results are not significant; more examples are required for the generalisation to be valid.

Authors' contributions

O.O. was the project leader, coordinated the research and collected the field data in Nigeria. C.N. performed the laboratory analysis of the field data in Australia.

References

1. Heesom D, Mahdjobi L. Effect of grid resolution and terrain characteristics on data from DTM. J Comput Civil Eng. 2001;15(2):137–143. http://dx.doi.org/10.1061/(ASCE)0887-3801(2001)15:2(137)

2. Yanalak M. Effect of gridding gethod on digital terrain model profile data based on scattered data. J Comput Civil Eng. 2003;1(58):58–67. http://dx.doi.org/10.1061/(ASCE)0887-3801(2003)17:1(58)

3. Karaborka H, Baykanb OK, Altuntasa C, Yildza F. Estimation of unknown height with artificial neural network on digital terrain model. Vol. XXXVII, Part B3b. Beijing: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; 2008.

4. Li J, Heap AD. A review of spatial interpolation methods for environmental scientists. Geoscience Australia Record. 2008/23. Canberra: Geoscience Australia; 2008. Available from: http://www.ga.gov.au/webtemp/image_cache/GA12526.pdf

5. Vapnik V. Statistical learning theory. Berlin: Springer; 1998.

6. Smola AJ, Scholkopf B. A tutorial on support vector regression. NeuroCOLT2 technical report NC2-TR-1998-030 [document on the Internet]. c1998 [cited 2015 Aug 31]. Available from: http://www.svms.org/regression/SmSc98.pdf

7. Gunn S. Support vector machines for classification and regression. ISIS technical report. Southamption: University of Southampton; 1998.

8. Basak D, Pal S, Patranabis DC. Support vector regression. Neural Inform Process Lett Rev. 2007;11:203–224.

9. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: Mozer M, Jordan M, Petsche T, editors. Advances in neural information processing systems 9. Cambridge, MA: MIT Press; 1997. p. 155–161.

10. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation and signal processing. In: Mozer M, Jordan M, Petsche T, editors. Advances in neural information processing systems 9. Cambridge, MA: MIT Press; 1997. p. 281–287.

11. Bhardwaj N, Langlois R, Zhao G, Lu H. Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res. 2005;33:6486–6493. http://dx.doi.org/10.1093/nar/gki949

12. Ripley BD. Spatial statistics. New York: John Wiley & Sons; 1981.

13. Isaaks EH, Srivastava RM. Applied geostatistics. New York: Oxford University Press; 1989.

14. Webster R, Oliver MA. Sample adequately to estimate variograms of soil properties. J Soil Sci. 1992;43:177–192. http://dx.doi.org/10.1111/j.1365-2389.1992.tb00128.x

15. Almeida CM, Gleriani JM, Castejon EF, Soares-Filho BS. Using neural networks and cellular automata for modelling intra-urban land-use dynamics. Int J Geogr Inform Sci. 2008;22(9):943–963. http://dx.doi.org/10.1080/13658810701731168

16. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press; 1986. p. 318–362.

17. Anderson JA. An introduction to neural networks. Cambridge, MA: MIT Press; 1995.

18. Chauvin Y, Rumelhart DE, editors. Backpropagation: Theory, architectures and applications. Hillsdale, NJ: Erlbaum; 1995.

19. Wang F. The use of artificial neural networks in geographical information systems for agricultural land suitability assessment. Environ Planning A. 1994;26:265–284. http://dx.doi.org/10.1068/a260265

Testing and implementation of a transportable and robust radio-element mapping system

AUTHOR:
Jacques Bezuidenhout[1]

AFFILIATION:
[1]School for Science and Technology, Stellenbosch University, Faculty of Military Science, Stellenbosch, South Africa

CORRESPONDENCE TO:
Jacques Bezuidenhout

EMAIL:
jab@ma2.sun.ac.za

POSTAL ADDRESS:
Faculty of Military Science, Stellenbosch University, Private Bag X2, Saldanha 7395, South Africa

DATES:
Received: 08 Oct. 2014
Revised: 04 Dec. 2014
Accepted: 06 Feb. 2015

KEYWORDS:
gamma ray spectroscopy; scintillation detector; in-situ measurements; naturally occurring radionuclides; geographical information systems

HOW TO CITE:
Bezuidenhout J. Testing and implementation of a transportable and robust radio-element mapping system. S Afr J Sci. 2015;111(9/10), Art. #2014-0350, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20140350

Gamma ray spectroscopy has been successfully applied as a survey tool in the fields of morphology, geology and mineral exploration. Gamma ray surveys are regularly done at ground level, which frequently requires transecting remote and unforgiving environments. Thus a need for the development of a transportable, robust and portable gamma ray detection system was identified. In addition to collecting radiation data, such a system was required to also provide the geographic position of the data and allow for various analyses tools to be utilised in the field. These functions were achieved by integrating a USB-driven scintillation detector with a field tablet and creating software to control acquisition and analyses of radiation data, as well as logging position. The system was tested in different geographical locations under different modes of transport. The instrument was tested by employing several different methods of data analysis in order to extract natural nuclide condensations. The consistency in the obtained data demonstrated the reliability of the instrument in the different environments. The system also successfully replicated previous radio-element survey findings and provided information on several geographical phenomena, including information on the geology, paved road structure and beach sediment characteristics.

Introduction

Radio-element mapping is utilised in various fields of research at different locations across southern Africa. These research locations are frequently difficult to access and environmental conditions tend to be extreme. Such areas include stretches of remote coastlines, deserts and mountains. Access to the areas therefore poses a great challenge to researchers who employ radiation-measuring techniques. Mapping of radionuclide concentrations has in the past required the time-intensive practice of sediment sampling or stationary in-situ measurements, both of which limit the number of sampling points and usually require post processing.

Transportable in-situ gamma ray spectrometry measurement techniques have been developed by several organisations[1-3] in order to measure larger areas more effectively. This motivated the development of a unique, cost-effective and transportable measuring system suitable for the local geographical environment in southern Africa. Various means of transportation of gamma ray measuring systems are possible and include aircraft, motor vehicle and even on foot. However, the mode of transportation must be carefully selected, taking into consideration the research requirements and the geographical factors of the research area. Most of these modes of transportation subsequently require a robust and modular system that is capable of multiplatform deployment. The system needs to have an extended battery capability but also must be lightweight to ensure measurements can be conducted on foot. The system also requires a global positioning system (GPS) to record position while surveying in remote study areas.

The most significant naturally occurring radionuclides are uranium (^{238}U), thorium (^{232}Th) and potassium (^{40}K), of which potassium is by far the most abundant and fluctuating.[4] Distinctive concentrations of these naturally occurring radionuclides usually indicate important geophysical phenomena which must be thoroughly investigated while on location. On-site indication of such geographical phenomena helps to guide the system operator to ensure sufficient field data are recorded that can later be used for further processing and investigation. The system must therefore provide the operator with real-time information on changes in the naturally occurring radionuclides in order to allow for the operator to modify data collection accordingly.

In-situ gamma ray spectra can be analysed through various methods[1,5-7] in order to more accurately extract nuclide concentrations. The acquisition of these spectra is managed by software and some of this software provides access and control over the methods that are employed for data analysis. The ideal is that the researcher should have the option of choosing between the preferred methods of analysis, or even be able to employ novel methods. The measuring systems should preferably have the functionality to allow the researcher to choose a method while on location.

A transportable in-situ gamma ray measurement system was consequently developed, taking all of the above-mentioned requirements into account. The system was tested in various geographical environments and the results were compared to the known physical factors of the test locations.

Methods

Measuring system

The measuring system consisted of a NaI(Tl) scintillation detector (Rexon Components Inc., NAI 3.0 PX 3.0 / 3.0 IV, Beachwood, NJ, USA), a digital multichannel analyser (MCA), a rugged tablet PC with an on-board GPS and real-time analysis software that controlled the entire system (Figure 1). The NaI(Tl) detector (76.2 mm x 76.2 mm) was coupled to the MCA and sealed in a padded case to protect the instruments from mechanical shock and dust. The scintiSPEC® MCA (http://gs.flir.com/) that is produced by FLIR® (Solingen, Germany) has a USB connection that acts as the power source for operation and allows data transfer. The combination of NaI(Tl) detector and

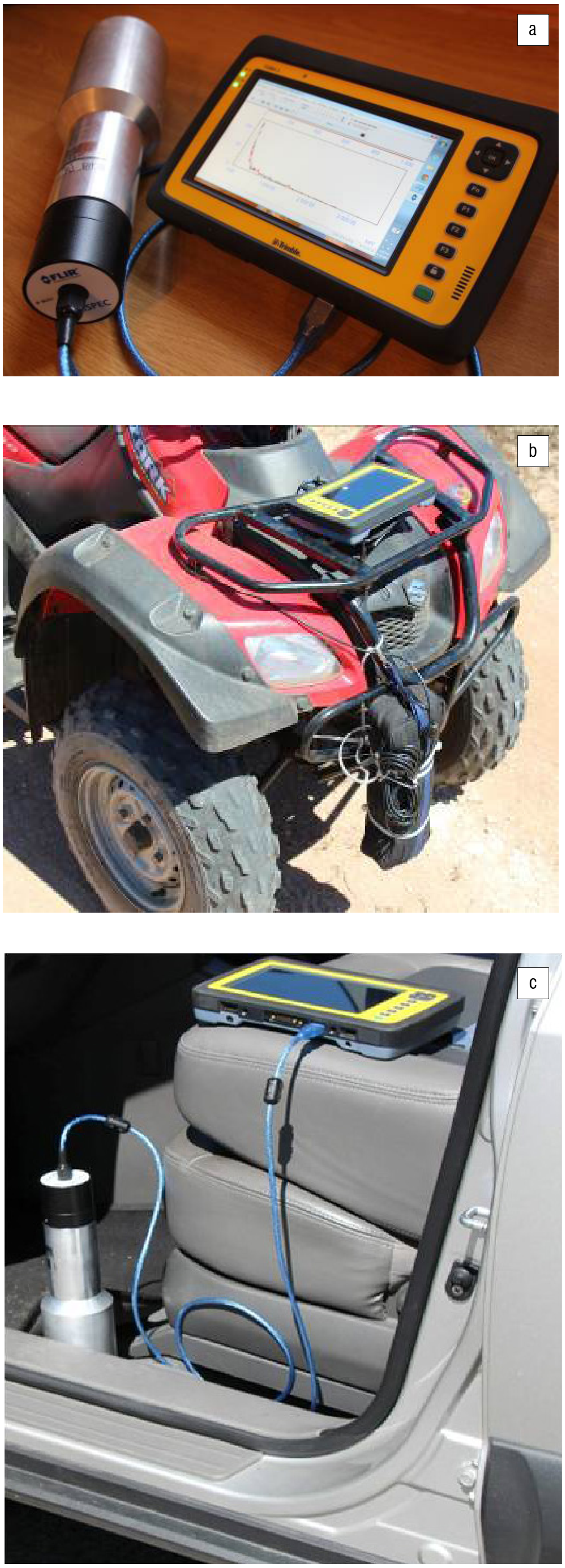


Figure 1: Photographs of the portable gamma ray detection system. (a) The system consisted of a 76.2 mm x 76.2 mm NaI(Tl) scintillation detector, a scintiSPEC multichannel analyser and a Trimble Yuma rugged tablet with an inbuilt GPS. The system mounted (b) on a quad motorcycle and (c) in a motor vehicle.

