Self-awareness and mental toughness in tennis athletes

Stormwater harvesting to improve water security in urban areas

Household income diversification in the poorest South African provinces

Witwatersrand wetland reduces water pollution from acid mine drainage

Dual investment needed for higher education in South Africa

SOUTH AFRICAN Journal of Science

JANUARY/FEBRUARY 2017

eISSN: 1996-7489

ASSAf ACADEMY OF SCIENCE OF SOUTH AFRICA

volume 113
number 1/2

JANUARY/FEBRUARY

eISSN: 1996-7489

SOUTH AFRICAN

Journal of Science

volume 113

number 1/2

Leader

What could scientists do about 'post-truth'?
John Butler-Adam 1

News & Views

South Africans pioneer heat transfer technology for conversion of waste to energy
Diane Hildebrandt, Xiaojun Lu & Thato Maphoto 2

Commentary

Survivorship of spekboom (*Portulacaria afra*) planted within the Subtropical Thicket Restoration Programme
Anthony J. Mills & Ashley Robson 3

Book Review

Marion Island half a century ago: A glimpse into an earlier era of sub-Antarctic exploration
Don A. Cowan 6

Creating critical conversations on higher education curricula in South Africa
Irma Eloff 7

From the NCHE to #FeesMustFall: An incomplete but important story of a difficult journey
Ahmed C. Bawa 8

Review Article

Educational investment towards the ideal future: South Africa's strategic choices
Suellen Shay 10

Research Article

Estimation of household income diversification in South Africa: A case study of three provinces
Jabulani Mathebula, Maria Molokomme, Siyanda Jonas & Charles Nhemachena 16

Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems
Febe de Wet, Neil Kleynhans, Dirk van Compernolle & Reza Sahraeian 25

Hassina Mouri
Department of Geology,
University of Johannesburg

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

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

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

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

Correspondence and enquiries
sajs@assaf.org.za

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

Cover caption
Can self-awareness improve athletic performance? In an article on page 50, Cowden explores whether competitive tennis players who are more self-aware tend to be mentally tougher.

Antifungal actinomycetes associated with the pine bark beetle, *Orthotomicus erosus*, in South Africa
Zander R. Human, Bernard Slippers, Z. Wilhelm de Beer, Michael J. Wingfield & Stephanus N. Venter 34

Soil fertility constraints and yield gaps of irrigation wheat in South Africa
Nondumiso Z. Sosibo, Pardon Muchaonyerwa, Lientjie Visser, Annelie Barnard, Ernest Dube & Toi J. Tsilo 41

On the mental toughness of self-aware athletes: Evidence from competitive tennis players
Richard G. Cowden 50

Osteopathology and insect traces in the *Australopithecus africanus* skeleton StW 431
Edward J. Odes, Alexander H. Parkinson, Patrick S. Randolph-Quinney, Bernhard Zipfel, Kudakwashe Jakata, Heather Bonney & Lee R. Berger 56

Attenuation of pollution arising from acid mine drainage by a natural wetland on the Witwatersrand
Marc S. Humphries, Terrence S. McCarthy & Letitia Pillay 63

Research Letter

Stormwater harvesting: Improving water security in South Africa's urban areas
Lloyd Fisher-Jeffes, Kirsty Carden, Neil P. Armitage & Kevin Winter 72

What could scientists do about 'post-truth'?

The *Oxford English Dictionary*'s Word of the Year for 2016 is 'post-truth', which they say is:

> *an adjective defined as 'relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief',*

to which they add this example:

> *In this era of post-truth politics, it is easy to cherry-pick data and come to whatever conclusion you desire.*

A recent article in the *Economist* points out that politicians have lied for as long as there has been formal politics and politicians. Few would disagree that scientists have lied (or dissembled or misled) for as long as there has been codified science. Whether in the form of prevailing, dominant ideology (for example, the sun goes around the earth and Lysenkoism) or personal ideology (vaccines cause autism; HIV does not lead to AIDS), science is not lacking in deceitfulness and claims made in the absence of seriously demonstrable truths.

The difference between 'dishonesty' (for whatever political or scientific end – or both, of course) and the era of 'post-truth' is, however, clear and important to understand. Dishonesty and lies are readily challenged and shown for what they are – even if belatedly and tragically, as in the case of the assertions that Iran had a vast store of weapons of mass destruction, or the claim that vaccinations cause autism. In the era of post-truth, lies are accepted as, and become, widely accepted truths – with a vast majority of people, in most instances, not only accepting the truth of what is not true but propagating it and disseminating it widely. So, for example, Barack Obama is not a citizen of the USA, the South African Public Protector does not have legal authority, or the President of South Sudan is a man innocent of war crimes. These and similar clearly deniable assertions, however, are asserted and disseminated, and become widely accepted. They also become the basis for action.

The *Economist* put it this way:

> *Post-truth has also been abetted by the evolution of the media. The fragmentation of news sources has created an atomised world in which lies, rumour and gossip spread with alarming speed. Lies that are widely shared online within a network, whose members trust each other more than they trust any mainstream-media source, can quickly take on the appearance of truth. Presented with evidence that contradicts a belief that is dearly held, people have a tendency to ditch the facts first.*

Sadly, the post-truth (sometimes called the 'post-factual') era has played a serious, dishonest and corrosive role in the turbulence that has come close to placing the survival of South African universities and their recognition in the international world of higher education at risk. Without trivialising the complexity of the demonstrations, some students have simply denied the economic fallacies implied in fee-free higher education, while some academics have fallen prey to the emotive rather than the demonstrable realities that have underpinned many of the issues that characterise the wide range of demands made by protesting student groups.

The implications and dangers of post-truthism are, however, much wider than those confronting South African higher education. And they are compounded by the anti-intellectual atmosphere being experienced in many parts of the world. Higher education institutions are attacked and threatened even in countries that have a democratic system and the risk is even greater in countries that do not have democratic systems.

For universities, and their academics, to counter post-truth they must have credibility, which makes the challenge a double one: to have trustworthiness, and to provide the hard data that call the lie to emotion-based beliefs.

Ole Petter Ottersen, Rector of the University of Oslo, recently wrote: 'Universities have to re-establish a respect for objective truth and powerful arguments – through our educational programmes and through our public outreach'.

To do so is not an easy task – universities have to establish, again, respect for objective truth and convincing arguments. Perhaps one of the most essential parts of the task would include respect for the process of the rigorous review of research findings in order to weed out potential deceits and dishonesties that serve to cast doubts on the academic project. But the task needs more than that. It will have to include academics being seriously engaged in public outreach and debates, and a willingness to present proven realities to counter 'post-factual' positions. An example of this is the challenge that scientists will face in order to take a stand against the views of Donald Trump's Secretary for the Environment who scoffs at the evidence for anthropogenic global warming – an undertaking that will face, additionally, the united efforts of the powerful fossil fuel industries. Closer to home, hard evidence about the dangers of fracking, and of the inappropriateness of a multibillion rand nuclear solution to our electric power needs, will be required.

Facing – and facing down – positions grounded on emotion-based fallacies is not an easy stance to assume, and will, in all likelihood, become more difficult in an increasingly populist world. Yet it is a duty that universities and their scientists cannot afford to neglect.

HOW TO CITE: Butler-Adam J. What could scientists do about 'post-truth'? S Afr J Sci. 2017;113(1/2), Art. #a0195, 1 page. http://dx.doi.org/10.17159/sajs.2017/a0195

South Africans pioneer heat transfer technology for conversion of waste to energy

AUTHORS:
Diane Hildebrandt[1]
Xiaojun Lu[1]
Thato Maphoto[2]

AFFILIATIONS:
[1]Material and Process Synthesis Research Unit, College of Science, Engineering and Technology, University of South Africa, Pretoria, South Africa
[2]The Riverbed Agency, Johannesburg, South Africa

CORRESPONDENCE TO:
Thato Maphoto

EMAIL:
thato@theriverbed.co.za

KEYWORDS:
waste; liquid fuel; reactor; Fischer–Tropsch

HOW TO CITE:
Hildebrandt D, Lu X, Maphoto T. South Africans pioneer heat transfer technology for conversion of waste to energy. S Afr J Sci. 2017;113(1/2), Art. #a0191, 1 page. http://dx.doi.org/10.17159/sajs.2017/a0191

In order to develop Africa and assist communities, particularly those with limited or no access to energy, more environmentally responsible and sustainable ways to supply energy need to be found. Transportation fuels and electricity are critical components of the energy mix – making our recent invention for improved reactor and process performance a potential game-changer.

While developing countries face a myriad of challenges, some of the most pressing ones they often face are access to energy and simultaneously reducing greenhouse gas emissions. These two needs conflict if fossil fuels are used. With this in mind, we started looking at renewable resources as an answer to energy problems – more specifically in terms of how municipal, agricultural and industrial wastes are potentially valuable resources, given that carbon and energy are locked in these wastes.

As a result of our research, we have developed a process that uses municipal waste, manure, woodchips from wood factories and forest waste and converts these waste products into fuel and electricity. One ton of forest waste can be converted to about one barrel of synthetic fuel. As such, we could produce around 40 million barrels per year of synthetic crude oil from the waste that is currently going to landfills in South Africa.

Together with our team, we invented an intensified heat transfer method for fixed-bed reactors that will improve the efficiency of this waste-to-energy technology on a small to medium scale. The invention is the culmination of the team's collective knowledge and experience over the past 20 years.

As a starting point, we wanted a system that was simple and robust to protect the catalyst. The invention, being patented as a 'tubular fixed bed reactor with heat pipe for internal heat removal for Fischer–Tropsch synthesis' improves the performance of the reactor which is the heart of the process. Our objective was to improve the efficiency of the reactor by reducing hotspots so that the catalyst life and online time can be longer, resulting in catalyst cost reduction.

Catalysts are components vital to convert feedstocks to fuel. The longer they last and the more efficiently they work, the smoother and less expensive production is likely to be. One of the most common ways in which catalysts are damaged or destroyed is when the reactor overheats during the conversion process. Based on this, we specifically looked at the stability of the catalyst and analysed the temperatures in the bed.

Although the reactor is a key piece of the invention, the whole process is important. The reactor makes the process as simple and cheap as possible. As such, the reactor is part of the development of the whole process for the waste-to-energy technology. The process uses the Fischer–Tropsch reaction to produce a synthetic fuel, which can then be separated and upgraded to produce marketable products including fuel and chemicals.

As part of the Unisa team, we have been involved in numerous synthetic fuel projects such as the Golden Nest pilot plant in China, the Linc Energy in Australia, and a GTL plant in Houston, Texas amongst others. We are looking forward to seeing the full impact of our work across the sector internationally – bringing communities that much closer to sustainability and a cleaner environment.

Survivorship of spekboom (*Portulacaria afra*) planted within the Subtropical Thicket Restoration Programme

AUTHORS:
Anthony J. Mills[1,2]
Ashley Robson[2]

AFFILIATIONS:
[1]Department of Soil Science, Stellenbosch University, Stellenbosch, South Africa
[2]C4 EcoSolutions, Cape Town, South Africa

CORRESPONDENCE TO:
Anthony Mills

EMAIL:
mills@sun.ac.za

KEYWORDS:
planting protocols; investment; micro-basin; cost–benefit analysis

HOW TO CITE:
Mills AJ, Robson A. Survivorship of spekboom (*Portulacaria afra*) planted within the Subtropical Thicket Restoration Programme. S Afr J Sci. 2017;113(1/2), Art. #a0196, 3 pages. http://dx.doi.org/10.17159/sajs.2017/a0196

Through the Subtropical Thicket Restoration Programme (STRP), about 21.5 million cuttings of spekboom (*Portulacaria afra*) were planted over the period 2004–2016 in the Addo Elephant National Park, Great Fish River Nature Reserve and the Baviaanskloof Nature Reserve. This planting includes a large experiment of 330 quarter-hectare plots in which 14 different planting treatments were used.[1] These experimental plots, known as the 'thicket-wide plots', comprised 200 000 cuttings, with the remaining 21.3 million cuttings planted out in what were called the 'large-scale plantings'. Some of the large-scale plantings were replanted with cuttings – a procedure referred to as blanking. The positioning and number of cuttings used in each blanking operation was not recorded and consequently the surviving cuttings in any particular landscape within the large-scale plantings cannot be aged accurately. Notwithstanding the limitation of many sites in the large-scale plantings made up of cuttings planted in different years, we saw value in monitoring survivorship of cuttings in random plots within the large-scale plantings, simply to determine the likely outcomes of the South African government's investment in planting 21.5 million cuttings over the past 12 years.

In June and November 2015 we collected survivorship data in large-scale plantings from 47 plots in Addo Elephant National Park and 17 plots in Great Fish River Nature Reserve (Figures 1 and 2). We used the STRP database hosted by the Gamtoos Irrigation Board in Patensie (Eastern Cape) to identify appropriate areas for sampling across a range of topography and geology. At each plot (20 m by 20 m) we counted all living cuttings and estimated survivorship using the assumption that each plot had originally contained 100 cuttings. This assumption was based on the standard STRP planting protocol of planting cuttings 2 m apart, i.e. 2500 cuttings per hectare. It should be noted, however, that depending on the rockiness of a particular landscape, the distance between cuttings – and consequently the original number of cuttings in each of our study plots – would have varied.

The data show that survivorship in the large-scale plantings is extremely variable, ranging from 0 to 93%, with a mean of 28% across all 64 plots sampled (Table 1a). Geographical reasons for this variation were not evident in our data set (Table 1b,c; Figure 3). A generalised linear model showed, for example, that geology, aspect, elevation and slope were not related to survivorship.

To better inform planting protocols of future restoration efforts, we suggest that future studies examine the effects of inter alia soil temperature, soil water content and quality of planting operations on cutting survivorship. Importantly, the future monitoring of large-scale plantings should be undertaken in such a way that the effects of blanking can easily be taken into account in analyses of cutting survivorship. Lastly, permanent monitoring plots should be established in some of the large-scale plantings immediately after planting to ensure that accurate baseline data on the number of cuttings planted in a particular plot are captured.

A new planting protocol (Figure 4) that has proved successful in Camdeboo National Park is the planting of cuttings in bunches in trenches or micro-basins (Taplin B 2016, personal communication, May 5). This protocol ostensibly results in rainwater harvesting in the depressions which increases the rate of growth of cuttings relative to individual cuttings planted outside of depressions. If the dense clusters of spekboom cuttings ultimately form vigorous patches of mature plants that expand outwards in all directions – as is evident in some photographic records (Hoffman T 2016, personal communication, June 22) and old restoration sites[2] – the number of micro-basins excavated per hectare could be reduced to 25 to 50, as opposed to the current protocol of 2500 holes per hectare.

The average survivorship of 28% of the 21.3 million cuttings planted to date by the STRP means that the likely current legacy of the programme is ~6 million surviving spekboom cuttings. Based on results from old restoration sites[2,3], many of these cuttings will in time form large spekboom clumps which will – where herbivore stocking densities are appropriate – continue to expand for decades to come. The end result will consequently be a new matrix in which other species of thicket plants can establish.[4] Assuming that 5 million of the 6 million surviving plants will over the rest of the 21st century grow to establish thicket patches of ~4 m in diameter, based on a conservative 25-mm outward spread per annum (i.e. a 50-mm increase in diameter of the thicket patch per annum), ~7000 ha of thicket will have been restored by 2100 through an investment totalling ~ZAR100 million. Given the considerable benefits of restored compared with degraded thicket in terms of soil quality[3,5], infiltration of rainwater[6], carbon sequestration[7] and herbivore carrying capacity[8], this investment by the South African public is likely to be deemed worthwhile by future generations. To reach such a conclusion, however, a comprehensive analysis of the costs and benefits in terms of public goods (e.g. contribution to baseflow in rivers) and private goods (e.g. tourism and wildlife) over the ensuing decades would be required. Such an analysis would ideally track the change in value of the restored thicket through time and would assist government as well as the private sector to take informed decisions on investments in the upscaling of thicket restoration.

Acknowledgements

We thank Stephan Coetzee, Adele Cormac, Zurelda le Roux, Mohammed Kajee and Julia Baum for technical contributions to the manuscript.

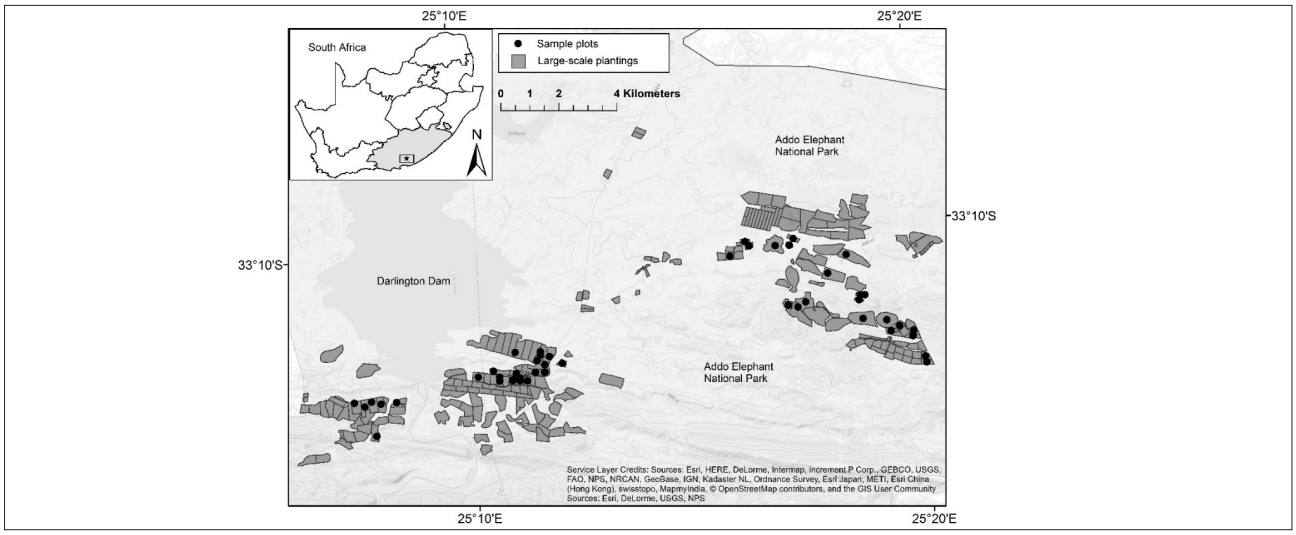


Figure 1: Sample plots and large-scale plantings in Addo Elephant National Park in the Eastern Cape, South Africa.

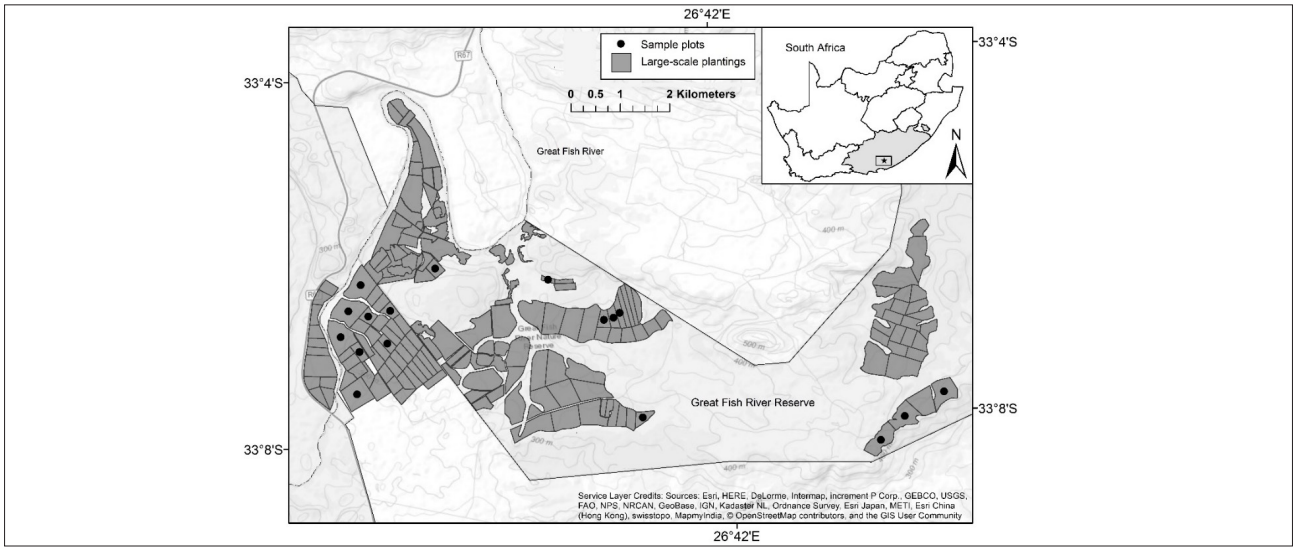


Figure 2: Sample plots and large-scale plantings in the Great Fish River Nature Reserve in the Eastern Cape, South Africa.

Table 1: Spekboom cutting survivorship (%) for different (a) sites, (b) geology types and (c) aspects

a

Site	*n*	Mean	Median	s.d.
Addo Elephant National Park	47	31	28	24
Great Fish River Nature Reserve	17	20	16	15
Combined	64	28	24	23

b

Site	Dwyka			Ecca		
	n	Mean	s.d.	*n*	Mean	s.d.
Addo Elephant National Park	29	35	28	18	25	14

c

Site	Flat			North-facing			West-facing		
	n	Mean	s.d.	*n*	Mean	s.d.	*n*	Mean	s.d.
Addo Elephant National Park	19	24	20	28	36	25	–	–	–
Great Fish River Nature Reserve	7	18	16	7	25	15	3	13	6
Combined	26	22	19	35	34	24	–	–	–

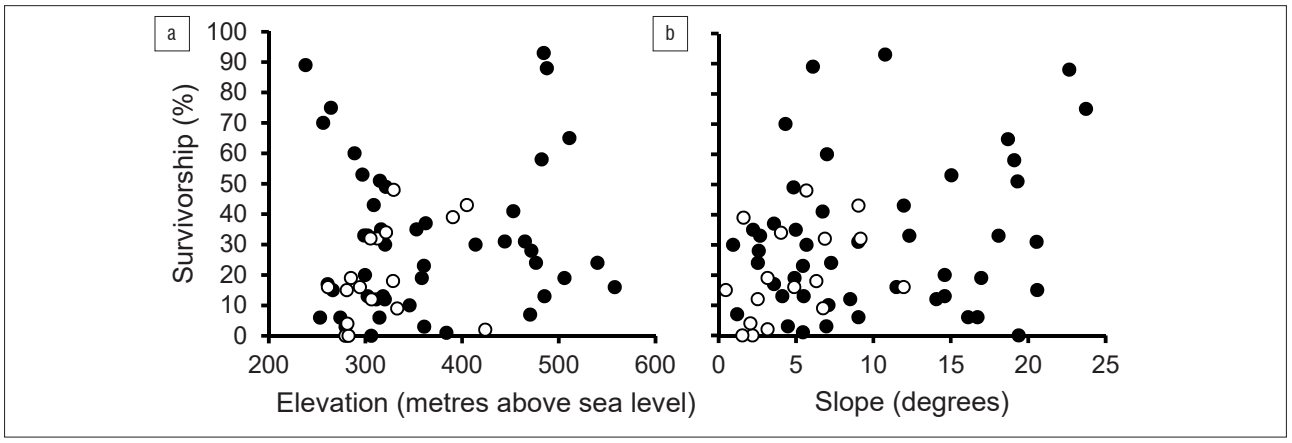


Figure 3: Spekboom cutting survivorship in relation to (a) elevation and (b) slope in Addo Elephant National Park (solid circles) and Great Fish River Nature Reserve (open circles).

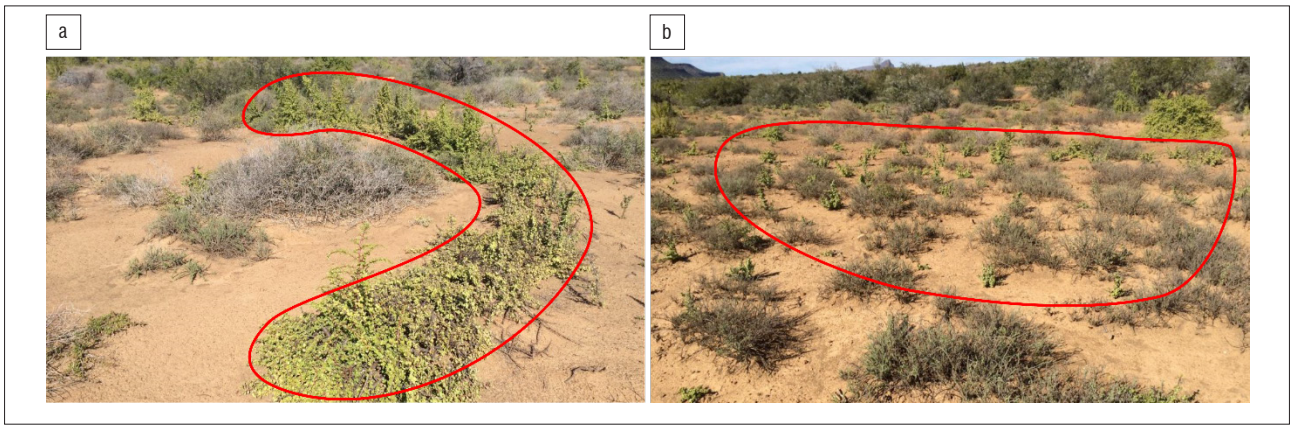


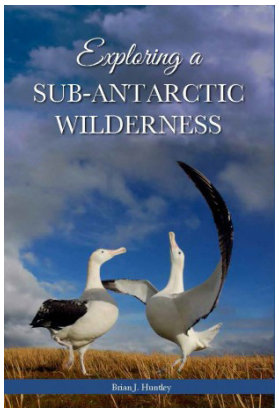
Figure 4: Comparison of spekboom growth after ~5 years after planting at Camdeboo National Park: (a) in dense clusters in a trench and (b) as single cuttings.

References

1. Mills AJ, Van der Vuyer M, Gordon IJ, Patwardhan A, Marais C, Blignaut J, et al. Prescribing innovation within a large-scale restoration programme in degraded subtropical thicket in South Africa. Forests. 2015;6:4328–4348. https://doi.org/10.3390/f6114328
2. Mills AJ, Cowling RM. Rate of carbon sequestration at two thicket restoration sites in the Eastern Cape, South Africa. Restor Ecol. 2006;14:38–49. https://doi.org/10.1111/j.1526-100X.2006.00103.x
3. Mills AJ, Fey MV. Transformation of thicket to savanna reduces soil quality in the Eastern Cape, South Africa. Plant Soil. 2004;265(1):153–163. https://doi.org/10.1007/s11104-005-0534-2
4. Van der Vyver ML, Cowling RM, Mills AJ, Difford M. Spontaneous return of biodiversity in restored subtropical thicket: *Portulacaria afra* as an ecosystem engineer. Restor Ecol. 2013;21:736–744. https://doi.org/10.1111/rec.12000
5. Mills AJ, Cowling RM, Frey MV, Kerley GIH, Lechmere ORG, Sigwela A, et al. Effects of goat pastoralism on ecosystem carbon storage in semi-arid thicket, Eastern Cape, South Africa. Austral Ecol. 2005;30(7):797–804. https://doi.org/10.1111/j.1442-9993.2005.01523.x
6. Van Luijk G, Cowling RM, Riksen MJPM, Glenday J. Hydrological implications of desertification: Degradation of South African semi-arid subtropical thicket. J Arid Environ. 2013;91:14–21. https://doi.org/10.1016/j.jaridenv.2012.10.022
7. Mills AJ, Cowling RM. How fast can carbon be sequestered when restoring degraded subtropical thicket? Restor Ecol. 2014;22:571–573. https://doi.org/10.1111/rec.12117
8. Stuart-Hill GC, Aucamp AJ. Carrying capacity of the succulent valley bushveld of the Eastern Cape. Afr J Range Forage Sci. 1993;10:1–10. https://doi.org/10.1080/10220119.1993.9638314

Marion Island half a century ago: A glimpse into an earlier era of sub-Antarctic exploration

BOOK TITLE:
Exploring a sub-Antarctic wilderness: A personal narrative of the first biological and geological expedition to Marion and Prince Edward Islands 1965/1966



AUTHOR:
Brian J. Huntley

ISBN:
9780620705219 (softcover)

PUBLISHER:
Antarctic Legacy of South Africa, Stellenbosch; ZAR250

PUBLISHED:
2016

REVIEWER:
Don A. Cowan

AFFILIATION:
Centre for Microbial Ecology and Genomics, University of Pretoria, Pretoria, South Africa

EMAIL:
don.cowan@up.ac.za

HOW TO CITE:
Cowan DA. Marion Island half a century ago: A glimpse into an earlier era of sub-Antarctic exploration. S Afr J Sci. 2017;113(1/2), Art. #a0192, 1 page. http://dx.doi.org/10.17159/sajs.2017/a0192

My first thoughts on this book were unkind: 'It's a diary, so don't expect too much'. Yet this book is so much more than a diary.

As an historical document, this text has much of interest. There are detailed descriptions of many of the physical and biological aspects of Marion Island (and neighbouring Prince Edward Island), with sections as diverse as climate and animal behaviour, all supported by the author's own photographs and even some data figures. Modern and future Marion Island explorers will be able to see the evidence of changes in the locations, numbers and distributions of plant and animal populations in this intervening period. Sadly, some of the evidence of previous human activities seen by Huntley and his colleagues on Marion Island will disappear in the intervening years.

At first, the reader may find the structure of the book somewhat bamboozling, principally because the author sections the text into six different themes, such as 'Teamwork on a remote island' and 'Getting down to the research programme', each of which is supported by a selection of his diary notes and prefaced with explanatory comments by the author. This style does provide some subject-related focus, but breaks any attempt to establish a continuous temporal sequence. What I missed was the 'between the lines' evidence of the change in the author from 'newby' to 'old hand'.

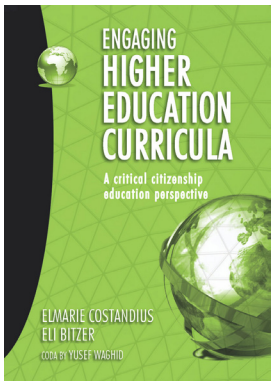
The lay reader will also struggle with the place names. As a diary, virtually every daily record includes the names of the places visited during the author's daily marches. While this attests to his amazing level of strength and fitness, the lay reader may struggle to place these 'treks' into an 'island mind map'. As a reader who has visited Antarctica regularly but (sadly) never been to Marion, I found it difficult to keep track of the locations mentioned, even though one can refer to the map of Marion Island presented inside the front cover of the book.

But this criticism is minor compared to the pervading sense, throughout the text, of the wonderful audacity and integrity of the author and his colleagues. I was constantly reminding myself that the author had, at the time of the expedition, only just graduated with a BSc degree (and was keenly anticipating an honours degree in the year ahead). Yet the youthful Brian Huntley demonstrates amazing depths of experience, resilience, dedication, physical strength and scientific acumen. Working under extremely difficult conditions (references to rain, sleet, snow and/or gale force winds appear on virtually every page of the book), he and his colleagues nevertheless undertook extensive biological, geological and meteorological surveys of much of the 29 000-ha island. It is perhaps not surprising that, 50 years later, Professor Brian Huntley, now retired, is a highly respected member of the national and international academic community.

So, who will benefit from and enjoy this book the most? The answer is obvious: other Marion Island explorers, who can see what Huntley describes in their own mind's eye, will find the book a fascinating – and even emotional – experience. Not that the book is written in an emotional style: the terse diary notations are fitting for a diary scribbled while the author lay prone in a damp tent pitched on rough ground. But the reader need not belong to this rather elite group to enjoy this text. Any natural history or outdoor enthusiast will be able to engage with the young Brian Huntley as he struggles across razor-sharp lava flows carrying a 70-lb pack in gale force winds with a hint of sleet.

Creating critical conversations on higher education curricula in South Africa

BOOK TITLE:
Engaging higher education curricula



AUTHORS:
Elmarie Costandius and Eli Bitzer

ISBN:
9781920689698 (softcover)

PUBLISHER:
SUN MeDIA Stellenbosch;
ZAR200

PUBLISHED:
2015

REVIEWER:
Irma Eloff

AFFILIATION:
Faculty of Education, University of Pretoria, Pretoria, South Africa

EMAIL:
Irma.Eloff@up.ac.za

HOW TO CITE:
Eloff I. Creating critical conversations on higher education curricula in South Africa. S Afr J Sci. 2017;113(1/2), Art. #a0193, 1 page. http://dx.doi.org/10.17159/sajs.2017/a0193

This excellent publication, *Engaging Higher Education Curricula,* could not have come at a more opportune time. As South African tertiary institutions grapple with more than a year's worth of student protests, frequent campus closures and extreme measures to ensure that examinations continue, the conceptual understandings of higher education curricula are expanding. This publication sets the tone for critical conversations that need to be had within universities in the country now and in the years to come.

A key strength of the text is that it argues against the utilitarian view of higher education curricula that has permeated South African universities in the last two decades. Higher education has increasingly engaged in the discourse of tertiary qualifications as 'instruments' to be used for specific purposes or professions. As such, the implicit and expansive notion of a university education has fallen by the wayside. The text advocates for a broader view and also leaves room for deeper interrogation of the purposes of higher education curricula. This view reaches beyond performativity.

In conjunction with sensitising the reader to the utilitarian view of higher education curricula, the text also speaks to the presumptuous notion of viewing students as 'consumers'. The moment a narrow view of students as mere 'consumers' of higher education is taken, a ripple effect of harmful reductionist effects is activated. This text can be commended for taking on the notion of students as consumers as it engages critical issues around higher education curricula. It creates a space in which the notion of what it means to be a student can be actively explored.

Overall, the text admirably problematises the concept and adjacent constructs around higher education curricula. It points to identity issues. It dives deeply into the emotional aspects of curricula. It propagates the importance of disruptive learning spaces. It highlights the interactive dynamics between 'teacher' and 'student'. It locates the discussion about higher education curricula against the historical background of the country in which it is unfolding. It is also frank about the challenges in higher education curricula and indeed dedicates an entire chapter to these challenges. It also effectively identifies the tensions between localisation of curricula and the importance of internationalisation. Most importantly, it recognises the changing nature of knowledge systems in the world today.

The text forefronts the integral role of the 'student voice' in higher education curricula. As universities contemplate responses to the calls to decolonise the curricula, the student voice within these discussions will be imperative. The creation of inclusive curricula, that retain historical accountability, whilst at the same time forge new narratives, will be critical in the next few years at South African universities.

Two areas in which the text could perhaps have been fleshed out slightly more, relate to the views of the past and also the views on marginalisation. Firstly, while higher education curricula necessitate a recognition of the past, they also have to entrench a hopeful view of the future. As such, the text could perhaps have done more in terms of future perspectives for higher education curricula in South Africa. In a sense, when history remains the primary departure point for the ways in which curricula are developed and criticised, the possibilities for future curricula are inadvertently limited. Higher education curricula can be crucial vehicles for creating alternative futures – futures that are free from the vestibules of the past.

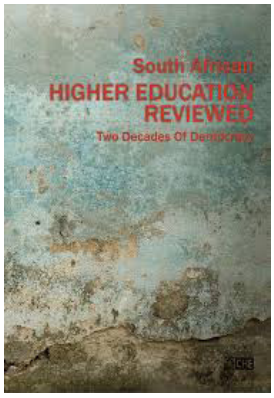
The example on art education in the book provides some perspective in this regard. It is heartening that the departure point for the chapter is the United Nations Millennium Development Goals and the Earth Charter Initiative. (It should be noted that the Millennium Development Goals have since been replaced by the Sustainable Development Goals, but it is recognised that this project most probably took place during a period when they were still in effect). It is equally encouraging that one of the identified themes from the project is 'hope for the future', which emerged from the participants in the study. This theme perhaps underlines the importance of higher education curricula that entreaty the future. If we want to overcome our history, what is it that we want? What is the future we want? It should not only be the past that we do not want. The emphasis on the history of South Africa is evident throughout the text. Perhaps the text could have included more on the future of South Africa.

Secondly, marginalisation is presented in the text as a fairly static concept. Even though marginalisation is seen as an ongoing process that is the result of our collective apartheid past, the text does seem to assume that 'marginalised groups' have remained the same over a period of more than 40 to 50 years. I would argue that marginalisation is a much more fluid notion and that it changes significantly over time. In the same way that the *ways* in which marginalisation are affected may change over time, so do the groups that are marginalised. Groups change and 'groupings' change. Corrective measures are implemented to address past injustices. In the very moments in which measures are taken, new groups are formed and new marginalisation may occur. Are the 'groups' that were marginalised at the dawn of our democracy, the same 'groups'? Have new 'groups' perhaps emerged? Who are the 'silent groups' that may be marginalised but for which marginalisation is not yet recognised? Marginalisation needs to be viewed as a sinuous, fluctuating phenomenon and acknowledged as such.

In conclusion, this book is a very important text in the higher education landscape in South Africa. It provides insightful perspectives on criticality and engagement in higher education curricula. It embraces the complexities and it elevates the debates in the field beyond the pragmatics.

From the NCHE to #FeesMustFall: An incomplete but important story of a difficult journey

BOOK TITLE:
South African higher education reviewed: Two decades of democracy



EDITOR:
Council on Higher Education

ISBN:
9780994678543 (softcover)

PUBLISHER:
Council on Higher Education, Pretoria; open online

PUBLISHED:
2016

REVIEWER:
Ahmed C. Bawa

AFFILIATION:
Universities South Africa, Pretoria, South Africa

EMAIL:
ahmed.bawa@usaf.ac.za

HOW TO CITE:
Bawa A. From the NCHE to #FeesMustFall: An incomplete but important story of a difficult journey. S Afr J Sci. 2017;113(1/2), Art. #a0194, 2 pages. http://dx.doi.org/10.17159/sajs.2017/a0194

The Council on Higher Education's *South African Higher Education Reviewed: Two Decades of Democracy* has to be seen as an important injection into our understanding of the performance of the higher education system over the last 20 years and its impact on the challenges facing our society. It has been a period of much change, expansion, rationalisation and experimentation and it is important that the Council has produced this collection of well-researched articles.

This book comes into circulation at a time when the South African higher education system is being severely rocked by the student activism that has so defined the 2015 and 2016 academic years. In multiple voices and forms this activism has raised two major themes: the idea of affordable access to higher education and the need for the transformation of the nature of the system's knowledge project – captured by the compositely expressed idea of free, quality and decolonised education. We must, however, constantly remind ourselves that the kind of instability that we have seen in the last year has been a part of the higher education system for at least 15 years and is driven primarily by the challenge of access for the poor. These student actions were restricted mainly to those institutions that have attracted large numbers of poor students and were driven primarily by insufficient financial aid.

Coming, as it does, about 20 years after the transition to democracy, the time of arrival of the book may be considered both appropriate and inappropriate. On the one hand, it comes at a time when there is an important need to reflect on the role played by the higher education system in post-apartheid South Africa and to assess the state of the system 20 years since the National Commission on Higher Education (NCHE) and the production of what has to be seen as the excellent *Higher Education Act* of 1997. It may therefore be important that this set of studies was concluded and published before the onset of the #FeesMustFall and #RhodesMustFall campaigns so that they serve the purpose for which they were designed.

On the other hand, nowhere in its 380 pages does this review foresee the unfolding of the instability that has beset the system over the last year – an indication perhaps that its writers may have held the common view that the long-standing instability at the historically disadvantaged institutions was caused purely by poor governance rather than by deeply systemic 20-year-old funding illnesses taking hold of the sector. What this means in essence is that this set of articles and analyses of the system may have glossed over or else simply missed some of the most critical issues facing the sector. So one may ponder the overall value of the analyses that constitute this review.

Having said this, the logical perspective to adopt would be to assume that one of the outcomes of the National Commission on Higher Education process in 1995–1996 was the crafting of the Higher Education White Paper and the *Higher Education Act* of 1997 – both of which provided the impetus for creation of the single, unified university system with clearly defined and progressive governance principles. The purpose of this review then was to measure the performance of the composite, complex system against the policy determinations that emerged in that process. As such, this is a valuable set of studies that draws on data and encompasses interesting and nuanced reflections.

The first chapter written by Denyse Webbstock does as it states: it provides an overview of the sector. It contains a very substantial amount of data and interesting interpretations of them. What it does not do is set the scene or form a foundation for the focus of the rest of the review on the key challenges facing higher education. In other words, it does not quite create sufficient tension relating to key issues facing the sector as a way to galvanise a discursive approach to addressing them. Examples of such tensions would be that between institutional autonomy and accountability, enrolment growth and funding, articulation and differentiation, and so on. Even so, the relevant chapters do address the issues of regulation, governance, teaching and learning in an interesting manner, focusing on the interplay between the institutions, the Department of Higher Education, the Council on Higher Education, the world of work, etc. There are important aspects of the sector – such as student services, student affairs and student housing – that are not covered and which form a vital area of engagement at our universities as we focus on the intellectual, social and emotional development of our students.

The student activism of 2015–2016 has brought to the fore, in many and varied ways, two key issues: the overall funding of the sector and the complex, interesting idea of a decolonised education. The first is discerned in quite substantial detail in the last chapter, which describes some of the key impacts and drivers of funding. The second escapes attention.

The review points out that the system has been through tremendous change in the last 20 years, with a number of institutional merges, a reshaping of the funding model to admit higher levels of state steering aimed at the broadening of access, improving success rates and increasing research outputs. This last chapter points out that while the block grant and steering funding grew at a rate of 4% per annum, this rate was smaller than that at which enrolments grew – squeezing into negative growth the subsidy per full-time equivalent. This final chapter does raise the need for state subsidy in the form of the block grant to grow at a higher rate. What it does not do is explore the impact of the chronic underfunding of higher education over the last 20 years. An inevitable outcome of this underfunding is above-inflation increases in tuition fees that ultimately result in the unaffordability of higher education to the vast majority of South African families, thereby undermining one of the most important roles of universities – which is to contribute to the creation of a more equal society.

Student demands for 'quality decolonised education' is a call for the re-imagination of the knowledge project of our universities. This demand also is not new. While the NCHE recognised this call as a key part of the sectoral

transformation agenda, it linked the matter to more instrumentalist notions of the role of knowledge in the economy, in nation building, etc. The chapter on 'Teaching and Learning' addresses a number of important issues relating to how teaching and learning may be improved, but it shies away from any major discussion about the nature of the curriculum.

This review is an important addition to the literature on South African higher education. It suffers somewhat from stylistic variation as one progresses through it. And it does have serious gaps – the most important of which for me is the one regarding student development. What it does do is provide us with an opportunity to reflect on how we ought to think about the next 10 years of higher education in South Africa.

Educational investment towards the ideal future: South Africa's strategic choices

AUTHOR:
Suellen Shay[1]

AFFILIATION:
[1]Centre for Higher Education Development, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Suellen Shay

EMAIL:
suellen.shay@uct.ac.za

DATES:
Received: 28 July 2016
Revised: 24 Oct. 2016
Accepted: 28 Oct. 2016

KEYWORDS:
higher education; Brazil; USA; curriculum; equity; flexible degree

HOW TO CITE:
Shay S. Educational investment towards the ideal future: South Africa's strategic choices. S Afr J Sci. 2017;113(1/2), Art. #2016-0227, 6 pages. http://dx.doi.org/10.17159/sajs.2017/20160227

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
None

Although there has been rapid expansion of higher education around the globe, such expansion has not resulted in a more equitable system. Drawing on the work of Nancy Fraser, equity in higher education is conceptualised as 'parity of participation' and includes both equity of access and outcomes. The tensions between expansion and equity are illustrated by comparing South Africa's equity challenges with those of Brazil and the USA. Focusing on South Africa's critical choices, four scenarios or possible futures are provided to illustrate some of the trade-offs and strategic choices. The main argument is that if South Africa's higher education system continues to expand without a concomitant investment in the effectiveness of teaching and learning, it will not achieve the policy goals of equity of access and outcomes. Furthermore the investment needs to be strategically targeted to interventions that can serve as systemic levers of change for reducing drop-out rates and improving graduation rates. To this end, over the next decade the state needs to prioritise an investment in an undergraduate curriculum more 'fit for purpose'. The investment needs to be in curriculum reform that normalises different levels of foundational provision, identifies and removes curriculum obstacles that delay or impede graduation, and provides opportunities for 'breadth' for all students, not only those who come from privileged backgrounds.

Significance:

- If South Africa's higher education system continues to expand without a concomitant investment in the effectiveness of teaching and learning, it will not achieve the policy goals of equity of access and outcomes.

Introduction

Much has been written about the rapid expansion of higher education over the past 50 years, which has been characterised as a shift from elite to massified systems, and there is a great deal of commentary on the relationship between this expansion and the goals of a more equitable higher education system.[1] The assumption may be that as the system expands, it will become more accessible to groups who have traditionally been excluded. Whilst there have been some gains in 'widening participation', the overall picture is that this unprecedented global growth has not resulted in a more equitable higher education system. The gains of expansion have not translated into gains for equity. Piketty argues that unequal access to higher education is one of the most important problems that states will face in the 21st century.[2(p.340)] I will argue, however, that access is not South Africa's most pressing problem.

The relationship between growth and equity is a tension that runs through South Africa's policy discourse across its decades of democracy where the state has advocated for the need to simultaneously address the imperatives of increasing access and improving success, particularly for those who have been historically under-represented in higher education.[3-5] Meanwhile the student protests of 2015/2016 – from #RhodesMustFall to #FeesMustFall – have put immense pressure on the state to expand access through 'fee-free' education and to this end a presidential Fees Commission of Inquiry has been set up to explore the feasibility of 'fee-free higher education and training'. The outcomes of this commission could have a profound impact on the future of higher education in South Africa; once again the sector is at a critical crossroad. If equity remains a policy goal there are strategic choices that need to be made with the inevitable trade-offs.

The main argument is that if South Africa's higher education system continues to expand without a concomitant investment in the improvement of its undergraduate completion rates, it will not achieve the policy goals of equity of access and outcomes. The state needs to prioritise over the next decade an investment in an undergraduate curriculum more 'fit for purpose'. The argument proceeds in four parts. Firstly, I conceptualise what a more equitable higher education system might look like. Secondly, I briefly explore the equity challenges of two other higher education systems: those of Brazil and the USA. Thirdly, four scenarios – or possible futures – are offered to illustrate the trade-offs and strategic choices. And finally, a proposal is made for the kind of educational investment required over the next decade.

Conceptualising a more equitable higher education

I borrow from political theorist Nancy Fraser's multidimensional framework of justice[6] to conceptualise equity in higher education: what is the goal, what are the obstacles that stand in the way of this goal, and what are the mechanisms that would enable progress towards this goal?

Fraser[6] defines justice as 'parity of participation'. She writes: 'Justice requires social arrangements that permit all to participate as peers in social life'[6(p.73)]. What would it mean for 'all to participate as peers' in higher education? 'Parity of participation' has a double meaning. Firstly, parity of participation means that the chances of an academically capable student gaining access to higher education are not predetermined by their background. Academic capability recognises that there are many talented school-leavers who have academic potential but because of poor educational provision will not be academically eligible. 'Background' includes prior schooling, socio-economic status, geographical location (e.g. urban or rural), language or any other feature that makes

up the relevant social and cultural capital that students bring with them to university. This 'parity of participation' points to equity of access. Secondly, parity of participation means that the same student's chances of successfully completing a degree are not primarily determined by that same set of variables. This 'parity of participation' points to equity of outcome. Thus a more equitable higher education system is one that ensures that a student's background does not predetermine their chance of gaining access to and successfully completing a qualification. This concept of equity is more far-reaching than those that focus only on access.[7]

According to Fraser[6], full participation requires dismantling institutionalised obstacles that prevent some people from participating on a par with others. Fraser's elaborated framework identifies a number of obstacles. For the purposes of this argument I focus on the economic or more generally 'resource' obstacles (for an elaboration of Fraser's conceptual framework see Shay and Peseta[8]). Fraser argues that people can be 'impeded from full participation by structures that deny them the resources they need in order to interact with others as peers'[6]; for example, a potential student may be denied access because of the cost of the application fee or lack of Internet facilities to complete the online application form. These would be examples of 'distributive injustice' in which an academically capable student is barred from access because of a lack of resources.

There is another kind of distributive injustice – a failure to gain epistemic access or access to powerful knowledge. The notion of 'epistemological access' – a term first coined by Morrow[9] – distinguishes between formal or physical access to higher education and *meaningful* access to the knowledge goods.[10] Morrow argues that if one of the key purposes of higher education is to produce knowledgeable citizens then it follows that one of its core functions must be to give students access to disciplinary knowledge. As the South African Department of Higher Education and Training (DHET) puts it:

> *The true meaning of transformation* [is] *when all students entering the system have a reasonable chance of success and access to powerful forms of knowledge and practices that will enable them to enter the productive economy and improve their life chances and that of their families.*[11]

There are thus two kinds of resource obstacles: financial and epistemic. It is necessary to overcome both for full participation, for access and for success.

'Expansion without equity': A comparative perspective

It is instructive to compare South Africa's equity challenges with those of other higher education systems, in particular with respect to the trade-offs between expansion and equity. Brazil's system shares many common features with South Africa's: it is located in a highly unequal society and it is a middle-income economy, with a relatively low participation rate. The US system, in contrast, is situated in a developed economy with a massified, highly differentiated system.

One measure of expansion is 'participation rate', also called gross enrolment rate, which refers to the total enrolment (of all ages) expressed as a percentage of the 20–24–year-old age group in the population.[12] Measuring the success of a system is more complex. The indicator used is completion rate which measures effectiveness and efficiency, that is, producing the desired results with the optimal resources. This is the percentage of a given first-year student intake, or cohort, that graduates in minimum time.[12] The measure of equity of access is the percentage of under-represented groups of the total of those enrolled. The measure of equity of outcomes is the percentage of under-represented groups of the total of those who have completed, that is, representativity of the graduating class.[13(p.39)] It is very difficult to obtain comparable data across all of these measures. The findings below are based on the most reliable and up-to-date data available.

Consistent with the global trend, all three of these higher education systems have expanded significantly. In South Africa, the number of enrolments has nearly doubled from approximately half a million in 1994 to close to a million by 2014 – an increase in participation rate from 12% to 20% in 2013.[12] By 2013, Brazil's enrolments were at 7.3 million (a participation rate of 30%), with 75% of these enrolments in the private sector – a 64% increase in the public sector and a 95% increase in the private sector from 2003.[14] In contrast to the massified system of the USA (with an 88% participation rate in 2013)[15], Brazil's and South Africa's systems are more 'elite'. At the same time, given the high levels of inequality in public schooling, the Brazilian and South African systems are made up of a significant proportion of school-leavers who are underprepared for university level study.[16]

Consider South Africa's class of 2015 matriculants: of the total cohort of National Senior Certificate writers, 33% of learners wrote Mathematics and only half of those (129 481) scored above a pass of 30%; of the total cohort, 25% wrote Physical Science, of whom 59% (113 121) scored 30% or more.[17] These results are of concern in terms of both the size and quality of the pool with respect to the requirements of science-based programmes such as Engineering, Health Sciences and Commerce. The National Benchmark Test (NBT) results for the 2015 writers provide further evidence of this underpreparedness: of those who wrote the 2015 NBT Mathematics (56 500) only 10% achieved a score of 'proficient' (meaning that they would be expected to cope with regular mainstream provision), whereas 45% scored 'basic' (meaning they will have serious challenges with university-level Mathematics).[18] This finding provides compelling evidence that a significant proportion of South Africa's matriculants are not prepared for university-level study in the science-based fields. Put another way, the universities are not prepared for the students. Either way, there is a misalignment.

The question is, should South Africa and Brazil be aspiring to higher participation rates? The pervasive view in higher education policy discourse favours expansion, given the global shift from manufacturing to knowledge-based economies resulting in the need for more highly skilled graduates.[19,20] Some would argue that an expanded tertiary sector – whether through government policy or market-driven – will contribute to the reduction of inequality.[2] However, this 'more is better' view needs to be interrogated. One of its assumptions is that there is a sufficient supply of academically prepared school-leavers to fill the enrolment pool. The data for South Africa suggest otherwise. Until such time as the output of public schooling improves, expansion will increase the proportion of underprepared students, widen the 'articulation gap' between secondary and tertiary provision and could lead to both higher drop-out rates and poorer completion rates.

What are the implications of expansion for equity of access and outcomes? The doubling of enrolments in South Africa means that historically under-represented groups now make up the overall majority (83%),[12(fig. 3)] although, as will be discussed below, challenges to access still remain. Brazil's growth however tells a cautionary story: McCowan[21] describes Brazil's growth as 'expansion without equity'. Its public system is no-fee but academically highly competitive (ratio of 1 to 8 acceptance) and thus remains the preserve of the 'best prepared and well-off applicants'[22]. The private system is less competitive (ratio of 1 to 1.5) but financially inaccessible to those of lower socio-economic status.[23] To address this ratio, the state has instituted a range of redress policies in the form of quotas and the bonus model (adding extra points to the admissions score) which have resulted in a small increase in under-represented student populations.[23] Brazil's experience provides evidence that removing financial barriers to access does not by itself result in equity of access – academic unpreparedness may remain an obstacle.[16]

In terms of equity of outcomes, the overall theme emerging from these accounts is a concern that the gains in equity of access have not materialised into equity of outcomes. In South Africa, about a third of those enrolled will have dropped out in the first or second year and 40–50% will not graduate at all.[13] The inequalities are even more starkly evident in the comparison between 2008 3-year degree completion rates (for N+4) of black students (49%) and white students (68%).[11] Cohort completion rates are not available for Brazil but graduate rates point to a fairly inefficient system: in 2013, the public sector and private graduate rates were 10% and 14%, respectively (if all students graduate in 4 years,

the rate would be 25%).[14] Cooper[24] describes the transformational gains of post-apartheid growth as a 'skewed' and even a 'stalled' revolution. The same could be said of Brazil.

The USA's massified and highly differentiated system is one often held up as an example, and yet it faces serious equity challenges. A 2011 report on the US 4-year university system[25] shows that only 40% of US students completed a degree in 4 years, 56% by the end of 5 years, with an unlikely chance of completion after 6 years. Disaggregating these data by ethnicity indicates a significantly lower completion for African–American students: 21% in 4 years and 35% in 5 years. A graph of social inequality by college degree attainment shows that, by 2013, 70% of the US families in the top income quartile had completed a degree by the age of 24 years, nearly double the graduation rate from 1970. In contrast, only 9% of those in the bottom quartile had completed a degree – up from 6% in 1970.[15] As Piketty puts it, 'parents' income has become an almost perfect predictor of university access'[2(p.339)].

These comparisons are instructive for South Africa's key policy decisions. What are the implications of an expanding system (even if only moderate growth) for the goals of greater equity of access and outcome? Brazil's public system has expanded and has no financial obstacles to access and yet equity of access remains a challenge. As Marginson[1] argues, the USA may have the highest proportion of world-class universities but there are serious concerns about whether it is a world-class system – it is certainly not equitably serving its minority population. These comparisons illustrate some of the challenges of the equity/growth trade-offs and point to some strategic choices that need to be considered.

Future scenarios and strategic choices

Scenarios are 'stories about how the future might unfold' for a particular organisation or sector.[26(p.7)] They are provocative and plausible stories. Scenario thinking begins by identifying a range of 'forces of change' both internal and external to the sector which may impact on its future. These forces 'combine in different ways to create a set of diverse stories about how a future could unfold'[26(p.8)]. The goal of scenario thinking is to inform discussion and debate about strategic choices.

A range of internal and external factors contribute to South Africa's inequitable higher education system, not least of all the legacy of apartheid. There is a recognition that the inequality cannot be fully addressed until conditions both inside and outside of the education system are repaired. The state recognises the need for investment in the whole education 'pipeline': as a result there has been increased attention to early childhood development and primary and secondary schooling. There are also calls for a more differentiated post-secondary sector including strong vocational and technical training.[5,27]

In terms of the critical forces internal to higher education, South Africa's low GDP growth suggests that there is unlikely to be substantial additional state funding for higher education. In addition to the significant funding implications emerging from the Fees Commission, the state is committing close to ZAR1 billion per annum from 2017 to 2020 as 'ear-marked' funding to support a coherent national programme for addressing transformational imperatives relating to equity and quality in the university system. All of these deliberations are happening against the backdrop of varying degrees of financial crises across the higher education sector as a result of years of decreasing block grant subsidy. This sets the stage for a highly contested set of competing choices that will profoundly impact on the next phase of higher education in South Africa.

For the purpose of scenario planning, these 'resource obstacles' are translated into continua of resource choices. The one choice is the extent to which the state increases financial aid to students. On the one end, financial aid is increased to make higher education more affordable (+), on the other end financial aid is frozen or decreased and higher education is essentially for those who can afford it (-). The other choice is the extent to which the state supports measures to improve the effectiveness of teaching and learning – what I will refer to as its 'educational investment'. On the one end, the state's investment is high (+) and on the other end the investment is low (-). These two resource continua result in a matrix of four possible future scenarios (Figure 1).

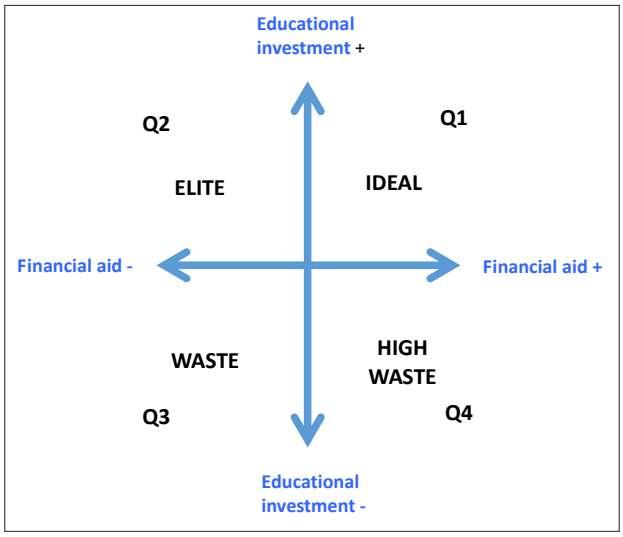


Figure 1: Possible future scenarios for South African higher education.

All of these scenarios involve assumptions. For the purposes of this exercise it is assumed that the students admitted are academically capable. It is also assumed that the state's educational investment yields improvements in the effectiveness of teaching and learning resulting in better retention and increased graduation rates. Another assumption is that there is no substantial additional state funding; that any increases come from reductions elsewhere – either from the higher education budget or from other areas of public spending. All these assumptions are debatable, as are the scenarios which they produce. This is the point of scenario thinking.

The top right quadrant (Q1) represents the 'ideal future'. In this scenario, the state increases to the extent that it can afford both its contribution to financial assistance and its investment in the effectiveness of teaching and learning. From the point of view of the students, irrespective of their socio-economic status, they are admitted and there is a high probability that they will successfully complete. From the point of view of the system, because the resource obstacles have been removed (both financial and academic) there is an increase in participation rate and there is equity of access and outcomes. This 'ideal future' is in fact South Africa's official future enshrined in policy since the 1997 White Paper[3]: increasing participation rates, equity of access and equity of outcomes in an efficient system.

The top left quadrant (Q2) is the 'elite future'. The state's contribution to financial aid is reduced. From the point of view of the student, if they can afford higher education, they will gain formal access. Given the likelihood of a reasonably good public or private schooling background and the state's educational investment, students are likely to successfully complete. From a system point of view, the reduced state funding for financial aid would result in a low participation rate with low equity of access and outcomes for those of socio-economically disadvantaged groups. Over time, given the demography of South Africa and the growth of the black upper-middle class, this system would be racially diverse with a black majority. The system would be reasonably efficient and increasingly dependent on private funding, which would result in a highly elite higher education system comprising the upper-middle class which, in time, would become racially diverse.

Quadrants 3 and 4 (Q3 and Q4) are both 'waste futures' with no educational investment made in improving the effectiveness of teaching and learning, and hence poor completion rates. The difference in the two quadrants is the state provision of financial aid. In Q3, the state freezes or reduces its current contribution to financial assistance. From the point of view of the students, if they can afford higher education, they will gain formal access. Their chances of succeeding will largely be determined by the quality of schooling. There is no equity of access, no equity of outcomes and poor efficiency.

In Q4, the state increases its contribution to financial aid. The participation rate increases, there is improved equity of access but completion rates do not rise (and in fact may decline), so there is virtually no equity of outcomes. From the students' point of view, irrespective of their socio-economic status, they will gain formal access. But given increased levels of under-preparedness and no investment in improving the effectiveness of teaching and learning, there is a high probability that they will not successfully complete. This is a highly inefficient system as it admits a significantly larger pool of students who are less well-prepared for university study.

Scenarios are theoretical reductions of a much more complex set of variables. Further debate requires probing and nuancing: can the state afford Q1 under the current economic climate? Unless there is substantial additional state funding, the system cannot expand, especially when this expansion is at the expense of poor students whose chance of completion is unlikely. While an extreme version of the 'elite future' scenario is, I would argue, politically and socially unacceptable, there may be possibilities at different points of the continua. Given that the system is by definition already 'elite' (with a 20% participation rate), the priority for existing state funding in Q2 is expansion not through enrolments but through graduations. As noted earlier, increasing the state's financial contribution to ensure 'fee-free' higher education is the rallying cry of the #FeesMustFall movement. There is an obvious appeal to this scenario, particularly as a political rallying cry. The pros and cons of this demand are not discussed here, but suffice it to say that a single focus on financial access will not guarantee equity of outcomes. The Brazilian case should be instructive. While these proposals are controversial, there is likely to be consensus that Q1 remains South Africa's goal and it should be evident that Q3 and Q4 are not desirable.

I would propose that South Africa's current system sits somewhere in Q3. The investment of the state in the past 20 years has produced an expanded system with greater equity of access, but is far from achieving equity of outcomes. The result is an inequitable and inefficient system. If there is an increase in financial assistance to talented but underprepared poor students, without a substantial educational investment to improve completion rates, South Africa's future trajectory is towards Q4 – a 'high waste future'. South Africa's higher education system seems to be precariously balanced between two 'future scenarios' of waste, neither of which will achieve its policy goals. Unless the state's GDP contribution to higher education is increased, it cannot afford Q1 in the short term. This expansion would be at the expense of poor students whose chance of completion is unlikely. While an extreme version of the 'elite future' scenario is, as argued, politically and socially unacceptable, a more nuanced version of the 'elite future' might be strategic in the short to medium term in order to achieve the 'ideal future': this is a capped-growth system that contributes to some improvement in equity of access but invests significantly in equity of outcomes.

What will enable the shift from Q3/Q4 to Q1? The state must invest in carefully targeted and monitored educational interventions that improve the effectiveness of teaching and learning.

Educational investment in systemic improvement

There is a significant body of scholarship on student retention. Tinto's model[28] for student persistence is seminal. [See Rooney[29] for a review of Tinto and subsequent modelling of student persistence]. These models point to a number of key determinants that influence whether students graduate or not. They recognise, on one hand, the influence of students' background variables (age, schooling, ethnicity, gender) and on the other hand, a range of variables within the institution: academic, environmental and social. These models propose a complex interplay between students' 'commitments' (including the resources they bring with them) and institutional conditions that explain the extent to which students successfully integrate and ultimately succeed.

There is an urgent need for these models to be tested in the South African context in order to better understand the causes of student drop-out/retention. A local study[29] in one historically advantaged institution provides sobering evidence of the extent to which background variables still profoundly shape success. The study concluded that being white, ineligible for financial aid and proficient in English, and having attended a top public or private school and obtaining good high school grades increased the likelihood of graduating. On the other hand, men who are on financial aid, non-English speaking, who attended poorly resourced schools and achieve low school grades are more likely to be academically excluded.[29(p.ii)] These are interesting, sobering but not surprising, findings. The challenge is: what are the enabling institutional conditions that can mitigate some of these determinants? What institutional commitment will enable the shift from Q3 to Q1?

There is no one answer or solution to this problem. There is a significant body of scholarship and practice on how to improve the quality of teaching and learning. The proposal which follows in no way denies a wide range of curricular, pedagogical and assessment interventions that can impact positively on student success. The focus here is on interventions in which state resources can be leveraged for *systemic* change that specifically contributes to equity of access and outcomes. While the state is to be commended for its investment in ear-marked funding to improve the effectiveness of teaching and learning, these resources need to be accompanied by a clear vision and plan – both at state and institutional level – for the optimal educational investment. It may not simply be more investment of the same or even a 'scaling up' of existing initiatives. A well-conceptualised strategy for educational change is required.

It was noted early that cohort completion data can shed light on discrepancies between curriculum intention and reality. The question must be asked: for whom is the curriculum working? I focus on completion data for the 3-year undergraduate degree obtained from DHET.[11(Tables 22–36)] These data were chosen given that the formative degree is the most common pathway to postgraduate/doctoral study. The minimum completion rate is 3 years (N) but an additional year (N+1) is not necessarily a problem. It may mean that a student failed a course or made some change to their curriculum (or added a major) that prolonged their degree. Beyond this additional year (N+2/3/4), a student has most likely failed multiple courses – a situation that is no longer optimal or efficient for the student or the state.

In terms of equity of access, the first observation is that, although the past two decades have seen significant achievements in terms of increasing equity of access, the data show that, with respect to the 3-year degree, black, coloured and Indian students are under-represented and white students are still over-represented. Of the total number of students who enrolled for a 3-year degree in 2008 (48 076), black students constituted 50%, coloured students 8%, Indian students 10% and white students 32%. The cohort completion data for specific qualification (e.g. Bachelor of Science) is not disaggregated by race, but it is likely that black students are even further under-represented in the science, technology, engineering and mathematics (STEM) areas. Addressing this problem will require greater intentionality in the recruitment of black students for degree studies, admission policies which are sufficiently flexible to admit talented but underprepared students and placement through sound diagnostic testing into the appropriate curriculum, including extended curriculum programmes.

In terms of equity of outcomes, two further observations can be made: one is the high drop-out rate across all the race groups including white students. Across all race groups by year 2, 20% have dropped out and by year 3, 25%. This is a significant loss in a system with a low participation. The second observation in terms of equity of outcomes is the low completion rate in N+1. Out of the total of those enrolled for 3-year degrees, only 36% graduate by N+1. Completion rate for black students by N+1 is 28%, coloured students 28%, Indian 32% and white students 50%. The evidence that half of white students and significantly more than half black, coloured and Indian students are taking 5 or more years to complete a 3-year degree and approximately one third have dropped out altogether would suggest that our current curriculum needs to be reviewed. Some educational investment needs to be made towards an undergraduate degree more fit for its purpose. The Council on Higher Education's (CHE) flexible degree[13] was proposed to address

this problem. The National Development Plan[5] argued for the need to extend STEM degrees to 4 years by redesigning first year to make it more accessible. The Ministry has supported neither of these proposals. I argue that the key principles informing the CHE proposal still hold and should serve to inform the priorities for educational investment.

The first proposal is that a fit-for-purpose curriculum will address the 'articulation gap' as a *systemic* problem. The 3-year degree data suggest that under-preparedness for university level study is a majority rather than a minority phenomenon. This should come as no surprise given that all the available data on the undergraduate enrolment pool – schooling background, National Senior Certificate results, NBT results and cohort performance – point to the need to reconceptualise the assumptions which inform the undergraduate entry-level 'norm', in order to cater for a more diversely prepared incoming cohort. This must be the sector's most urgent transformation priority.

In terms of addressing this gap, South African higher education has three decades of experience to draw on from both the successes and failures of its extended curriculum programmes (ECP). Overall the achievements of the ECPs have been to increase equity of access by admitting students (particularly to historically advantaged universities) who might not otherwise have been admitted, and secondly, to increase retention beyond first year. A study of nine ECP programmes across four institutions[30] found that seven out of nine of these programmes had year 1 to year 2 progression rates that exceeded the mainstream year 1 to year 2 progression. This suggests that these ECPs are successful in setting strong epistemic foundations for students. Over the decades staff involved in teaching on these programmes have developed a deep expertise in innovative entry-level curriculum (and pedagogical) interventions that can be drawn on and extended to 'mainstream' curriculum development.

Thus the first proposal for a more fit-for-purpose is to 'normalise' foundational provision; it should be conceptualised for the majority of South Africa's incoming students. The data suggest that approximately one third of enrolled students (those who graduate by N+1) may not need foundational provision and they could be exempted. (The racial composition of this group will vary depending on the institution but in most institutions this group would be racially diverse and in some it would be predominantly black). The rest, however, would benefit from either some foundational courses – for example, a foundational mathematics or physics or academic literacy course – or a full foundational programme. Space needs to be made in mainstream curricula for the required range of foundational provision to be offered to all who would benefit therefrom. Diagnostic instruments such as the NBTs can be used to place incoming students according to diverse levels of preparedness as is standard practice in other parts of the world.

However, these interventions alone are not enough. Expanding and normalising 'foundational provision' may not yield the desired results unless there is further curriculum change 'upstream'. Studies on ECP performance provide evidence of a general pattern of poor completion rates: despite some successes the gains of the foundational provision are not sustained through to completion.[13,29-31] What is required is a thorough review of the 'epistemic obstacles' which students face beyond first year that result in high failure. Thus the second proposal for a curriculum 'more fit for purpose' is to identify those key 'high risk' courses or combinations of courses across the degree which delay or impede graduation for a significant proportion of the students. This curriculum development is also an opportunity to ensure that the discipline-specific academic literacies are pulled through from first year to more senior years.

The third proposal for a more 'fit-for-purpose' curriculum is one that provides opportunities for 'breadth'. The previous two proposals address disciplinary 'depth', that is, acquiring adeptness in at least one discipline. Increasingly around the globe higher education institutions are embarking on large-scale curriculum reform to produce graduates with a wider skill set than higher education has traditionally produced. The CHE proposal calls for the formation of a particular kind of graduate through 'broadening the curriculum to include learning that is professionally and socially important in the contemporary world … and that lays the foundations for critical citizenship'[13(p.19)]. This 'breadth' would include key graduate attributes, opportunities for electives outside the discipline and the promotion of interdisciplinary thinking. While this may not on the surface appear to have direct impact on equity of outcomes, the reality is that students from privileged schooling do often experience curriculum breadth through additional electives, majors and extracurricular opportunities. These options have significant spin-offs for their employability opportunities. A curriculum fit-for-purpose will ensure that this breadth of experience is an expectation and outcome for all.

To summarise the main proposals: a curriculum more fit-for-purpose will address the 'articulation gap' as a *systemic* problem by normalising different levels of foundational provision to support the majority of capable students who either drop out or take unacceptable time to complete. To further support this outcome, the second proposal is to identify key 'high risk' combinations of courses across the degree that delay or impede graduation for a significant proportion of the students. The third proposal is a curriculum that provides opportunities for 'breadth' for all students, not only those who come from privileged backgrounds. The challenges of this kind of curriculum review are often more political than educational. There is extensive expertise both locally and internationally to support this curriculum development work. What is harder to find is the vision, leadership and political will for change.

Conclusion

Despite the significant gains of the past two decades South African higher education risks perpetuating or worsening its current waste scenario unless there is significant educational investment into improving the effectiveness of teaching and learning. Furthermore, the investment needs to be strategically directed at interventions that can serve as systemic levers of change that lead to reduced drop-out rates and improved graduation rates, especially for black and coloured students. It needs to be noted that this investment will have significant spin-offs on some of South Africa's other systemic challenges, such as expanding the pool of postgraduate students and the next generation of academic staff.[19] Indeed these cohorts cannot increase unless there is a more fit-for-purpose undergraduate curriculum.

If these proposals have merit, a next immediate step would be to set up a national collaborative research and development project funded by the DHET. The goal of this project would be to inform a 5–10-year curriculum review starting with the Bachelor of Science. This priority is not because science is more important than the social sciences and humanities but because it is the best place to start. Science-based courses are gateways to other fields of study: commerce, engineering and health sciences. Improved performance in science-based subjects will have a positive knock-on effect on a number of other qualifications.