scintiSPEC® MCA was used in a previous experiment and proved to be suitable for in-situ measurements under local conditions.[8] The Trimble® Yuma (Yuma 2, Sunnyvale, CA, USA), a rugged tablet PC (http://www.trimble.com/), was chosen because of its on-board GPS and extended battery life.

The system settings and spectrum acquisition were controlled by the winTMCA32® software (which has 1024 channels) that is also produced by FLIR®. The program code was developed in the winTMCA32 software to analyse the spectra and obtain geographical locations, as well as export these data in a format compatible with the geographic information system (GIS). The winTMCA32 code directly acquired the geographic coordinates from the on-board GPS via a virtual communications port. The code also extracted the time-corrected counts from the various energy windows and combined the radiation data with the positions data. The code finally stored all hardware settings, in-situ spectra and the extracted results in files. The result files for these trials were read into and interpreted by Quantum GIS software.

The modified detector system was tested on three modes of transport. The system was initially mounted on a quad motorcycle with the detector fitted to the front metal structure of the motorcycle, 200 mm from the ground. The Yuma rugged tablet PC was mounted on the front carrier to make it easily accessible to the operator (Figure 1). The system was also carried on foot with the detector held at a height of 200 mm above the ground. Thirdly, the system was fitted inside a motor vehicle (Figure 1) and driven for several hours over distances of more than 300 km. The tablet PC was always mounted or carried in such a way that the operator had a clear view of the screen, as the system continuously provides radiation, position and timing information.

Calibrations

Energy calibrations were performed before and after each period of measurements. The calibrations were performed in the range 0.2–2.7 MeV by using anthropogenic nuclides and natural environmental spectra. The following nuclides and associated gamma ray emissions were used for the energy calibration: ^{214}Pb (351.3 keV), ^{137}Cs (661.7 keV), ^{60}Co (1173.2, 1332.5 keV), ^{40}K (1460.8 keV), ^{214}Bi (1764.5 keV) and ^{208}Tl (2614.5 keV). The symmetry assumptions and corrections described by McCay et al.[9] were adopted for the calibrations and trial surveys. All concentrations were recorded as relative values during acquisition in the test trials of the system.

Measurements and analyses

The ^{40}K nuclide emits a single gamma ray with an energy of 1460.8 keV and this decay alone was used to obtain the relative concentrations of natural potassium in an area. The nuclide ^{40}K has an abundance of 0.0117% of the weight of naturally occurring potassium.[10] The ^{40}K emission is commonly very strong in natural spectra, as a result of the abundance of potassium in nature. An active stabilisation function in the winTMCA32 software was used to correct for energy drift, which results from temperature changes while sites are surveyed. The 1460.8 keV emission of ^{40}K was chosen as centroid for stabilisation and the fine gain was automatically adjusted by the winTMCA32 software to correct for any drift from the peak.

Thorium (^{232}Th) and uranium (^{238}U) concentrations are typically determined from the decay of daughter nuclides in the respective decay chains of these elements. Various cascades of gamma rays are emitted by the daughters of uranium and thorium, consequently resulting in several convoluted peaks in the recorded gamma ray spectra. This result is especially true when NaI(Tl) detectors, which have poor resolution power, are utilised for measurements. Thus several of the gamma ray cascades of the daughters of uranium and thorium are superimposed. However, because of the low cost and high effectiveness of NaI(Tl) scintillation detectors, they continue to be the preferred choice for in-situ measurements.

Five counting windows, or regions of interest (ROIs), positioned on the six gamma ray emissions were used to extract the thorium and uranium concentrations. These ROIs were 238.6 keV (^{212}Pb), 351.9 keV (^{214}Pb), 583.2 keV (^{208}Tl), 609.3 keV (^{214}Bi), 1764.5 keV (^{214}Bi) and 2614.5 keV (^{208}Tl). The daughter nuclide from which the gamma ray emissions originates is indicated in brackets after each energy value. The proximity of the ^{208}Tl (583.2 keV) and the ^{214}Bi (609.3 keV) peaks resulted in a ROI with a combined count for ^{238}U and ^{232}Th. These combined counts in the convoluted peak, with the centroid at 596.0 keV, were also extracted. The combined counts in this ROI are useful as a result of the relationship that regularly exists between thorium and uranium concentrations in rock and soil.[11-13] All the ROIs that were used to extract the count rates of the naturally occurring nuclides are illustrated in Figure 2.

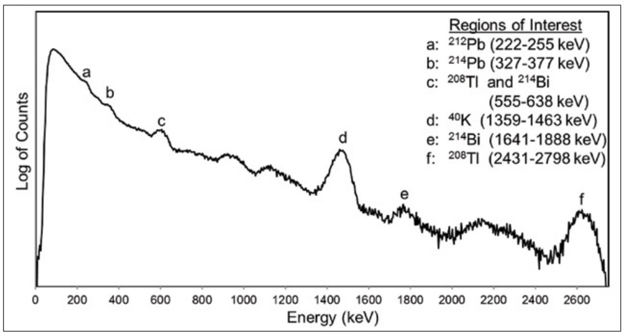


Figure 2: An in-situ gamma ray spectrum that indicates the six counting windows or regions of interest that were used to extract the relative concentrations of potassium, thorium and uranium.

The count rates in each ROI do not directly translate to the absolute concentrations for the various naturally occurring radionuclides. The counts do, however, provide a reasonable indication of the concentrations of the different radionuclides and the count rates were therefore inferred as concentrations. For the purposes of this article the concentrations will consequently be expressed as relative concentrations. These relative concentrations for potassium, uranium and thorium were subsequently classified and grouped with the help of GIS software. The potassium, uranium and thorium concentrations were then colour graded in red, blue and green, respectively. Darker shades of the colours indicate higher concentrations of the specific element. The graded colour symbols were then overlaid on Google Earth images, with the help of Quantum GIS software. These images are shown in Figures 3 to 8.

Results and discussion

The measuring system was initially mounted on a quad motorcycle and tested in a nature reserve close to the town of Saldanha on the west coast of South Africa. This reserve was selected based on the availability of radiation data from previous studies[8,14] as well as the unique geological features of the area. The nature reserve is characterised by multiple granite outcrops and soils in the reserve are also mainly granite based, as there is little input of material from elsewhere. Thus the reserve provides a test area dominated by granite, which has high levels of natural radioactivity; granite-rich areas consequently provide interesting radiation characteristics.[15,16]

The route started in a light residential zone with tar roads and continued along a dirt track into the nature reserve. The relative potassium, uranium and thorium concentrations are plotted in Figures 3, 4 and 5, respectively. The highest levels of potassium were measured in an area that was previously investigated and this area is indicated by the dashed ovals on the figures. This previous study investigated anthropogenic disturbances that took place during World War II[8] and found high potassium concentrations in this area. The higher potassium concentrations of the test trials that are demonstrated in the dashed oval in Figure 3 therefore correlated with these previous findings.

The relative uranium concentrations in Figure 4 were extracted according to a method that was proposed by Bezuidenhout[7] in which the 351.9 keV ROI of the ^{214}Pb daughter nuclide was utilised for analyses. A similar method was adapted for the extraction of thorium by employing the 238.6 keV ROI of the ^{212}Pb daughter nuclide (Figure 5). The highest levels of uranium and thorium on this route were measured on newly constructed tar roads in the small settlement. This is clearly visible in the northern section of the tracks in Figures 4 and 5.

Comparing Figure 4 with Figure 5 it is clear that the variations of the relative uranium concentrations correspond well with that of thorium. This relation between the uranium and thorium concentrations is not unusual for areas dominated by igneous geology.[11,17] The relative uranium and thorium concentrations in the area where anthropogenic disturbances took place are also notably different from that on the rest of the trial in the reserve (see dashed ovals in Figures 4 and 5). These higher uranium and thorium concentrations of the test trials also correlated with the findings of similarly elevated levels in the previous study.[8]

The system was also tested on the east coast of South Africa, close to the town of Amanzimtoti. As vehicles are not allowed on the beach, the system was carried on foot while measurements were conducted.

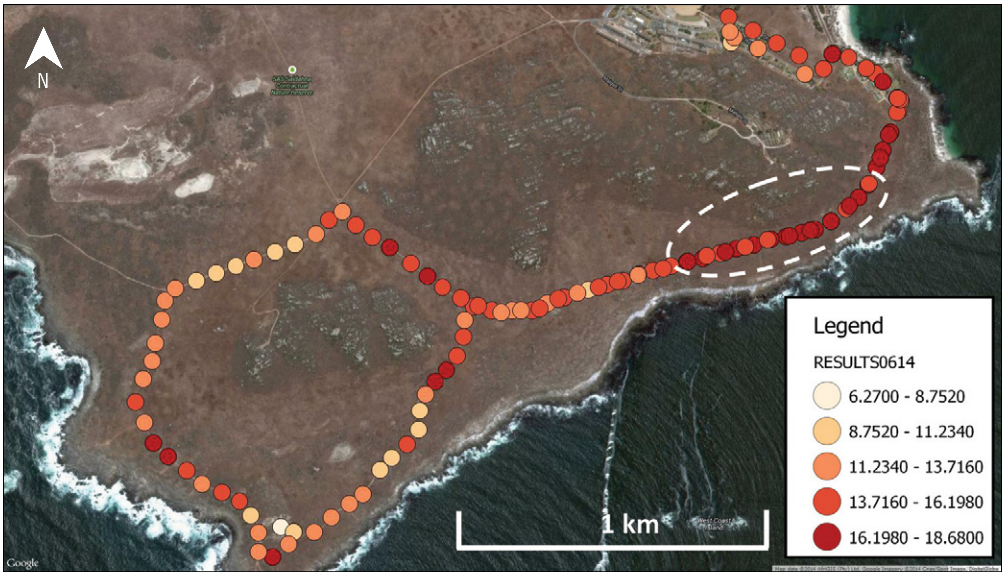


Figure 3: A Google Earth image indicating the relative potassium concentrations in the Saldanha Bay nature reserve, South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 1460.8 keV gamma ray emission.

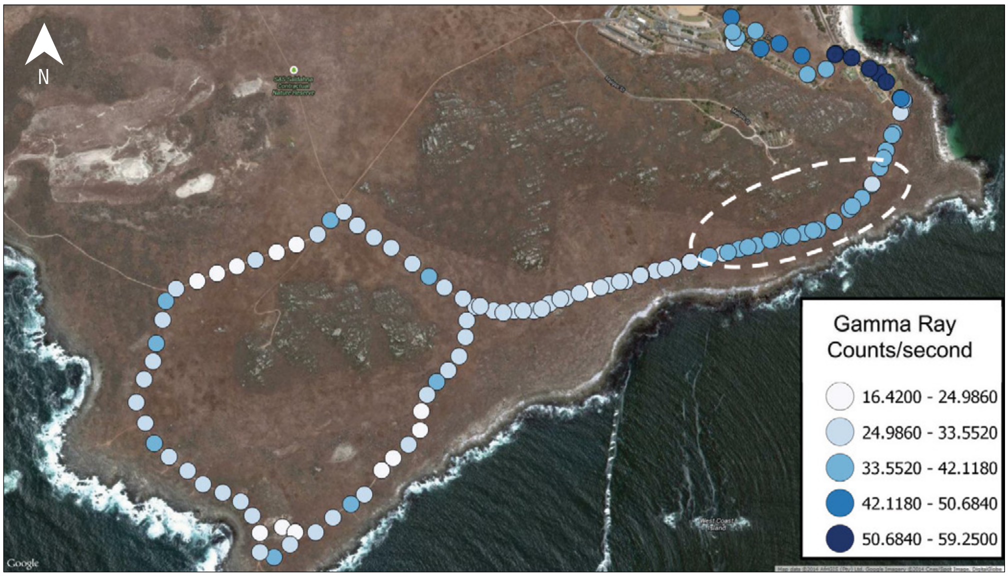


Figure 4: A Google Earth image indicating the relative uranium concentrations in the Saldanha Bay nature reserve, South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 351.9 keV gamma ray emission.

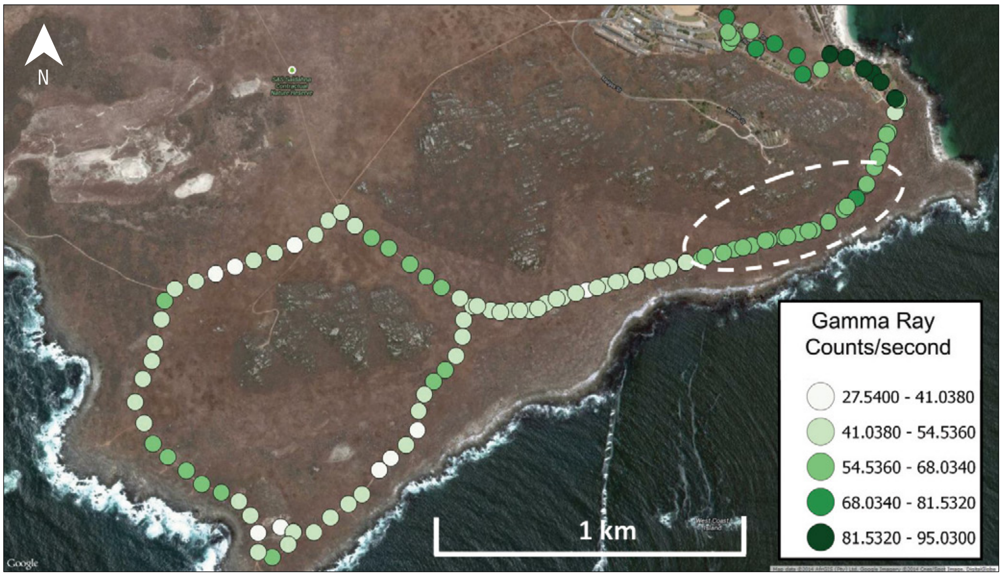


Figure 5: A Google Earth image indicating the relative thorium concentrations in the Saldanha Bay nature reserve, South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 238.6 keV gamma ray emission.

The relative potassium concentration results are presented in Figure 6. Concentrations varied noticeably along the beach; the sections where elevated potassium concentrations were measured coincided with outlets of storm water pipes. Organic materials are usually relatively rich in potassium as a consequence of the prevalence of water-soluble potassium chloride in nature. Potassium-rich deposits therefore most likely settle in the parts of the beach where storm water runs off to the ocean.

The beach was also measured in a transect perpendicular to the shoreline in order to investigate the variation of nuclide concentrations with relations to the distance from the surf line (see the line indicated by a dashed oval on the image). These measurements indicated low levels of potassium concentrations in the section of the coastline furthest from the surf line. Dunes and sand in these furthest areas are usually mainly influenced by aeolian processes, whereas sand closest to the breaker line is mainly affected by littoral processes. The different processes may have different impacts on the potassium concentrations in the sand. The relation between potassium concentrations and the sand sedimentation processes requires further investigation.

The measuring system was lastly mounted in a motor vehicle and measurements were done at higher speeds on dirt and tar roads in the west coast region of South Africa. The results for the relative uranium and thorium concentrations for one such route are plotted on Google Earth images in Figures 7, 8 and 9.

Figure 7 and Figure 9 show the relative concentrations that were extracted from the traditional ROIs that are associated with uranium and thorium, namely 1764.5 keV and 2614.5 keV, respectively. The 351.9 keV ROI associated with uranium and the 238.6 keV ROI associated with thorium were also used to determine the relative concentrations. Figure 8 shows the relative uranium concentrations that were extracted from the 351.9 keV ROI and the results correlate well with the relative uranium concentrations that were deduced from the traditional ROI (Figure 7). The relative concentrations that were extracted from the two ROIs that are associated with thorium also correlated well.

The higher count rates at the ROIs within the lower energy range of the spectra are mainly a result of detector efficiency characteristics. There are, however, gamma ray emissions from other daughters of the naturally occurring radionuclides that interfere with these ROIs. However, the good correlations between the various ROIs that were associated with uranium and thorium suggest that these interferences are negligible. The use of these non-traditional ROIs for analyses is mainly motivated by the substantially higher count rates compared with the count rates of the traditional ROIs. This finding supports the use of these non-traditional ROIs for in-situ analyses and measurements, which is of particular importance if dynamic surveys are conducted at higher speeds, resulting in shorter detector exposure times.

The highest levels of uranium and thorium were measured on recently constructed tar roads that carry high volumes of traffic. This observation is evident on the coloured overlay for straight roads on the western edges of the images in Figure 8 and Figure 9. Large but similar variations in uranium and thorium concentrations are also visible in the images. The variations of uranium concentrations correspond to that of thorium, for reasons similar to those discussed earlier in this section.

The lowest levels of uranium and thorium were measured on an old tar road and this road is indicated by dashed ovals on Figures 7, 8 and 9. It might be that the older roads were not constructed using the same techniques as those employed for the newer roads. For example, if less gravel was used in the construction, there would be a lower radiation signature. A further study is planned to investigate the correlation between the strength of tar roads and their radiation signatures.

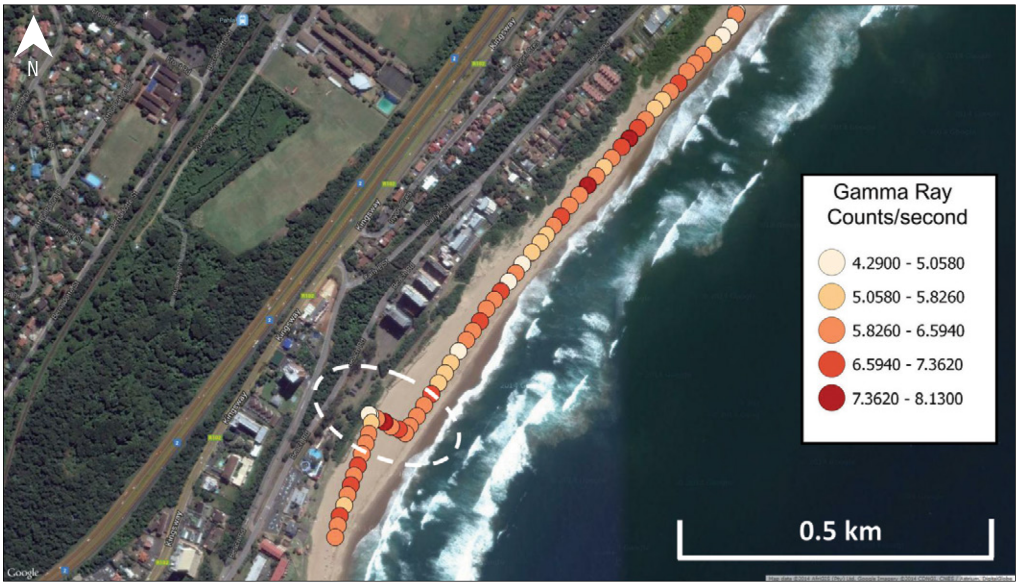


Figure 6: A Google Earth image indicating the relative potassium concentrations on the north beach of Amanzimtoti, South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 1460.8 keV gamma ray emission.