The collaborative project could commission, for example, some of the following areas of research – the first being research on student retention. This commission would develop a better understanding of why students fail. How many of the nearly 50% of the intake who fail to complete are academically excluded, financially excluded, or drop out in good academic and financial standing? And why? This investigation would include drilling down into the existing cohort studies for a better understanding of the obstacles to completion. The second area, emerging from findings of the 'retention' project, would be research on the 'obstacle' courses and course combinations that lead to poor completion rates or academic exclusion. There is already some momentum developing through the Kresge-funded Siyaphumelela project and its focus on data analytics and 'high risk' courses. The third area would be research on the 'articulation gap' between school exit competency and tertiary preparedness in key subjects of the Bachelor of Science, e.g. Mathematics and Physical Science. This could involve a detailed investigation of both the entry-level proficiency of the applicant pool across the sector using NBT data and the entry-level requirements for key first-year courses. The outcomes of this commissioned research – to be conducted over a 2- to 3-year period – would provide a data-informed sector-wide basis for systemic review of the Bachelor of Science that may or may not result in recommendations for a 4-year degree, but would certainly result in a more fit-for-purpose curriculum appropriate to the South African context.

Acknowledgements

I acknowledge Katherine Fulton, Sam Keen and Patricia de Jong for their inspiration and encouragement on the early ideas that informed this paper and Ian Scott for his careful read and valuable input on the final draft.

References

1. Marginson S. Higher education and inequality in Anglo-American societies: Twenty-five years of a fair chance for all. In: Harvey A, Burnheim C, Brett M, editors. Student equity in Australian higher education. Singapore: Springer Science + Business Media; 2016. p. 165–182. http://dx.doi.org/10.1007/978-981-10-0315-8_10
2. Piketty T. Capital in the twenty-first century. Cambridge, MA: The Belknap Press of Harvard University Press; 2014. http://dx.doi.org/10.4159/9780674369542
3. South African Department of Education (DOE). Education White Paper 3: A programme for the transformation of higher education. Pretoria: DOE; 1997.
4. South African Department of Higher Education and Training (DHET). White Paper for post-school education and training: Building an expanded, effective and integrated post-school system. Pretoria: DHET; 2013.
5. National Planning Commission. National development plan vision 2030. Pretoria: Ministry of the Presidency; 2011.
6. Fraser N. Reframing justice in a globalizing world. New Left Review. 2005;36:69–88.
7. McCowan T. The growth of private higher education in Brazil: Implications for quality. J Educ Policy. 2004;19(4):453–472. http://dx.doi.org/10.1080/0268093042000227492
8. Shay S, Peseta T. A socially just curriculum reform agenda. Teach Higher Educ. 2016;21(4):361–366. http://dx.doi.org/10.1080/13562517.2016.1159057
9. Morrow W. Bounds of democracy: Epistemological access in higher education. Cape Town: Human Sciences Research Council; 2009.
10. Muller J. Every picture tells a story: Epistemological access and knowledge. Educ as Change. 2014;18(2):255–269. http://dx.doi.org/10.1080/16823206.2014.932256
11. South African Department of Higher Education and Training (DHET). 2000-2008 First time entering undergraduate cohort studies for public higher education institutions. Pretoria: DHET; 2016.
12. South African Council on Higher Education (CHE). Vital stats public higher education 2013. Pretoria: CHE; 2015.
13. South African Council on Higher Education (CHE). A proposal for undergraduate curriculum reform in South Africa: The case for a flexible curriculum structure. Pretoria: CHE; 2013.
14. Schwartzman S. Massification, equity and quality: Challenges of higher education in Brazil – Period analysis 2009–2013. In: Brunner J, Villalobos C, editors. Educacion superio policies in Ibero-America [Policies of higher education in Latin America]. Santiago: Ediciones Universidad Diego Portales; 2015 p. 199–243. Spanish.
15. Marginson S. The worldwide trend to high participation higher education: Dynamics of social stratification in inclusive systems. High Educ. 2016;72:413–434. http://dx.doi.org/10.1007/s10734-016-0016-x
16. Swchwartzman S. Student quotas in Brazil: The policy debate. Int Higher Educ. 2009;56:11–13.
17. South African Department of Basic Education (DBE). 2015 National Senior Certificate Examination: Examination report. Pretoria: DBE; 2015.
18. NBTP National report: 2015 intake cycle. Cape Town: University of Cape Town; 2015.
19. Cloete N. For sustainable funding and fees, the undergraduate system in South Africa must be restructured. S Afr J Sci. 2016;112(3/4), Art. #a0146, 5 pages. http://dx.doi.org/10.17159/sajs.2016/a0146
20. Pillay P. Research and innovation in South Africa. In: Swchwartzman S, Pinheiro R, Pillay P, editors. Higher education in the BRICS countries: Investigating the pact between higher education and society. Dordrecht: Springer; 2015. p. 463–485. http://dx.doi.org/10.1007/978-94-017-9570-8_23
21. McCowan T. Expansion without equity: An analysis of current policy on access to higher education in Brazil. High Educ. 2007;53:579–598. http://dx.doi.org/10.1007/s10734-005-0097-4
22. De Magalhaes Castro M. Higher education policies in Brazil: A case of failure in market regulation. In: Swchwartzman S, Pinheiro R, Pillay P, editors. Higher education in the BRICS countries: Investigating the pact between higher education and society. Dordrecht: Springer; 2015. p. 271–289. http://dx.doi.org/10.1007/978-94-017-9570-8_14
23. Neves CE. Demand and supply for higher education in Brazil. In: Swchwartzman S, Pinheiro R, Pillay P, editors. Higher education in the BRICS countries: Investigating the pact between higher education and society. Dordrecht: Springer; 2015. p. 73–96. http://dx.doi.org/10.1007/978-94-017-9570-8_5
24. Cooper D. Social justice and South African university student enrolment data by 'race', 1998-2012: From 'skewed revolution' to 'stalled revolution'. High Educ Quart. 2015;69(3):237–262. http://dx.doi.org/10.1111/hequ.12074
25. DeAngelo L, Franke R, Hurtado S, Pryor J, Tran S. Completing college: Assessing graduation rates at four-year institutions. Los Angeles, CA: Higher Education Research Institute, UCLA; 2011.
26. Scearce D, Fulton K. What if? The art of scenario thinking for nonprofits. Emeryville, CA: Global Business Network; 2004. Available from: http://www.monitorinstitute.com/downloads/what-we-think/what-if/What_If.pdf
27. Bawa AC. Righting an inverted pyramid: Managing a perfect storm. Alternation Special Edition. 2013;9:25–45.
28. Tinto V. Dropout from higher education: A theoretical synthesis of recent research. Rev Educ Res. 1975;45(1):89–125. http://dx.doi.org/10.3102/00346543045001089
29. Rooney C. Using survival analysis to identify the determinnats of academic exclusion and graduation in three faculties at UCT. Cape Town: University of Cape Town; 2015.
30. Shay S, Wolff K, Clarence-Fincham J. New generation extended curriculum programmes: Report to DHET. Cape Town: University of Cape Town; 2016.
31. Smith L. Measuring the impact of educational interventions on the academic performance of black academic development students. South Afr Rev Educ. 2012;18(1):85–113.

Estimation of household income diversification in South Africa: A case study of three provinces

AUTHORS:
Jabulani Mathebula[1]
Maria Molokomme[2]
Siyanda Jonas[2]
Charles Nhemachena[3]

AFFILIATIONS:
[1]Department of Agricultural Economics and Animal Production, University of Limpopo, Polokwane, South Africa
[2]Economic Performance and Development, Human Sciences Research Council, Pretoria, South Africa
[3]International Water Management Institute, Pretoria, South Africa

CORRESPONDENCE TO:
Jabulani Mathebula

EMAIL:
jabulani.hazel@gmail.com

DATES:
Received: 09 Mar. 2016
Revised: 05 July 2016
Accepted: 15 Aug. 2016

KEYWORDS:
income diversification; livelihood diversification; poverty; unemployment; inequality

HOW TO CITE:
Mathebula J, Molokomme M, Jonas S, Nhemachena C. Estimation of household income diversification in South Africa: A case study of three provinces. S Afr J Sci. 2017;113(1/2), Art. #2016-0073, 9 pages. http://dx.doi.org/10.17159/sajs.2017/20160073

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
None

We estimated household income diversification in settlement types of the poorest provinces in South Africa – the Eastern Cape, Limpopo and KwaZulu-Natal. We obtained data from the 2010/2011 Income and Expenditure Survey from Statistics South Africa and Wave 3 data from the National Income Dynamics Study. We used the number of income sources, the number of income earners and the Shannon Diversity Index to estimate income diversification in the study provinces. The results show that households in the traditional and urban formal areas diversified income sources to a greater extent than households in urban informal and rural formal settlements. The varied degrees of income diversification in the three provinces suggest that targeted policy initiatives aimed at enhancing household income are important in these provinces.

Significance:

- Indices yet to be used in South Africa were used in the analysis of StatsSA data to understand income diversification.
- Poverty is mostly concentrated in the traditional areas and urban informal areas.
- Households in the traditional areas and urban informal areas derive livelihood mostly from social transfers and remittances, whereas those in the urban formal areas derive income from business, labour income and financial capital returns.

Introduction and problem statement

Livelihood diversification is increasingly seen as one of the pathways for poverty reduction and economic growth in sub-Saharan Africa.[1] Asset, activity and income diversification characterises the livelihood strategies of many rural communities in sub-Saharan Africa. Income diversification refers to an increase in the number of sources of income, or the balance between different sources.[2] There are multiple motives which prompt households or individuals to diversify assets and income-earning activities. These motives include the need to increase income to provide a sufficient livelihood[3] – and to reduce risk and even out consumption – because reliance on one source of income increases the risk of destitution and prevents achievement of economies of scope.[4] Thus, a household with two sources of income would be more diversified than a household with only one source[5], and a household with two income sources, each contributing half of the combined total, would be more diversified than a household with two sources with one income accounting for 90% of the total[6]. Income diversification is a component of livelihood diversification, which is a process by which households construct a diverse portfolio of activities and social support capabilities in order to improve their living standards and manage risk.[6]

Barrett et al.[2] further classifies motives for diversification into 'pull and push' factors. Pull factors are those which are related to risk reduction, whereas push factors are related to a need to expand the line of production in order to produce complementary goods. Unemployment is one of the pull factors which somehow compels households to diversify livelihood activities for the provision of sustenance. In South Africa, households or individuals diversify incomes to overcome the consumption challenges made worse by the high unemployment rate, which is amongst the highest in the world, with over 25% of the labour force unemployed.

The challenges of unemployment in South Africa are also compounded by high levels of poverty and inequality. South Africa is one of the countries which have a high Gini coefficient – the third highest in Africa after Namibia and the Seychelles.[7] Since the advent of democracy in South Africa in 1994, various policies have been implemented to address these challenges. These include the Reconstruction and Development Programme (RDP) of the early 1990s; the Growth, Employment and Redistribution (GEAR) strategy in 1996; and the Accelerated and Shared Growth Initiative for South Africa (ASGISA) in 2005. A recent strategy is the National Development Plan (NDP) Vision 2030 – South Africa's long-term socio-economic development roadmap. In spite of all these strategies, South Africa seems not to have achieved the intended objectives of reducing poverty and inequality. It is, therefore, important to understand the livelihood strategies which households adopt in their struggle for survival and to improve incomes.

Studies on income diversification in developing countries, particularly in Africa, have concentrated on income diversification of households who are already participating in agriculture and seek to diversify within, and outside, agriculture.[8-10] There is scarce literature, however, on income diversification and the variation of income sources in South Africa.

Efforts to measure income diversification in southern African countries include[6,11-13] the work of Chitiga-Mabugu et al., on which this paper builds, who analysed the profile of poverty in the nine provinces of South Africa. Chitiga-Mabugu et al.[13] used the Foster, Greer and Thorbecke family of poverty indices to measure important indicators of poverty (incidence, depth and severity of poverty). Although the study used poverty incidence to present poverty by income sources, one limitation was that household income diversification and the degree (scatteredness) of diversification was not explored.

Therefore, this study contributes to the existing knowledge of income diversification by analysing household income diversification and the degree of diversification in the three poorest provinces of South Africa, as identified in the study by Chitiga-Mabugu et al.[13]: the Eastern Cape, Limpopo and KwaZulu-Natal. We used the number of income sources (NIS) and the number of income earners (NYE) to estimate household income diversification, and the Shannon Equitability Index (SEI) to account for the degree of diversification. These indices, to the best of our knowledge, are yet to be used in South Africa to understand income diversification.

Review of income diversification studies

Livelihood diversification is a process involving the maintenance and continuous variation of a highly diverse portfolio of activities over time in order to secure survival and improve standards of living. Livelihood diversification has been coupled to the diversification of rural economies. The diversification of rural economies in sub-Saharan Africa has followed a different trajectory from those in Asia and Europe,[14] but this does not necessarily mean that it has not taken place. Over the years, the diversification of small-holder rural economy in sub-Saharan Africa has been underpinned by household and livelihood diversification. Hilson[15] traces smallholder agriculture and rural household diversification patterns over a period of structural adjustment, during which households experienced immense suffering. He argues that, during this time, a delicate balance between agriculture and off-farm activities existed.

It is largely within the context of smallholder rural economy diversification that the patterns of household and individual diversification became visible, and received scholarly attention. Patterns of rural livelihood diversification are characterised by the variation of activities which can be categorised by sector (farm activities and non-farm activities, or agricultural activities and non-agricultural activities); by function (wage employment activities and self-employment, depending on how labour is compensated); and by location (on-farm and off-farm activities, depending on where the activity takes place).[14-16] In addition, Hosu and Mushunje[17] highlight that on-farm diversification, such as a combination of crop and livestock, can raise incomes and mitigate against risk.

It is clear that rural households avert risk and respond to shock through diversification of their livelihoods. Rural households initially engaged in diversification of their income sources as a coping or risk aversion strategy and to accumulate wealth or assets to reduce household level uncertainty.[15] The motives for livelihood diversification can be characterised as push (e.g. risk aversion or coping strategy) or pull factors (e.g. wealth accumulation strategy). Loison[14] further categorises motives for diversification as survival-led and opportunity-led diversification. Survival-led diversification is mainly driven by push factors and occurs when poorer rural households engage in low-return activities to ensure survival, reduce vulnerability or avoid falling deeper into poverty. Opportunity-led diversification is mainly driven by pull factors and it occurs when wealthier rural households engage in high-return non-farm activities, with accumulation objectives, in order to increase household income by maximising returns from their assets.

A number of studies has focused on the diversification between agricultural activities and non-agricultural activities of smallholder farmers. A key assumption of these studies is that rural communities are agriculturally based economies. In spite of the appeal of this assumption, the viability of smallholder farming has decreased and by default has pushed unskilled labour to the non-agricultural sector.[18,19] A decrease in the size of farms as well as the inability to produce a sufficient crop yield for the market place increases pressure on households to shift to participation in non-agriculture.[15-19] Coupled with land constraints resulting from increased population concentration, this situation has led to the development of varied patterns of diversification strategies under new settlement typologies.

Empirical studies on income diversification

Measuring household income diversification is important in a number of ways. It facilitates the comparison of urban and rural household income sources,[13] the understanding of income diversification of poor and better-off households, and the elucidation of the underlying factors influencing household income diversification. Empirical literature explores the concept of household income diversification from a number of different perspectives, each with varying findings.

Schwarze and Zeller[20] examined two aspects of income diversification in Indonesia. The first aspect was diversification as a shift away from agricultural activities, and the second was diversification as an increasing mix of income-generating activities. They used the SEI to measure income diversification and the Tobit model to analyse the determinants of income diversification. The results of their study showed that the degree of participation in agricultural activities and non-agricultural activities differs. Wealth was found to increase diversification outside agriculture, and income of poor households seemed to be generated from different sources and to be evenly distributed between the sources.

Ersado[6] examined changes in and welfare implications of income diversification in Zimbabwe using the NIS which is a relatively easy measure of income diversification. The findings showed that households with a more diversified income base were better able to withstand the unfavourable impacts of policy changes and could more easily weather shocks. However, the weakness of NIS as a measure is the assumption that if there are adult members in the household, the number of sources of income increases.[13] The study by Ersado[6] addressed this limitation by using the NYE instead of the number of adults in a household. Ersado[6] then used the inverse of the Herfindahl Index to calculate the scatteredness of income sources.

Fausat[9] examined the determinants of income diversification in rural farming households in Nigeria. The study used multiple regression analysis to analyse the determinants of income diversification among farming households in Borno State. Fausat[9] estimated the impact of age of the respondent, education level of the household head, ownership of assets, household size, access to loans and marital status on income diversification. It was expected that the educational level of the household head, ownership of assets and age would have positive relationships with income diversification, whereas access to loans, household size and marital status would have a negative relationship with household consumption, age and ownership of assets, when conformed to the expected outcome. On the contrary, household size, access to loans and marital status did not predict the theoretical postulations.

In another study in Nigeria, Adebayo et al.[10] applied the Tobit regression model to identify determinants of income diversification among farm households. They regressed socio-economic variables on the income diversification index. The results showed that non-farm income was a major determinant of the income diversification strategy of farm households. The coefficient of education was positive, showing that a high level of education raises income diversification. An increase in farm size will, other factors being equal, generate additional income. Conversely, a farming household is likely to reduce other non-farm activities. Membership of cooperatives also increases income diversification because it increases access to credit.

The role of Civil Society Organisations in livelihood diversification in South Africa was assessed by Chitiga-Mabugu et al.[21] The Civil Society Organisations participated in six income-generating activities: agricultural production (crops and livestock), agricultural wage employment, non-agricultural wage employment, non-farm enterprises, social transfers, and non-labour employment. These activities were important in providing additional benefits which included contributing to reducing poverty, improving the well-being as well as empowerment of the communities, self-reliance and community development.

A study conducted by Alemu[22], which identified dominant livelihood activities in South Africa, is the most relevant for use as a baseline of livelihood activities in South Africa. Unlike in the previous studies, Alemu made use of a more recent data set (2009 General Household Survey) to calculate the dominance of livelihood activities; a first-order stochastic dominance test was applied and multinomial logistic regression was used to identify factors constraining household entry into high-income earning activities. The livelihood activities were ranked in order of their

dominance: only non-farm wage earners, farm and non-farm wage earners, farm and non-farm non-wage earners, pensioners, only non-farm non-wage earners, remittances and social grants. The study found that various factors – such as age and gender of the household head, human capital and social infrastructure – influenced the chances of entry into high-income earning activities.

Similar studies were conducted in the Eastern Cape and Limpopo Provinces by Perret et al.[11,12] These studies were undertaken in communities of the former Transkei in the Eastern Cape, and in the communities of Ga-Makgato and Sekgopo in Limpopo Province, with the objective of understanding the different livelihood systems people develop over time. The results in the Eastern Cape confirmed that diversity was a major trait of local livelihood systems, in which pensions and remittances were major sources of income and farming contributed to income for only a small proportion of households. The majority of households in Limpopo benefitted from social grants in the form of childhood allowances and old-age pensions. Fewer households benefitted from employment wages in Sekgopo than in Makgato, while a small proportion in Sekgopo benefitted from farming. There was a dramatic drop in the number of households benefitting from remittances and farming income in Limpopo, when compared to the study conducted previously by Barber[23] in another two communities in this province.

Livelihood diversification has been researched internationally and in other parts of southern Africa; however, there is minimal evidence on the variation of diversification into different income sources. This study adds to existing knowledge by showing the differences in the levels of diversification in four settlement types (urban formal, urban informal, traditional and rural) in three provinces in South Africa.

Methodology, data and variables

Empirical model for measuring household income diversification

Studies on income diversification have adopted measures used in various disciplines to evaluate the scatteredness of individual or household income sources. Block and Webb[24] used the inverse of the Herfindahl Index to calculate income diversification. The Herfindahl Index is a measure of market concentration. Ersado[6] also adopted the Herfindahl Index to elaborate on the scatteredness of household income sources.

The Gini coefficient is also used to calculate income diversification. This measure is mostly used in income distribution studies. The Gini coefficient measures the area under the Lorenz curve as a complementary proportion of the area that would be captured were the variable (e.g. assets, activities, income) perfectly equally distributed. So, a value of zero represents perfect equality in income distribution studies. The disadvantage of the Gini coefficient is its computational complexity.[25] Zhao and Barry[25] employed numerical integration techniques to derive a reasonably accurate discrete approximation to the true Gini coefficient.

The advantage of the Herfindahl Index, in comparison to the Gini coefficient, is its computational simplicity. The Herfindahl Index is the sum of squared shares where *i* is income sources and *S* represents shares. Other studies have used measures equivalent to the Herfindahl Index, like the SEI and Simpson Index. The Simpson Index was adopted from agronomy and geology studies, and is simply the sum of squared levels divided by the squared total. The Simpson Index is the same as the Herfindahl Index and the SEI as they also estimate the evenness of the incomes.

To measure income diversification in the three provinces, the NIS – a relatively easy to use index – was used. The NIS involves accounting for the actual household incomes from various sources. Despite the simplicity of measurement, it has been criticised for its arbitrariness. For instance, it assumes that households with more economically active adults would have more income sources.[14] To overcome this weakness, we used the number of per capita sources and the number of household members. Ersado[6] also used these approaches in similar settings.

To measure the degree of household income diversification (scatteredness), we applied the SEI, a commonly used measure of diversification, which is derived from the Shannon Diversity Index (SDI). This index is used in biodiversity studies to reflect how many different types of species are in a data set, and simultaneously takes into account how evenly the basic entities (such as individuals) are distributed among those types.[26] The SDI (*H*) is expressed as follows:

$$H_{income} = -\sum_{i=1}^{s} [(incshare_i).\ln(incshare_i)], \qquad \text{Equation 1}$$

where *S* is the number of income sources and $incshare_i$ is the share of income from activity *i* in total household income. SEI takes into account the evenness of the income sources, with the values 0 and 1 representing complete evenness. Based on this index *H*, the SEI (*E*) is calculated as:

$$E = \left(\frac{H_{income}}{\sum_{i=1}^{s} \left(\frac{1}{s} . \ln \left(\frac{1}{s} \right) \right)} \right) X100, \qquad \text{Equation 2}$$

where the denominator is the maximal possible SDI and *E* ranges from 0 to 100 and reflects the percentage share of the actual income diversification in relation to the maximal possible diversity of income.

The measures of income diversification can also be classified into dimensions. Zhao and Barry[25], in their endeavour to identify income diversification measures which better represent rural household income diversification in China, noted that diversification can be divided into one-dimensional and two-dimensional measures. One-dimensional measures comprise counts of the number of business activities or evaluate changes in the volumes of different divisions, whereas two-dimensional measures consider both the number of areas of activities and their relative volumes of turnover.

Both one-dimensional and two-dimensional measures were used in this study. NIS and NYE are one-dimensional measures and the SDI is a two-dimensional measure because it goes beyond counting the income sources to including shares from each source. These measures were used to support the assessment of income diversification of both poor and better-off households.

Data sources and variables

The Income and Expenditure Survey (IES) of 2010/2011, produced by Statistics South Africa (StatsSA)[27], was used to estimate income diversification in the provinces of Limpopo, the Eastern Cape and KwaZulu-Natal. The IES's primary objective was to provide relevant statistical information on household consumption expenditure patterns that inform the updating of the Consumer Price Index (CPI) basket of goods and services. Moreover, the IES also encompasses the individual incomes and household characteristics which were used in this study. Diary and recall methodology was employed in the collection of the data. The sample size was 31 419 dwelling units in 2010/2011. Table 1 gives a description of the variables used in this study. Questions 1.6 and 1.7 from the IES were mostly used in the classification of the variables and calculations.

Two data files from IES were merged prior to the analysis: person information and person income. A total of 96 281 persons was recognised across all nine provinces. Limpopo, Eastern Cape and KwaZulu-Natal contributed 42 312 to this total. As the unit of analysis was a 'household', the data were reshaped to represent household data. The realised households were 25 328 and the three provinces constituted 10 264 households.

The main data source was IES; however, because of challenges related to the structure of some questions and responses in the data source, such as lack of continuous income responses, it was difficult to measure the degree of income diversification in the respective provinces. To address this challenge, we used the Wave 3 data set from the National Income Dynamic Study to estimate the degree of diversification, because of its richness in continuous income data. The weakness of this data set is that it was not possible to disaggregate to settlement type because of differences with that of the IES 2010.

Table 1: Sources of income and their descriptions

Variables (income sources)	Source of variables	Definition of variables
Business	Statistics South Africa	Net profit from business or professional practice/activities or commercial farming; royalties and income from letting of fixed property
Labour income	Statistics South Africa	Salaries and wages
Subsistence farming	Statistics South Africa	Income from subsistence agricultural production
Financial capital return	Statistics South Africa	Interest received and/or accrued on deposits, loans, savings certificates, dividends on shares other than building society shares and regular receipts from pension from previous employment and pension from annuity funds
Social transfers	Statistics South Africa	Social welfare grants including old-age pension
Remittances	Statistics South Africa	Alimony, maintenance and similar allowances from divorced spouse, family and non-household members
Other income	Statistics South Africa	Unspecified income

Source: StatsSA[27]

Settlement types

The four settlement types distinguished in the study are defined as follows:

1. Urban formal – non-metropolitan urban areas that include secondary and tertiary towns[28], for example Nelspruit and Polokwane.
2. Urban informal – settlements on the peri-urban fringe[27], for example Soweto and Gugulethu.
3. Traditional areas (former homelands) – areas that were created during the apartheid era to house black populations to prevent them from living in urban areas[27], for example Transkei and Venda.
4. Rural areas – sparsely populated areas in which people farm or are dependent on natural resources, including dispersed villages and small towns. These areas can also include larger settlements from the former homelands, which are dependent on migratory labour and remittances as well as government grants for survival[27], for example Hlankomo and Mdeni.

Poverty profile of the study provinces

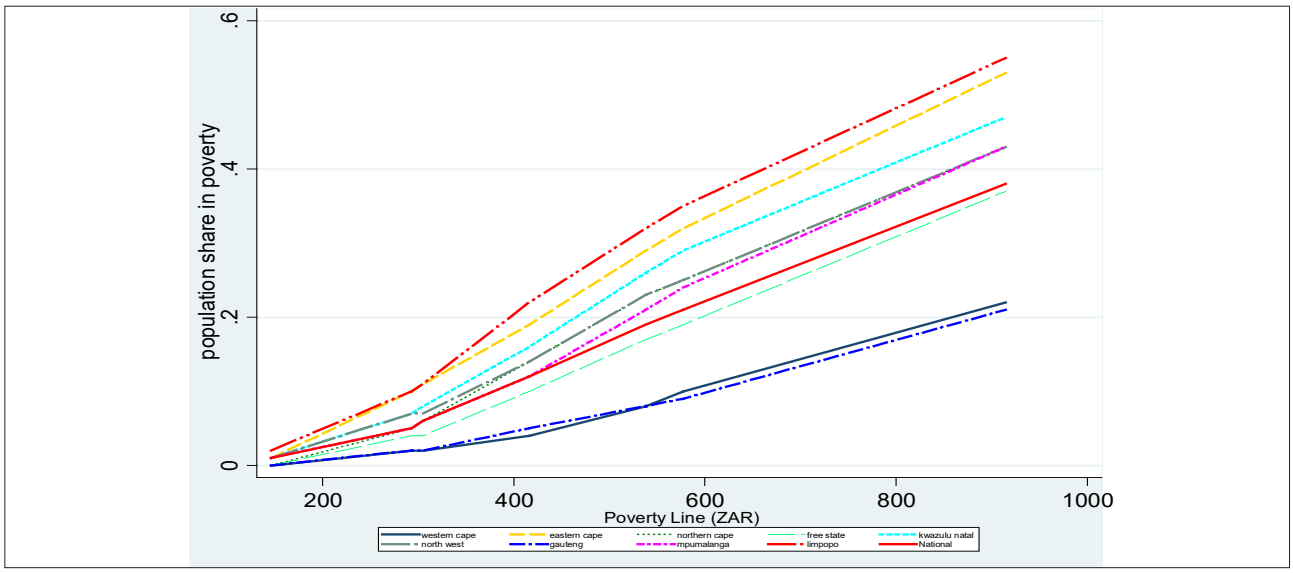
StatsSA released inflation adjusted poverty line types to be used for poverty measurements in the country in 2012. These poverty lines are the food poverty line and lower and upper bound poverty lines. Table 2 shows the poverty lines applicable in South Africa.

Table 2: Poverty lines in South Africa

Poverty line type	Value†
Food poverty line	ZAR305 per person per month
Lower bound poverty line	ZAR416 per person per month
Upper bound poverty line	ZAR577 per person per month

†Published in March 2009

The Human Sciences Research Council's report on the state of poverty and its manifestation in South Africa[13] – which is the premise of this study – applied the upper bound poverty line to analyse the state of poverty in the nine provinces of South Africa. Figure 1 plots the poverty incidence for South Africa and the nine provinces by poverty line. The figure indicates that, irrespective of the choice of poverty line, poverty comparisons across provinces remain consistent. Only three provinces have a poverty incidence below the national average for all the poverty lines. These provinces are Gauteng, Western Cape and Free State, in order of increasing poverty incidence. The poorest provinces are Limpopo, the Eastern Cape and KwaZulu-Natal, in order of decreasing poverty incidence.



Source: State of poverty and manifestation in the nine provinces of South Africa, Human Sciences Research Council

Figure 1: Poverty incidence sensitivity to poverty lines in 2010.

Table 3 shows conditions of poverty in the settlement types of the three provinces. In Limpopo, 39% of poor households reside in traditional areas and 37% reside in rural formal settlements. In the Eastern Cape, again 39% reside in traditional areas, followed by 36% residing in urban informal areas. In KwaZulu-Natal, 46% reside in traditional areas followed by 35% in urban informal areas. These results show that poverty occurs mostly in traditional and urban informal areas.

Table 3: Poverty by settlement type in Limpopo, Eastern Cape and KwaZulu-Natal

Province	Settlement type	Poverty incidence	Poverty intensity	Poverty severity
Limpopo	Urban formal	0.13	0.04	0.04
	Urban informal	0.08	0.03	0.03
	Traditional area	0.39	0.15	0.15
	Rural formal	0.37	0.12	0.12
Eastern Cape	Urban formal	0.24	0.09	0.09
	Urban informal	0.36	0.12	0.12
	Traditional area	0.39	0.14	0.14
	Rural formal	0.14	0.04	0.04
KwaZulu-Natal	Urban formal	0.11	0.03	0.03
	Urban informal	0.35	0.11	0.11
	Traditional area	0.46	0.16	0.16
	Rural formal	0.29	0.10	0.10

Source: State of poverty and manifestation in the nine provinces of South Africa, Human Sciences Research Council

Table 4 presents the distribution of income sources in the households. The results show that 53% of households in urban formal settlements in Limpopo relied on income from business and labour (including wage and salaries). In addition, 60% of these households relied on financial capital return. In the traditional areas, 82% and 75% of the households relied on social transfers and remittances, respectively. Households across all the settlement types seemed to rely less on subsistence farming than on other income sources.

In the Eastern Cape, 83% of households in urban formal settlements relied on income from financial capital return. The results also indicated that 53% of households in urban formal settlements relied on income from business activities and labour, while 50% relied on income from subsistence farming. In traditional areas in the Eastern Cape, 54% and 48% of households relied on social transfers and remittances, respectively. These percentages are lower than those of households in Limpopo for the same sources.

In urban formal settlements in KwaZulu-Natal, 50% of households earned income from business and labour activities and 77% and 82% of households reported income from financial capital returns and other income, respectively. In traditional settlements, 48% and 46% of households reported income from social transfers and remittances, respectively. In rural settlements of KwaZulu-Natal, 12% of households reported income from financial capital return.

In the rural areas of both the Eastern Cape and Limpopo Provinces, income earned from other sources was less than 11%.

Results and discussions

The main income sources of members of households in all three provinces are shown in Table 5. In urban formal settlements in Limpopo, the main sources of income of the household members were: social transfers (86%), financial capital return (45%) and labour income (38%). In contrast, in the urban informal settlements, the main sources of income were: subsistence farming (88%), other income (74%) and financial capital return (55%). In the traditional settlements, the main income sources were: remittances (83%), business (74%) and labour income (56%). In rural areas, social transfers and labour income were the highest sources of income at 4.24% and 2.8%, respectively.

Table 4: Distribution of income sources across the three provinces

Province	Settlement type	Income source						
		Business	Labour income	Subsistence farming	Financial capital return	Social transfers	Remittances	Other income
Limpopo (n=3.306)	Urban formal	53.23	52.94	0	60	13.1	21.81	12.5
	Urban informal	4.84	4.81	0	0	2.07	1.91	0
	Traditional area	36.02	36.36	100	40	82.07	74.59	87.5
	Rural formal	5.91	5.88	0	0	2.76	1.69	0
Eastern Cape (n=3.333)	Urban formal	52.81	52.49	50	82.61	40	45.27	20
	Urban informal	7.82	7.71	0	0	4.88	4.92.	10
	Traditional area	34.72	35.07	50	17.39	54.15	48.48	70
	Rural formal	4.65	4.73	0	0	0.98	1.32	0
KwaZulu-Natal (n=3.625)	Urban formal	49.8	49.7	0	76.92	41.21	40.17	81.82
	Urban informal	18.73	18.76	0	0	6.06	10.37	9.09
	Traditional area	21.12	21.16	0	11.54	48.48	45.57	9.09
	Rural formal	10.36	10.38	0	11.54	4.24	3.94	0

Source: StatsSA[27]

Table 5: Main source of income brought by household members in the households

	Urban formal	Urban informal	Traditional area	Rural formal
Limpopo				
Business	24.62	0.51	74.37	0.51
Labour income	38.17	3.20	55.83	2.80
Subsistence farming	12.50	87.50	0.00	0.00
Financial capital return	44.62	55.38	0.00	0.00
Social transfers	85.81	8.96	0.99	4.24
Remittances	15.18	1.40	83.07	0.35
Other income	26.09	73.91	0.00	0.00
Eastern Cape				
Business	0.20	19.69	71.22	8.88
Labour income	0.20	19.69	71.22	8.88
Subsistence farming	25.00	75.00	0.00	0.00
Financial capital return	70.86	0.66	27.81	0.66
Social transfers	89.28	8.69	0.16	1.87
Remittances	24.74	1.72	73.20	0.34
Other income	25.71	5.71	68.57	0.00
KwaZulu-Natal				
Business	51.60	11.60	33.09	3.70
Labour income	0.12	28.54	62.99	8.36
Subsistence farming	15.38	84.62	0.00	0.00
Financial capital return	68.82	1.08	23.66	6.45
Social transfers	73.36	17.41	0.20	9.02
Remittances	34.40	6.00	58.00	1.60
Other income	53.13	6.25	40.63	0.00

Source: StatsSA[27]

In the Eastern Cape, the main sources of income reported by individuals in in urban formal settlements were: social transfers (89%), financial capital return (71%) and other income (26%). Similar to Limpopo Province, in the urban informal settlements, the main source of income was subsistence farming (75%). In traditional settlements, the main sources of income were remittances (73%), labour income and business (both 71%) and other income (69%). In rural areas, business and labour income contributed the most to income, at 8.8% each.

Households in urban formal settlements in KwaZulu-Natal reported the following main income sources: social transfers (73%), financial capital return (69%), other income (53%) and business (52%). In the urban informal settlements, subsistence farming was again reported as the main income source (85%). The main sources of income reported in the traditional settlements were: labour income (63%), remittances (58%) and other income (41%). In rural areas, social transfers contributed 9.02% to total household income.

The results from the three provinces indicate that high proportions of individuals from urban formal settlements received income from social transfers (73% to 89%). In the traditional areas, the main source of income reported across the three provinces was remittances, while in the urban informal settlement types, subsistence farming was reported as the main source of income. It is important to note that because the incomes earned from these sources could be insufficient to provide household necessities, these households are most likely to diversify incomes to complement the main source of income (mostly earned by the household head).

Table 6 presents the results based on the NIS in households. The largest proportion of households with no source of income was reported in the traditional settlement type (68%) for Limpopo Province; while in the Eastern Cape and KwaZulu-Natal Provinces, households reporting no source of income represented 53% and 48%, respectively, of households in urban formal settlements, and 40% and 36%, respectively, of households in traditional settlements. Furthermore, the largest proportion of households with at least two income sources was for households in traditional settlements in Limpopo (60%) and households in urban formal areas in the Eastern Cape (58%) and KwaZulu-Natal (51%). Some households in traditional areas in all three provinces had diversified into three income sources, with a few in urban formal areas also diversifying into three income sources. Households which diversified into four sources of income were mostly in traditional areas, followed by those in urban formal areas in Limpopo Province, and (equally) by both urban formal and urban informal areas in KwaZulu-Natal and the Eastern Cape. This diversification could be driven by the small incomes from various sources and high level of poverty, especially in the traditional settlement type, which pushes households to diversify their income sources. Households in the rural formal and urban informal areas did not diversify income as much as households in traditional and urban formal areas. Perret et al.[11] also found diversification in the Eastern Cape Province among poor households.