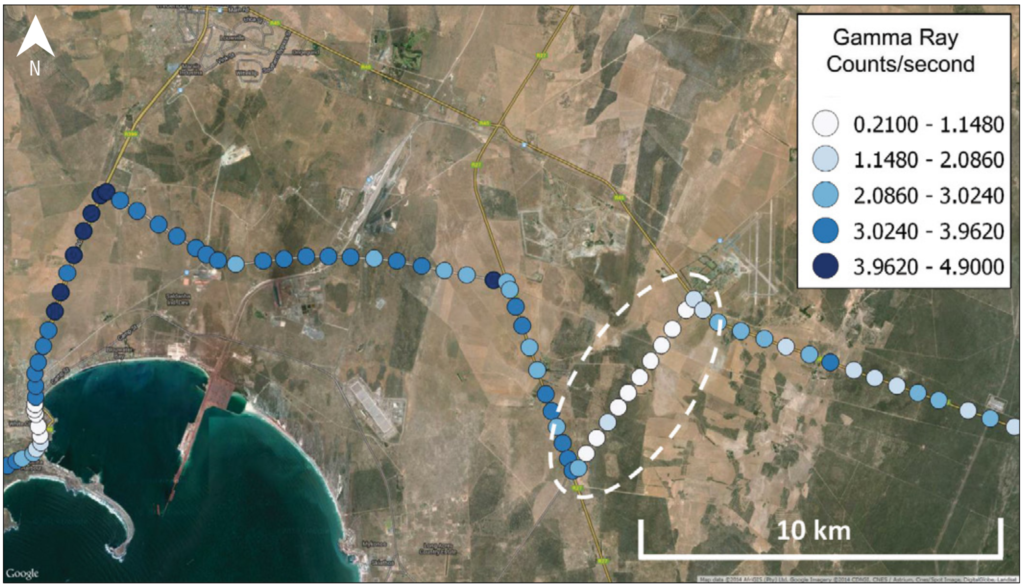


Figure 7: A Google Earth image indicating the relative uranium concentrations on tarred roads of the West Coast of South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 1764.5 keV gamma ray emission.

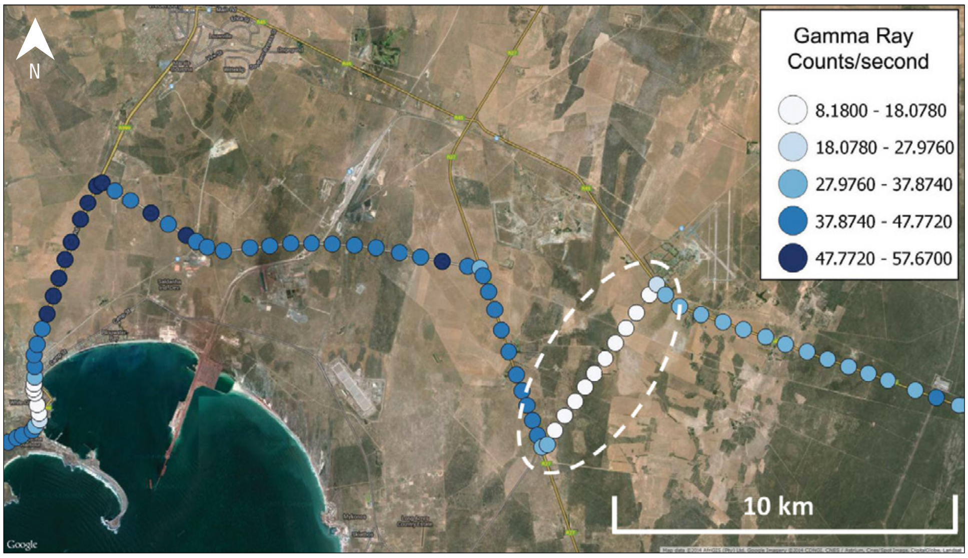


Figure 8: A Google Earth image indicating the relative uranium concentrations on tarred roads of the West Coast of South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 351.9 keV gamma ray emission.

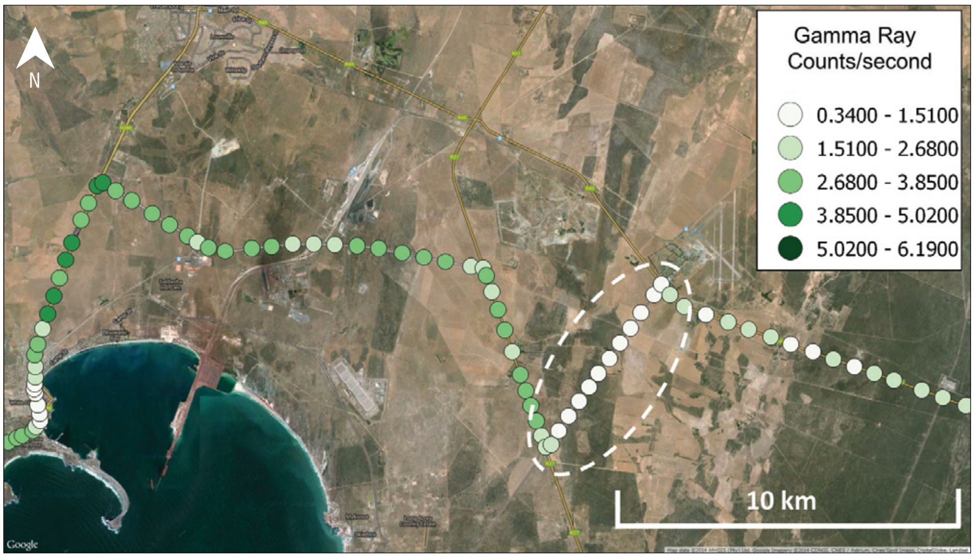


Figure 9: A Google Earth image indicating the relative potassium concentrations on tarred roads of the West Coast of South Africa. The colour graded overlay displays the gamma ray counts in the region of interest around the 2614.5 keV gamma ray emission.

The measurements that we conducted at higher speeds from the motor vehicle gave good and consistent results, supporting the use of the system in such a manner and the use of the alternative ROIs for uranium and thorium concentration analyses.

Conclusions

The necessity arose for a robust and highly transportable gamma ray field survey system that could be used in geographically remote areas in southern Africa. Various new technical advances provided the opportunity for the development of such a unique field measuring system. The required specifications for the measuring system were achieved by integrating a NaI(Tl) scintillation detector with a rugged Trimble tablet PC and an on-board GPS. The system is managed in real time by analyses software. The system acquires gamma ray spectra, extracts radionuclide concentrations and finally interpolates data to provide radionuclide concentrations and produce maps while on location. The system was tested in different geographical settings and by different means of transport which included a motor vehicle, a quad motorcycle and by carrying the system on foot. Novel analyses methods were also developed in order to extract nuclide concentrations from the in-situ acquired spectra. The results from different geographical areas displayed variation in radiation well and also reproduced results from previous studies. There was also consistency across results that were obtained using the different analyses methods. These findings support the viability of various means of in-situ transport as well as the different in-situ analyses methods.

It is planned that this system, utilising multiple platforms of transport, will be used in several future studies in southern Africa. One potential study is to further map nuclide concentrations in coastal sediment[18] along the South African and Angolan coastlines in order to develop a model for coastal sediment transportation. Additionally, another investigation will look at the relationship between radiation signatures of tar roads and the road quality. The development of this transportable radiation measuring system makes these and other studies both cost and time effective.

Acknowledgements

I thank the National Research Fund of South Africa and Stellenbosch University for providing the funding for various aspects of this research.

References

1. Tyler AN. High accuracy in situ radiometric mapping. J Environ Radioact. 2004;72:195–202. http://dx.doi.org/10.1016/S0265-931X(03)00202-9
2. Nilsson JMC, Östlund K, Söderberg J, Mattsson S, Rääf C. Tests of HPGe- and scintillation-based backpack γ-radiation survey systems. J Environ Radioact. 2014;135:54–62. http://dx.doi.org/10.1016/j.jenvrad.2014.03.013
3. Déjeant A, Bourva L, Sia R, Galoisy L, Calas G, Phrommavanh V, Descostes M. Field analyses of 238U and 226Ra in two uranium mill tailings piles from Niger using portable HPGe detector. J Environ Radioact. 2014;137105–112. http://dx.doi.org/10.1016/j.jenvrad.2014.06.012
4. Van der Graaf ER, Koomans RL, Limburg J, De Vries K. In situ radiometric mapping as a proxy of sediment contamination: Assessment of the underlying geochemical and -physical principles. App Rad Iso. 2007;65(5):619–633. http://dx.doi.org/10.1016/j.apradiso.2006.11.004
5. Hendriks PHGM, Limburg J, De Meijer RJ. Full-spectrum analysis of natural γ-ray spectra. J Environ Radioact. 2001;53:365–380. http://dx.doi.org/10.1016/S0265-931X(00)00142-9
6. Chiozzi P, De Felice P, Fazio A, Pasquale V, Verdoy M. Laboratory application of NaI(Tl) γ-ray spectrometry to studies of natural radioactivity in geophysics. App Rad Iso. 2000;53:127–132. http://dx.doi.org/10.1016/S0969-8043(00)00123-8
7. Bezuidenhout J. Measuring naturally occurring uranium in soil and minerals by analysing the 352 keV gamma-ray peak of 214Pb using a NaI(Tl)-detector. App Rad Iso. 2013;80:1–6. http://dx.doi.org/10.1016/j.apradiso.2013.05.008
8. Bezuidenhout J. Mapping of historical human activities in the Saldanha Bay military area by using in situ gamma ray measurements. Scientia Militaria. 2012;40(2):89–101.
9. McCay T, Harley TL, Younger PL, Sanderson DCW, Cresswell AJ. Gamma-ray spectrometry in geothermal exploration: State of the art techniques. Energies. 2014;7(8):4757–4780. http://dx.doi.org/10.3390/en7084757
10. National Institute of Standards and Technology, Physical Measurement Laboratory. Atomic weights and isotopic compositions with relative atomic masses [database on the Internet]. No date [cited 2015 Aug 24]. Available from: http://www.nist.gov/pml/data/comp.cfm
11. Montes ML, Mercader RC, Taylor MA, Runco J, Desimoni J. Assessment of natural radioactivity levels and their relationship with soil characteristics in undisturbed soils of the northeast of Buenos Aires province, Argentina. J Environ Radioact. 2012;105:30–39. http://dx.doi.org/10.1016/j.jenvrad.2011.09.014
12. Abbady AGE, El-Arabi AM, Abbady A. Heat production rate from radioactive elements in igneous and metamorphic rocks in Eastern Desert, Egypt. App Rad Iso. 2006;64:131–137. http://dx.doi.org/10.1016/j.apradiso.2005.05.054
13. Abbady AGE. Evaluation of heat generation by radioactive decay of sedimentary rocks in Eastern Desert and Nile Valley, Egypt. App Rad Iso. 2010;68:2020–2024. http://dx.doi.org/10.1016/j.apradiso.2010.03.023
14. Bezuidenhout J. Using GIS to estimate background radiation levels. PositionIT Magazine. 2013 April/May;65–73.
15. Llope WJ. Activity concentrations and dose rates from decorative granite countertops. J Environ Radioact. 2011;102:620–629. http://dx.doi.org/10.1016/j.jenvrad.2011.03.012
16. Baranwal VC, Sharma SP, Sengupta D, Sandilya MK, Bhaumik BK, Guin R, et al. A new high background radiation area in the geothermal region of Eastern Ghats Mobile Belt (EGMB) of Orissa, India. Rad Meas. 2006;41:602–610. http://dx.doi.org/10.1016/j.radmeas.2006.03.002
17. Abbady AGE, El-Arabi AM, Abbady A. Heat production rate from radioactive elements in igneous and metamorphic rocks in Eastern Desert, Egypt. App Rad Iso. 2006;64:131–137. http://dx.doi.org/10.1016/j.apradiso.2005.05.054
18. Thereska J. Natural radioactivity of coastal sediments as tracer in dynamic sedimentology. Nukleonika. 2009;54(1):45–50.

An LC-MS/MS based survey of contaminants of emerging concern in drinking water in South Africa

AUTHORS:
Christiaan Odendaal[1]
Maitland T. Seaman[1]
Gabre Kemp[2]
Huibreght E. Patterton[2]
Hugh-George Patterton[2]

AFFILIATIONS:
[1]Centre for Environmental Management, University of the Free State, Bloemfontein, South Africa
[2]Department of Microbial, Biochemical and Food Biotechnology, University of the Free State, Bloemfontein, South Africa

CORRESPONDENCE TO:
Hugh-George Patterton

EMAIL:
hpatterton@sun.ac.za

POSTAL ADDRESS:
Institute for Wine Biotechnology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

DATES:
Received: 14 Nov. 2014
Revised: 15 Feb. 2015
Accepted: 15 Feb. 2015

KEYWORDS:
pharmaceuticals; personal care products; mass spectrometry; multiple reaction monitoring; ToxNet

HOW TO CITE:
Odendaal C, Seaman MT, Kemp G, Patterton HE, Patterton HG. An LC-MS/MS based survey of contaminants of emerging concern in drinking water in South Africa. S Afr J Sci. 2015;111(9/10), Art. #2014-0401, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20140401

Advances in many analytical techniques allow the detection of compounds in water at very low concentrations (ng/L), which has facilitated the identification of many compounds in drinking water that went previously undetected. Some of these compounds are contaminants of emerging concern (CECs), which is broadly defined as any chemical or microorganism that is not currently being routinely monitored but has recently been identified as being present in the environment, and that may pose health or ecological risks. CECs can include pharmaceuticals, personal health care products and pesticides. Some CECs can act as endocrine disruptors, interfering with the normal functioning of the human endocrine system, potentially influencing foetal and child development. Although the level of many of these compounds are orders of magnitude below known acute toxicity levels, the health impact of long term exposure at low levels is mostly unknown. In this study, we present the results of a national survey over four seasons of potential CECs in the drinking water of major South African cities. The contaminants most often detected were the related herbicides atrazine and terbuthylazine, and the anticonvulsant and mood-stabilising drug, carbamazepine. The levels of these CECs were well below maximum levels proposed by the World Health Organization and the US Environmental Protection Agency. However, the range of CECs detected in drinking water, and seasonal and geographic variability in CECs levels, warrant a more frequent screening programme.

Introduction

Advances in analytic technologies allow the identification of chemical compounds at exceedingly low concentrations (10^{-9} g/L) in drinking water.[1] This permits the identification of compounds which, until recently, were undetectable in water. These compounds fall into broad categories, including pesticides, pharmaceuticals and personal care products. Because we are only now becoming aware of the presence of these chemicals in drinking water, most of these compounds are not included in routine monitoring programmes. Although these compounds are generally present at concentrations several orders of magnitude below established acute toxicity levels, the effect of long-term exposure to very low concentrations of these compounds on human health and development is not known. This is particularly relevant to pharmaceutical contaminants, which are designed to be physiologically active at very low concentrations. Furthermore, some of these compounds interfere with the human endocrine system (endocrine disruptors), which may result in severe developmental defects with exposure of foetuses or infants during critical developmental windows. There is therefore a pressing need to investigate the potential health impacts of these compounds in drinking water, collectively known as contaminants of emerging concern (CECs).[2,3]

The US Geological Survey undertook several national reconnaissance studies, including a 1999/2000 programme, in which samples were analysed from 139 streams across 30 states in the USA.[4] A wide range of chemicals present in residential, industrial, and agricultural wastewaters was found to occur at low concentrations in streams in the United States. The chemicals detected included human and veterinary drugs, natural and synthetic hormones, detergent metabolites, plasticisers, insecticides and fire retardants. One or more of these chemicals was found in 80% of the streams sampled. In a national groundwater study by the US Geological Survey, samples from 47 wells in 18 states were analysed for 65 chemicals.[5] A profile of chemical pollutants similar to that observed in streams was found, although the contaminants were generally present at much lower levels. In another US Geological Survey study of untreated drinking water from 25 groundwater and 49 surface water sites in 25 states, pharmaceuticals, plasticisers and fire retardants were detected.[6] Taken together, these studies provided valuable baseline information on the presence of CECs and other compounds in the US water system, and provides a valuable frame for further toxicity and public health impact studies.

The list of CECs is extensive, and includes sucralose, antimony, siloxanes, musks, nanomaterials, perfluorooctanoic acid, perfluorooctane sulphonate and other perfluorinated compounds, pharmaceuticals, hormones and hormone-active compounds, collectively known as endocrine disrupting compounds, drinking water disinfection by-products, sunscreens/UV filters, brominated flame retardants, benzotriazoles, naphthenic acids, cyanobacterial toxins, perchlorate, dioxane, pesticides and pesticide degradation products, and microorganisms, including viruses.[7]

Generally, organisations involved in water health and safety monitor CECs based on available technologies, known occurrence and health impacts.[8,9] A technique that is currently widely used to monitor CECs is high performance liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS).[7]

Here we report the first national survey of CECs in the drinking water of major South African cities. The survey includes a qualitative screen for approximately 700 compounds, as well as the quantitation of three critical compounds identified in the qualitative screen, atrazine, terbuthylazine and carbamazepine. Atrazine is a herbicide used for the control of broadleaf weeds in the maize, sorghum and sugar cane agricultural industries. Epidemiological studies showed a correspondence between elevated atrazine levels in drinking water and low sperm volume and motility[10], foetal growth defects, including restriction[11], small-for-gestational-age[12] and intrauterine growth retardation[13], foetal gastroschisis[14] and increases in limb reductions (upper and lower), hypospadias and epispadias, cryptorchidism, and spina bifida[15]. Terbuthylazine is a general, broad-spectrum pre- or post-emergence herbicide used in agriculture. Terbuthylazine was shown to cause an increase in DNA damage in cultured mammalian cells at concentrations equivalent to the occupational exposure limits.[16] Carbamazepine is a therapeutic used as an anticonvulsant and a mood-stabilising drug. While it was reported that epilepsy patients who receive carbamazepine therapy during pregnancy delivered babies with an increased rate of congenital anomalies such as neural tube defects, and cardiovascular and urinary tract anomalies[17], no epidemiological studies on the presence of carbamazepine at low concentrations in drinking water have been published to date.

Materials and methods

Reagents and materials

High purity (>98%) chemical standards for atrazine and carbamazepine were purchased from Sigma Aldrich (St. Louis, MO, USA), while terbuthylazine and deuterated-atrazine were purchased from Dr Ehrenstorfer (Augsburg, Germany). Stock solutions for each standard were prepared in methanol (1 µg/L). High-performance liquid chromatography (HPLC) grade methanol (MeOH), acetonitrile (ACN), formic acid and ammonium hydroxide were purchased from Sigma Aldrich.

Ultra-pure water (18 mΩ) was prepared with a Milli-Q purification device (Millipore, Billerica, MA, USA) and used in all experiments.

Method development and validation

The quantitative method was developed according to the Food and Drug Agency guidelines for method validation.[18]

Sampling

Samples (1 L) were collected in amber glass bottles from water treatment plants (WTPs) in Cape Town, Port Elizabeth, Durban, Pietermaritzburg, Johannesburg, Pretoria and Bloemfontein during months in each of the four seasons (February, April, July and October 2012), as well as from residential taps in Bloemfontein south and Bloemfontein north, supplied by two different reservoirs. Confidentiality agreements were entered into with the WTPs to not disclose the identity of the individual plants. Samples were collected and stored at 4 °C until analysis, usually within 24–48 h.

Solid phase extraction

Sample preparation involved compound extraction and reconstitution in 1 mL of H_2O / 0.1% formic acid. Solid phase extraction is still the preferred approach of extraction, because it produces higher yields than liquid/liquid extraction, can be automated and significantly reduces preparation time.[7,19] Milli-Q water fortified with CEC standards was used to optimise solid phase extraction parameters. Different solid phase extraction cartridges with varying sorbent characteristics were analysed to identify the cartridge with the best recovery.