Table 6: Distribution of household by the number of income sources (NIS)

NIS	Limpopo				Eastern Cape				KwaZulu-Natal			
	Urban formal	Urban informal	Traditional area	Rural formal	Urban formal	Urban informal	Traditional area	Rural formal	Urban formal	Urban informal	Traditional area	Rural formal
0	26.83	1.63	67.89	3.66	52.43	7.03	40.00	0.54	47.77	12.96	36.03	3.24
1	9.05	0.74	89.85	0.37	31.36	2.77	65.72	0.15	25.55	6.24	64.59	3.63
2	34.74	2.61	60.33	2.32	57.69	5.45	34.64	2.21	50.73	11.24	33.61	4.42
3	16.21	4.14	77.24	2.41	46.78	10.17	40.00	3.05	31.68	16.75	48.43	3.14
4	23.08	0.00	76.92	0.00	6.67	6.67	86.67	0.00	21.43	21.43	50.00	7.14
Total	21.84	1.89	74.66	1.61	45.31	4.86	48.51	1.33	40.14	10.34	45.56	3.95

Source: StatsSA[27]

Table 7: Distribution of household by number of income earners

Income sources	Urban formal			Urban informal			Traditional area			Rural formal			Total
	Children	Youth	Over 35	Children	Youth	Over 35	Children	Youth	Over 35	Children	Youth	Over 35	
Limpopo													
0	30.95	23.59	7.71	3.07	1.86	0.10	0.18	0.13	29.23	1.86	1.01	0.30	100
1	1.69	7.80	13.91	0.08	0.80	0.88	15.84	58.06	0.12	0.24	0.56	0.00	100
2	0.05	17.02	21.21	1.25	1.36	0.05	22.19	34.48	0.00	1.03	1.36	0.00	100
3	6.31	9.46	1.58	2.84	29.02	48.58	0.00	0.00	0.00	0.95	1.26	0.00	100
4	7.69	15.38	0.00	0.00	23.08	53.85	0.00	0.00	0.00	0.00	0.00	0.00	100
Eastern Cape													
0	0.12	0.10	40.39	14.91	8.84	2.69	0.23	0.14	28.25	2.25	1.47	0.61	100
1	4.49	13.78	45.13	0.25	1.62	3.80	1.87	27.93	0.07	0.37	0.12	0.56	100
2	25.00	38.00	2.00	3.00	9.00	20.00	0.00	0.00	0.00	1.00	2.00	0.00	100
3	14.55	33.64	3.03	6.67	15.15	24.24	0.00	0.00	0.00	1.21	1.52	0.00	100
4	0.00	11.76	0.00	5.88	29.41	52.94	0.00	0.00	0.00	0.00	0.00	0.00	100
KwaZulu-Natal													
0	0.07	0.06	23.97	21.61	16.11	3.49	0.16	0.10	20.78	7.81	4.37	1.47	100
1	0.90	13.10	34.41	0.06	4.32	5.36	1.29	34.86	0.07	1.87	3.74	0.00	100
2	0.08	24.41	29.18	6.94	5.07	13.46	17.05	0.00	0.00	2.21	1.60	0.00	100
3	14.84	16.67	9.82	7.99	20.55	26.48	0.00	0.00	0.00	1.83	1.83	0.00	100
4	0.00	21.43	7.14	14.29	7.14	42.86	0.00	0.00	0.00	0.00	7.14	0.00	100

Source: StatsSA[27]

The proportions of income earners across households by age group are presented in Table 7. In traditional settlements in Limpopo, 58% of youth relied on one source of income. In urban formal settlements in Limpopo, 24% of the youth had no income compared to 8% for those above 35 years old. This finding is not surprising as there is a relatively higher youth unemployment in South Africa. In the urban informal settlements in Limpopo, 49% of those above 35 years of age and 29% of the youth had three income sources. In traditional areas in Limpopo, 34% of the youth had two income sources, while 29% of those above 35 years had no income. As a largely rural province, subsistence farming is one of the livelihoods of those above 35 years of age.

The Eastern Cape has one of the highest unemployment rates (38%). In urban formal settlements in the Eastern Cape, 38% of the youth had two income sources, while a striking 2% of those above 35 years had two income sources. A similar trend is observed in traditional areas, where 28% of those below 35 years had at least one source of income and 28% of those above 35 years had no income.

In urban formal settlements of KwaZulu-Natal, 29% of those above 35 years had two sources of income, and of those who were below 35 years, 24% had two sources of income. A similar trend was observed in urban informal settlements with 26% of those above 35 years having three sources of income compared to 21% of the youth.

Table 8: Shannon Equitability Index

Income source	Income from activities (ZAR)	Share of income	Shannon Diversity Index	Shannon Equitability Index
Limpopo				
Labour income	1 750 629.64	56.00	-0.325	69.21
Subsistence farming	23 562.47	0.75	-0.037	
Financial capital return	514 308.85	16.45	-0.297	
Social transfers	646 048.00	20.67	-0.326	
Remittances	186 367.54	5.96	-0.168	
Other	5015.00	0.16	-0.010	
Total	3 125 931.50	100.00	1.163	
Eastern Cape				
Labour income	1 714 791.82	47.90	-0.353	40.25
Subsistence farming	33 029.13	0.92	-0.043	
Financial capital return	677 090.64	18.91	-0.315	
Social transfers	920 350.00	25.71	0.066	
Remittances	209 816.30	5.86	0.003	
Other	25 227.00	0.70	-0.035	
Total	3 580 304.89	100.00	0.676	
KwaZulu-Natal				
Labour income	4 712 026.90	53.82	-0.333	68.99
Subsistence farming	46 043.89	0.53	-0.028	
Financial capital return	1 482 333.91	16.93	-0.301	
Social transfers	2 088 271.00	23.85	-0.342	
Remittances	409 879.92	4.68	-0.143	
Other	16 731.00	0.19	-0.012	
Total	8 755 286.61	100.00	1.159	

Source: Calculated using Wave 3, National Income Dynamics Study 2010

The results of the SEI analyses are shown in Table 8. Total household income ranged from ZAR3 million in Limpopo to nearly ZAR9 million in KwaZulu-Natal in 2010. Labour income constituted a high share of the total household income in all three provinces at 56%, 48% and 54% in Limpopo, the Eastern Cape and KwaZulu-Natal, respectively. Social transfers were the second source of income which contributed significantly to the total household income in all three provinces, at 21%, 26% and 24% in Limpopo, the Eastern Cape and KwaZulu-Natal, respectively. There was a small share of income from subsistence farming for households in all the provinces. A plausible reason for this could be that these households sold smaller portions of their produce.

The SDI was calculated as it was a prerequisite to measure the degree of income diversification using SEI. The SDI reached 1 in the Limpopo and KwaZulu-Natal Provinces, indicating that the household income was evenly distributed across the six sources of income. In the Eastern Cape, the SDI was 0.676, indicating less evenness than in Limpopo and KwaZulu-Natal.

The SEI, which ranges from 0 to 100, was then calculated. The SEI increases with the number of income sources. The SEI was 69% in Limpopo and KwaZulu-Natal and 40% in the Eastern Cape; households in the Limpopo and KwaZulu-Natal therefore generally diversified around the portfolio of activities in Table 8, more so than households in the Eastern Cape.

The SEI also illustrates the evenness of incomes; therefore 69% of household income in Limpopo and KwaZulu-Natal was evenly distributed across the sources of income investigated, whereas only 40% of the income was evenly distributed across the income sources in the Eastern Cape.

These results are similar to the findings of Schwarze and Zeller[20] who also used the SEI. Their results illustrated that poor households tended to have more income sources and a more even distribution of income among these sources. Perret[11] found that diversity was a major trait of local livelihood systems in the Eastern Cape. Households mostly relied on pension and remittances but also pursued other sources of incomes for supplementary purposes.

Conclusion and policy recommendations

We analysed household income diversification and the degree of diversification in the three poorest provinces of South Africa – the Eastern Cape, Limpopo and KwaZulu-Natal. Specifically, we analysed income diversification of households in different settlement types in each of the provinces and measured the degree of income diversification in these provinces. Data for empirical estimates were obtained from the 2010/2011 IES from StatsSA and Wave 3 data from the National Income Dynamic Study. NIS, NYE and the SDI were calculated to estimate income diversification.

The Chitiga-Mabugu et al.[13] report on which this paper builds, highlighted that poverty was concentrated in the traditional areas and urban informal areas. It was illustrated in this paper that households in traditional areas derive their livelihood mostly from social transfers and remittances, while those in the urban formal areas derive income from business, labour and financial capital returns. It is crucial that government interventions that aim at creating employment and enhancing the incomes of households focus on the rural areas of these provinces. Schwarze and Zeller[20] revealed that wealth increases the likelihood of income diversification. We confirmed these findings by revealing that households in the urban formal settlements follow those in the traditional area settlements in terms of diversifying.

The social wage policy of government which provides social wages (such as old-age pensions and child support grants) seems to have played an important role as a source of income for most households in the traditional and urban informal areas. These sources of income are not enough, however, as these households turn to diversify livelihoods into agriculture and off-farm activities. To address this, in an effort to achieve economic restructuring and poverty alleviation, government should increase its momentum in the provision of incentives to households in these settlement types to assist them in venturing into businesses, most especially in the provision of financial and skills development support to small, medium and micro enterprises (SMMEs).

Acknowledgements

This paper builds on the research project funded by the National Development Agency in 2014 entitled 'The state of poverty and its manifestation in the nine provinces of South Africa'.

Authors' contributions

J.M. and C.N. worked on the original report and conceptualised the paper; M.M. and S.J. participated in data analysis and editing of the manuscript.

References

1. World Bank. World development report 2008. Washington DC: The World Bank; 2007.
2. Barrett CB, Reardon T, Webb P. Nonfarm income diversification and household livelihood strategies in rural Africa: Concepts, dynamics, and policy implications. Food Policy. 2001;26(4):315–331. http://dx.doi.org/10.1016/s0306-9192(01)00014-8
3. Minot N, Epprecht M, Thi T, Anh T, Trung Q. Income diversification and poverty in the northern uplands of Vietnam. Washington DC: International Food Policy Research Institute; 2006.
4. Meyer W, Mollers J, Buchenrieder G. Does non-farm income diversification in Northern Albania offer an escape from rural poverty? [document on the Internet]. c2008 [cited 2016 Dec 19]. Available from: http://nbn-resolving.de/urn:nbn:de:gbv:3:2-11375
5. Ibrahim H, Rahman SA, Envulus EE, Oyewole SO. Income and crop diversification among farming households in a rural area of north central Nigeria. J Trop Agric Food Environ Extension. 2009;8(2):84–89.
6. Ersado L. Income diversification in Zimbabwe: Welfare implications from urban and rural areas. Washington DC: World Bank; 2003.
7. Keeton G. Inequality in South Africa. Journal of the Helen Suzman Foundation. 2014; 74:26–31.
8. Awoniyi OA, Salman KK. Non-farm income diversification and welfare status of rural household in South West Zone of Nigeria. Washington DC: International Food Policy Research Institute; 2008.
9. Fausat AF. Income diversification determinants among farming household in Konduga, Borno state, Nigeria. Acad Res Int. 2012; 2(1):555–561.
10. Adebayo CO, Akogwu GO, Yisa ES. Determinants of income diversification among farm households in Kaduna State: Application of Tobit regression model. Prod Agric Technol J. 2012;8(2):1–10.
11. Perret SR. Poverty and diversity of livelihood systems in post-apartheid rural South Africa: Insights into local levels in the Eastern Cape Province. Pretoria: Department of Agricultural Economics, Extension and Rural Development, University of Pretoria; 2001.
12. Perret S, Anseeuw W, Mathebula N. Poverty and livelihoods in rural South Africa. Investigating diversity and dynamics of livelihoods: Case studies in Limpopo. Number 05/01. Battle Creek, MI: Kellogg's Foundation; 2005.
13. Chitiga-Mabugu M, Ngepah N, Nhemachena C, Motala S, Mathebula J, Mupela E. The state of poverty and its manifestation in the nine provinces of South Africa. Johannesburg: National Development Agency; 2014.
14. Loison SA. Rural livelihood diversification in sub-Saharan Africa: A literature review. J Dev Stud. 201551(9):1125–1138. http://dx.doi.org/10.1080/00220388.2015.1046445
15. Hilson G. Farming, small-scale mining and rural livelihoods in sub-Saharan Africa: A critical overview. Extractive Ind Soc. 2016;3(2):547–563. http://dx.doi.org/10.1016/j.exis.2016.02.003
16. Barrett C. Rural poverty dynamics: Development policy implications. Paper presented at: International Conference for Agricultural Economists; 2003, Aug 17–23; Durban, South Africa.
17. Hosu S, Mushunje A. Optimizing resource use and economics of crop-livestock integration among small farmers in semiarid regions of South Africa. Agroecol Sustain Food Syst. 2013;37(9):985–1000. http://dx.doi.org/10.1080/21683565.2013.802755
18. Jayne TS, Chamberlin J, Headey DD. Land pressures, the evolution of farming systems, and development strategies in Africa: A synthesis. Food Policy. 2014;48:1–17. http://dx.doi.org/10.1016/j.foodpol.2014.05.014
19. Headey D, Jayne TS. Adaptation to land constraints: Is Africa different? Food Policy. 2014;48:18–33. http://dx.doi.org/10.1016/j.foodpol.2014.05.005
20. Schwarze S, Zeller M. Income diversification of rural households in Central Sulawesi, Indonesia. Quar J Int Agric. 2005;44(1):61–74.
21. Chitiga-Mabugu M, Nhemachena C, Karuaihe S, Motala S, Tsoanamatsie N, Mashile L. Civil society participation in income generating activities in South Africa. Johannesburg: National Development Agency; 2013.
22. Alemu ZG. Livelihood strategies in rural South Africa*:* Implications for poverty reduction. Foz do Iguacu: Zarihun Gudeta Alemu; 2012.
23. Barber BM, Odean T, Zheng L. Out of sight, out of mind: The effects of expenses on mutual fund flows. Journal of Business. 2005;78(6):2095–2119.
24. Block S, Webb P. The dynamics of livelihood diversification in post-famine Ethiopia. Food Policy. 2001;26(4):333–350. http://dx.doi.org/10.1016/S0306-9192(01)00015-X
25. Zhao J, Barry PJ. Implications of different income diversification indexes: The case of rural China. Econ Bus Lett. 2013;2(3):13–20. http://dx.doi.org/10.17811/ebl.2.1.2013.13-20
26. Tuomisto HA. Diversity of beta diversities: Straightening up a concept gone away. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography. 2010;33:2–22. http://dx.doi.org/10.1111/j.1600-0587.2009.05880.x
27. Statistics South Africa (StatsSA). Income and expenditure of households 2010/2011. Pretoria: StatsSA; 2012. Available from: http://www.statssa.gov.za/publications/P0100/P01002011.pdf
28. Statistics South Africa (StatsSA). Investigation into appropriate definitions of urban and rural areas for South Africa: Discussion document. Report no. 03-02-20. Pretoria: StatsSA; 2001.

Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems

AUTHORS:
Febe de Wet[1]
Neil Kleynhans[2]
Dirk van Compernolle[3]
Reza Sahraeian[3]

AFFILIATIONS:
[1]Human Language Technologies Research Group, Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa
[3]Center for Processing Speech and Images, Department of Electrical Engineering, University of Leuven, Leuven, Belgium

CORRESPONDENCE TO:
Febe de Wet

EMAIL:
fdwet@csir.co.za

DATES:
Received: 04 Feb. 2016
Revised: 31 May 2016
Accepted: 24 Aug. 2016

KEYWORDS:
acoustic modelling; Afrikaans; Flemish; automatic speech recognition

HOW TO CITE:
De Wet F, Kleynhans N, Van Compernolle D, Sahraeian R. Speech recognition for under-resourced languages: Data sharing in hidden Markov model systems. S Afr J Sci. 2017;113(1/2), Art. #2016-0038, 9 pages. http://dx.doi.org/10.17159/sajs.2017/20160038

ARTICLE INCLUDES:
✓ Supplementary material
× Data set

FUNDING:
Fund for Scientific Research of Flanders; National Research Foundation (South Africa); South African Department of Arts and Culture: Programme of Collaboration on HLT.

For purposes of automated speech recognition in under-resourced environments, techniques used to share acoustic data between closely related or similar languages become important. Donor languages with abundant resources can potentially be used to increase the recognition accuracy of speech systems developed in the resource poor target language. The assumption is that adding more data will increase the robustness of the statistical estimations captured by the acoustic models. In this study we investigated data sharing between Afrikaans and Flemish – an under-resourced and well-resourced language, respectively. Our approach was focused on the exploration of model adaptation and refinement techniques associated with hidden Markov model based speech recognition systems to improve the benefit of sharing data. Specifically, we focused on the use of currently available techniques, some possible combinations and the exact utilisation of the techniques during the acoustic model development process. Our findings show that simply using normal approaches to adaptation and refinement does not result in any benefits when adding Flemish data to the Afrikaans training pool. The only observed improvement was achieved when developing acoustic models on all available data but estimating model refinements and adaptations on the target data only.

Significance:

- Acoustic modelling for under-resourced languages
- Automatic speech recognition for Afrikaans
- Data sharing between Flemish and Afrikaans to improve acoustic modelling for Afrikaans

Introduction

Speech interfaces to different types of technology are becoming increasingly more common. Users can use their voice to search the Internet, control the volume of their car radio or dictate. However, this possibility is only available to users if the required technology exists in the language they speak. Automatic speech recognition (ASR) technology already exists and is regularly used by speakers of American English, British English, German, Japanese, etc. The development of ASR systems requires substantial amounts of speech and text data. While such resources are readily available for a number of languages, the majority of the languages that are spoken in the world can be classified as under-resourced, i.e. the resources required to create technologies like ASR do not exist or exist only to a limited degree. Researchers in the field of speech technology development for under-resourced languages are investigating various possibilities to address this challenge and to establish resources and technologies in as many languages as possible.

One of the strategies that has been explored is to fast-track progress in under-resourced languages by borrowing as much as possible – in terms of both data and technology – from well-resourced languages. Here we report on an investigation on data sharing between Afrikaans – an under-resourced language – and Flemish – a well-resourced language. The approach was focused on the exploration of model adaptation and refinement techniques associated with hidden Markov model (HMM) based speech recognition systems to improve the benefit of sharing data. The focus was specifically on the use of currently available techniques, some possible combinations and the exact utilisation of the techniques during the acoustic model development process.

Most of the techniques that are used in language and speech technologies are based on statistical methods. These methods require substantial amounts of data for a reliable estimation of the statistical parameters that are used to model the language, either in its written or spoken form. The required amounts often exceed what is available for resource-scarce languages.[1] The restricted resources that are available for these languages can be supplemented with resources from other languages, especially from those for which extensive resources are available. We investigated different possibilities to improve acoustic modelling in an under-resourced language, Afrikaans, by using data from a well-resourced language, Flemish. The techniques that were investigated include bootstrapping Afrikaans models using Flemish data as well as individual and combined model adaptation techniques.

Specifically, our aim throughout was to improve the performance of Afrikaans acoustic models by adding the Flemish data using various model adaptation and refinement approaches. As we focused on the model level, we utilised maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation as well as a combination of these adaptation techniques. In addition, heteroscedastic linear discriminant analysis (HLDA) and speaker adaptive training (SAT) acoustic model refinements were investigated in terms of sharing acoustic data. The purpose of investigating these techniques – described in later sections – is to determine whether these methods are sufficient in our data sharing scenario.

Background

Some of the approaches to data combination that have been reported on in the literature include cross-language transfer[2], cross-language adaptation[3], data pooling[2,4] as well as bootstrapping[5]. However, results as well as

conclusions vary between studies and seem to be highly dependent on the amount of data that is used and the specific modelling task investigated. Some studies report small gains under very specific conditions.

In a study by Adda-Decker et al.[6] in which no acoustic data were available for the target language (Luxembourgish), English, French and German data sets were used to train a multilingual as well as three monolingual ASR systems. Baseline models for Luxembourgish were subsequently obtained by using the International Phonetic Alphabet associations between the Luxembourgish phone inventory and the English, French and German phone sets. (A phone is the smallest discrete segment of sound in a stream of speech). Results showed that the language identity of the acoustic models has a strong influence on system performance with the German models yielding much better performance than the French or English ones. The acoustic data that were available for Luxembourgish were not enough to train a baseline system. It was therefore not possible to compare the performance of the German models with models trained on the target language.

Positive results were reported for multilingual acoustic modelling when only a small amount of training data was available for Dari, Farsi and Pashto.[7] MAP adaptation of the multilingual models to the individual target languages yielded a 3% relative improvement in word error rate compared to the corresponding monolingual models. However, as more data were added during training for the individual languages, the monolingual models overtook their multilingual counterpart very quickly in terms of recognition performance – given equal amounts of training data and the same number of model parameters.

Van Heerden et al.[4] found that simply pooling data for closely related languages resulted in improvements in ASR phone accuracies. They grouped languages according to expert knowledge of language families – Nguni and Sotho. The generally observed trend was that adding one to two languages gave slight improvements in accuracy – however, this trend was not observed for Sepedi. In addition, for the majority of cases, adding a third language to the training pool resulted in a decrease in accuracy (except for isiZulu). On average, each language contained about 7 h of audio training data, thus 14 h and 21 h of training data indicated improvement.

Niesler[8] investigated the possibility of combining speech data from different languages spoken in a multilingual environment to improve the performance of ASR systems for the individual languages. The systems were all HMM based. The recognition performances of language-specific systems for Afrikaans, South African English, isiXhosa and isiZulu were compared with that of a multilingual system based on data pooling as well as data sharing by means of decision-tree clustering. The clustering process was modified to allow for language-specific questions. Data from different languages could therefore be shared at HMM state level. The results of the study showed that the multilingual acoustic models obtained using this data sharing strategy achieved a small but consistent improvement over the systems that were developed for the languages individually or by just pooling the data.

Kamper et al.[9] performed several data sharing experiments on accented English audio data collected in South Africa. They specifically considered the accents of South African English defined in the literature: Afrikaans English, Black South African English, Cape Flats English, White South African English and Indian South African English. Overall they found that their multi-accent modelling approach outperformed accent-specific and accent-independent acoustic models. To create the multi-accent acoustic models, a modified decision-tree state cluster approach was used when accent-specific questions could be asked, which allowed the sharing of data across accents at the HMM state level. This approach is similar to that of Niesler[8] except accent questions were used instead of language-specific questions. Of interest, was the analysis of the proportions of data shared at the state level. It was found that the optimal phone and word operating points were different and that the amount of data shared at these points also differed – 33% and 44%, respectively.

A current popular trend for data sharing is to make use of deep neural networks (DNNs) for robust feature extraction, for which gains have been observed even for unrelated languages. Approaches mainly focus on bottleneck features with different network architectures and optimisations. Some examples of the bottleneck feature approach are described in Veselý et al.[10] (language-independent bottleneck features), Zhang et al.[11] (multilingual stacked bottleneck features), Nguyen et al.[12] (multilingual shifting deep bottleneck features) and Vu et al.[13] (multilingual DNNs cross-language transfer). Once the features are extracted they are fed through to a Gaussian mixture model (GMM)/HMM or Kullback–Leibler divergence based HMM (KL-HMM) system, where normal ASR techniques are applied. It is difficult to interpret how exactly the DNNs are combining the different data and what effective operation is being applied to the data, but it does seem that the DNNs are applying a necessary feature normalisation.[14] In line with this feature processing, there is great scope for improvement at the feature level as shown in intrinsic spectral analysis combination investigation.[15]

Monolingual acoustic modelling for Afrikaans has been investigated previously using a conventional Mel frequency cepstral coefficient (MFCC) based HMM system and broadcast news data[16] as well as using intrinsic spectral analysis in combination with a broadband, monolingual Afrikaans corpus[15].

In a study on resource and technology transfer between closely related languages, a case study was conducted for Dutch and Afrikaans. The distance between Afrikaans and other West Germanic languages and dialects was quantified in terms of acoustically weighted Levenshtein distances.[17] The results identified Dutch and Flemish as well-resourced, donor languages for the development of language and speech technology in Afrikaans, especially in terms of supplying background data for acoustic modelling (cf. Box 1). These results were confirmed by a series of experiments that investigated the possibility of improving acoustic modelling for Afrikaans by using Dutch, Swiss German and British English as background data in Tandem and KL-HMM ASR systems. The best results were obtained when Dutch was used as out-of-language background data.[18]

Box 1: Closeness

> In the context of statistical modelling 'closeness' is defined in terms of the acoustic distances between the languages. Phonetic and lexical overlap can also be taken into consideration to determine 'closeness'. Historical and linguistic considerations may be related to but are not always reflected in objective measures such as acoustic distance.

We report on an attempt to improve acoustic modelling for Afrikaans (as an example of a resource-scarce language) by borrowing data from Flemish (as an example of a well-resourced language). Flemish was chosen as the donor language because we had access to previously developed ASR systems for Flemish as well as the relevant data. It was also decided to start with Flemish rather than a combination of Flemish and Dutch as previous studies have shown that the two languages have distinctive acoustic properties and that better recognition results are obtained if they are first modelled separately and then combined.[19]

A previous study on this topic investigated the use of multilayer perceptrons, KL-HMMs and subspace Gaussian mixture models (SGMMs) and used Dutch as a donor language.[20] The systems based on SGMMs achieved the best monolingual as well as multi-lingual performance. When the models were trained on Dutch data and adapted using the Afrikaans data, the SGMM systems also yielded the best results. Overall, the results showed that Dutch/Afrikaans multilingual systems yield a 12% relative improvement in comparison with a conventional HMM/GMM system trained only on Afrikaans.

The literature review sketches a domain in which many approaches have been explored to enable speech recognition performance gains for under-resourced languages through data sharing, but the results are quite varied. In summary, our research reported here investigates the possibility of combining Flemish and Afrikaans data at the model level using model adaptation (MLLR and MAP) and refinement (HLDA and SAT) techniques as well as combinations thereof. Although the DNN and intrinsic spectral analysis feature approaches have yielded success, this investigation will not focus on these.

Data

In this study, Flemish was used as an example of a well-resourced language and Afrikaans as an example of a closely related but under-resourced language. The Flemish and Afrikaans speech data and pronunciation dictionaries are described in this section.

Box 2: Standard and 'less standard' varieties

The data sets that were used in this study were designed to include the standard varieties of the relevant languages. For most languages it is difficult – sometimes to the point of being controversial – to define exactly what a 'standard variety' is. The Flemish data correspond to radio news bulletins. Extreme varieties of a language are usually not used for news broadcasts, although we did not confirm this supposition in terms of internationally accepted news broadcasting standards. The National Centre for Human Language Technology Afrikaans data set has a 70:30 ratio of urban versus rural accents. The 'less standard' varieties of the language are usually spoken in rural rather than urban areas. Although 'less standard' varieties could therefore be present in the data, their properties are bound to be dominated by those of the more standard variety which constitutes the majority of the data.

Flemish resources

The Spoken Dutch Corpus – Corpus Gesproken Nederlands (CGN)[21] – is a standard Dutch database (cf. Box 2) that includes speech data collected from adults in the Netherlands and Flanders. The corpus consists of 13 components that correspond to different socio-situational settings. In this study only Flemish data from component 'O' were used. This component of the database contains phonetically aligned read speech. These data were chosen for the development of the Flemish acoustic models because read speech is carefully articulated and the corresponding phone models present a 'best case scenario' of the acoustics in the language. For instance, words and phones are not affected by the co-articulation effects that typically occur in more spontaneous speech. Component 'O' includes about 38 h of speech data recorded at 16 KHz and produced by 150 speakers.

For the purposes of the current investigation the data set was divided into training and test sets as follows: 8 (4 male, 4 female) speakers were randomly chosen for the evaluation set, corresponding to about 2 h of audio data. From the remaining 36 h, 10 h of training data were randomly selected. The training set was selected to match the size of the set of unique Afrikaans prompts described in the next section. Matching training sets were used to avoid CGN data from dominating the acoustic models.

The CGN dictionary uses 48 phones, including silence. In the cross-lingual experiments, the set was reduced to 38 phonemes using knowledge-based phonetic mapping. The mapping that was used is provided in Appendix 1 of the supplementary material. Nomenclature is given in Appendix 2 of the supplementary material.

Afrikaans resources

The Afrikaans speech data that were used in this study were taken from the National Centre for Human Language Technology (NCHLT) speech corpus.[22] The development of the corpus was funded by the South African Department of Arts and Culture with the aim of collecting 50–60 h of transcribed speech for each of the 11 official South African languages. The Afrikaans set contains data collected from 210 (103 male, 107 female) speakers. The set includes about 52 h of training data and a predefined test set of almost 3 h.

During data selection for this study, an analysis was made of the type (i.e. the unique set of words) and token (i.e. the set of words) counts in the Afrikaans data set. The values for the training and test sets are summarised in the first row of Table 1. These values indicate that only 20% of the recorded utterances in the training set are unique. This figure relates to about 10 h of unique training data and 2.2 h of unique evaluation data. The unique data subset statistics are shown in the second row of Table 1 (Type frequency 1).

If each unique token is allowed to occur a maximum of five times, the training set size increases to 37.1 h and the evaluation set to 2.7 h. Row 3 in Table 1 (Type frequency 5) shows the data subset statistics for this data selection criterion.

Table 1: Summary of the National Centre for Human Language Technology Afrikaans data

	Training set			Test set		
	Types	**Tokens**	**Duration**	**Types**	**Tokens**	**Duration**
All data	12 274	61 413	52.2 h	2513	3002	2.7 h
Type frequency 1	12 274	12 274	10.6 h	2513	2513	2.2 h
Type frequency 5	12 274	44 538	37.1 h	2513	3002	2.7 h

From Table 1, we observed quite a large drop in training data amount when limiting the data by uniqueness or frequency of occurrence. Subsequently, the effect on ASR performance was investigated given the various training data subsets. The ASR systems were set up according to a standard configuration – MFCCs, first- and second-order derivatives, tristate left-to-right triphone models – and were built using the hidden Markov toolkit (HTK).[23] Cepstral mean and variance normalisation was applied at the speaker level.

The ASR systems were evaluated using the predefined NCHLT evaluation set as well as two additional Afrikaans corpora. The first corpus was a text-to-speech data set while the second was a broadcast news-style data set created by recording radio news broadcasts from *Radio Sonder Grense*, a local Afrikaans radio station.[16] System performance was measured in terms of phone recognition accuracy, defined as:

$$Accuracy = 100 - \left(\frac{S+D+I}{N} \times 100\right)\%, \qquad \text{Equation 1}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the total number of phones in the reference.

The results of the various evaluations are summarised in Table 2. As expected, the ASR performance drops as less data are used to develop the acoustic models. Based on the NCHLT and radio broadcast data, even though there is about a 10% absolute drop in accuracy (on average) between the unique and all data sets, the ASR performance is still quite high for the unique data set given that only 20% of the training data were used. This result probably means that the full and unique data sets represent more or less the same data properties.

The text-to-speech results show very little variation for the three different sets of acoustic models. This result may be because of the nature of the corpus: it contains speech from a single speaker and the sentences are phonetically balanced. As a consequence, the data do not contain as much variation as a multi-speaker corpus such as the radio broadcast data. The specific set of training data does not seem to influence the match between the acoustic models and the single speaker in the text-to-speech corpus.

Table 2: Phone accuracy results for different sets of training data

	NCHLT	**Text to speech**	**Radio**
All data	86.24	75.39	65.81
Type frequency 1	75.04	75.19	57.87
Type frequency 5	85.21	75.28	61.04

NCHLT, National Centre for Human Language Technology

Method

Several techniques related to model adaptation and refinement and the application to data sharing were used: MLLR, MAP, SAT and HLDA. The application of the techniques is discussed in terms of data sharing.

Maximum likelihood linear regression

Maximum likelihood linear regression (MLLR), proposed by Leggetter and Woodland[24] for speaker adaptation, provides a means to update acoustic models without having to retrain the parameters directly. The technique estimates a set of linear-regression matrix transforms that are applied to the mean vectors of the acoustic models. Their initial speaker adaptation implementation performed mean-only adaptation.

Gales and Woodland[25] extended the framework to include variance adaptation. Generally, a cascaded approach is used, in which mean adaptation is applied first and then the variance transformation is applied. Another form of the MLLR transformation is the constrained MLLR transformation (CMLLR). In this approach, a joint transform is estimated in which the aim is to transform the mean and variance simultaneously. To do so, the transform is applied directly to the data vectors and not to the means and variances.

The MLLR adaptation technique utilises a regression class tree to ensure robust transformation parameter estimation. The regression class tree defines a set of classes that contain similar acoustic models that allow data to be shared amongst similar acoustic classes. The tree is developed by using a centroid splitting algorithm[23] that can be used to automatically create the user-specified number of classes, but in this study only a single class or phone-specific classes were defined. This limitation was introduced by the HTK HLDA implementation that makes use of a single class. In terms of data sharing, the adaptation process can be used to adapt acoustic models to better fit a specific language. Here we view the languages as different speakers or channels. In this scenario, we could pool the data to increase the training data amount and then utilise MLLR to adapt these models to statistically fit the target language better.

Maximum a posteriori

Gauvain and Lee[26] proposed the use of a MAP measure to perform parameter smoothing and model adaptation. The MAP technique differs from maximum likelihood estimation by including an informative prior to aid in HMM parameter adaptation. The results for speaker adaptation showed that MAP successfully adapted speaker-independent models with relatively small amounts of adaptation data compared to the maximum likelihood estimation techniques. However, as more adaptation data became available, MAP and maximum likelihood estimation yielded the same performance. In this adaptation scenario, the speaker-independent models served as the informative priors, whereas in the experiments conducted in this study, the donor language will serve as the informative prior. Similar to the MLLR data sharing scenario, MAP can be used to adapt the acoustic models to a target language. The acoustic models trained on the pooled data serve as the prior.

Acoustic model adaptation

Under certain circumstances, as shown in Van Heerden et al.[4], simply pooling speech data (combining language resources such as data and dictionaries) into a larger training set can lead to an improvement in the results. There is no guarantee, however, that an improvement in the system accuracies will be observed and if the data amounts for the target language are small, then the donor language could possibly dominate the acoustic space. Therefore, in a resource-constrained environment, a better approach may be to adapt, using a relatively small amount of data.

MAP and MLLR are commonly used to perform speaker and environment adaptation and it is fairly simple to make use of these to perform language or dialect adaptation. It has been shown previously that simply applying MLLR and MAP to data sharing does not yield improvements.[20] However, there are many points in the acoustic model development pipeline at which these techniques can be inserted and they can be used either in isolation or in certain combinations. Thus one focus of the experimental investigation is to establish which combination of adaptation techniques could produce an improvement in overall ASR accuracy and at what point during the acoustic model development it should be applied.

Acoustic model refinement

Most current ASR systems make use of HLDA and SAT to improve the overall accuracies, which in the HTK-style development cycle are applied during the last stage of model refinement. HLDA estimates a transform that reduces the dimension of the feature vectors while trying to improve class separation. The main purpose of SAT is to produce a canonical acoustic model set by using transforms to absorb speaker differences and thus create a better speaker independent model.

As these techniques are applied as last stage refinements, there are a few possibilities that can be investigated with respect to data sharing. In terms of HLDA, a donor language can be used to develop acoustic models and the target language data used to estimate the feature dimension reduction transform. For SAT, as the transforms are absorbing speaker differences, and the language or dialect used creates acoustic differences, this approach could help create an acoustic model set better suited for the target language.

Experimental set-up

For all experiments we used 10 h of randomly selected CGN data and 10 h of NCHLT data for acoustic model development and transformation estimation. The NCHLT data correspond to the set of unique utterances described above. The developed ASR systems are evaluated on the corresponding 2.2-h subset of the NCHLT evaluation data (see Tables 1 and 2). Our aim throughout was to improve the performance of NCHLT acoustic models by adding the CGN data using various model adaptation and refinement approaches.