Before extraction, cartridges were equilibrated with 6 mL pure MeOH. After equilibration, samples were loaded at a flow rate of approximately 6 mL/min. After samples were loaded, cartridges were washed with 6 mL of ultrapure water. Extracts were eluted into 6 mL tubes using 2 mL of MeOH and 2 mL of acetonitrile. Eluates were evaporated using a Savant SC 210A Speedvac concentrator with a Thermo RVT 4104 refrigerated vapour trap. Extracts were reconstituted in 1 mL of H_2O / 0.1% formic acid and suspended using a vortex (Velp Scientifica, Italy) as well as by sonication (Branson, USA).

LC-MS/MS analysis

The analysis was performed on an HPLC (Agilent 1200) linked to a 3200 QTRAP hybrid triple quadrupole mass spectrometer (AB SciEx, Framingham, MA, USA). The HPLC was fitted with a 3-µm Gemini-NX-C18 110-Å (150 x 2 mm) column (Phenomenex, CA, Torrance, USA). Formic acid (0.1% v/v) in water (solvent A) and formic acid (0.1% v/v) in MeOH (solvent B) were used as elution solvents for positively charged analytes. Negatively charged analytes were separated in NH_3OH (0.1% v/v) in water (solvent A) and NH_3OH (0.1% v/v) in MeOH (solvent B).

Analytes were detected and quantified using multiple reaction monitoring using precursor and two fragment transitions for each of the analytes.[20,21] The *m/z* values used are shown in Table 1. Multiple reaction monitoring provides increased selectivity and reduces the likelihood of spectral interferences.

Results and discussion

Initial screening

We performed an initial LC-MS/MS analysis of drinking water from Bloemfontein and Johannesburg to obtain an insight into the range of CECs present in drinking water in South African cities. We made use of a MS/MS fragmentation library of approximately 700 compounds (see Supplementary table 1 online). The result of this initial screen is shown in Table 2.

A review of the frequency of occurrence, coupled with toxicity data and community health impact from epidemiological studies,[22] where available, suggested that atrazine, terbuthylazine and carbamazepine posed the highest public health risk to the South African water consumer. For this reason it was decided, apart from the general screening of drinking water for CECs, to also quantitate atrazine, terbuthylazine and carbamazepine in all collected water samples. In the absence of an established method, we needed to develop a robust protocol for the quantitation of these three CECs by LC-MS/MS. Method selectivity, accuracy and precision, as well as analyte recovery and stability are generally essential parameters to consider in method development and validation.[18]

Table 1: Precursor and fragment *m/z* values

	Precursor *m/z*	Fragment 1 *m/z*	Fragment 2 *m/z*
Atrazine	216.0	174.1	104.0
Terbuthylazine	230.0	174.1	104.0
Carbamazepine	237.1	194.2	192.1

Table 2: Preliminary screening of contaminants of emerging concern in drinking water

Analyte	Bloemfontein						Johannesburg		Occurrence (%)
	Jan 2010	Oct 2010	Jan 2011	May 2011	Jul 2011	Jul 2011	Dec 2010	Jun 2011	
Amphetamine	•	•		•	•			•	63
Atrazine		•	•			•	•	•	63
Carbamazepine	•	•	•		•		•		63
Diphenylamine								•	13
Imidacloprid								•	13
Metolachlor			•	•	•	•	•	•	75
Oxadixyl		•							13
Simazine								•	13
Tebuthiuron				•	•	•	•	•	63
Telmisartan		•							13
Terbuthylazine	•	•	•	•	•	•	•	•	100

A solid circle indicates that the stipulated compound was identified in the water sample.

Method validation

Calibration curve

A calibration curve was determined by measuring the MS ion count over a concentration range of 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, 1×10^{-3}, 5×10^{-3}, 1×10^{-2}, 5×10^{-2}, and 1×10^{-1} µg/L for each of atrazine, terbuthylazine and carbamazepine. The representative calibration curve of atrazine is shown in Figure 1. Comparable results were obtained for terbuthylazine and carbamazepine (data not shown).

The limit of detection, lower limit of quantification and upper limit of quantification were determined for each of the three CECs using the MS spectra in the concentration range $5\times10^{-5} - 1\times10^{-1}$ μg/L. The limit of detection and lower limit of quantification were determined at signal-to-noise ratios of 3 and 10, respectively.[23-25]

The upper limits of quantification were defined as the highest concentration of analyte detectable with reasonable precision and accuracy.[18,24,26] The lower limit of quantification, upper limit of quantification, recoveries, coefficient of variance and maximum contaminant levels are shown in Table 3. An internal standard, deuterated atrazine, was added at 1×10^{-1} μg/L before solid phase extraction. The same concentration of internal standard was injected into each of the vials ($5\times10^{-5} - 1\times10^{-1}$ μg/L), and was used during quantification of atrazine and terbuthylazine.

Selectivity and crosstalk

The selectivity of a method can be verified by establishing the absence of analyte peaks in a blank sample at the determined elution time for that analyte.[18] The absence of crosstalk is shown by detecting comparable concentration for an analyte in a sample containing the single analyte compared to a sample containing a mixture of different, possibly interfering, analytes. To establish the selectivity and absence of crosstalk in our quantitation protocol, three vials were filled with 50 ng/L atrazine, terbuthylazine or carbamazepine, and a fourth vial was filled with a mixture that contained 50 ng/L of each of atrazine, terbuthylazine and carbamazepine. It was particularly important to demonstrate the absence of crosstalk for atrazine and terbuthylazine, because the *m/z* values of the two major fragments were identical (Table 1). The single analytes showed no significant difference compared to that of the mixture of three analytes in three independent repetitions of the experiment (paired *t*-test, confidence interval = 99%). Similarly, no analyte could be detected in sample blanks. The results are shown in Figure 2. Comparisons of mean analyte peak areas of a single analyte and in a mixture revealed no significant difference.

Accuracy and precision

The precision and accuracy of the quantitation protocol was demonstrated by determining the concentration of each of the three analytes in standard samples of 5×10^{-2} μg/L, a concentration in the intermediate range between the lower limit of quantification and upper limit of quantification. In all cases, the coefficient of variance was less than 15% and the bias less than 20% (see Table 3), within the prescribed limits.[18]

Presence and seasonal variation of CECs in drinking water

Drinking water samples were taken at seven WTPs in major cities in South Africa at a point before the water entered the reticulation system. The samples were extracted on a solid phase cartridge, eluted, and analysed by LC-MS/MS. The precursor *m/z* as well as the *m/z* values of two major fragments were compared to a library of compounds (see Supplementary table 1 online). Compounds were identified where the precursor and well as both fragment *m/z* values could be matched to a library entry. The combined results of the screening of the seven drinking water samples are shown in Table 4. Atrazine, terbuthylzine and carbamazepine were detected in more than 60% of the drinking water samples. The seasonal distribution of atrazine fitted with its agricultural use as herbicide for summer crops. Carbamazepine, an anticonvulsant that is also prescribed for treatment of bipolar disorder, was present at a steady level in more than 70% of the samples. Cinchonidine, which is used in the chemical synthesis industries, was detected in almost 90% of the samples. Diphenylamine, which was present in about 40% of the samples, has wide application, including as an anti-scalding agent for fruit. The antifungal fluconazole and herbicides hexazinone and metolachlor were present in approximately 16% of the samples, with the latter present exclusively in the summer, most likely as a result of its agricultural application. Phenytoin, an anticonvulsant drug prescribed under the trademark ‘Epanutin’ in South Africa, was present in drinking water throughout the year. The antibacterial agent, sulphisomidine, was present in 18% of the samples. The herbicides, terbuthiuron and terbuthylazine, were consistently present in drinking water throughout the year. Interestingly, ephedrine, used as a decongestant

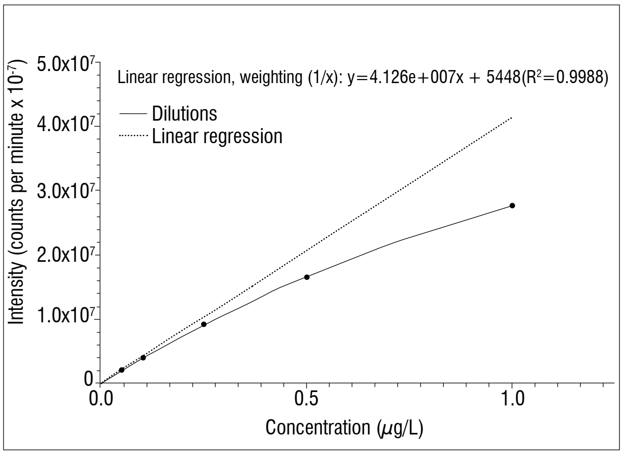


Figure 1: Calibration curve for atrazine.

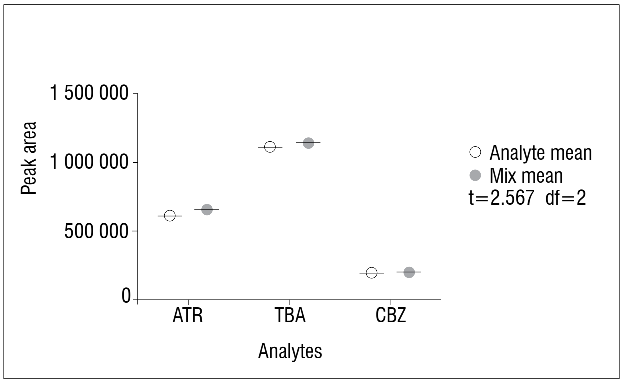


Figure 2: Comparisons of mean analyte peak areas of a single analyte and in a mixture revealed no significant differences.

Table 3: Measures of optimised measurement method

Analyte	Linearity (R^2-value)	Lower limit of quantification (μg/L)	Upper limit of quantification (μg/L)	Recovery†	Precision† (% coefficient of variance)	Accuracy† (% bias)
Atrazine	0.99880	0.00010	0.10000	103%	2%	3%
Terbuthylazine	0.99860	0.00005	0.10000	103%	3%	3%
Carbamazepine	0.99000	0.00005	0.10000	120%	1%	20%

†Measured at 5×10^{-2} μg/L

and bronchodilator, was observed only in the winter, consistent with its expected increased medical use. Enilconazole, an antifungal agent widely used in the growing of citrus fruit, was observed only in autumn. Interestingly, we never detected any cyanobacterial microcystins, but had no information on the occurrence of upstream algal blooms.

Having established the frequency of occurrence of a range of pesticides and therapeutic compounds in metropolitan drinking water, it was decided to quantitate the levels of atrazine, terbuthylazine and carbamazepine, as these three compounds were present at very high frequency and were also associated with significant public health risks.

Quantitation of three critical CECs in drinking water

The drinking water samples, treated as before, were separated by reverse phase HPLC and quantitated by multiple reaction monitoring on a hybrid triple quadrupole mass spectrometer using the developed method described above. This procedure involved the integration of the ion count during elution of a compound from the HPLC column, with concomitant confirmation of the identity of the compound by the presence of peaks at the correct precursor and major transition fragment *m/z* values. The peak area was used to deduce the concentration from the standard curve of each of the three compounds of interest. The concentrations are tabulated in Supplementary table 2 online.

The guideline value proposed by the World Health Organization (WHO) for atrazine is 100 mg/L[27], whilst the maximum contaminant level stipulated by the US Environmental Protection Agency (EPA) is 3 mg/L[8]. Figure 3 indicates that the highest level of atrazine recorded during the one year survey was more than an order of magnitude below the maximum contaminant level set by the EPA. The level of atrazine was

Table 4: Seasonal screening and analyte occurrence (%) at all sampling sites: Cape Town, Port Elizabeth, Durban, Pietermaritzburg, Johannesburg, Pretoria and Bloemfontein

Analytes	Summer (%)	Autumn (%)	Winter (%)	Spring (%)	Average annual occurrence (%)
2-deoxyguanosine	0%	0%	14%	0%	4%
Atrazine†	86%	71%	29%	57%	61%
Benzocaine	0%	0%	0%	14%	4%
Carbamazepine†	71%	71%	57%	86%	71%
Cinchonidine	86%	86%	100%	71%	86%
Cinchonine	0%	0%	0%	14%	4%
Diphenylamine	14%	43%	0%	100%	39%
Enilconazole	0%	14%	0%	0%	4%
Ephedrin	0%	14%	14%	0%	7%
Flecainide	0%	14%	0%	0%	4%
Fluconazole	14%	29%	14%	14%	18%
Hexazinone	14%	14%	14%	14%	14%
Imidacloprid	0%	0%	0%	14%	4%
Metazachlor	0%	14%	0%	0%	4%
Metolachlor	71%	0%	0%	0%	18%
Minoxidil	0%	14%	0%	0%	4%
Nalidixicacid	0%	0%	14%	0%	4%
Paracetamol	0%	14%	0%	0%	4%
Phenytoin	29%	57%	29%	43%	39%
Sebuthylazine-desethyl	14%	0%	0%	0%	4%
Simazine	0%	14%	0%	0%	4%
Sulphisomidine	29%	29%	0%	14%	18%
Tebuthiuron	71%	57%	57%	43%	57%
Telmisartan	14%	71%	0%	29%	29%
Temazepam	0%	14%	0%	0%	4%
Terbumeton	0%	14%	0%	0%	4%
Terbuthylazine†	86%	86%	86%	100%	89%
Thiabendazole	0%	14%	0%	0%	4%

†*Contaminants of emerging concern that were quantitated in this study.*

consistently high throughout the year in Johannesburg, compared to the average value recorded for all the samples. Interestingly, high atrazine values were also recorded in tap water in Bloemfontein in the autumn and spring, even though low levels were recorded at the WTP at the same times. This suggested that the concentration of atrazine may vary very sharply, and that a much higher sampling frequency is required to accurately determine its variation over time.

The guideline value proposed for terbuthylazine by the WHO is 7 mg/L.[27] The EPA has no set maximum contaminant level for terbuthylazine.[8] Referring to Figure 3, it is seen that the highest recorded concentration for terbuthylazine in drinking water (Pretoria, autumn) is at least an order of magnitude less that the WHO guideline value. Johannesburg, again, showed a consistently high level of terbuthylazine throughout the year, compared to the other WTPs.

The maximum contaminant level for the pharmaceutical carbamazepine was set at 12 mg/L.[28] The highest level of carbamazepine detected in drinking water (see Figure 3) was significantly less than this level. Interestingly, the level of this anti-epileptic and mood-stabilising drug was consistently high throughout the year in Bloemfontein, compared to the average national level. Particularly high levels were recorded in

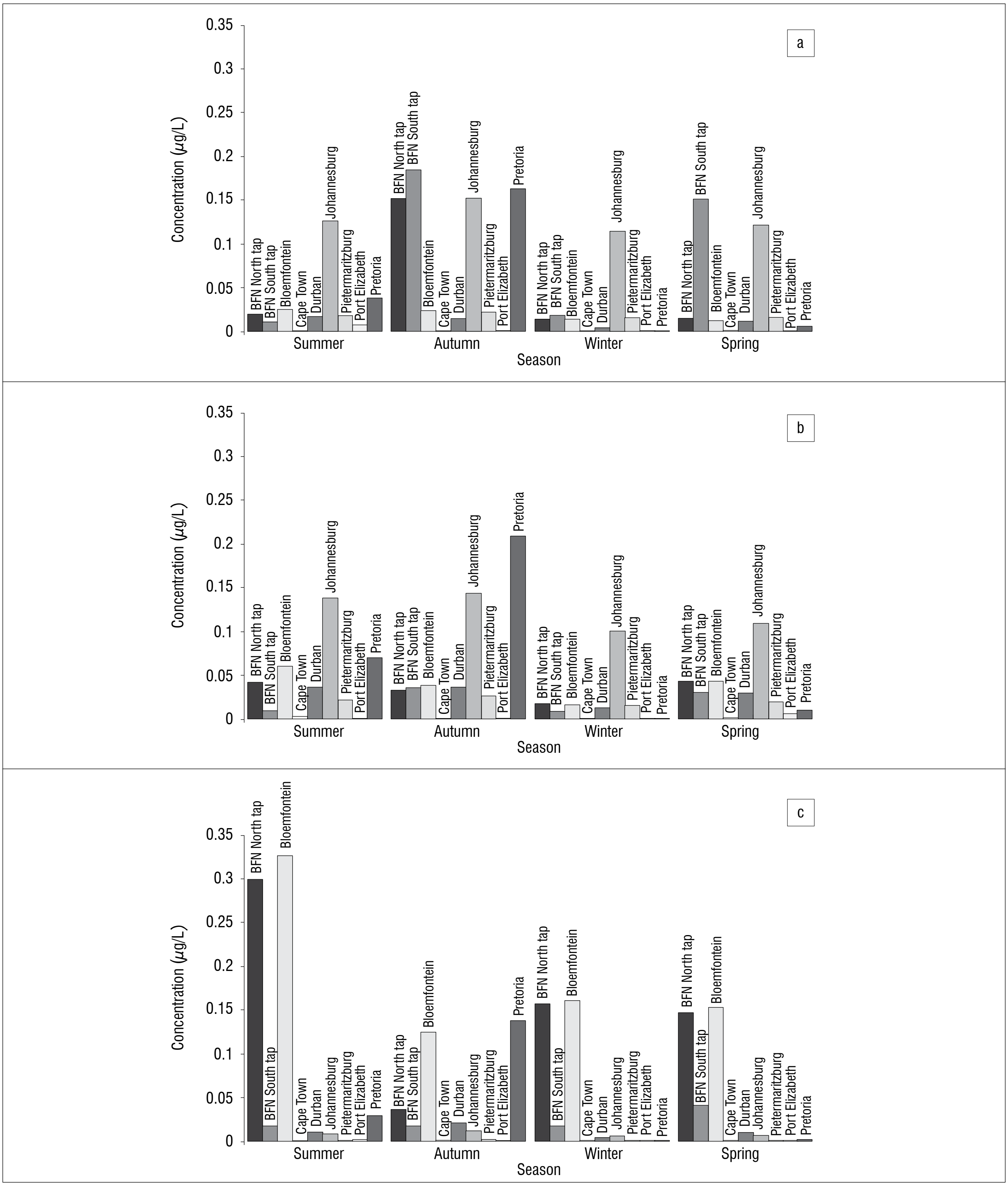


Figure 3: Concentration of (a) atrazine, (b) terbuthylazine and (c) carbamazepine – three major contaminants of emerging concern – in the drinking water of major South African cities.

the summer (Figure 3). We again observed a discordance between the carbamazepine concentrations recorded at the WTP and in tap water in Bloemfontein in the autumn. This result also suggests significant concentration spikes, indicating a need for a high sampling frequently to obtain a reliable insight into the level of this CEC in drinking water.

Conclusion

During this analysis, a method was developed to determine atrazine, terbuthylazine and carbamazepine quantities in drinking water. A qualitative analysis identified 29 potential CECs (Table 4). Importantly, the critical CECs identified during preliminary analyses were also part of the subsequent qualitative list of CECs. Quantification of atrazine, terbuthylazine and carbamazepine revealed no immediate health risks, since all concentrations were below the published thresholds.

Although the concentration levels were below published maximum contaminant level thresholds, the range of CECs routinely detected in drinking water, and the large geographical and seasonal variability that we observed, suggest that a qualitative survey and quantitation of select CECs should be performed more frequently to have a current view of the presence of levels of CECs in drinking water that may impact on human health. Also, with an increase in the pressure on water health as this resource in increasingly being utilised, the introduction of such a CEC monitoring programme becomes essential to ensure the production of healthy and safe drinking water for the consumer.

Acknowledgements

We thank the Water Management in Water-scarce Areas and the Advanced Biomolecular Research clusters at the University of the Free State for their contributions to this project. This study was supported by a grant from the Water Research Commission (to H.G.P.).