Baseline speech recognition system

The baseline speech recognition system was developed using a standard HTK recipe. The audio files were parameterised into 39 dimensional MFCC features – 13 static, 13 delta and 13 delta-delta. These include the MFCC 0th coefficient. Cepstral mean normalisation was applied. The acoustic models were systematically developed, starting from mono phone models, expanding the mono phone models to context-dependent triphone models and finally consolidating this model set to tied-state triphone models. A three state left-to-right HMM topology was used for each acoustic model set. A phone-based question state-tying scheme was employed to develop the tied-state models. Lastly, a mixture incrementing phase was performed to better model the state distributions – eight mixture Gaussian mixture models were used for each HMM state.

Acoustic model adaptation

The first set of experiments focused on MLLR and MAP adaptation. Block diagrams illustrating the different experimental set-ups are provided in Figures 1 to 5. The following experiments were performed:

- **Baseline NCHLT:** Baseline NCHLT acoustic models were developed on the 10-h Afrikaans NCHLT data. No adaptations were applied.
- **Language CMLLR transforms:** Starting from the baseline NCHLT system, two language-based (Afrikaans on NCHLT and Flemish on CGN) CMLLR transforms were estimated using the baseline acoustic models and the separate 10-h NCHLT and 10-h CGN data. Phone-specific transforms were estimated using the phone-defined regression class tree. Once the corpus-specific transforms were estimated, the baseline acoustic models were updated using two iterations of maximum likelihood training. Both 10-h training sets were used for this update but the specific language CMLLRs were applied to the corresponding training set. The NCHLT CMLLR was applied during evaluation.
- **Retrain using language CMLLR transforms:** The language CMLLR transform generated by the 'Language CMLLR transforms' experiment was used to develop a new acoustic model set using both the 10-h NCHLT and 10-h CGN. The normal baseline training procedure was modified to incorporate the CMLLR transforms

which were used throughout the training cycle. This meant that, at each model estimation iteration, the language-specific CMLLRs were applied when updating with the corresponding training data set. During evaluation, the estimated NCHLT CMLLR transform was applied.

- **Retrain using language CMLLR transforms with MAP:** Starting with the system developed in the 'Retrain using CMLLR transforms' experiment, one final step was added to the acoustic model development cycle: two iterations of MAP adaptation were performed using the 10-h NCHLT data only. The NCHLT CMLLR transform was applied during evaluation.
- **AutoDac training approach:** For this approach, acoustic models were developed using the best method described in Kleynhans et al.[27] Initially, only the 10-h NCHLT data were used to develop the acoustic models until the state-tying phase. Then, for the last phase, mixture incrementing, the 10-h CGN data were added to the training data pool and the Gaussian densities were estimated on all the data. No CMLLR transforms or MAP adaptation were used.

Acoustic model refinement

In this experimental set-up, two additional steps were added to the acoustic model development training cycle: HLDA and SAT. Both the HLDA and SAT use a global regression tree (all states pooled into a single node). Note that no language-dependent MAP or MLLR adaptation was applied. The HLDA ASR systems appended 13 delta-delta-delta coefficients to the baseline MFCCs, which increased the feature dimension to 52. An HLDA transform was estimated using a global transform, which was then used to transform the 52-dimensional feature vectors to 39 dimensions. For SAT, a global CMLLR transform was used to model each speaker's characteristics. The following acoustic model refinement experiments were defined:

- **NCHLT HLDA-SAT:** Baseline acoustic models were developed using the 10-h NCHLT, followed by HLDA and SAT model refinements.
- **NCHLT+CGN HLDA-SAT:** Baseline acoustic models were developed using both the 10-h NCHLT and 10-h CGN data sets, and then applying the HLDA and SAT model refinements using all the training data.
- **NCHLT+CGN+NCHLT HLDA-SAT:** For this training set-up, baseline acoustic models were developed on both the 10-h NCHLT and 10-h CGN training data sets. The HLDA and SAT transformations were estimated using the 10-h NCHLT training data only.

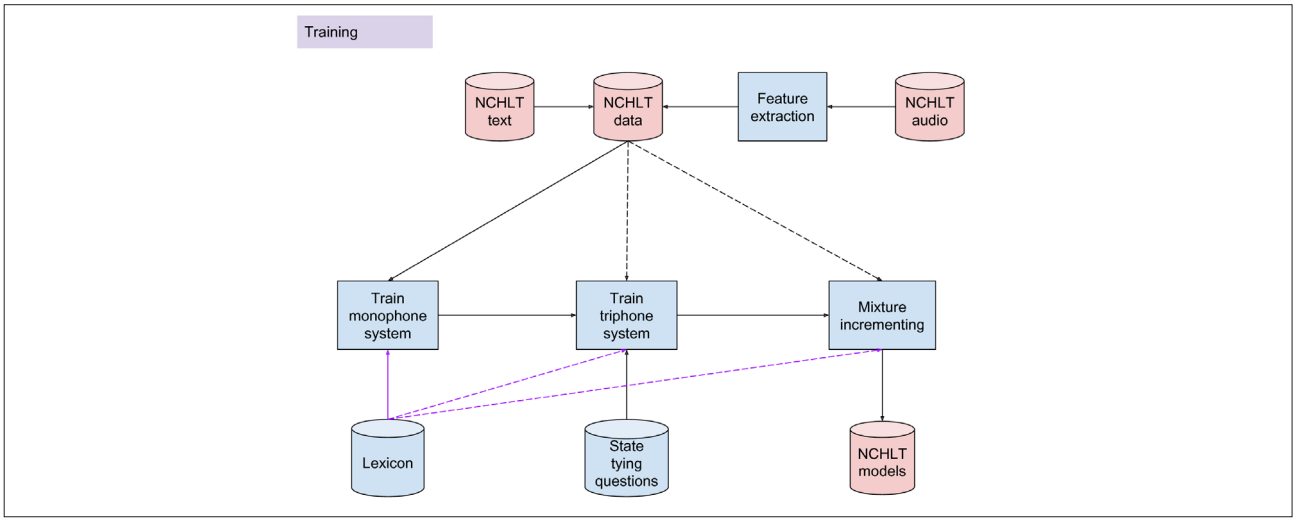
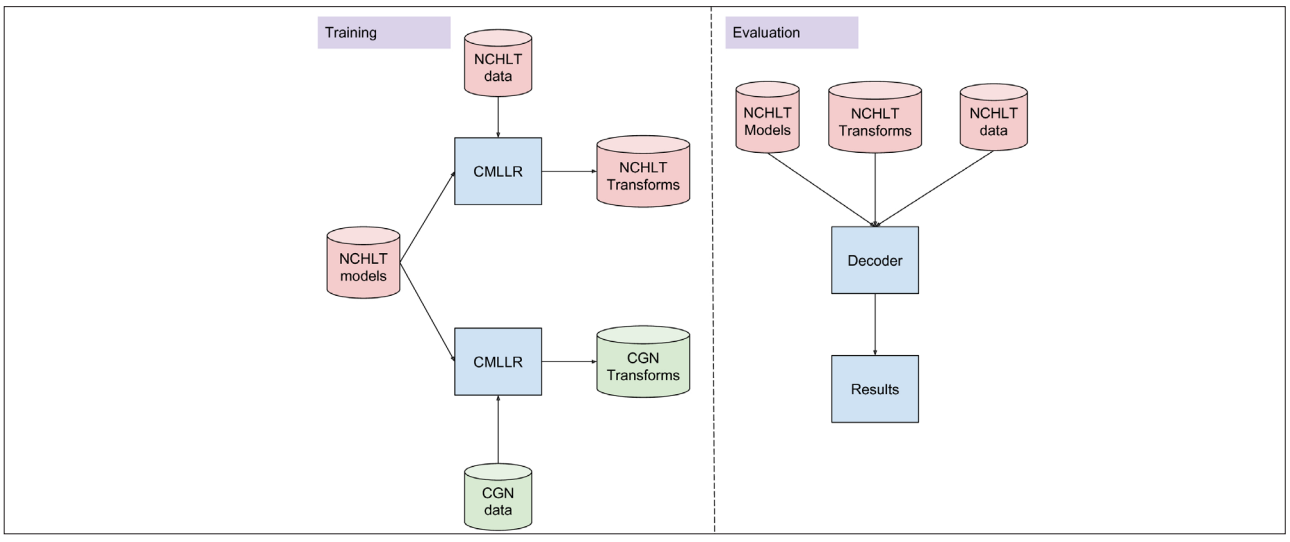
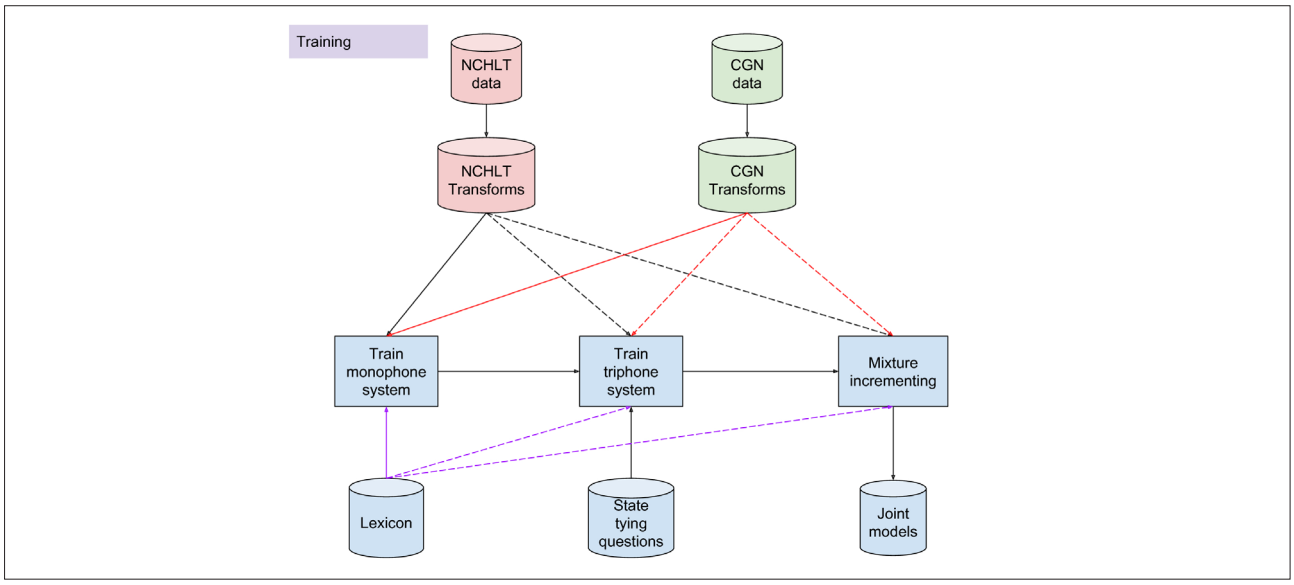


Figure 1: Baseline National Centre for Human Language Technology (NCHLT) training scheme.



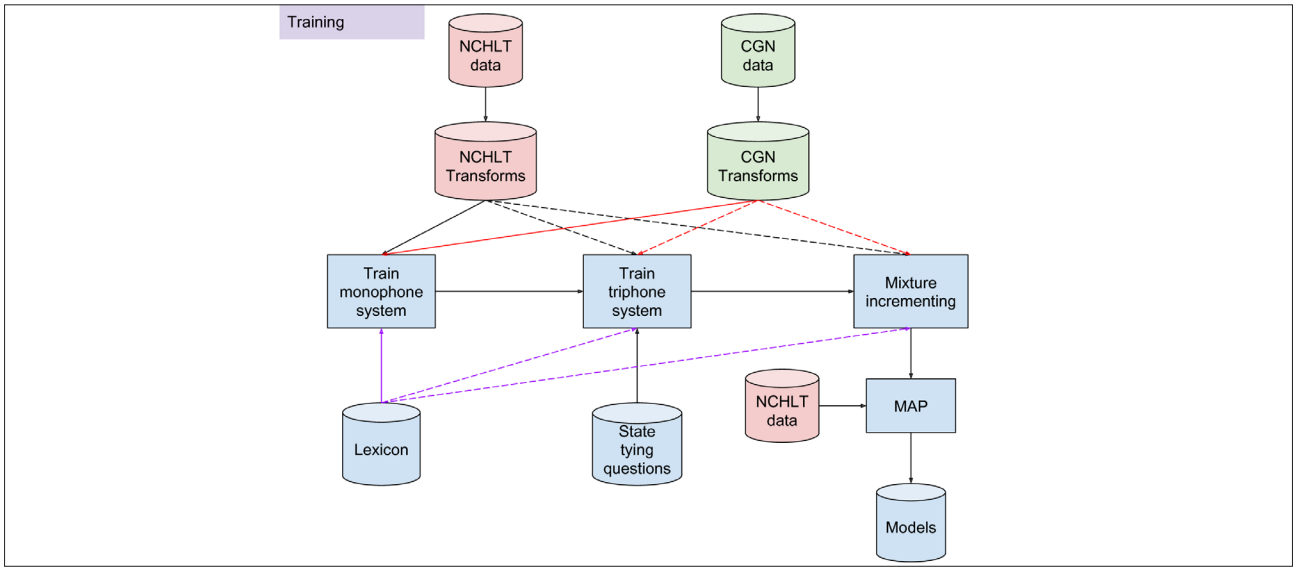
NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 2: Language constrained maximum likelihood linear regression (CMLLR) training scheme.



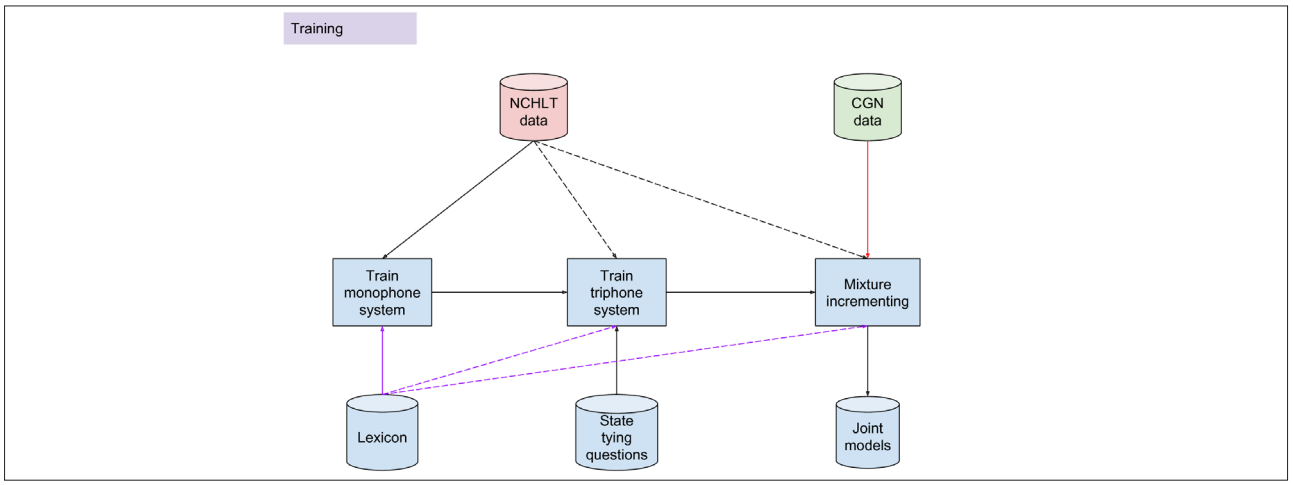
NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 3: Retrain using language constrained maximum likelihood linear regression transform training scheme.



NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 4: Retrain using language constrained maximum likelihood linear regression transforms with maximum a posteriori (MAP) training scheme.



NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Figure 5: AutoDac training scheme.

Metrics

The ability of the different system configurations to model the training data accurately was measured in terms of the accuracy with which the test data could be decoded. Phone recognition accuracy was calculated according to Equation 1 and correctness values were derived as follows:

$$Correctness = \left(\frac{C}{N} \times 100\right)\%, \quad \text{Equation 2}$$

where C is the number of correctly recognised phones and N is the total number of phones in the reference.

Results

Experimental results are presented for CMLLR and MAP adaptation as well as HLDA plus SAT combinations. System performance is quantified in terms of phone recognition accuracy and correctness.

Acoustic model adaptation

Table 3 provides an overview of the results that were obtained using different data sets and model adaptation combinations. The first row in the table represents the performance of the baseline system without any data sharing or model adaptation.

Table 3: Correctness and accuracy results for various automatic speech recognition data sharing set-ups

	Correctness (%)	Accuracy (%)
Baseline NCHLT	78.77	71.17
Language CMLLR transforms	75.81	68.83
Retrain using language CMLLR transforms	78.02	71.82
Retrain using language CMLLR transforms with MAP	78.87	71.31
AutoDac training approach	75.69	68.41

NCHLT, National Centre for Human Language Technology; CMLLR; constrained maximum likelihood linear regression; MAP, maximum a posteriori

Unfortunately, the results in Table 3 show that none of the adaptation and training schemes provide an improvement in ASR performance, when adding CGN data to the NCHLT training data. This is in line with the results reported by Imseng et al.[20] for a similar experiment using a smaller corpus of telephone data. It would seem that both CMLLR and MAP provide insufficient mechanisms to effectively combine data from different sources in the context of cross-language data sharing.

Acoustic model refinement

The performance of the systems in which the models were refined by applying HLDA and SAT is captured in Table 4. Comparing the first row in Table 4 with the corresponding row in Table 3 shows that the application of HLDA and SAT results in a substantial improvement in both phone accuracy and correctness. When the CGN data are added to the training data, the performance decreases. However, the best result is obtained when the acoustic model set is developed on the combined data but the HLDA and SAT are estimated on the 10-h NCHLT data only. This finding may suggest that these transforms are sensitive to language-specific data. The HLDA in effect estimates a projection from a higher dimensional space to a lower one. Thus, a better projection, in terms of class separation, might be estimated on the target data only – in this case, the NCHLT data. For SAT, the single global CMLLR transforms may be insufficient to fully absorb the speaker and channel characteristics; therefore the acoustic model set is not in the best canonical form. Further tests on HTK are not possible as this is a software limitation.

Table 4: Correctness and accuracy results for heteroscedastic linear discriminant analysis (HLDA)- and speaker adaptive training (SAT)-based data sharing automatic speech recognition set-ups

	Correctness (%)	Accuracy (%)
NCHLT HLDA-SAT	85.71	79.66
NCHLT+CGN HLDA-SAT	84.37	78.33
NCHLT+CGN+NCHLT HLDA-SAT	86.89	81.07

NCHLT, National Centre for Human Language Technology; CGN, Corpus Gesproken Nederlands

Discussion

To investigate why only a single improvement was observed over the different experiments, the state-tying process was analysed as this process determines the manner in which acoustic data are shared. HTK makes use of the question-based tying scheme described by Young et al.[28]: initially all acoustic states are grouped into a single root class and then a process to split the nodes is run by 'asking' left and right context questions – all triphones that have the same left or right phone are removed from the pool and the change in pool log-likelihood is captured. The question that results in the greatest change in score is selected and a new node is created that contains all the triphones described by the question. The pre-split node contains all other triphones. The process is continued until a user-defined stopping criterion is met.

Tracking which question is used to split the data pools (create nodes) can give an indication of when the data between the two languages are shared: if language-specific questions are used to split the nodes early on in the state-tying process then no real cross-language data sharing is occurring. To perform the state-tying tracking, a modified, but similar, version of the HTK implementation was developed in which language-specific questions could be used to split the acoustic data pools. Table 5 shows the level at which a language question was used to split the data.

Table 5: The percentage of phones for which the language question was used to split the data during state tying

	State 2	State 3	State 4
First question	69.44	91.67	63.89
First or second question	86.11	91.67	77.78

The values in Table 5 show that, for the majority of cases, the best reduction in overall data pool log-likelihood can be achieved by splitting the data into language-dependent paths. The central context makes use of the language split question to partition the data, in over 91% of the cases for the very first question. This finding is significant as the central context state generally consumes the majority of the speech frames when compared to the start and end states. This result shows that minimal data sharing would occur if the system had a choice and may point to a data artefact – such as channel or environment – which prevents data sharing between the CGN and NCHLT corpora. Further investigation is needed to establish the mechanisms that are inhibiting data sharing and their relative contributions. Possible sharing prevention mechanisms could be: grammar, channel and environment. As positive pooling results were reported by Van Heerden et al.[4] and all experiments were conducted on the same corpus, channel may be a strong candidate. In this instance 'channel' refers to all the factors that could influence the acoustic properties of the speech signals, e.g. the acoustic environment in which the data were recorded and the recording equipment.

Table 5 shows that cross-language data sharing is clearly not taking place to the same extent as reported by Mandal et al.[7] The low data sharing rates are also in contrast to the results presented

by Kamper et al.[9], in which 33% and 44% sharing was seen across accents for phone and word optimal results, respectively, and by Niesler[8] where 20% sharing was measured across language at optimal system performance. For these investigations, data sharing resulted in improved system performance but it is not clear if a positive correlation exists between the percentage of data shared among clusters and the eventual ASR performance.

It could be argued that the acoustic differences between Afrikaans and Flemish are bigger than those observed between the various English accents investigated in the Kamper et al.[9] study. However, the majority of the sounds could be expected to differ to at least the same extent as the languages studied by Niesler because they are from the same language families, as are Afrikaans and Flemish. They are also similar from an acoustic point of view, as are the languages that were investigated in this study. It should be kept in mind that both Kamper et al.[9] and Niesler conducted experiments within the same corpus. Acoustic factors – other than those caused by differences between accents and languages, such as channel and environment effects – could therefore not have influenced their results. This strengthens the possibility that the lack of data sharing in the present study could probably be a result of cross-corpus rather than cross-language artefacts.

Imseng et al.[18] showed that a systematic improvement in phone performances was observed for in-domain phones that had relatively small data amounts. Thus, it would seem that we should rather target states that may need out-of-language data to improve the distribution modelling.

Conclusion

While the idea of data sharing makes sense intuitively – increase the amount of training data for robust density estimation – realising a performance gain in ASR accuracy is difficult to achieve within the context of HMM-based ASR. From the experimental results obtained in this study, using standard MAP and MLLR techniques to enable data sharing did not provide phonetic recognition performance gains. These MAP and MLLR results are in line with those presented by Imseng et al.[20] In addition, the various alternative training strategies also failed. Thus, the standard MAP, MLLR and our various training strategies are not sufficient for data sharing when simply pooling the data.

Surprisingly, the NCHLT+CGN+NCHLT HLDA-SAT experiment managed to achieve a better phone error rate; however, the baseline NCHLT+CGN HLDA-SAT did not yield a gain. The improved result may imply that the combined data are useful but the Afrikaans-specific HLDA projection and SAT acoustic model adjustment are required. This has similarities to some DNN data sharing approaches in which pre-training is performed on many languages but final network parameter optimisations are performed on the target language only.

Recent results from SGMM and DNN experiments show much more potential for data sharing between languages and should be pursued rather than MAP and MLLR. One possible line of research would be to use SGMM for data sharing but rather than pooling all the data, only include data for low occurrence phones, as suggested by results reported in Imseng et al.[18]

Acknowledgements

This research was supported by the South African National Research Foundation (grant no. UID73933), the Fund for Scientific Research of Flanders (FWO) under project AMODA (GA122.10N) as well as a grant from the joint Programme of Collaboration on HLT funded by the Nederlandse Taalunie and the South African Department of Arts and Culture.

Authors' contributions

F.D.W. and D.V.C. conceptualised and led the project on acoustic modelling for under-resourced languages; F.D.W., D.V.C., R.S. and N.K. were responsible for conceptual contributions and experimental design; F.D.W. and D.V.C. designed the phone mapping between Flemish and Afrikaans; R.S. and N.K. performed the experiments; D.F.W. and N.K. prepared the manuscript; D.V.C. is R.S.'s PhD promotor.

References

1. Besacier L, Barnard E, Karpov A, Schultz T. Automatic speech recognition for under-resourced languages: A survey. Speech Commun. 2014;56:85–100. http://dx.doi.org/10.1016/j.specom.2013.07.008
2. Schultz T, Waibel A. Multilingual and cross lingual speech recognition. In: DARPA Workshop 1998: Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding; 1998 February 08–11; Lansdowne, VA, USA. Lansdowne, VA: NIST; 1998. p. 259–262.
3. Schultz T, Waibel A. Language independent and language adaptive large vocabulary speech recognition. In: ICSLP 1998: Proceedings of the 5th International Conference on Spoken Language Processing; 1998 November 30 – December 04; Sydney, Australia. p. 1819–1822.
4. Van Heerden C, Kleynhans N, Barnard E, Davel, M. Pooling ASR data for closely related languages. In: SLTU 2010: Proceedings of the 2nd Workshop on Spoken Languages Technologies for Under-resourced languages; 2010 May 03–0; Penang, Malaysia. Penang: SLTU; 2010. p. 17–23.
5. Schultz T, Waibel A. Language-independent and language-adaptive acoustic modelling for speech recognition. Speech Commun. 2001;35(1):31–51. http://dx.doi.org/10.1016/S0167-6393(00)00094-7
6. Adda-Decker M, Lamel L, Adda G. A first LVCSR system for Luxembourgish, an under-resourced European language. In: LTC 2011: Proceedings of 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics; 2011 November 25–27; Poznań, Poland. Poznań: LTC; 2011. p. 47–50.
7. Mandal A, Vergyri D, Akbacak M, Richey C, Kathol A. Acoustic data sharing for Afghan and Persian languages. In: ICASSP 2011: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2011 May 22–27; Prague, Czech Republic. IEEE; 2011. p. 4996–4999. http://dx.doi.org/10.1109/ICASSP.2011.5947478
8. Niesler T. Language-dependent state clustering for multilingual acoustic modelling. Speech Commun. 2007;49(6):453–463. http://dx.doi.org/10.1016/j.specom.2007.04.001
9. Kamper H, Mukanya FJM, Niesler T. Multi-accent acoustic modelling of South African English. Speech Commun. 2012;54(6):801–813. http://dx.doi.org/10.1016/j.specom.2012.01.008
10. Veselý K, Karafiát M, Grézl F, Janda M, Egorova E. The language-independent bottleneck features. In: SLT 2012: Proceedings of the Spoken Language Technology Workshop; 2012 December 02–05; Miami, FL, USA. Miami, FL: SLT; 2012. p. 336–341. http://dx.doi.org/10.1109/slt.2012.6424246
11. Zhang Yu, Chuangsuwanich E, Glass J. Language ID-based training of multilingual stacked bottleneck features. In: Interspeech 2014: Proceedings of the International Speech Communication Association; 2014 September 14–18; Singapore. p. 1–5.
12. Nguyen QB, Gehring J, Muller M, Stuker S, Waibel A. Multilingual shifting deep bottleneck features for low-resource ASR. In: ICASSP 2014: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2014 May 04–09; Florence, Italy. p. 5607–5611. http://dx.doi.org/10.1109/ICASSP.2014.6854676
13. Vu NT, Imseng D, Povey D, Motlicek P, Schultz T, Bourlard H. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: ICASSP 2014: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2014 May 04–09; Florence, Italy. p. 7639–7643. http://dx.doi.org/10.1109/ICASSP.2014.6855086
14. Sahraeian R, Van Compernolle D, De Wet F. Under-resourced speech recognition based on the speech manifold. In: Interspeech 2015: Proceedings of the International Speech Communication Association; 2015 September 06–10; Dresden, Germany. p. 1255–1259.
15. Sahraeian R, Van Compernolle D, De Wet F. On using intrinsic spectral analysis for low-resource languages. In: SLTU 2014: Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages; 2014 May 14–16; St Petersburg, Russia. p. 61–65.
16. De Wet F, De Waal A, Van Huyssteen GB. Developing a broadband automatic speech recognition system for Afrikaans. In: Interspeech 2011: Proceedings of International Speech Communication Association; 2011 August 27–31; Florence, Italy. p. 3185–3188.
17. Heeringa W, De Wet F, Van Huyssteen GB. Afrikaans and Dutch as closely-related languages: A comparison to West Germanic languages and Dutch dialects. Stellenbosch Papers in Linguistics Plus. 2015;47:1–18. http://dx.doi.org/10.5842/47-0-649

18. Imseng D, Bourlard H, Garner PN. Boosting under-resourced speech recognizers by exploiting out of language data – Case study on Afrikaans. In: SLTU 2012: Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-resourced Languages; 2012 May 07–09; Cape Town, South Africa. Cape Town: SLTU; 2012. p. 60–67.

19. Despres J, Fousek P, Gauvain J, Gay S, Josse Y, Lamel L, et al. Modeling northern and southern varieties of Dutch for STT. In: Interspeech 2009: Proceedings of the International Speech Communication Association. 2009 September 06–10; Brighton, UK. p. 96–99.

20. Imseng D, Motlicek P, Bourlard H, Garner PN. Using out-of-language data to improve an under-resourced speech recognizer. Speech Commun. 2014;56:142–151. http://dx.doi.org/10.1016/j.specom.2013.01.007

21. Oostdijk N. The spoken Dutch corpus: Overview and first evaluation. In: LREC 2000: Proceedings of the Second International Conference on Language Resources and Evaluation; 2000 May 31 – June 02; Athens, Greece. p. 887–894.

22. Barnard E, Davel MH, Van Heerden C, De Wet F, Badenhorst J. The NCHLT speech corpus of the South African languages. In: SLTU 2014: Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages; 2014 May 14–16; St Petersburg, Russia. p. 194–200.

23. Young S, Evermann G, Gales M. The HTK book [document on the Internet]. c2009 [cited 2016 Jan 12]. Available from: http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf

24. Leggetter CJ, Woodland PC. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput Speech Lang. 1995;9(2):171–185. http://dx.doi.org/10.1006/csla.1995.0010

25. Gales MJF, Woodland PC. Mean and variance adaptation within the MLLR framework. Comput Speech Lang.1996;10(4):249–264. http://dx.doi.org/10.1006/csla.1996.0013

26. Gauvain JL, Lee CH. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans Speech Audio Process. 1994;2(2):291–298. http://dx.doi.org/10.1109/89.279278

27. Kleynhans N, Molapo R, De Wet F. Acoustic model optimisation for a call routing system. In: PRASA 2012: Proceedings of the 23rd Meeting of the Pattern Recognition Association of South Africa; 2012 November 29–30; Pretoria, South Africa. p. 165–172.

28. Young SJ, Odell JJ, Woodland PC. Tree-based state tying for high accuracy acoustic modelling. In: HLT 1994: Proceedings of the Workshop on Human Language Technology; 1994 March 08–11; Plainsboro, NJ, USA. p. 307–312. http://dx.doi.org/10.3115/1075812.1075885

Antifungal actinomycetes associated with the pine bark beetle, *Orthotomicus erosus*, in South Africa

AUTHORS:
Zander R. Human[1]
Bernard Slippers[2]
Z. Wilhelm de Beer[1]
Michael J. Wingfield[1]
Stephanus N. Venter[1]

AFFILIATIONS:
[1]Department of Microbiology and Plant Pathology, Forestry and Agriculture Biotechnology Institute, University of Pretoria, Pretoria, South Africa
[2]Department of Genetics, Forestry and Agriculture Biotechnology Institute, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Wilhelm de Beer

EMAIL:
Wilhelm.debeer@fabi.up.ac.za

DATES:
Received: 18 July 2016
Revised: 29 Aug. 2016
Accepted: 31 Aug. 2016

KEYWORDS:
Streptomyces; Ophiostomatales; *Pinus*; mutualism; antibiotics

HOW TO CITE:
Human ZR, Slippers B, De Beer ZW, Wingfield MJ, Venter SN. Antifungal actinomycetes associated with the pine bark beetle, *Orthotomicus erosus*, in South Africa. S Afr J Sci. 2017;113(1/2), Art. #2016-0215, 7 pages. http://dx.doi.org/10.17159/sajs.2017/20160215

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
DST-NRF Centre of Excellence in Tree Health Biotechnology; National Research Foundation (South Africa); University of Pretoria

Actinomycete bacteria are often associated with insects that have a mutualistic association with fungi. These bacteria are believed to be important to this insect–fungus association as they produce antibiotics that exclude other saprophytic fungi from the immediate environment. The aim of this study was to investigate the presence of potentially protective actinomycetes associated with *Orthotomicus erosus,* an alien invasive pine bark beetle, in South Africa. This bark beetle and its relatives have an association with Ophiostomatales species which are often the only fungi found in the bark beetle galleries. We hypothesised that antibiotic-producing actinomycetes could be responsible for the paucity of other fungi in the galleries by producing compounds to which the *Ophiostoma* spp. are tolerant. Several actinomycetes in the genus *Streptomyces* and one *Gordonia* sp. were isolated from the beetle. Interestingly, most isolates were from the same species as actinomycetes associated with other pine-infesting insects from other parts of the world, including bark beetles and the woodwasp *Sirex noctilio*. Most actinomycetes isolated had strong antifungal properties against the selected test fungi, including *Ophiostoma ips*, which is the most common fungal symbiont of *Orthotomicus erosus.* Although the actinomycetes did not benefit *Ophiostoma ips* and the hypothesis was not supported, their sporadic association with *Orthotomicus erosus* suggests that they could have some impact on the composition of the fungal communities present in the bark beetle galleries, which is at present poorly understood.

Significance:

- Discovery of four putative undescribed *Streptomyces* spp. with antibiotic potential
- First record of the introduction of actinomycete bacteria with pine-infesting insects into South Africa
- Actinomycetes from South Africa group with undescribed *Streptomyces* spp. from pine-infesting insects of North America

Introduction

The European bark beetle *Orthotomicus erosus* (Curculionidae: Scolytinae) is an introduced pine-infesting pest in South Africa.[1] It typically infests stressed or dying trees and introduces blue stain fungi that invade the sapwood and depreciate the timber value.[1,2] The blue stain fungus *Ophiostoma ips* (Ascomycota: Ophiostomatales) is the dominant associate of *O. erosus* in South Africa, but several other related fungi co-occur with this species in the beetle galleries.[3] Although *O. ips* consistently co-occurs with *O. erosus* at varying frequencies[3,4], it is not a serious pathogen to living pine trees[5] and its role as symbiont remains uncertain, as is the case with most ophiostomatoid fungi associated with conifer-infesting bark beetles[6]. Although the fresh bark beetle galleries represent an environment rich in nutrients and other growth substrates, it is remarkable that this niche is seldom overgrown with common mould fungi.

The presence of primarily *Ophiostoma* spp. and their relatives and the lack of contaminating moulds in the galleries of the beetles has raised the question as to the factors that increase the fitness of fungi commonly associated with the insect, over other fungi expected to be found in these environments. One possibility is that antibiotic-producing actinomycetes could play a role in this symbiotic relationship. In this regard, actinomycetes are the most important producers of antibiotics[7] with more than 100 000 antibiotic compounds estimated to be produced by members of the genus *Streptomyces*[8]. The formation of heat and desiccation-resistant spores is also a common feature of these bacteria[7] and the hydrophobicity of their spores can facilitate their transport[9]. All these features could be important in their association with arthropods such as insects and mites.[10]

There are various symbiotic communities in which insects exploit actinomycetes to produce metabolites for protection.[10-13] Examples include attine ants (Attini: Formicidae) that have co-evolved with actinomycetes in the genus *Pseudonocardia* to protect their food source against a parasite.[11] The ants cultivate a basidiomycete fungus that is used for nutrition,[11] but the fungal garden can be parasitised by another fungus (*Escovopsis* spp.), thus threatening the survival of the entire colony. Secondary metabolites produced by the actinomycetes residing on the ants' integuments protect the crop by inhibiting the growth of *Escovopsis.*[11,12] Actinomycete–insect symbioses also occur with the southern pine beetle, *Dendroctonus frontalis* (Curculionidae: Scolytinae), in its native environment in the USA.[13] Survival of larvae in the galleries of these beetles is negatively impacted by *Ophiostoma minus*, a fungal symbiont of mites that competes with the fungal mutualist, an *Entomocorticium* sp., of the beetle. *Streptomyces* symbionts in the mycangium of *D. frontalis* produce antibiotics that inhibit the growth of *O. minus,* whereas the mutualistic fungus is tolerant to the antibiotics.[13]

The aim of this study was to isolate and identify putative actinomycete symbionts from the invasive *O. erosus* in South Africa, and to determine whether they have antifungal properties that might be important in this niche. We hypothesised that actinomycete symbionts of *O. erosus* produce antifungal compounds, similar to cycloheximide that is known to have broad antifungal effects except on *Ophiostoma* spp. and their relatives.[14,15] We expect that these compounds will negatively affect the fitness of potentially competing saprophytic fungi from the galleries.

Materials and methods

Bacterial isolation

Orthotomicus erosus galleries were collected from dead *Pinus patula* trees in the Lothair plantation, Mpumalanga Province, South Africa. In total, 40 beetles were removed from these galleries and individually crushed in sterilised 10% phosphate-buffered saline solution (PBS). Three tenfold serial dilutions were prepared for each sample using 10% PBS.

An aliquot of 100 μL of each dilution was inoculated onto chitin agar[16] in duplicate, supplemented with antibiotics (cycloheximide 5 mg/L and nystatin 10 000 units/L).[12] These plates were incubated for approximately 30 days at 28 °C during which they were inspected daily for growth. Isolates presumed to be actinomycetes based on their morphology were selected and inoculated onto yeast malt extract glucose agar (YMEA) – consisting of 1% malt extract (Biolab Diagnostics, Johannesburg, South Africa), 0.4% yeast extract (Oxoid, Hampshire, England), 0.4% D-glucose (Merck Chemicals, Johannesburg, South Africa) and 0.12% bacteriological agar (Biolab Diagnostics)[12] – and incubated at 28 °C until sufficient growth had occurred.

DNA sequencing

Fifteen isolates were collected and DNA was extracted using a Quick-gDNA™ MiniPrep kit (Zymo Research, Orange, CA, USA). The 16S rRNA gene was amplified and partially sequenced using the primers pA and pH previously designed by Edwards et al.[17] Subsequently, the *trpB* (tryptophan synthase β-subunit), *rpoB* (RNA polymerase β-subunit) and *gyrB* (DNA gyrase β-subunit) genes were amplified for eight of the strains grouping with other pine associated *Streptomyces* isolates. Amplification of the *trpB* and *rpoB* genes followed the methods of Guo et al.[18] and that for the *gyrB* gene followed the method of Rong et al.[19] All reactions were performed on a Veriti™ Thermal Cycler (Applied Biosystems, Foster City, CA, USA) using Super-Therm Taq polymerase (Southern Cross Biotechnology, Cape Town, South Africa). Polymerase chain reaction (PCR) products were verified using agarose gel electrophoresis and purified using *E. coli* exonuclease I and alkaline phosphatase.

Sequencing of the PCR products for 16S rRNA, *gyr*B, *rpo*B, *trp*B was performed using the ABI BigDye Terminator v3.1 (Applied Biosystems) following the protocols previously described.[17-19] Precipitation of sequencing reactions was done through sodium acetate precipitation and the sequencing products were analysed on an ABI 3130 sequence analyser (Applied Biosystems).

A BLASTN[20] search was performed to identify the closest matching sequences in GenBank[21]. A search was also performed against the Ribosomal Database Project (RDP)[22] using the Seqmatch platform. Similar sequences were downloaded for phylogenetic analyses. Sequences representing the closely related type strains were also obtained.

Phylogenetic analyses

To determine the relationship between sequences obtained and the published reference sequences, a phylogenetic tree based on the 16S rRNA sequence alignment was constructed employing a maximum likelihood analysis. The alignment was made using the online version of MAFFT version 6.[23] In addition, a concatenated alignment consisting of the *gyrB, rpoB* and *trpB* genes of a selected number of isolates were made using SequenceMatrix.[24] The appropriate nucleotide substitution model was selected for both sets of genes using JModeltest version 2.1.1.[25,26] Phylogenetic tree construction was performed using the maximum likelihood approach in PhyML version 3.0.[25] The models used were TIM3+I+G (16S rRNA) and GTR+I+G (*gyrB, trpB* and *rpoB*).[25,26] Trees were visualised using Mega 5.05.[27]

Dual-plate bioassay challenges

A preliminary assay was performed for all 15 actinomycete isolates to serve as a selection step for further antifungal assays. Four different isolates were inoculated onto the four quadrants of YMEA plates and these were incubated for 2 weeks. These test plates were then inoculated with a *Trichoderma* sp. by placing a plug, 15 mm in diameter, at the centre of the pre-inoculated plate. Any isolates showing antifungal activity were subjected to further in-vitro assays.

Bioassay challenges were done for the selected isolates following the approach of Cafaro and Currie[12]. *Streptomyces* isolates were inoculated (10 mm in diameter) onto a 90-mm Petri dish containing YMEA.[12] These plates were incubated for 21 days. Three fungal species were chosen for use in bioassays. These three species were the most common fungal associate of *O. erosus* (*O. ips*), a common saprophyte (*Trichoderma* sp.) and a commonly occurring pine endophyte (*Diplodia sapinea*). *D. sapinea* and *Trichoderma* spp. are regularly isolated from pine wood and are potential competitors of *O. ips* (Table 1). Isolates were obtained from the Culture Collection of the Forestry and Agricultural Biotechnology Institute (FABI) of the University of Pretoria. A single 15-mm fungal plug was inoculated at the edge of the 21-day-old plates and incubated at 25 °C until sufficient growth on control plates was observed. Two repeats were performed for each pairing. Plates were examined and the average zone of inhibition measured for all the bioassay challenges.

Table 1: Results of bioassays in which actinomycete isolates were tested for their ability to inhibit growth of *Trichoderma* sp. (a saprophyte), *Diplodia sapinea* (an endophyte) and *Ophiostoma ips* (a fungal symbiont)

Actinomycete isolate	Fungal isolate		
	***Trichoderma* sp.**	***Diplodia sapinea***	***Ophiostoma ips***
BCC1197	+++	+++	+++
BCC1193	++	++	+++
BCC1195	++	++	+++
BCC1194	++	++	+++
BCC1189	++	++	+++
BCC1191	++	++	+++
BCC1188	++	++	+++
BCC1190	++	++	+++
BCC1204	+	++	+++
BCC1196	++	+	++
BCC1198	+	+	++

Inhibition zones: +++, 15 mm; ++, 10 mm; +, 5 mm

Following the above-mentioned challenges, the beetle symbiont fungus *O. ips*, and one of the bacteria (isolate BCC1988), were simultaneously inoculated on YMEA plates. Isolate BCC1988, representative of the most commonly recurring phylogenetic group associated with the beetle, was inoculated at the centre of 90-mm Petri dishes and a single 15-mm *O. ips* plug was inoculated at the edge of the same plate. This plate was incubated at 25 °C until sufficient growth had been observed and the result was recorded. This trial was repeated.

Results

Isolates and DNA sequence based identifications

Fifteen actinomycete isolates were obtained from the 40 *O. erosus* individuals collected in this study (Table 2; Figure 1). These isolates were obtained from 11 different beetles, with one isolate representative of each actinomycete taxon selected from each beetle. Partial 16S rRNA sequences were obtained for all 15 isolates. Isolates were all initially identified based on the best matches for the 16S rRNA gene sequences in GenBank and the RDP database. Based on these data, all but one isolate (a species of *Gordonia*) belonged to the genus *Streptomyces*.

Isolates were deposited in the Bacterial Culture Collection (BCC) of FABI, and 16S rRNA (Table 2) and protein-coding gene sequences (Table 2) were deposited in NCBI GenBank.

Phylogenetic analysis

16S rRNA

According to RDP Seqmatch, 8 of the 14 isolates initially identified as *Streptomyces* spp. had sequences most similar to the sequence of the type strain of *S. ambofaciens*. In the 16S rRNA based phylogenetic analysis, these isolates grouped in a single clade together with isolates from other pine-infesting insects. This clade had 94% bootstrap support (Figure 1). Another isolate from the southern pine beetle[13,28] associated with this group, but the grouping was not well supported. None of the type strains' sequences formed part of this clade.

Two of the remaining isolates clustered closely with *S. sanglieri* and *S. atratus* in a well-supported group (99%). A further two isolates were related to the latter isolates but were clearly separated from the initial clade and formed a separate clade. Of the remaining isolates, BCC1197 grouped most closely with the type strain sequence (*Streptomyces alni*), but was well separated and there was no clear bootstrap support for their association.

Table 2: *Streptomyces* isolates for which sequences were produced in this study (GenBank accession numbers in bold type) and reference sequences generated in previous studies (GenBank accession numbers in normal type)

			Gene region		
Species	**Isolate numbers***	**16S rRNA**	***trp*B**	***rpo*B**	***gyr*B**
Streptomyces phaeoluteichromatogenes^T	NRRL B-5799	AJ391814	**HG423654**	**HG423678**	**HG423666**
S. misionensis^T	CBS 885.69	EF178678	**HG423655**	**HG423679**	**HG423667**
S. ambofaciens^T	NRRL ISP-5053	AB184182	**HG423656**	**HG423681**	**HG423668**
S. lienomycini^T	NRRL B-16371	AJ781353	**HG423657**	**HG423683**	**HG423669**
S. rubrogriseus^T	NRRL B-16375	AB184681	**HG423658**	**HG423684**	**HG423672**
S. collinus^T	NRRL B-5412	AB184123	**HG423659**	**HG423680**	**HG423671**
S. levis^T	NRRL B-16370	AB184670	**HG423660**	**HG423682**	**HG423670**
S. janthinus^T	CBS 909.68	AB184851	**HG423661**	**HG423685**	**HG423673**
S. albidoflavus^T	CBS 416.34	AB184255	FJ406450	FJ406439	FJ406417
S. cinereorectus[1]	NRRL B-16360	AB184646	EF661795	EF661774	EF661732
Streptomyces sp. SA3ActG	SA3ActG[29]	HM235477	NZ_ADXA01000162	NZ_ADXA01000004	NZ_ADXA01000012
Streptomyces sp. SPB078 [13,28]	SPB078[10,25]	EU798708	NZ_GG657742	NZ_GG657742	NZ_GG657742
Streptomyces sp. SPB074	SPB074[28]	EU798707	NZ_GG770539	NZ_GG770539	NZ_GG770539
Streptomyces sp.	BCC1191	**HG423693**	**HG423662**	**HG423689**	**HG423675**
Streptomyces sp.	BCC1192	**HG423694**	**HG423663**	**HG423688**	**HG423677**
Streptomyces sp.	BCC1195	**HG423697**	**HG423664**	**HG423686**	**HG423676**
Streptomyces sp.	BCC1188	**HG423690**	**HG423665**	**HG423687**	**HG423674**
Streptomyces sp.	BCC1194	**HG423696**	**KM031100**	**KM031094**	**KM031096**
Streptomyces sp.	BCC1189	**HG423691**	**KM031103**	**KM031095**	**KM031099**
Streptomyces sp.	BCC1193	**HG423695**	**KM031101**	**KM031093**	**KM031098**
Streptomyces sp.	BCC1190	**HG423692**	**KM031102**	**KM031092**	**KM031097**
Streptomyces sp.	BCC1196	**HG423703**			
Streptomyces sp.	BCC1197	**HG423702**			
Streptomyces sp.	BCC1198	**HG403701**			
Streptomyces sp.	BCC1200	**HG403700**			
Streptomyces sp.	BCC1203	**HG423699**			
Streptomyces sp.	BCC1204	**HG423698**			

^{T}Type strains

**NRRL, Northern Regional Research Laboratory culture collection, maintained by the USDA Agricultural Research Service, Peoria, Illinois, USA; CBS, Centraalbureau Voor Schimmelcultures, Utrecht, the Netherlands; BCC, Bacterial Culture Collection, Forestry and Agriculture Biotechnology Institute, University of Pretoria, Pretoria, South Africa. References to the publications in which three unnamed isolates (SA3ActG, SPB078, SPB074) from private collections were studied, are provided at the isolate numbers.*

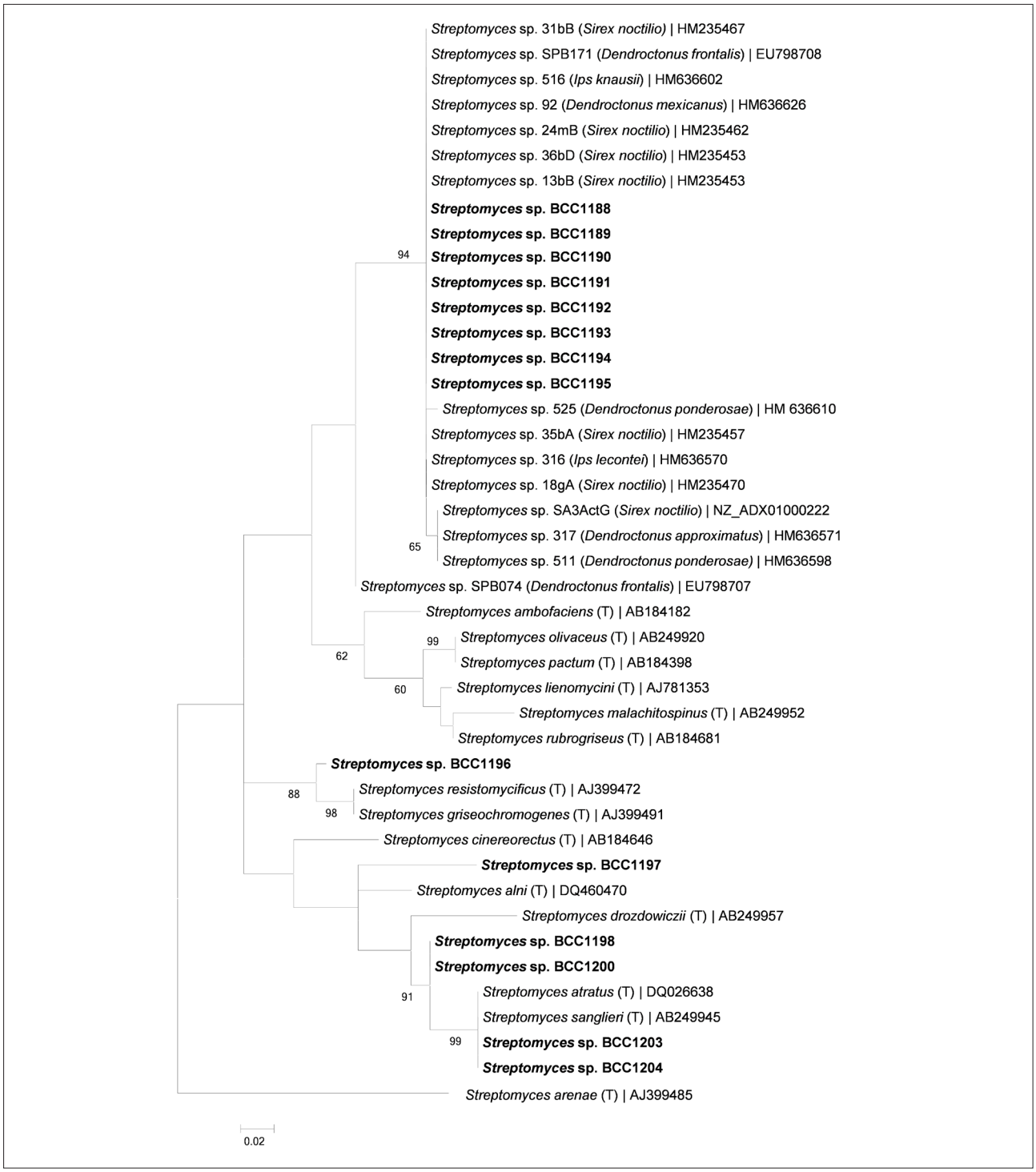


Figure 1: Maximum likelihood tree representing the 16S rRNA gene of all isolates from this study (in bold type) with closest matching type strains and isolates from other pine-infesting insects. Type strains are indicated by (T) and host names are included for sequences from *Streptomyces* spp. from pine-infesting insects. *Streptomyces arenae* was used as the outgroup.

Although separate from the type strains of *S. griseochromogenes* and *S. resistomycificus,* BCC1196 formed a well-supported (88%) cluster with these type strain sequences (Figure 1).

Multi-locus sequence analysis

All *Streptomyces* spp. isolates from pine-infesting insects, including eight isolates from this study, formed a clade with 100% bootstrap support in the multi-gene phylogeny (Figure 2). This clade was split between a branch consisting of a single isolate from *D. frontalis* (*Streptomyces* spp. SPB074)[13,28] and another clade that contained two branches, one with the eight isolates from this study and the other a clade with isolates from *Sirex noctilio* (*Streptomyces* sp. SA3ActG)[29] and *D. frontalis* (*Streptomyces* spp. SPB074) [28]. All of these branches were well supported. The type strain sequence matching most closely to the larger clade, including all isolates from pine-infesting insects, was *S. albidoflavus*.

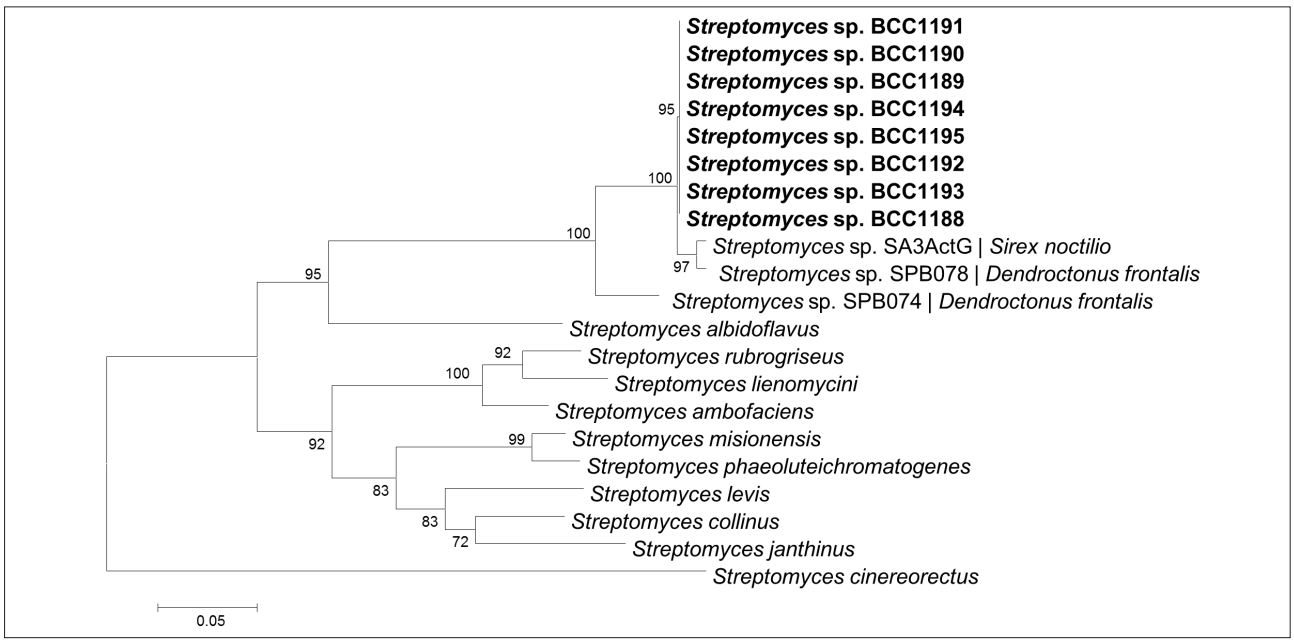


Figure 2: Maximum likelihood tree representing concatenated nucleotide sequence alignments of the *gyrB*, *rpoB* and *trpB* genes of nine *Streptomyces* type strains, eight isolates from this study (bold type) and three *Streptomyces* strains isolated from other pine-infesting insects, as retrieved from the literature.[10,25,26] *Streptomyces cinereorectus* was used as the outgroup.

Dual-plate bioassay challenges

In preliminary antifungal assays, 11 of the 15 cultures were found to have moderate to strong inhibitory effects on the *Trichoderma* sp. These 11 isolates were used in the subsequent in-vitro antifungal assays (Table 1). All of these actinomycete strains inhibited the three fungal species *D. sapinea*, *Trichoderma* sp. and *O. ips*, but to varying degrees. Of the three fungi, *O. ips* was the most strongly inhibited (Table 1).

The phylogenetically related isolates had similar levels of activity against the test fungi (Table 1). The isolates most similar to *S. ambofaciens* (BCC1988, BCC1989, BCC1990, BCC1991, BCC1992, BCC1993, BCC1994, BCC1995) all displayed moderate to strong (6–10 mm) levels of inhibition against both the *Trichoderma* sp. and *D. sapinea*. These isolates had even higher levels of inhibition when tested against *O. ips*. Most other isolates with antifungal activity had moderate to strong inhibitory activity against *Trichoderma* sp. and *D. sapinea*, with a higher or very strong activity against *O. ips*. Isolate BCC1197 had very strong inhibitory activity against all test fungi.

When isolate BCC1188, representing the group of most common actinomycete isolates, and *O. ips* were simultaneously inoculated on fresh growth medium, in contrast to the previous assay in which *O. ips* was inhibited, fungal growth occurred until they came into close contact (Figure 3). Furthermore, living fungal material could still be isolated from the edges of the *O. ips* culture despite inhibition, showing that the fungus had not been killed by the actinomycete.

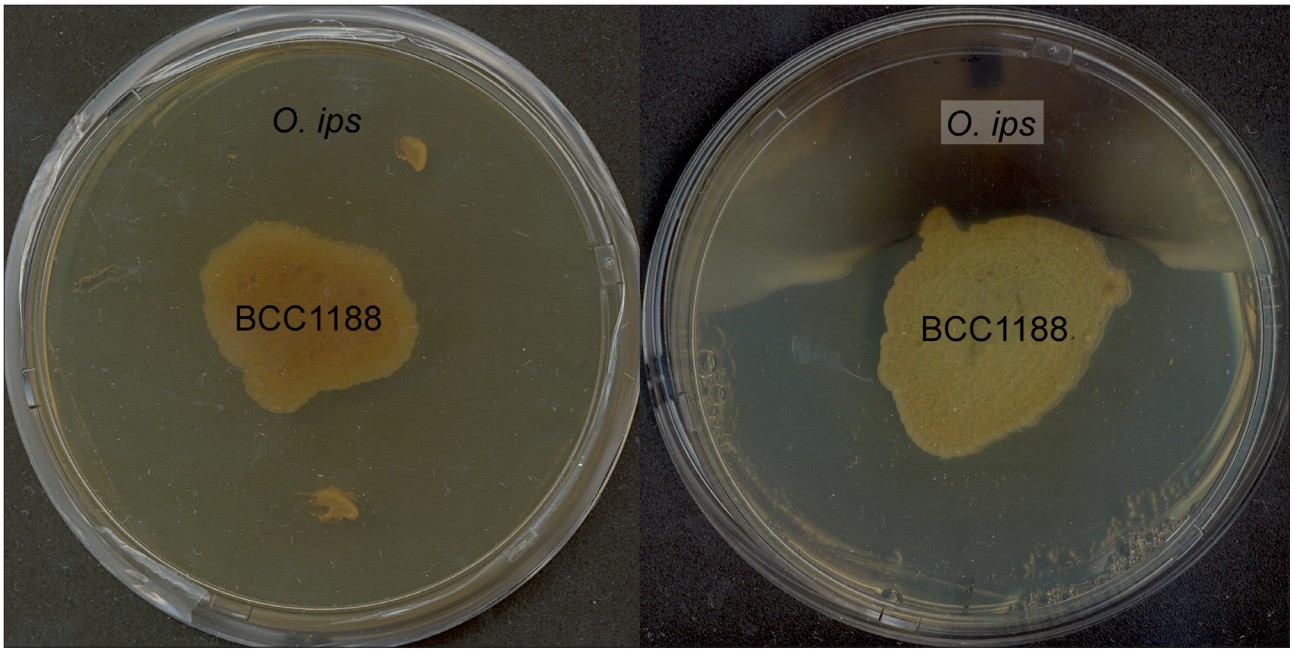


Figure 3: Bioassay challenge with isolate BBC1188 and *Ophiostoma ips* simultaneously inoculated (right) and bacteria inoculated 2 weeks before fungi (left) on yeast malt extract agar. This figure illustrates how a fungal isolate can grow uninhibited with a bacterial culture when inoculated at the same time.

Discussion

In this study, 15 actinomycete isolates were collected from adult *O. erosus* beetles that infest *Pinus* spp. in South Africa. These bacteria were identified as actinomycetes based on colony morphology and comparisons of the 16S rRNA sequence data. The majority of these isolates represented *Streptomyces* spp. Although relatively few isolates of actinomycetes were recovered during this preliminary study, these bacteria appear to frequently encounter *O. erosus*. This is the first time that members of the actinomycetes have been reported from this or any other tree-infesting bark beetle in South Africa.

Based on the 16S rRNA phylogeny, one group of bacteria was consistently isolated from *O. erosus*. Comparison of the eight strains included in this group revealed that they grouped within one of the three clades of *Streptomyces* spp. that were identified by Hulcr et al.[30] This clade also included a strain isolated from the pine-infesting beetle *D. frontalis*[13,28] and cellulose degrading *Streptomyces* spp. associated with a pine-infesting siricid wasp, *Sirex noctilio*[30,31]. Further analysis based on several housekeeping genes showed that isolates from *S. noctilio* and *D. frontalis* from the USA were closely related with the isolates from *O. erosus* in South Africa. These isolates most likely represent the same species. This lineage is also associated with another isolate from *D. frontalis*, and they most likely share a common ancestor. The clade formed by our isolates and those from *S. noctilio* and *D. frontalis* was identified in another study.[32] Several previously reported isolates from pine-infesting insects[13,28,30,31] were grouped into two clades, based on their core genomes. One of these clades, containing a single isolate from both *S. noctilio*[31] and *D. ponderosae* had remarkable lignocellulose digestion capacity. Another group, containing the exact isolates from *S. noctilio*[31] and *D. frontalis*[13] used in our phylogeny, had significantly less lignocellulose hydrolytic capabilities. The data suggest that this undescribed *Streptomyces* species has a strong association with insects associated with pine trees and that it could be a common inhabitant in this niche. According to Book et al. [32], one of their clades of *Streptomyces* isolates is well adapted to thrive and utilise the abundant lignocellulosic substrates in the pine tree environment. The exact niche for members of the clade containing isolates from this study remains unclear, but our findings suggest that they are common and often encountered by pine-infesting insects, although a strong biological association with *O. erosus* is precluded by their low frequency.

The low frequency at which the actinomycetes were isolated in this study corresponded with the findings of Hulcr et al.[30] who found that *Streptomyces* associates of North American bark beetles occur at low frequency. These low frequencies preclude definite conclusions regarding specific interactions between beetles and actinomycetes. The low frequency of isolation also suggests that this association is not essential for the beetles and fungi involved. However, our results suggest that it is most likely not a completely random association. Wider sampling, throughout the life cycle of the beetle and using more sensitive techniques (e.g. next-generation sequencing), will be required to conclude on the true frequency of interaction between these organisms. One possible scenario is that *Streptomyces* spores are more numerous on beetles when emerging from galleries and less abundant on beetles at the end of their life cycle – which is when they were sampled in this study. Contaminating bacteria from galleries could also preclude successful isolation of slower-growing actinomycetes.

The bioassays to test the potential effect of the *Streptomyces* spp. on fungi in *O. erosus* galleries showed that several of these bacteria have antifungal properties. The selection of test fungi used for the assay included a common saprophyte (*Trichoderma* sp.), an endophyte and opportunistic pathogen of *Pinus* spp. (*D. sapinea*), and the fungal symbiont of *O. erosus*. The levels of inhibition varied amongst test strains, ranging from weak to very strong. The most frequently isolated strains were able to inhibit all test fungi, including *O. ips*, the most common fungal symbiont to *O. erosus*. Previous studies on insect–fungus associated actinomycetes have suggested that beneficial fungi should be inhibited to a lesser extent than parasitic or other saprobic fungi.[11-13] The beneficial fungal associate of the southern pine beetle is weakly inhibited compared with the parasitic *O. minus*.[13] This is also commonly believed to be the case in fungus-growing ants, in which the observed inhibition against *Escovopsis* spp. is higher than that against the mutualistic basidiomycetes[11,12], although some have suggested that the beneficial cultivar is also harmed[33]. This result suggests that *Streptomyces* isolates collected in this study are unlikely to be associates of *O. erosus,* but may be linked to this beetle through another common partner, such as pine trees or mites.

Although it did not appear that the isolated *Streptomyces* spp. directly benefitted *O. ips*, it remains possible that they play some role in the ecology of these fungi and the associated beetles. For example, the fungal symbionts of the beetles such as *O. ips* are inoculated into the newly formed galleries at the time of infestation, either directly from the beetle's exoskeleton or with the help of mites.[34,35] These fungi become established and dominate the niche, and it is likely that contaminating saprophytes enter the niche only at a later stage. If the antibiotic-producing Actinobacteria are introduced at the same time as the fungal associates, there would be sufficient opportunity for the fungus to establish itself and penetrate the wood before widespread colonisation of the bacteria. However, once the bacteria are established and producing antibiotics in the galleries, these would then be protected against possible harmful saprophytes that are expected to enter later. Simultaneous inoculation of *Streptomyces* spp. and the fungal symbiont on medium showed that *O. ips* can initially colonise large amounts of the resource and grow to the edge of the bacterial colony, before inhibition is seen. The results might suggest that *O. ips* can survive, while other saprophytes subsequently introduced may be inhibited completely. However, as it is not explicitly known that *O. ips* is beneficial to its bark beetle symbionts, the possibility exists that it is a mite associate that has no beneficial effects for the beetle, or that it might even be detrimental to beetle fitness and development. Therefore, partial inhibition of *O. ips* does not necessarily equate to having an impact on the survival of *O. erosus*.

This study represents the first investigation of actinomycetes associated with insects in South Africa and we have shown that *Streptomyces* spp. are occasional symbionts of *O. erosus* in this country. Several of the isolates formed part of a group of symbionts associated with bark beetles and a pine-infesting woodwasp in North America.[13,30] This finding suggests some link between this *Streptomyces* species and the *Pinus* environment, which deserves further investigation. This species could have entered South Africa with *Pinus* planting stock or, given that they are apparently common to other pine-infesting bark beetles, it is likely that they entered South Africa with these insects. Future work should investigate the presence of similar *Streptomyces* spp. on other insects associated with *Pinus* spp. across different geographical ranges. The specific role in the galleries of *O. erosus* and the biology of this bark beetle should also be surveyed using culture-independent methods.

Authors' contributions

S.N.V., M.J.W., Z.W.d.B. and B.S. conceptualised the research. Z.R.H., S.N.V. and Z.W.d.B. conducted the experiments and analysed the data. All authors contributed to the interpretation of the results, and the writing and editing of the manuscript.

References

1. Tribe GD. Phenology of *Pinus radiata* log colonization and reproduction by the European bark beetle *Orthotomicus erosus* (Wollaston) (Coleoptera: Scolytidae) in the south-western Cape province. J Entomol Soc S Afr. 1990;53:117–126.
2. Hurley BP, Hatting HJ, Wingfield MJ, Klepzig KD, Slippers B. The influence of *Amylostereum areolatum* diversity and competitive interactions on the fitness of *Sirex* parasitic nematode *Deladenus siricidicola*. Biol Control. 2012;61:207–214. http://dx.doi.org/10.1016/j.biocontrol.2012.02.006
3. Zhou XD, De Beer ZW, Wingfield BD, Wingfield MJ. Ophiostomatoid fungi associated with three pine-infesting bark beetles in South Africa. Sydowia. 2001;53(2):290–300.
4. Romón P, Zhou X, Iturrondobeitia JC, Wingfield MJ, Goldarazena A. *Ophiostoma* species (Ascomycetes: Ophiostomatales) associated with bark beetles (Coleoptera: Scolytinae) colonizing *Pinus radiata* in northern Spain. Can J Microbiol. 2007;53(6):756–767. http://dx.doi.org/10.1139/W07-001

5. Zhou XD, De Beer ZW, Wingfield BD, Wingfield MJ. Infection sequence and pathogenicity of *Ophiostoma ips*, *Leptographium serpens* and *L. lundbergii* to pines in South Africa. Fungal Divers. 2002;10:229–240.

6. Six DL, Wingfield MJ. The role of phytopathogenicity in bark beetle-fungus symbioses: A challenge to the classic paradigm. Annu Rev Entomol. 2011;56(1):255–272. http://dx.doi.org/10.1146/annurev-ento-120709-144839

7. Lechevalier HA, Lechevalier MP. Biology of actinomycetes. Annu Rev Microbiol. 1967;21:71–100. http://dx.doi.org/10.1146/annurev.mi.21.100167.000443

8. Watve MG, Tickoo R, Jog MM, Bhole BD. How many antibiotics are produced by the genus *Streptomyces?* Arch Microbiol. 2001;176:386–390. http://dx.doi.org/10.1007/s002030100345

9. Ruddick SM, Williams ST. Studies on the ecology of actinomycetes in soil. V. Some factors influencing the dispersal and adsorbtion of spores in soil. Soil Biol Biochem. 1972;4:93–100. http://dx.doi.org/10.1016/0038-0717(72)90046-6

10. Kaltenpoth M. Actinobacteria as mutualists: General healthcare for insects? Trends Microbiol. 2009;17:529–535. http://dx.doi.org/10.1016/j.tim.2009.09.006

11. Currie CR, Scott JA, Summerbell RC, Malloch D. Fungus-growing ants use antibiotic producing bacteria to control garden parasites. Nature. 1999;398:701–704. http://dx.doi.org/10.1038/19519

12. Cafaro MJ, Currie CR. Phylogenetic analysis of mutualistic filamentous bacteria associated with fungus-growing ants. Can J Microbiol. 2005;51:441–446. http://dx.doi.org/10.1139/w05-023

13. Scott JJ, Oh DC, Yuceer MC, Klepzig KD, Clardy J, Currie CR. Bacterial protection of a beetle-fungus mutualism. Science. 2008;322:63. http://dx.doi.org/10.1126/science.1160423

14. Whiffen AJ. The activity in vitro of cycloheximide (actidione) against fungi pathogenic to plants. Mycologia. 1950;42:253–258. http://dx.doi.org/10.2307/3755437

15. Harrington TC. Ecology and evolution of mycophagous bark beetles and their fungal partners. In: Vega FE, Blackwell M, editors. Ecological and evolutionary advances in insect-fungal associations. New York: Oxford University Press; 2005. p. 257–291.

16. Hsu SC, Lockwood JL. Powdered chitin agar as a selective medium for enumeration of actinomycetes. Appl Microbiol. 1975;29:422.