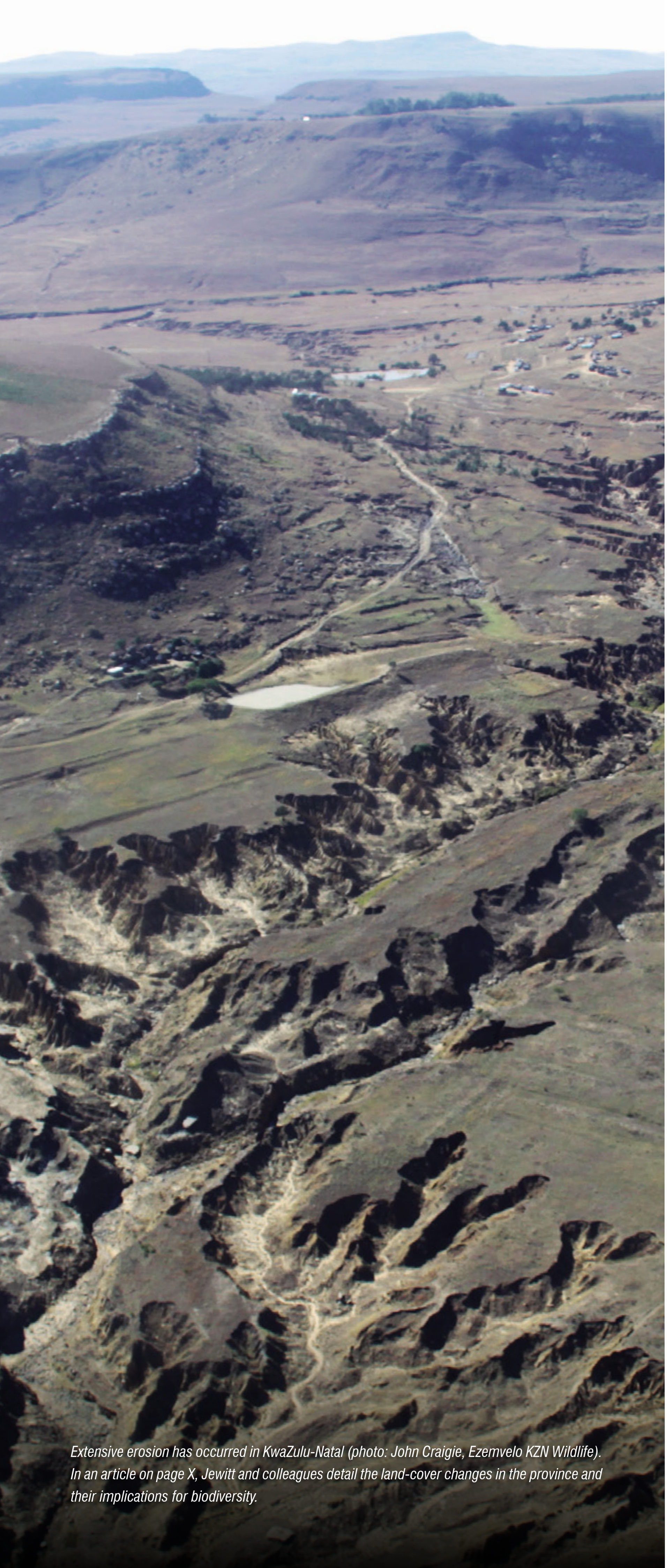
Authors' contributions

C.O. performed the sample preparation and LC-MS/MS analysis; M.T.S. performed data analysis and contributed to writing the manuscript; H.E.P. performed data analysis and contributed to writing the manuscript; G.K. performed LC-MS/MS analysis, data analysis and contributed to writing the manuscript; H.G.P. managed the research project, performed data analysis and contributed to writing the manuscript.

References

1. Fatta-Kassinos D, Ledin A. Editorial. Environ Pollut. 2010;158:3016. http://dx.doi.org/10.1016/j.envpol.2010.07.008
2. US Environmental Protection Agency. Contaminants of emerging concern [homepage on the Internet]. No date [cited 2015 Feb 15]. Available from: http://water.epa.gov/scitech/cec/
3. US Environmental Protection Agency. Basic information on CCL and regulatory determinations [homepage on the Internet]. No date [cited 2015 Feb 15]. Available from: http://www2.epa.gov/ccl/regulatory-determination-3
4. Koplin DW, Furlong ET, Meyer MT, Thurman EM, Zuagg SD, Barber LB, et al. Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999–2000: A national reconnaissance. Environ Sci Technol. 2002;36:1202–1211. http://dx.doi.org/10.1021/es011055j
5. Barnes KK, Kolpin DW, Furlong ET, Zaugg SD, Meyer MT, Barber LB. A national reconnaissance of pharmaceuticals and other organic wastewater contaminants in the United States – I: Groundwater. Sci Total Environ. 2008;402:192–200. http://dx.doi.org/10.1016/j.scitotenv.2008.04.028
6. Focazio MJ, Kolpin DW, Barnes KK, Furlong ET, Meyer MT, Zaugg SD, et al. A national reconnaissance for pharmaceuticals and other organic wastewater contaminants in the United States – II: Untreated drinking water sources. Sci Total Environ. 2008;402:201–216. http://dx.doi.org/10.1016/j.scitotenv.2008.02.021
7. Richardson SD. Water analysis: Emerging contaminants and current issues. Anal Chem. 2009;81:4645–4677. http://dx.doi.org/10.1021/ac9008012
8. Environmental Protection Agency. National primary drinking water regulations. Report EPA816-F-09-004. Washington, DC: Environmental Protection Agency; 2009.
9. World Health Organization. The WHO recommended classification of pesticides by hazard and guidelines to classification: 2009. Geneva: World Health Organization; 2010.
10. Swan SH. Semen quality in fertile US men in relation to geographical area and pesticide exposure. Int J Androl. 2006;29:62–68. http://dx.doi.org/10.1111/j.1365-2605.2005.00620.x
11. Chevrier C, Limon G, Monfort C, Rouget F, Garlantézec R, Petit C, et al. Urinary biomarkers of prenatal atrazine exposure and adverse birth outcomes in the Pelagie birth cohort. Environ Health Pers. 2011;119:1034–1041. http://dx.doi.org/10.1289/ehp.1002775
12. Ochoa-Acuna H, Frankenberger J, Hahn L, Carbajo C. Drinking-water herbicide exposure in Indiana and prevalence of small-for-gestational-age and preterm delivery. Environ Health Persp. 2009;117:1619–1624. http://dx.doi.org/10.1289/ehp.0900784
13. Munger R, Isacson P, Hu S, Burns T, Hanson J, Lynch CF, et al. Intra-uterine growth retardation in Iowa communities with herbicide-contaminated drinking water supplies. Environ Health Persp. 1997;105:308–314. http://dx.doi.org/10.1289/ehp.97105308
14. Waller SA, Paul K, Peterson SE, Hitti JE. Agricultural-related chemical exposures, season of conception, and risk of gastroschisis in Washington State. Am J Obstet Gynecol. 2010;202(3), Art. #241, 6 pages. http://dx.doi.org/10.1016/j.ajog.2010.01.023
15. Davis JA. Environmental atrazine exposure and congenital malformations in New York State: An ecologic study. Birth Defects Res A Clin Mol Teratol. 2005;73:926.
16. Mladinic M, Zeljezic D, Shaposhnikov SA, Collins AR. The use of FISH-comet to detect c-Myc and TP 53 damage in extended-term lymphocyte cultures treated with terbuthylazine and carbofuran. Toxicol Lett. 2002;211:62–69. http://dx.doi.org/10.1016/j.toxlet.2012.03.001
17. Matalon S, Schechtman S, Goldzweig G, Ornoy A. The teratogenic effect of carbamazepine: A meta-analysis of 1255 exposures. Reprod Toxicol. 2002;16:9–17. http://dx.doi.org/10.1016/S0890-6238(01)00199-X
18. Food and Drug Administration. Guidance for industry bioanalytical method validation. Silver Spring, MD: Food and Drug Administration; 2001. Available from: http://www.fda.gov/downloads/Drugs/Guidances/ucm070107.pdf.
19. Sigma-Aldrich. Guide to solid phase extraction. Bulletin 910. St. Louis: Sigma-Aldrich; 1998. Available from: http://www.sigmaaldrich.com/Graphics/Supelco/objects/4600/4538.pdf.
20. Cox DM, Zhong F, Du M, Duchoslav E, Sakuma T, Mcdermott JC. Multiple reaction monitoring as a method for identifying protein post-translational modifications. J Biomol Tech. 2005;16:83–90.
21. Botitsi HV, Garbis SD, Economou A, Tsipi DF. Current mass spectrometry strategies for the analysis of pesticides and their metabolites in food and water matrices. Mass Spectrom Rev. 2010;30:907–939. http://dx.doi.org/10.1002/mas.20307
22. National Library of Medicine. TOXNET: Toxicology data network. Washington, DC: National Library of Medicine; 2013. Available from: http://toxnet.nlm.nih.gov/.
23. European Union. Commission decision of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results. Brussels: European Union; 2002. Available from: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32002D0657
24. Bennet G, Hale S, Scifres J, Wasko M. Laboratory operations and quality assurance manual. Atlanta, GA: Environmental Protection Agency; 2011.
25. Van Iterson RA. A guide to validation in HPLC. Emmen: Drenthe College; 2011. Available from: http://parasshah.weebly.com/uploads/9/1/3/5/9135355/hplc_validation_pe.pdf.
26. Shah VP, Midha KK, Findlay JWA, Hill HM, Hulse JD, Mcgilveray IJ, et al. Bioanalytical method validation – A revisit with a decade of progress. Pharm Res. 2000;17:1551–1557. http://dx.doi.org/10.1023/A:1007669411738
27. World Health Organization. Guidelines for drinking-water quality. Geneva: World Health Organization; 2011.
28. Washington Suburban Sanitary Commission. Emerging contaminants. Washington, DC: Washington Suburban Sanitary Commission; 2013. Available from: https://www.wsscwater.com/files/live/sites/wssc/files/water%20quality/2013wqr.pdf

Note: This article is supplemented with online only material.



Extensive erosion has occurred in KwaZulu-Natal (photo: John Craigie, Ezemvelo KZN Wildlife). In an article on page X, Jewitt and colleagues detail the land-cover changes in the province and their implications for biodiversity.

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