17. Edwards U, Rogall T, Blöcker H, Ernde M, Böttger E. Isolation and direct complete nucleotide sequence determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res. 1989;17:7843–7853. http://dx.doi.org/10.1093/nar/17.19.7843

18. Guo Y, Zheng W, Rong X, Huang Y. A multilocus phylogeny of the *Streptomyces griseus* 16S rRNA gene clade: use of multilocus sequence analysis for streptomycete systematics. Int J Syst Evol Microbiol. 2008;58:149–159. http://dx.doi.org/10.1099/ijs.0.65224-0

19. Rong X, Guo Y, Huang Y. Proposal to reclassify the *Streptomyces albidoflavus* clade on the basis of multilocus sequence analysis and DNA–DNA hybridization, and taxonomic elucidation of *Streptomyces griseus* subsp. *solvifaciens*. Syst Appl Microbiol. 2009;35:7–18. http://dx.doi.org/10.1016/j.syapm.2009.05.003

20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2

21. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2005;33:D34–D38. http://dx.doi.org/10.1093/nar/gki063

22. Maidak BL, Cole JR, Lilburn TG, Parker CTJ, Saxman PR, Farris RJ, et al. The RDP-II (Ribosomal Database Project). Nucleic Acids Res. 2001;29:179–206. http://dx.doi.org/10.1093/nar/22.17.3485

23. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: Improvement in accuracy of multiple sequence alignments. Nucleic Acids Res. 2005;33:511–518. http://dx.doi.org/10.1093/nar/gki198

24. Vaidya G, Lohman DJ, Meier R. SequenceMatrix: Concatenation software for the fast assembly of multi-gene datasets with character sets and codon information. Cladistics. 2010;27:121–123. http://dx.doi.org/10.1111/j.1096-0031.2010.00329.x

25. Guindon S, Gascuel O. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol. 2003;52:696–704. http://dx.doi.org/10.1080/10635150390235520

26. Darriba D, Taboada GL, Doallo R, Posada D. jModeltest 2: More models, new heuristics and parallel computing. Nat Methods. 2012;9:772. http://dx.doi.org/10.1038/nmeth.2109

27. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA 5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731. http://dx.doi.org/10.1093/molbev/msr121

28. Blodgett JAV, Oh D-C, Cao S, Currie CR, Kolter R, Clardy J. Common biosynthetic origins for polycyclic tetramate macrolactams from phylogenetically diverse bacteria. Proc Natl Acad Sci USA. 2010;107(26):11692–11697. http://dx.doi.org/10.1073/pnas.1001513107

29. Takasuka TE, Book AJ, Lewin GR, Currie CR, Fox BG. Aerobic deconstruction of cellulosic biomass by an insect-associated *Streptomyces*. Sci Reports. 2013;3, Art. #1030, 10 pages. http://dx.doi.org/10.1038/srep01030

30. Hulcr J, Adams AS, Raffa KF, Hofstetter RW, Klepzig KD, Currie CR. Presence and diversity of *Streptomyces* in *Dendroctonous* and sympatric bark beetle galleries across North America. Microbial Ecol. 2011;61:759–768. http://dx.doi.org/10.1007/s00248-010-9797-0

31. Adams AS, Jordan MS, Adams SM, Suen G, Goodwin LA, Davenport KW, et al. Cellulose degrading bacteria associated with the invasive woodwasp *Sirex noctilio.* ISME J. 2011;5:1323–1331. http://dx.doi.org/10.1038/ismej.2011.14

32. Book AJ, Lewin GR, McDonald BR, Takasuka TE, Doering DT, Adams AS, et al. Cellulolytic *Streptomyces* strains associated with herbivorous insects share a phylogenetically linked capacity to degrade lignocellulose. Appl Environ Microbiol. 2014;80(15):4692–4701. http://dx.doi.org/10.1128/AEM.01133-14

33. Sen R, Ishak HD, Estrada D, Dowd SE, Hong E, Mueller UG. Generalized antifungal activity and 454-screening of *Pseudonocardia* and *Amycolatopsis* bacteria in nests of fungus-growing ants. Proc Natl Acad Sci USA. 2009;106(42):17805–17810. http://dx.doi.org/10.1073/pnas.0904827106

34. Moser JC, Perry TJ, Solheim H. Ascospores hyperphoretic on mites associated with *Ips typographus*. Mycol Res. 1989;93:513–517. http://dx.doi.org/10.1016/S0953-7562(89)80045-0

35. Klepzig KD, Moser JC, Lombardero MJ, Ayres MP, Hofstetter RW, Walkinshaw CJ. Mutualism and antagonism: Ecological interactions among bark beetles, mites and fungi. In: Jeger MJ, Spence NJ, editors. Biotic interactions in plant-pathogen associations. Wallingford: CAB International; 2001. p. 237–268. http://dx.doi.org/10.1079/9780851995120.0000

Soil fertility constraints and yield gaps of irrigation wheat in South Africa

AUTHORS:
Nondumiso Z. Sosibo[1,2]
Pardon Muchaonyerwa[2]
Lientjie Visser[1]
Annelie Barnard[1]
Ernest Dube[1]
Toi J. Tsilo[1,3]

AFFILIATIONS:
[1]Agricultural Research Council – Small Grain Institute, Bethlehem, South Africa
[2]Soil Science, School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[3]Life and Consumer Sciences, University of South Africa, Pretoria, South Africa

CORRESPONDENCE TO:
Toi Tsilo

EMAIL:
tsilot@arc.agric.za

DATES:
Received: 15 May 2016
Revised: 30 June 2016
Accepted: 06 Sep. 2016

KEYWORDS:
tillage; wheat yield potential; yield gap analysis; conservation agriculture

HOW TO CITE:
Sosibo NZ, Muchaonyerwa P, Visser L, Barnard A, Dube E, Tsilo TJ. Soil fertility constraints and yield gaps of irrigation wheat in South Africa. S Afr J Sci. 2017;113(1/2), Art. #2016-0141, 9 pages. http://dx.doi.org/10.17159/sajs.2017/20160141

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
National Research Foundation (South Africa); South African Agency for Science and Technology Advancement; Winter Cereal Trust

South Africa currently faces a wheat (*Triticum aestivum* L.) crisis as production has declined significantly over the past few years. The objective of this study was to explore opportunities for improving yields in intensive irrigated wheat production systems of South Africa through analyses of yield gaps, soil fertility constraints and conservation agriculture practices. The study was conducted in the major irrigation wheat production areas across four geographical regions: KwaZulu-Natal, eastern Highveld, warmer northern and cooler central. Actual yield (*Ya*) based on long-term yield data ranged from 5.99±0.15 t/ha to 8.32±0.10 t/ha across different geographical regions. The yield potential (*Yp*) ranged from 7.57 t/ha to 11.45 t/ha. Yield gaps (*Yp*–*Ya*) were in the range of 1.58–3.13 t/ha. Yields could be increased by 26–38% through closing yield gaps. On 88.37% and 13.89% of the fields in the KwaZulu-Natal and warmer northern regions, respectively, there was strong evidence of the practise of conservation agriculture, but none in the other regions. On 42.31% of irrigated wheat fields, soil organic carbon was below 1% at a soil depth of 0–20 cm. Fields in which conservation tillage was practised had double the soil organic carbon of conventionally tilled fields (2.15±0.10% versus 1.02±0.05%), but greater acidity and phosphorus deficiency problems. Sustainable approaches for addressing phosphorus deficiency and acidity under conservation tillage practices need to be sought, especially in the KwaZulu-Natal region.

Significance:
- Opportunities for improving wheat yields in South Africa need to be explored to address the wheat crisis.
- Sustainable approaches for addressing phosphorus deficiency and acidity of soil under conservation tillage practices need to be sought, especially in the KwaZulu-Natal region.

Introduction

South Africa's wheat (*Triticum aestivum* L.) production has declined progressively from 2.5 million tonnes, produced on 974 000 ha in 2002, to approximately 1.7 million tonnes, produced on 500 000 ha in 2013.[1] The country is therefore increasingly reliant on imports of wheat to sustain domestic demand. A decline in land area under wheat suggests producer disinterest in wheat production in South Africa, because of the low profitability of the crop.[2,3] Much of the wheat production area is being lost to other economically important crops such as maize (*Zea mays* L.) and soybean (*Glycine max* L.) as the country has limited land and water resources for expansion of the crop production area. Therefore, in search of solutions for increasing wheat production, the focus has not only been on how to return some land area to wheat, but also on how to immediately and realistically improve yields on current production lands.

Irrigation is an effective tool for increasing yield potential on cropped lands in South Africa; currently, irrigation wheat covers approximately 21% of the total wheat production area, but produces 41% of the crop.[1] The irrigation wheat area in South Africa is divided into four main geographical regions: (1) the cooler central irrigation region in the Free State and Northern Cape Provinces, (2) the warmer northern irrigation region in the North West, Limpopo and Gauteng Provinces, (3) the Highveld region in Mpumalanga and the Free State and (4) the KwaZulu-Natal region. The yield potential of irrigation wheat in South Africa is increasing progressively because of improvements in the genetic yield potential of cultivars, pest and disease resistance as well as technological advancements that enable producers to improve crop management.[4] Hence, in recent years, researchers and industry agronomists conducting cultivar trials in South Africa have documented potential yields of up to 12 t/ha under controlled field experiments.[5] When these yields are compared with the national average yield of approximately 6 t/ha, it appears that there may be opportunity for improving wheat yield in some production areas of South Africa through refinements of crop and resource management strategies.

Yield gaps refer to the difference between attainable yields and actual yields, and are caused by poor crop management practices.[6,7] Therefore, yield gap analysis could be an effective policy framing device for addressing the yield challenge in the ailing South African wheat sector. According to Armour et al.[8], the environmental and management circumstances that enable the production of a 15 t/ha wheat crop are a combination of cultivar and sowing date that lead to grain growing through the solar radiation peak, cool but sunny grain filling conditions and, most importantly, attention to agronomic detail so that no growth constraints occur.

Nutrient demand and removal inevitably increases as producers intensify crop production and target higher yields, which suggests that it is critical for producers to refine soil fertility management practices in improving yield, production efficiency and profitability. Poor nutrient management appears as the most frequently reported yield limiting factor in intensive crop production systems.[9-13] A policy document of the Food and Agriculture Organization of the United Nations on constraints to food production across the world identified high nutrient removal in irrigation crop production as a major cause of deterioration in soil fertility in developing countries.[14] As a result, application rates of inorganic fertilisers have increased, in order to meet the increased nutrient demands. These high rates of inorganic fertiliser may negatively affect soil properties such as soil pH and organic carbon, resulting in reduced

soil fertility and productivity. Meanwhile, there is no record of studies carried out to determine the extent to which poor soil fertility constrains the production capacity of irrigation wheat producers in South Africa. Hence, research and development projects aimed at resuscitating the wheat sector may not be aligned well to farmer priorities.

Soil organic matter is important for improving soil fertility, crop yields and the efficient cycling of nutrients within the system. Soil organic carbon (SOC) content is commonly used as an index of soil organic matter, and sandy soils with less than 1% SOC are prone to structural destabilisation and crop yield reduction.[15] At such SOC levels, it may not be possible to obtain the potential wheat yields, irrespective of soil type.[16] There is agreement in the current literature about the proposition that conservation agriculture (CA) is a sustainable way of managing SOC in such a way that soil structure and fertility is well sustained for the future. The three principles of CA are no-tillage or conservation tillage, crop rotation and a permanent crop residue cover.[17] The need to manage fertiliser efficiently for the success of CA has been recently proposed as a fourth principle.[18] The adoption of CA in Africa is reported as slow, as highlighted in the case of resource poor smallholder producers.[17,19] There is a general preference to use crop residues as fodder and not as soil cover during winter.[19]

Intensive irrigation systems combining winter wheat and a summer crop (usually maize or soybean), whereby producer's harvest eight or more crops in the course of 5 years are common in South Africa. Winter wheat production provides an opportunity for maximising the benefits of CA through provision of continuous soil cover in such systems. The objective of the current study was to explore opportunities to improve wheat yields in the intensive irrigation systems of South Africa through analyses of yield gaps, soil fertility constraints and CA practices.

Materials and method

Description of study sites

The study covered the major irrigation wheat production regions of South Africa which are the cooler central, warmer northern, eastern Highveld and KwaZulu-Natal regions as shown in Figure 1. Wheat producers within each of these geographical regions have broadly similar resource bases, enterprise patterns and constraints. The cooler central region is arid, with average annual temperatures ranging from 15 °C to 31 °C and predominantly deep, loamy oxidic soils[20] which are ideal for irrigation; average rainfall varies between 200 mm and 715 mm annually. In the warmer northern irrigation region, the climate is semi-arid, with average monthly temperatures ranging between 18 °C and 32 °C; the average annual rainfall varies between 200 mm and 600 mm and the region has oxidic soils. The Highveld region has a semi-arid climate and receives an average rainfall of 200 mm to 500 mm annually; mean monthly temperatures range between 14 °C and 26 °C and the area is dominated by plinthic soils. Irrigation wheat in KwaZulu-Natal is mostly produced around Bergville and Winterton (Figure 1), at high altitude areas with highly weathered and well-drained oxidic soils. The climate of KwaZulu-Natal is sub-humid and warm, with average temperatures ranging between 15 °C and 32 °C and average annual rainfall of 600–1000 mm.

Calculation of yield gaps

The Agricultural Research Council – Small Grain Institute (ARC–SGI) of South Africa conducts an annual National Wheat Cultivar Evaluation Programme (NWCEP) to evaluate and characterise all commercial wheat cultivars in the major production areas under farmers' cultivation practices. The NWCEP uses four to eight test locations for each geographical region annually. Test sites are systematically selected in such a way that they are representative of all the production conditions in the geographical region of interest.

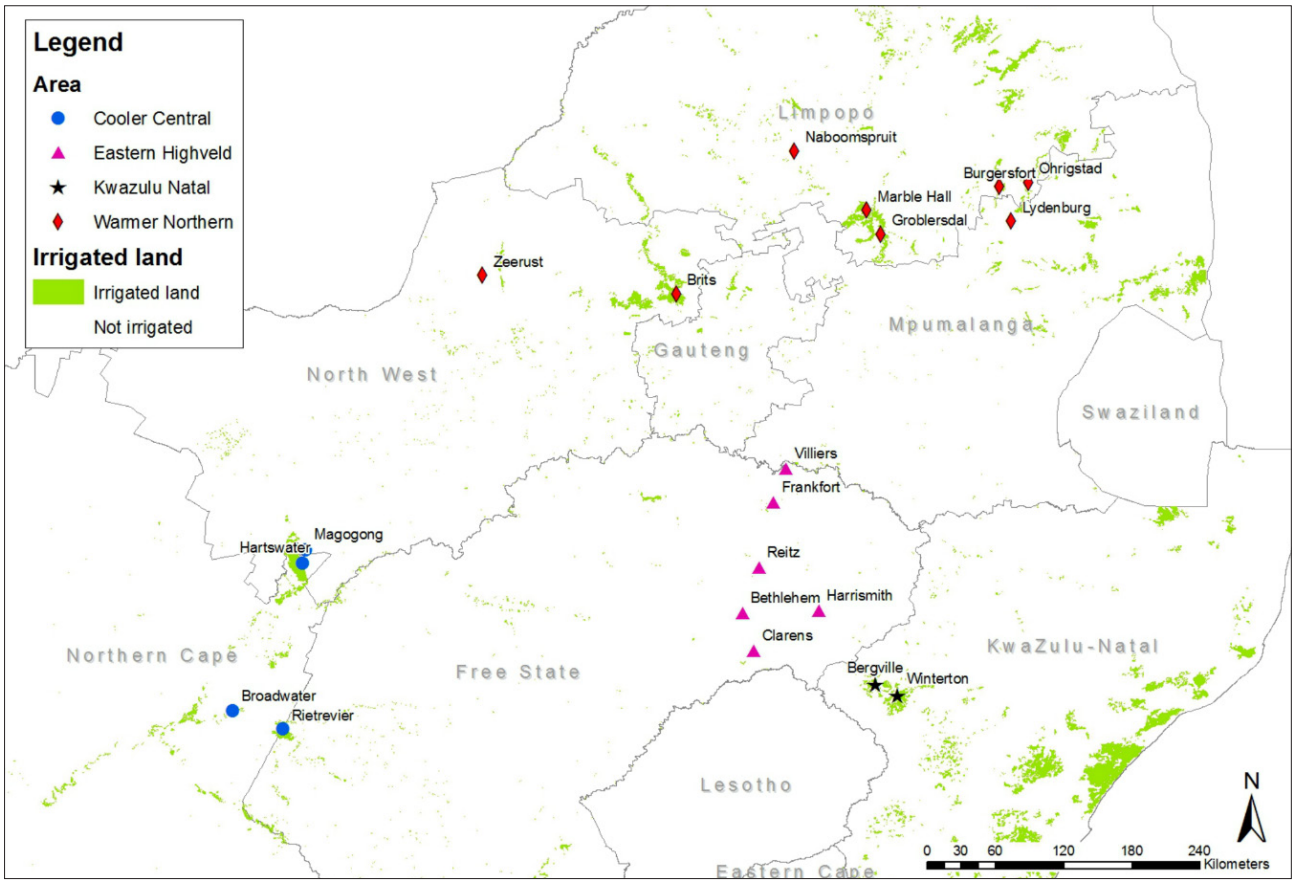


Figure 1: The major irrigation wheat production regions of South Africa.

A randomised complete block design is used for trial layout. All trials are planted inside wheat farmer's fields in line with the farmer's crop management practices with regard to tillage practices, seed rates, weed control, fertiliser application, irrigation scheduling, pest and disease control as well as planting date. Therefore, ARC–SGI has archives of reliable wheat yield data from production systems of South Africa. These data were useful for analysing yield gaps.

The NWCEP includes between 20 and 30 commercial wheat cultivars. Poorly performing cultivars are consistently replaced with newly released cultivars in the programme. Since its inception over three decades ago, nearly all experimental entries – except the yield and quality check cultivar called Buffels – have changed through the evaluation period. The grain yield data of Buffels provided a standard measure of farmer yields. Grain yield data of Buffels under the NWCEP are available from wheat production guidelines that were published annually by the ARC-SGI for the entire period (2009–2014) and online from the ARC. The yield gap (*Yg*) is the difference between yield potential (*Yp*) and actual yield (*Ya*), i.e. $Yg = Yp - Ya$. Data from on-farm trials such as those of the NWCEP can provide a robust estimate of *Yp* for a given location under a specific set of management practices provided that the trials are replicated over many years.[21]

Recent reviews of methods for assessing yield gaps with a global relevance[6,22] provided some guidelines for properly estimating *Yp* based on maximum yields achieved among a sizable sample of farmers in a region of interest. Based on these reviews, 5 years' data from the most recent period is considered adequate for estimates of *Ya* in favourable, high-yielding environments such as irrigated systems. The upper (95th) percentile of farmer yield data is also recommended as an ideal approach for calculating *Yp*, based on the assumption that in any given production system of many farmers, it is likely for a few progressive farmers to come quite close to the *Yp* through best cultivation practices.[6,22] Before the analysis of Buffels yield data to determine mean yields and variance components, the data were validated to check and remove outliers. Summary statistics for means, standard deviations and percentiles of the combined data were determined using GenStat® 17 statistical software.

Selection of irrigation wheat farms for soil fertility evaluation

The fields of producers who planted irrigation wheat during the 2015 season were used for the soil fertility evaluation. Representative producers for each of the geographical regions were identified in collaboration with the NWCEP. The geographical regions were further sub-divided into localities of interest where most irrigation wheat producers were concentrated. Producers were contacted and only those who gave permission for sampling on their wheat fields were considered in this study. A limitation of the purposive sampling procedure used in this study is that it excluded the fields of those irrigation wheat producers who were not willing to have their fields surveyed. It also excluded the fields of wheat producers who did not plant irrigation wheat during the 2015/2016 season. However, the results from this study may also be indicative of the conditions in the fields of these producers, as long as the soils are from the same parent materials and are managed the same way.

Different tillage systems were identified through observation of the fields. Within the context of the current study, conservation tillage fields were identified as those fields in which wheat was either planted directly into the previous crop's residues with no soil disturbance, or where there were signs of slight soil disturbance and about 30% of crop residues on the soil surface. Conventional tillage fields were those with signs of complete turning of soil and less than 30% or no residues on the soil surface. The residues of the crops which preceded wheat were used to identify the crop rotation system as either legume or non-legume. CA fields were those in which conservation or no-till was combined with a legume–wheat crop rotation system, assuming the wheat also served as a winter cover crop for permanent cover.

Soil sampling and analysis

Soil samples were collected from the 0–20 cm and 20–40 cm depths using a graduated auger, after clearing the litter layer. A simple random sampling procedure was used. Soil sampling was carried out from May to September 2015, after the emergence of wheat seedlings to ensure clear identification of wheat fields. At least 10 random samples were collected from each of the fields and bulked to form a composite sample. The samples were air dried (visible organic debris removed), ground (< 2 mm) and analysed at the ARC–SGI soil laboratory. The samples were analysed for electrical conductivity (EC; 1:1 soil to water suspension), pH (1:5 soil to 1 M KCl suspension), exchangeable acidity (1 M KCl), extractable P (Bray 1), exchangeable cations and extractable S (1 N NH_4OAc at pH 7) and extractable Zn (0.1 M HCl) using procedures of the Non-affiliated Soil Analysis Working Committee.[23] In addition to these analyses, organic C (Walkley–Black method)[24] and particle size distribution (hydrometer and sieve method)[25] were also determined. Using equivalent values ($cmol_c/kg$), cation exchange capacity (CEC; sum of exchangeable acidic [H and Al] and basic [Ca, Mg, K and Na] cations), acid saturation (ratio of exchangeable acidic cations to CEC), exchangeable sodium percentage (ESP; exchangeable Na to CEC) and Ca:Mg ratio were calculated. Nitrogen adequacy was determined through visual assessments of irrigation wheat crops at the flag leaf stage using a guide for field identification by Snowball and Robson[26].

The number of sites that were sampled varied across geographical regions, and the resulting soil fertility data were unbalanced, with both fixed (geographical regions, crop rotations, tillage systems, soil depth) and random (locations) effects. Therefore, a mixed model, the residual (or restricted) maximum likelihood (REML) algorithm was used to reliably estimate variance components.[27,28] The REML was performed using GenStat® 17 statistical software. Third-order interactions were not included. Conclusions regarding nutrient status were made through comparisons between soil test results and nutrient management guidelines for cereal crops.[29,30] The extractants used in the current study correspond to those used in the nutrient management guidelines.

Results

Actual yields, yield potentials and yield gaps

Actual yields for irrigated wheat ranged from 5.99±0.15 t/ha in the KwaZulu-Natal region to 8.32±0.10 t/ha in the cooler central region (Table 1). In agreement with *Ya*, *Yp* ranged from 7.57 t/ha in the KwaZulu-Natal region to 11.45 t/ha in the cooler central region. The resulting *Yg* range is therefore 1.58–3.13 t/ha, implying irrigation wheat yields could be increased by 26% to 38%.

Tillage and crop rotation practices

The majority (63.85%) of irrigation wheat producers who participated in the study practised conventional tillage, with 36.15% using conservation tillage (Table 2). Most (88.37%) of the producers in the KwaZulu-Natal region practised CA; that is conservation till combined with a legume–wheat crop rotation system, assuming the wheat also serves as a winter cover crop for permanent cover. In the warmer northern region, only 13.89% of the sampled farms practised conservation tillage with legume–wheat crop rotation. In the eastern Highveld and cooler central regions, all the farms (100%) practised conventional tillage. There were more farms practising a non-legume–wheat rotation than farms practising a legume–wheat rotation in the cooler central and warmer northern region. The overall adoption rate of CA was 33.08%.

Soil fertility variation across all irrigation wheat fields

Summary statistics for soil fertility parameters are presented in Table 3. There was considerable variation within each of these parameters as shown by the high coefficients of variation and the corresponding large difference between minimum and maximum values. However, over 95% of sampled farms had acceptable values for Ca (>150 mg/kg), Mg (>60 mg/kg), Zn (>1.5 mg/kg), S (>7.5 mg/kg), ESP (<10), Ca:Mg ratio (>1<15) and EC(<1 dS/m). Field observations of wheat crops showed that there was generally adequate N on wheat fields across the geographical regions. These parameters were therefore excluded from further analysis and the study only focused on those parameters that appeared to be limiting on a considerable fraction of farms, i.e. >10%.

Table 1: Actual yields, yield potentials and yield gaps (t/ha) for irrigated wheat in South Africa

Geographical region	*n*	Minimum	Maximum	Yield potential (Y*p*)	Actual yield (Y*a*)	Standard error	Yield gap (Y*g*)	Y*g*:Y*a* ratio	Coefficient of variation (%)
Cooler central	426	3.02	13.67	11.45	8.32	0.10	3.13	0.38	25.3
Warmer northern	176	3.46	11.10	8.84	6.59	0.11	2.25	0.34	21.6
Eastern Highveld	128	2.59	9.81	9.25	6.64	0.15	2.61	0.39	26.0
KwaZulu-Natal	38	3.04	7.65	7.57	5.99	0.22	1.58	0.26	22.3
Average	768	3.03	10.56	9.28	6.89	0.15	2.39	0.34	23.8

Table 2: Tillage and crop rotation practices of South African wheat producers who participated in the study

Geographical region	Tillage system	Crop rotation system	Rotation crops	Number of farms (*n*)
KwaZulu-Natal	Conventional tillage	Non-legume–wheat	–	0
		Legume–wheat	Soybean	1
	Conservation tillage	Non-legume–wheat	Maize	4
		Legume–wheat	Soybean	38
Cooler central	Conventional tillage	Non-legume–wheat	Maize, oats, cotton	21
		Legume–wheat	Groundnut, soybean	4
	Conservation tillage	Non-legume–wheat	–	0
		Legume–wheat	–	0
Warmer northern	Conventional tillage	Non-legume–wheat	Tobacco, maize	16
		Legume–wheat	Sugar bean, soybean	15
	Conservation tillage	Non-legume–wheat	–	0
		Legume–wheat	Soybeans	5
Eastern Highveld	Conventional tillage	Non-legume–wheat	Maize, potatoes	12
		Legume–wheat	Soybean, white bean	14
	Conservation tillage	Non-legume–wheat	–	0
		Legume–wheat	–	0

Table 3: Summary statistics for soil fertility parameters on irrigation wheat fields in South Africa

Parameter	Mean	Minimum	Median	Maximum	Coefficient of variation (%)	Crop requirement†
pH (KCl)	5.33	3.81	5.08	7.51	19.01	5.5–6.5
Acid saturation (%)	3.23	0.00	0.00	54.07	207.90	<8%
Calcium (mg/kg)	1056	64.80	235.90	11 770	125.80	>150
Magnesium (mg/kg)	301.80	14.04	172.60	2332	108.10	60–300
Potassium (mg/kg)	197.20	36.80	164.20	602.90	58.69	125–800
Phosphorus (mg/kg)	39.96	3.47	34.87	128.70	69.81	40–100
Sulphur (mg/kg)	24.21	0.76	18.01	122.10	88.59	>7.5
Zinc (mg/kg)	4.77	0.83	2.84	127.60	179.50	>1.5
Electrical conductivity (dS/m)	0.28	0.08	0.23	0.91	56.97	<1.0
Exchangeable sodium percentage	1.78	0.19	1.06	9.14	95.03	<10%
Ca:Mg ratio	2.47	1.05	2.44	5.56	34.78	2–15
Cation exchange capacity (cmolc/kg)	8.56	1.20	5.84	73.62	107.80	2–58
Soil organic carbon (%)	1.42	0.13	1.27	6.02	60.93	>1

†Based on Brady and Weil[29] and Horneck et al.[30]

These parameters were SOC, P and pH. SOC content was below 1% at 0–20 cm soil depth on 43.85% of the farms. Soil pH on more than 40% of the farms was below the recommended range of 5.5–6.5 for optimal wheat growth at 0–20 cm. For P, more than 30% of the farms had less than the minimum requirement of 40 mg/kg.

The SOC, pH and extractable P varied significantly ($p<0.001$) with different geographical regions and tillage systems, as shown in Table 4. KwaZulu-Natal (2.00±0.09%) had the highest level of SOC, followed by the warmer northern (1.65±0.14%), cooler central (0.84±0.08%) and eastern Highveld (0.82±0.07%) regions (Table 5). The eastern Highveld (56.08±4.53 mg/kg) and cooler central (49.30±2.77 mg/kg) regions had adequate P, but the warmer northern (36.65±3.65 mg/kg) and KwaZulu-Natal (27.49±2.04 mg/kg) regions showed potential deficiencies (Table 5). Mean soil pH of all the geographical regions was in the acidic range; pH was outside the acceptable range of 5.5–6.5 in the KwaZulu-Natal (pH 4.51±0.05) and eastern Highveld (pH 4.97±0.08) regions.

Conservation tillage fields (2.15±0.10%) had more SOC than conventional tillage fields (1.02±0.05%) but lower pH (4.51±0.06) than conventional tillage fields (5.82±0.08). The P content of conventional tillage fields was adequate (48.48±2.27 mg/kg) when compared to that of conservation tillage fields (25.58±1.92 mg/kg). Soil pH and P levels varied significantly ($p<0.001$) across crop rotation systems (Table 4). Rotation systems in which wheat was preceded by non-legumes had acceptable P levels (52.20±3.01 mg/kg) and higher soil pH (5.9) than those in which wheat was preceded by legumes, which had low P levels (31.70±1.81 mg/kg) and lower soil pH (4.95).

There was significantly more extractable P at a soil depth of 0–20 cm (45.57±2.54 mg/kg) than at 20–40 cm (34.36±2.28 mg/kg) (Table 4). Overall, there was also more SOC in the 0–20 cm soil layer (1.55±0.09%) than in the 20–40 cm soil layer (1.33±0.08%).

The geographical region and crop rotation interaction effect on soil pH and SOC was significant ($p<0.001$). The nature of the interactions is shown in Figure 2. The eastern Highveld and warmer northern regions had slightly lower soil pH for rotations in which wheat was preceded by a legume than when wheat was preceded by a non-legume. The rotations had similar pH results in the KwaZulu-Natal and cooler central regions. In KwaZulu-Natal, there was more SOC on non-legume–wheat crop rotations, whereas in the warmer northern region, the opposite was true.

Table 4: Significance of the fixed effects tested by chi-squared F-statistic (Wald statistic/d.f.) values in the overall REML analysis for the soil fertility parameters pH, Mg, K, P, S, Zn, Ca:Mg ratio, CEC, EC, ESP and SOC of irrigation wheat fields in South Africa

		***p*-value for various nutrient availability parameters**												
Source of variation	**d.f.**	**Zn**	**S**	**P**	**AS**	**Mg**	**K**	**EC**	**Ca**	**Ca:Mg**	**CEC**	**ESP**	**SOC**	**pH**
Geographical region	3	<0.001	0.135	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Tillage†	1	0.320	0.016	<0.001	0.503	0.672	0.644	0.141	0.270	0.168	0.329	<0.001	<0.001	<0.001
Crop rotation	1	0.668	0.209	<0.001	0.076	0.846	0.109	0.534	0.426	0.375	0.591	0.405	0.763	<0.001
Soil depth	1	0.063	0.487	<0.001	0.446	0.753	0.005	0.234	0.883	0.141	0.928	0.325	0.011	0.910
Geographical region × crop rotation	3	0.150	<0.001	0.283	0.229	0.839	0.491	0.958	0.291	0.383	0.369	<0.001	<0.001	0.009
Geographical region × soil depth	3	0.262	0.830	0.493	0.772	0.994	0.952	<0.001	0.998	0.352	0.998	0.899	0.008	0.699
Tillage × soil depth	1	0.549	0.731	0.696	0.923	0.902	0.913	0.018	0.972	0.941	0.938	0.990	0.073	0.890
Crop rotation × soil depth	1	0.214	0.554	0.939	0.561	0.940	0.868	0.217	0.900	0.871	0.888	0.803	0.866	0.849

AS, acid saturation; EC, electrical conductivity; CEC, cation exchange capacity; ESP, exchangeable sodium percentage; SOC, soil organic carbon

†Tillage interactions with crop rotation and geographical region were not considered in the analysis because not all tillage systems were represented in either crop rotations or geographical regions.

Table 5: Effects of geographical region on soil pH, phosphorus and organic carbon

Geographical region	**pH**	**Phosphorus (mg/kg)**	**Soil organic carbon (%)**
KwaZulu-Natal	4.51[d]	27.49[d]	2.00[a]
Eastern Highveld	4.97[c]	56.08[a]	0.82[c]
Warmer northern	6.32[a]	36.65[c]	1.65[b]
Cooler central	5.75[b]	49.30[b]	0.84[c]
p-value	<0.001	<0.001	<0.001
Standard error of difference	0.09	3.25	0.10

Values with different letters (a-d) in a column indicate significant differences at p*<0.05.*

Similar amounts of SOC were observed for legumes and non-legumes in wheat rotations in the cooler central and the eastern Highveld regions. Fields in KwaZulu-Natal had more SOC in the topsoil (0–20 cm) than in the subsoil (20–40 cm) (Figure 3). In the warmer northern, eastern Highveld and cooler central regions, similar SOC levels were observed in soil from both depths.

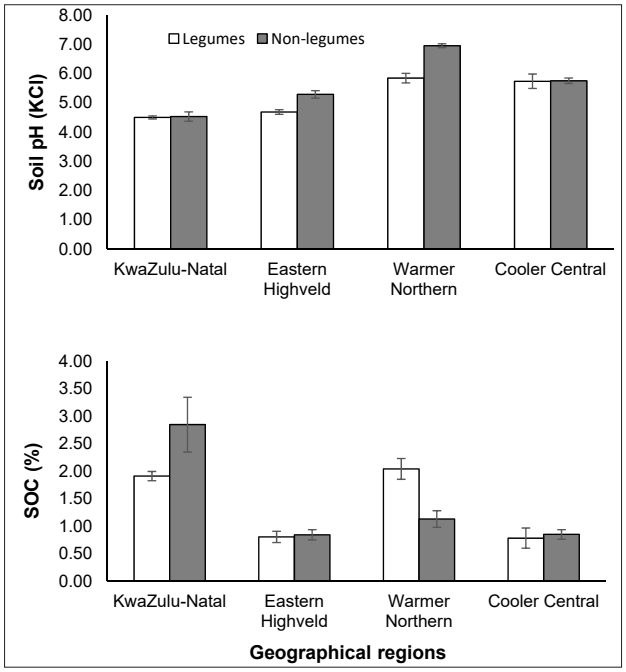


Figure 2: Soil pH and organic carbon (SOC) variation (mean±s.e.) with crop rotation across different geographical regions of South Africa.

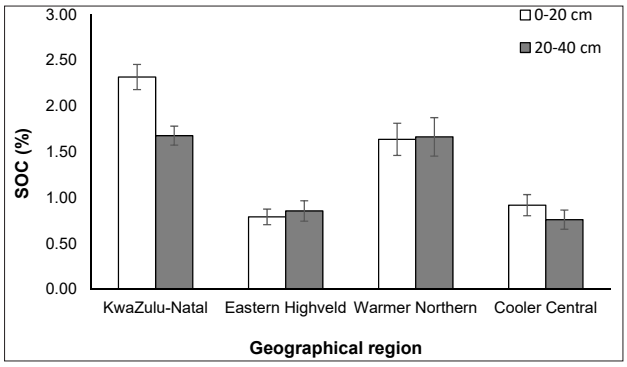


Figure 3: Soil organic carbon (SOC) variation (mean±s.e.) with soil depth in different geographical regions of South Africa.

Analysis of particle size distribution showed that there were differences in mean textural classes of soils in the geographical regions (Table 6). Soils in the cooler central region were predominantly sandy and those of the eastern Highveld region were classed as loamy sands. The KwaZulu-Natal and warmer northern regions had higher clay contents and were classified as sandy clay loam soil. There was, however, considerable variation in clay and silt content of the soils within geographical regions as shown by the high coefficients of variation. Linear correlation of SOC against soil clay content showed that there was no relationship between SOC and clay content ($r=0$) in the KwaZulu-Natal region, but all the other regions showed significant positive Pearson's correlations (Figure 4).

Discussion

This study contributed to our knowledge pool through quantifying yield gaps and investigating CA practices and soil fertility constraints of irrigated wheat fields in different production areas of South Africa. It has been shown, using actual farm data, that wheat production in South Africa can be increased by exploiting the available potential for increasing yields in various production areas. These yield gaps range from 1.58 t/ha to 3.13 t/ha, representing up to 38% of the yield potential. The findings are in agreement with Licker et al.[31] who stated that large yield gaps in grain production are concentrated in developing countries, and that poor crop management is the major cause of yield loss for grain crops. A yield potential of 13.67 t/ha which was calculated for the cooler central region is comparable to the world record for farm wheat yield of 16.52 t/ha, which was obtained in the United Kingdom.[32] The study also identified regions such as KwaZulu-Natal where the yield gap (1.58 t/ha) and yield potential (7.57 t/ha) are rather low, and efforts would probably need to be channelled towards strategies for increasing the *Yp*. Although spring wheat can tolerate high temperatures between 22 °C and 34 °C[33], cool and moist climate is the most ideal for growth of the currently recommended cultivars. It is most likely that temperature is one of the major limiting factors of wheat productivity in KwaZulu-Natal, where average monthly temperatures are in the range 15–32 °C. There is evidence suggesting that an increase in temperature of 1 °C above the optimal can reduce wheat yield by up to 50%.[34]

We also identified opportunities to improve soil fertility management on irrigated wheat fields. Most irrigation wheat producers who participated in the study practised conventional tillage and 43.85% of the sampled farms had less than 1% SOC. Kay and Angers[16] found that when the SOC is less than 1%, yield potential of a crop is limited on low clay soils. This finding could mean that nearly half of the irrigation wheat producers fail to achieve the yield potential of irrigation wheat on their farms because of low SOC, among other reasons. The high adoption rate of CA amongst irrigation wheat producers in KwaZulu-Natal (Table 2) is remarkable, considering that there was very low adoption of the technology in other regions. The No Till Club of KwaZulu-Natal, formed more than 15 years ago, may have played a huge role in the promotion of CA adoption in this region. Currently, about 130 commercial producers from KwaZulu-Natal are members of the No-Till Club, and the club provides a no-till training course to these producers and any other interested parties.

Table 6: Particle size distribution for wheat production soils in different geographical regions of South Africa

Geographical region	Number of farms	Clay		Sand		Silt		Textural class†
		Mean	CV%	Mean	CV%	Mean	CV%	
Cooler central	23	7.04	83.8	91.65	7.2	1.304	118.9	Sand
Eastern Highveld	26	11.62	62.3	86.96	10.2	1.423	175.8	Loamy sand
KwaZulu-Natal	42	23.98	26.5	71.07	11.4	4.952	75.4	Sandy clay loam
Warmer northern	35	21.91	59.6	75.03	20.3	3.057	102.5	Sandy clay loam

CV, coefficient of variation

†Based on the USDA textural triangle.

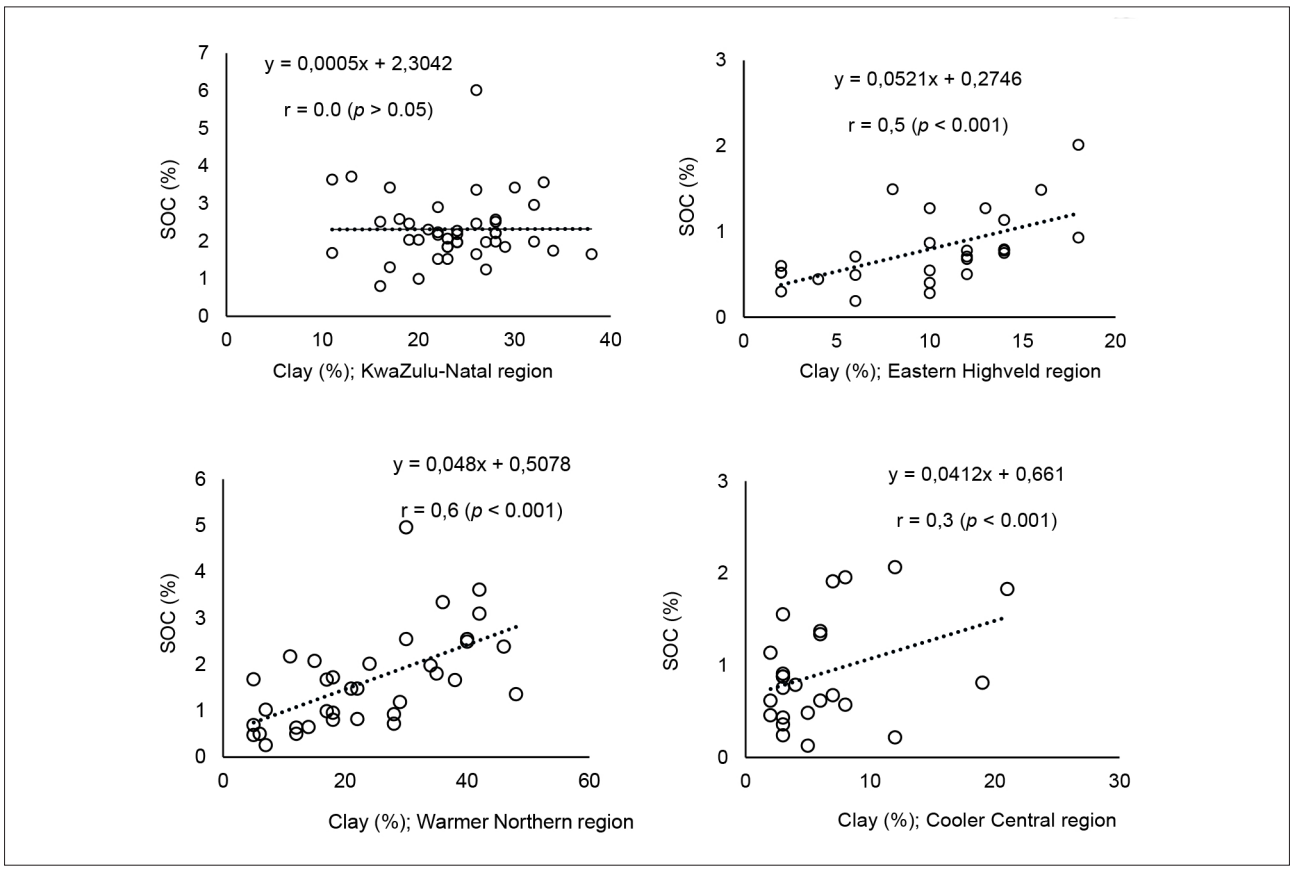


Figure 4: The relationship between soil organic carbon (SOC) and clay content on irrigation wheat fields in South Africa.

More effort is required by non-profit farming organisations to actively promote CA in the other regions. Dumanski et al.[35] pointed out that successful CA is achieved through community-driven development processes whereby local researchers, communities and producer associations identify and promote the best options for CA in their locations.

Another reason for wide-scale adoption of CA in KwaZulu-Natal could be the fact that the heavy soils of KwaZulu-Natal generally require heavier machinery and more fuel for tillage, and CA was an obvious attraction for reducing fuel and traction costs to the KwaZulu-Natal farmers. The heavy soils also compact easily when worked under wet conditions. The KwaZulu-Natal region is also warmer and wetter, such that producers have many options for increasing biomass to obtain the benefits of CA. Maize yields in excess of 10 t/ha are common in irrigation production systems of South Africa. The slow decomposition of the preceding summer crop's residues presents serious planting and crop emergence challenges for subsequent winter wheat, especially if the producers do not have the right planting equipment. Many producers resort to the plough to solve the problems, and this probably explains why most irrigation wheat farmers still practise conventional tillage.

However, soil acidity and P deficiencies were most severe in the KwaZulu-Natal region in comparison to other regions. It appears as if the wide-scale adoption of CA has not addressed soil acidity and P deficiency problems that may be inherent to this region. The KwaZulu-Natal region has well-weathered soils that are derived from dolerite.[20] Under a humid climate, soils tend to have excess sesquioxides[29] and the low P could possibly be attributed to the high fixation of P by sesquioxides in this region. Low soil pH and relatively high acid saturation are common in highly weathered soils. The accumulation of SOC in this region could have resulted in greater acidity, through degradation of SOM and mineralisation of N under the sub-humid conditions. It could also be deduced that producers in KwaZulu-Natal have always struggled with acidity and P deficiency. Therefore, they were better motivated to adopt CA, which is generally purported to increase P availability, as well as reduce acidity problems over the long term.[36] There is a need for dedicated research to refine the CA practice in KwaZulu-Natal to enhance P availability, reduce P stratification and reduce acidification of soils.

The pH of soil from more than 40% of the farms was below the recommended range of 5.5–6.5 for optimal wheat growth. Therefore, soil acidity could be a major concern in the fertility of irrigation wheat fields in South Africa. Soil pH was influenced by crop rotation, whereby there was lower soil pH on the legume–wheat crop rotations in comparison to non-legume–wheat crop rotations. The decrease in soil pH following legumes crop rotation (Figure 2) may be attributable to more rapid degradation of legume residues as a result of a favourable C:N ratio and the associated nitrification which has an acidifying effect.[37] The decrease in soil pH and increase in exchangeable acidity on the conservation tillage systems observed in this study may be attributed to SOC accumulation. The accumulation of SOC leads to a dissociation of humic material which contains carboxylic and phenolic groups. When these groups dissociate, H^+ is released which further reduces soil acidity. Accumulation of SOC could also result in more N and S mineralisation, thus increasing H^+ concentrations and lowering soil pH.[38,39] There was more plant available P on rotations in which wheat was preceded by non-legumes than in legume–wheat rotations. This finding may be because legumes degrade rapidly as a result of a low C:N ratio, and, after decomposition of legumes, mineralisation of N can occur (which has acidifying effects), thus resulting in P fixation and hence low P availability.

Zn varied significantly across the geographical regions, but means were generally within the acceptable range (>1.5 mg/kg) in all the regions (Table 3). These results appear contradictory to findings by Herselman[40] who reported that 91% of South African soils are Zn deficient as they contained <1.5 mg/kg. It should be noted that most N, P and K basal fertilisers that are used in South Africa by commercial

producers are fortified with at least 0.5% Zn, and continuous use of these fertilisers probably explains the general adequacy of Zn on these fields. Field observations showed that there was generally adequate N on wheat fields across the geographical regions, suggesting that the producers are managing N well. Much of the fertilisers promoted by fertiliser companies in South Africa are N-based, hence producers are more inclined to purchase these above others. Secondly, N deficiency symptoms on wheat are relatively easy to diagnose as a characteristic yellowing of the lower leaves. The high mobility of N also means that the deficiency can be corrected at any stage using split applications during crop growth.

Soils in the different irrigation wheat production regions of South Africa developed from various parent materials and are subjected to different climatic conditions, thus they are of different properties and texture. The KwaZulu-Natal and warmer northern regions are relatively warmer and wetter than the eastern Highveld and cooler central regions. Hence, the soils in KwaZulu-Natal and warmer northern regions are predominantly well weathered and fine textured whereas those of the eastern Highveld and cooler central regions are coarse textured.[20] Weathering of soils results in high Al and Fe oxides which enables SOC to exist as organo-oxide complexes,[29] which could protect SOC. In general, clay content is thought to be the most important rate modifier of SOC accumulation.[41] However, the higher SOC contents in the top soil (0–20 cm) of irrigation wheat fields in KwaZulu-Natal could be related to the level of adoption of CA in the region, as no significant relationship could be established between SOC and clay contents for this region (Figure 4).

The producers retain crop residues on their fields and the biomass is protected from rapid decomposition through reduced or no-tillage. The variations in SOC with geographical region and soil depth as observed in the study could also be the result of differences in predominant tillage systems in these regions. Under conservation tillage, SOC is concentrated on the topsoil because of a limited aggregate turnover[41,42], while in the conventional tillage system SOC is evenly distributed in the soil profile as a result of regular mixing of aggregates during tillage[43]. SOC is more stable in the no-tillage and conservation tillage fields compared to conventionally tilled fields because of the limited turnover of aggregates.[44-47] More action is needed to increase the awareness of wheat producers on the consequences of conventional tillage in terms of SOC losses, and hence soil fertility depletion. Wheat producers that practise conventional tillage could benefit from reducing soil disturbance if they convert to CA, as many already practise rotations, permanent soil cover and proper fertiliser management, which are requirements for CA.[18]

Soil organic matter was observed to vary with different crop rotation systems on different geographical regions. There was more SOC in the KwaZulu-Natal region when a non-legume crop rotation was practised and this could probably be attributed to slow decomposition of the non-legume residue under conservation till.[48] In the warmer northern region where most producers practise conventional tillage, more SOC was measured in a legume–wheat crop rotation. These results suggest that inclusion of legumes in rotation could be offsetting some of the negative effects of conventional tillage on SOC. These unexpected findings are partly in agreement with the observations of Corbeels et al.[17] and Naresh et al.[49] who investigated the extent of global CA adoption in resource-poor environments. They reported that legume-based crop rotations enrich soil fertility. Wheat producers in arid and semi-arid areas could benefit from adopting CA practices with legumes in rotation systems, while producers from wetter regions could benefit from including non-legumes in their rotation systems.

Conclusions

Irrigated wheat production in South Africa could be increased by closing large yield gaps in production regions; these yield gaps ranged from 1.58 t/ha to 3.13 t/ha, representing 26–38% of the yield potential. Poor soil fertility may be a major yield constraint in intensive irrigated wheat production systems. It is recommended that future studies must focus on sustainable approaches for effectively enhancing P availability and addressing pH problems under conservation till and legume–wheat rotations, especially in KwaZulu-Natal. More action is required in order to increase wheat producer awareness on the soil fertility benefits of CA in the eastern Highveld and cooler central areas. EC, ESP, N, K, Mg, S, Zn and Ca:Mg ratio were, however, acceptable on more than 90% of wheat fields. It is hoped that the knowledge generated in this study would be useful to policymakers and researchers in better orienting investments in research and development projects aimed at addressing the South African wheat production crisis.

Acknowledgements

We thank ARC–SGI technical staff for their assistance with the management of field trials and soil analysis. The Winter Cereal Trust and National Research Foundation of South Africa (Project TTK150717127405) are acknowledged for funding the 'Yield gaps analysis for irrigated wheat in South Africa' project, from which this study emerged.

Authors' contributions

N.Z.S. was the lead author and the MSc student responsible for soil data collection on the project. E.D. was the project leader who initiated the project under the 'Yield Gaps for Irrigated Wheat Production Systems of South Africa'. L.V. and A.B. were collaborators on the project who provided significant intellectual input as a soil scientist and crop physiologist, respectively. P.M. and T.J.T. gave significant scientific input on context and relevance of the project and also revised and refined the manuscript to its current format. P.M. was also the main MSc supervisor of N.Z.S.

References

1. South African Department of Agriculture, Forestry and Fisheries (DAFF). Agricultural statistics. Pretoria: DAFF; 2014.
2. Lemmer W, De Villiers J. Challenges in the wheat industry and the relevance of the import tariff. SA Grain. 2012;14:46–49.
3. Payne T. Maize production blossoms in SA. Mail and Guardian. 2012 May 18. Available from: http://mg.co.za/article/2012-05-18-maize-production-blossoms-in-sa/
4. Smit HA, Tolmay VL, Barnard A, Jordaan JP, Koekemoer FP, Otto VM, et al. An overview of the context and scope of wheat (*Triticum aestivum*) research in South Africa from 1983 to 2008. S Afr J Plant Soil. 2010;27:81–96. http://dx.doi.org/10.1080/02571862.2010.10639973
5. Agricultural Research Council Small Grain Institute (ARC–SGI). Wheat production guidelines: Production of the small grains in the summer rainfall area. Pretoria: South Africa; 2015.
6. Van Ittersum MK, Cassman KG, Grassini P, Wolf J, Tittonell P, Hochman Z. Yield gap analysis with local to global relevance – A review. Field Crops Res. 2013;143:4–17. http://dx.doi.org/10.1016/j.fcr.2012.09.009
7. Bryan BA, King D, Zhao G. Influence of management and environment on Australian wheat: Information for sustainable intensification and closing yield gaps. Environ Res Lett. 2014;9:44–45. http://dx.doi.org/10.1088/1748-9326/9/4/044005
8. Armour T, Jamieson PD, Nicholls A, Zyskowski R. Breaking the 15 t/ha wheat yield barrier. In: New directions for a diverse planet: Proceedings of the 4th International Crop Science Congress; 2004 September 26 – October 01; Brisbane, Australia. Erina, NSW: The Regional Institute Online Publishing; 2004.p. 1–4.
9. Neumann K, Verburg PH, Stehfest E, Müller C. The yield gap of global grain production: A spatial analysis. Agr Syst. 2010;103:316–326. http://dx.doi.org/10.1016/j.agsy.2010.02.004
10. Nadim MA, Awan IU, Baloch MS, Khan EA, Naveed K, Khan MA. Micronutrient use deficiency in wheat as affected by different application methods. Pakistan J Bot. 2013;45:887–892.
11. Rani YS, Jayasree G, Sai MVRS. Yield gap analysis using oryza 2000 model in two rice growing districts of Andhra Pradesh. Am Eur J Agr Env Sci. 2013;13:930–934.
12. Tittonell P, Giller KE. When yield gaps are poverty traps: The paradigm of ecological intensification in African smallholder agriculture. Field Crops Res. 2013;143:76–90. http://dx.doi.org/10.1016/j.fcr.2012.10.007

13. Affholder F, Poeydebat C, Corbeels M, Scopel E, Tittonell P. The yield gap of major food crops in family agriculture in the tropics: Assessment and analysis through field surveys and modelling. Field Crops Res. 2013;143:106–118. http://dx.doi.org/10.1016/j.fcr.2012.10.021

14. Food and Agricultural Organization (FAO). Nutrient management guidelines for some major field crops. Plant Nutrition and Environ Issues. 2006;11:301–302.

15. Howard PJA, Howard DM. Use of organic carbon and loss-on-ignition to estimate soil organic matter in different soil types and horizons. Biol Fert Soils. 1990;9:306–310. http://dx.doi.org/10.1007/BF00634106

16. Kay BD, Angers DA. Soil structure. In: Sumner ME, editor. Handbook of soil science. Boca Raton, FL: CRC Press; 1999. p. 229–276.

17. Corbeels M, De Graaff J, Ndah TH, Penot E, Baudron F, Naudin K, et al. Understanding the impact and adoption of conservation agriculture in Africa: A multi-scale analysis. Agr Ecosys Env. 2014;187:155–170. http://dx.doi.org/10.1016/j.agee.2013.10.011

18. Vanlauwe B, Wendt J, Giller KE, Corbeels M, Gerard B, Nolte C. A fourth principle is required to define conservation agriculture in sub-saharan Africa: The appropriate use of fertiliser to enhance crop productivity. Field Crops Res. 2014;155:10–13. http://dx.doi.org/10.1016/j.fcr.2013.10.002

19. Giller KE, Corbeels M, Nyamangara J, Triomphe B, Affholder F, Scopel E, et al. A research agenda to explore the role of conservation agriculture in African smallholder production systems. Field Crops Res. 2011;124:468–472. http://dx.doi.org/10.1016/j.fcr.2011.04.010

20. Fey M. Soils of South Africa. Cambridge, MA: Cambridge University Press; 2010. http://dx.doi.org/10.1017/CBO9780511782183

21. Cassman KG, Dobermann A, Walters DT, Yang H. Meeting cereal demand while protecting natural resources and improving environmental quality. Annu Rev Env Resour. 2003;28:315–358. http://dx.doi.org/10.1146/annurev.energy.28.040202.122858

22. Lobell DB, Cassman KG, Field CB. Crop yield gaps: Their importance, magnitudes, and causes. Annu Rev Env Resour. 2009;34:179. http://dx.doi.org/10.1146/annurev.environ.041008.093740

23. Non-Affiliated Soil Analysis Working Committee. Handbook of standard soil testing methods for advisory purposes. Pretoria: Soil Science Society of South Africa; 1990.

24. Nelson DW, Sommers LE. Total carbon, organic carbon, and organic matter. In: Sparks DL, Page AL, Helmke PA, Loeppert RH, Soltanpour PN, Tabatabai MA, et al., editors. Methods of soil analysis. Part 3: Chemical methods. SSSA Book Series no. 5. Madison, WI: SSSA and ASA; 1996. p. 961–1010. http://dx.doi.org/10.2136/sssabookser5.3.c34

25. Bouyoucos GJ. Hydrometer method improved for making particle size analyses of soils. Agron J. 1962;464–465. http://dx.doi.org/10.2134/agronj1962.00021962005400050028x

26. Snowball K, Robson AD. Nutrient deficiencies and toxicities in wheat: A guide for field identification. Mexico: CIMMYT; 1991.

27. Virk DS, Pandit DB, Sufian MA, Ahmed F, Siddique MAB, Samad MA, et al. REML is an effective analysis for mixed modelling of unbalanced on-farm varietal trials. Exp Agr. 2009;45:77–91. http://dx.doi.org/10.1017/S0014479708007047

28. Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM. GenStat for Windows. 15th ed. Hemel Hempstead: VSN International; 2013.

29. Brady NC, Weil RR. The nature and properties of soils. 13th ed. Upper Saddle River, NJ: Prentice Hall; 2008.

30. Horneck DA, Sullivan DM, Owen JS, Hart JM. Soil test interpretation guide. Corvallis, OR: Oregon State University Extension Service; 2011.

31. Licker R, Johnston M, Foley JA, Barford C, Kucharik CJ, Monfreda C, et al. Mind the gap: How do climate and agricultural management explain the 'yield gap' of croplands around the world? Global Ecol Biogeogr. 2010;19:769–782. http://dx.doi.org/10.1111/j.1466-8238.2010.00563.x

32. Guinness World Records. Wheat yield record [homepage on the Internet]. c2016. [cited 2016 Dec 06]. Available from: http://www.guinnessworldrecords.com/world-records/1/highest-wheat-yield

33. South African Department of Agriculture, Forestry and Fisheries (DAFF). Wheat production guidelines. Pretoria: DAFF; 2010.

34. You L, Rosegrant MW, Wood S, Sun D. Impact of growing season temperature on wheat productivity in China. Agr Forest Meteor. 2009;149:1009–1014. http://dx.doi.org/10.1016/j.agrformet.2008.12.004

35. Dumanski J, Peiretti R, Benetis J, Mcgarry D, Pieri C. The paradigm of conservation agriculture. Procedures of the World Association on Soil and Water Conservation. p. 58–62.

36. Dube E, Chiduza C, Muchaonyerwa P. High biomass yielding winter cover crops can improve phosphorus availability in soil. S Afr J Sci. 2014;110(3/4), Art. #2013-0135, 4 pages. http://dx.doi.org/10.1590/sajs.2014/20130135

37. Roosevelt FD. Understanding soil acidity. Magill, SA: Agricultural Bureau of South Australia; 2011. p. 1–8.

38. Zeng F, Ali S, Zhang H, Ouyang Y, Qiu B, Wu F, et al. The influence of pH and organic matter content in paddy soil on heavy metal availability and their uptake by rice plants. Environ Pollut. 2011;159:84–91. http://dx.doi.org/10.1016/j.envpol.2010.09.019

39. Mathew RP, Feng Y, Githinji L, Ankumah R, Balkcom KS. Impact of no-tillage and conventional tillage systems on soil microbial communities. Appl Environ Soil Sci. 2012;2012, Art. #548620, 10 pages. http://dx.doi.org/10.1155/2012/548620

40. Herselman JE. The concentration of selected trace metals in South African soils [dissertation]. Stellenbosch: Stellenbosch University; 2007.

41. Six J, Elliott ET, Paustian K. Soil macroaggregate turnover and microaggregate formation: A mechanism for C sequestration under no-tillage agriculture. Soil Biol Biochem. 2000;32(14):2099–2103. http://dx.doi.org/10.1016/S0038-0717(00)00179-6

42. Kay BD, Vandenbygaart AJ. Conservation tillage and depth stratification of porosity and soil organic matter. Soil Till Res. 2002;66:107–118. http://dx.doi.org/10.1016/S0167-1987(02)00019-3

43. Bot A, Benites J. The Importance of soil organic matter: Key to drought-resistant soil and sustained food production. FAO Soils Bulletin #80. Rome: Food and Agriculture Organization; 2005.

44. Six J, Elliott ET, Paustian K. Aggregate and soil organic matter dynamics under conventional and no-tillage systems. Soil Sci Soc Am J. 1999;63:1350–1358. http://dx.doi.org/10.2136/sssaj1999.6351350x

45. De Moraes Sá JC, Séguy L, Tivet F, Lal R, Bouzinac S, Borszowskei PR, et al. Carbon depletion by plowing and its restoration by no-till cropping systems in oxisols of subtropical and tropical agro-ecoregions in Brazil. Land Degrad Develop. 2015;26(6):531–543. http://dx.doi.org/10.1002/ldr.2218

46. Nascente AS, Li YC, Crusciol CAC. Cover crops and no-till effects on physical fractions of soil organic matter. Soil Till Res. 2013;130:52–57. http://dx.doi.org/10.1016/j.still.2013.02.008

47. Shrestha BM, Singh BR, Forte C, Certini G. Long-term effects of tillage, nutrient application and crop rotation on soil organic matter quality assessed by NMR spectroscopy. Soil Use Manage. 2015;31:358–366. http://dx.doi.org/10.1111/sum.12198

48. Chivenge PP, Murwira, HK, Giller KE, Mapfumo P, Six J. Long-term impact of reduced tillage and residue management on soil carbon stabilization: Implications for conservation agriculture on contrasting soils. Soil Till Res. 2007;94:328–337. http://dx.doi.org/10.1016/j.still.2006.08.006

49. Naresh RK, Gupta RK, Misra AK, Kumar D, Kumar V, Kumar V. Conservation agriculture for smallholder irrigation production: Opportunities and constraints of new mechanized seeding systems: A review. Int J Life Sci Biotech Pharma Res. 2014;3:2250–3137.

On the mental toughness of self-aware athletes: Evidence from competitive tennis players

AUTHOR:
Richard G. Cowden[1,2]

AFFILIATIONS:
[1]Institute of Psychological Wellbeing, North-West University, Potchefstroom, South Africa
[2]Discipline of Psychology, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Richard Cowden

EMAIL:
richardgregorycowden@gmail.com

DATES:
Received: 08 Apr. 2016
Revised: 15 July 2016
Accepted: 07 Sep. 2016

KEYWORDS:
self-reflection; self-insight; athletes; sport; South Africa

HOW TO CITE:
Cowden RG. On the mental toughness of self-aware athletes: Evidence from competitive tennis players. S Afr J Sci. 2017;113(1/2), Art. #2016-0112, 6 pages. http://dx.doi.org/10.17159/sajs.2017/20160112

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
None

This study examined the relationship between mental toughness (MT) and self-awareness in a sample of 175 male and 158 female South African tennis athletes (mean age = 29.09 years, s.d. = 14.00). The participants completed the Sport Mental Toughness Questionnaire and the Self-Reflection and Insight Scale to assess MT (confidence, constancy, control) and self-awareness (self-reflection and self-insight) dimensions, respectively. Linear regression indicated that self-insight (β=0.49), but not self-reflection (β=0.02), predicted global MT. Multivariate regression analyses were significant for self-reflection (η_p^2=0.11) and self-insight (η_p^2=0.24). Self-reflection predicted confidence and constancy (η_p^2=0.05 and 0.06, respectively), whereas self-insight predicted all three MT subcomponents (η_p^2=0.12 to 0.14). The findings extend prior qualitative research evidence supporting the relevance of self-awareness to the MT of competitive tennis athletes, with self-reflection and insight forming prospective routes through which athletes' MT may be developed.

Significance:

- Self-awareness attributes were predictive of higher levels of mental toughness among competitive tennis players.
- Dimensions of self-awareness may offer routes for developing athletes' mental toughness.

Introduction

Mental toughness (MT) is widely recognised as a fundamental attribute for attaining success in sport.[1] Mentally tougher athletes maintain performance levels during adversity; perceive pressure as a challenge and a catalyst for prospering; and maintain emotional, cognitive and behavioural control despite situational stressors.[2] Considering the appeal that these cognitive and behavioural signatures have to athletes, MT has become a prominent research area in the sport performance literature.[3]

Scholars' primary interest in MT is based on the capacity to acquire MT attributes through sport and non-sport developmental influences and experiences[4], as well as through psychological interventions[5]. However, determining the MT dimensions that may be taught and the most effective approaches to develop them requires resolutions to the current conceptual and operational disparities that exist. Some researchers contend that MT is a narrow personality trait that is situationally stable[5,6], whereas others suggest MT is state-specific and may fluctuate depending on the situation[7,8]. In addition to MT manifestation distinctions, these conceptualisations differ in the extent to which MT may be developed. However, in support of the mutual inclusivity of these perspectives, Gucciardi et al.[3] reported that a combination of intraindividual (i.e. within person) and interindividual (i.e. between person) differences may be attributed to the variability of MT. Accordingly, an athlete may display enduring patterns of MT across similar situations, but varied levels of MT across dissimilar situations.

Although the multidimensionality of MT has generally been supported[9], the type and quantity of constituents comprising MT remains unclear[10]. In addition to dimensional discrepancies between sport types,[11] within-sport MT differences have been found. For instance, Coulter et al.[12] reported that risk-taking is an integral MT component in soccer, whereas Thelwell et al.[13] indicated that MT in a soccer player involved affecting one's opponents. The characterisation of MT variations are reflected in the range of instruments that often diverge in the types of MT that are measured. To illustrate, affective intelligence is included as a subfactor on the Cricket Mental Toughness Inventory[14], but is not contained within the Australian Football Mental Toughness Inventory[15].

Although unequivocally determining the components that constitute MT is necessary, there are several components that are repeatedly referred to in the literature.[16] These components include confidence or self-belief; emotional and cognitive control; accepting, persevering and thriving through challenges; and commitment and determination.[2,17] Accordingly, MT refers to a collection of personal resources (inherent and developed) associated with athletes' pursuit of optimal athletic performance levels, irrespective of positive and negative situational demands.[18,19]

In the extant literature, considerable attention has been devoted towards examining the characteristics associated with MT. Commonly identified correlates of MT include effective coping, the use of self-talk, relaxation strategies and mental imagery.[20-22] Mentally tougher athletes have greater flow experiences (concentration, autotelism)[23], perceive stressors as less intense[24], and utilise performance- and mastery-approach achievement goals[25]. Collectively, MT is related to a number of positive psychological characteristics. However, self-awareness, also referred to as psychological self-mindedness, is one concept that has received limited quantitative MT research attention. Self-awareness represents the capacity to attend to, recognise and examine one's thoughts, physiological sensations, emotions and behavioural reactions, either as they occur or retrospectively.[26,27]

Although the self-awareness process is multifaceted and associated with an array of corollaries and self-directed attention areas[26], common conceptualisations encompass two primary components: engagement in self-reflection and the attainment of self-insight[28-30]. Self-reflection involves emotional, cognitive and behavioural self-introspection, whereas self-insight refers to clarifying and obtaining a deeper understanding of such experiences.[29]

Even though self-reflective activities may not automatically result in self-insight[31], self-awareness represents an important process for identifying and replacing maladaptive responses as well as establishing progress towards achieving positive psychobehavioural changes[28,32].

In sport, awareness of one's emotions has been linked to superior performance.[33] In particular, maintaining peak performance levels is at least partly dependent on the ability to recognise negative emotions and cognitions and effectively control or avoid the detrimental effects of such experiences.[34] With research supporting the emotional and cognitive control of mentally tough athletes[35], along with the understanding that MT is associated with positive performance outcomes[3], self-awareness attributes may be relevant to athletes' MT.

Recent qualitative research has posited the relevance of several forms of self-awareness (e.g. emotional and cognitive) in relation to MT. Bull et al.[25], for instance, qualitatively established *thinking clearly* (awareness, focus and control of thoughts) as an essential component of MT in elite cricket. Slack et al.[36] extended this finding to denote cognitive *awareness of own emotions* as indicative of mentally tough English Premier League football referees. There is also evidence to suggest that self-awareness promotes or facilitates heightened levels of MT[37] – a finding that supports early heuristic MT perspectives[38].

Taken together, these findings provide preliminary support for the applicability of self-awareness characteristics to the MT of athletes. However, prior MT studies have not specified what embodies self-awareness, and, despite recent qualitative findings, there is a dearth of knowledge about the role of emotional, cognitive and behavioural self-awareness in relation to MT. Therefore, the purpose of the current study was to explore the relationships between MT and self-awareness components (i.e. self-reflection and self-insight) in competitive tennis players. It was hypothesised that MT and each of its subcomponents would be significantly predicted by both (1) self-reflection and (2) self-insight.

Method

Participants

The participants were 175 male (mean(s.d.) age = 31.99(15.64) years) and 158 female (mean(s.d.) age = 25.89(11.12) years) tennis players competing at various levels: county club (n=58), local county tournament (n=21), university league (n=147), national tournament (n=76) and international tournament (n=31). The athletes had played tennis for a minimum of 5 years (mean(s.d.) of 17.13(12.27) years) and had engaged in tennis competition within 2 weeks prior to their participation in the study.

Materials

Mental toughness

The Sports Mental Toughness Questionnaire (SMTQ)[17], which comprises 14 Likert-type items rated from 1 ('not at all true') to 4 ('very true'), was used to ascertain MT. As a multidimensional measure of MT developed from the most common components of MT identified in the literature,[39] the SMTQ measures control, confidence and constancy. There are four control items (e.g. 'I am overcome by self-doubt'), six confidence items (e.g. 'I interpret potential threats as positive opportunities') and four constancy items (e.g. 'I take responsibility for setting myself challenging targets'). The subscales may be combined for a global measure of MT. The selection of the SMTQ was based on the demonstrated validity (i.e. factorial, divergent, discriminative) and reliability of the instrument reported in the initial validation study.[17] Subsequent studies have supported the convergent validity[40,41] and internal consistency of global MT.[42,43] In this study, Cronbach's alpha for global MT was 0.75.

With alpha inclined to underestimate internal consistency when fewer than 10 items are included on a scale, mean inter-item correlations are important for assessing scalar homogeneity.[44] According to Briggs and Cheek[45], mean inter-item correlation values with a range of 0.2–0.4 indicate appropriate item homogeneity. The internal consistency estimates (and mean inter-item correlations) for confidence, constancy and control were 0.64 (0.23), 0.56 (0.25) and 0.66 (0.33), respectively.

Self-awareness

The Self-Reflection and Insight Scale (SRIS)[29] was used to assess self-awareness. The SRIS comprises 20 Likert-type items (1 = 'strongly disagree', 6 = 'strongly agree') on two subscales: self-reflection (12 items) and self-insight (8 items). Self-reflection measures one's need for and engagement in self-evaluation (e.g. 'I frequently examine my feelings') and self-insight assesses the lucidity of one's thought, emotional and behavioural understanding (e.g. 'I usually know why I feel the way I do'). The SRIS has received construct, convergent and cross-cultural validity support[29,46] and both subscales have evidenced acceptable internal consistency and test-retest reliability[29,47-48]. Internal consistency for self-reflection in this study was 0.90, and alpha (and the mean inter-item correlation) for self-insight was 0.78 (0.30).

Procedure

Permission letters were obtained from relevant tennis organisations in order to acquire institutional ethical approval to conduct the study. Full ethical approval was subsequently granted by the University of KwaZulu-Natal Humanities and Social Sciences Research Ethics Committee (HSS/0740/013D). Suitable tennis tournaments were identified, and the organisers were approached in order to request permission to access the competitive tennis players. Various tennis tournament and club venues across South Africa were attended and the self-administered questionnaires were distributed in groups of approximately 5 to 10 athletes at a time, according to the players' availability. Informed consent was obtained prior to the players' participation, and all relevant Declaration of Helsinki principles were adhered to. A quiet and comfortable venue was established at each location for completion of the questionnaire. The inventories required approximately 15 to 20 minutes to complete; each player completed the SMTQ followed by the SRIS. The participants were requested to consider the extent to which each item applied to them, generally, in relation to their participation in competitive tennis.

Data analyses

Box-plot assessment revealed a small number of gross outliers on the global MT scale and subscales. These individual case values were removed before computing the analyses (see Table 1). Normality estimates (i.e. skewness and kurtosis) were within acceptable limits (i.e. ± 2)[49] for proceeding with parametric computations. Along with these estimates, the descriptive statistics for each variable and bivariate relationships are reported in Table 1. After satisfying the hypothesis testing assumptions associated with conducting parametric regression analyses (e.g. normality, homoscedasticity), multiple linear regression and multivariate regression were used to determine whether self-reflection and self-insight predicted global MT and each of the MT components, respectively. For significant multivariate analyses, a Bonferroni adjustment was applied to follow-up univariate p-values to preserve familywise alpha. An alpha value of 0.05 was used for each statistical test.

Results

Bivariate analyses

According to Cohen's[50] effect size standards, the correlations between global MT and self-reflection (r^2=0.02) and self-insight (r^2=0.25) were small and large, respectively (see Table 1). With the exception of control, which was not significantly associated with self-reflection (r^2 =0.00), the relationships between the MT subcomponents and self-reflection and insight were medium in effect size (r^2 =0.06 to 0.15).

Univariate and multivariate analyses

The multiple linear regression results indicated that self-insight (β=0.49, p<0.001, 95% CI [0.41, 0.57]), but not self-reflection (β=0.02, p=0.652, 95% CI [-0.09, 0.13]), significantly predicted global MT: $F(2, 327)=54.38$, $p<0.001$, $r^2=0.25$, 95% CI [0.17, 0.33].

Table 1: Normality estimates, descriptive statistics and bivariate relationships

Variable	Global mental toughness	Confidence	Constancy	Control	Self-reflection	Self-insight
Global mental toughness	–	0.74**	0.72**	0.72**	0.14*	0.50**
Confidence	–	–	0.41**	0.27**	0.24**	0.34**
Constancy	–	–	–	0.30**	0.24**	0.39**
Control	–	–	–	–	-0.07	0.37**
Self-reflection	–	–	–	–	–	0.25**
Self-insight	–	–	–	–	–	–
n	330	330	327	330	333	333
Mean (s.d.)	41.53 (4.66)	18.26 (2.25)	12.93 (1.71)	10.51 (2.26)	50.41 (10.12)	34.28 (5.80)
Skewness	0.15	0.23	-0.09	0.30	-0.17	-0.09
Kurtosis	-0.17	0.07	-0.54	-0.54	-0.30	-0.34

Note: *p<*0.05 (two-tailed);* **p<*0.001 (two-tailed)*

Multivariate regression revealed self-reflection, $F(3, 318)=12.87$, $p<0.001$, Wilk's $\Lambda=0.89$, $\eta_p^2=0.11$, 95% CI [0.05, 0.17], and self-insight, $F(3, 318)=33.43$, $p<0.001$, Wilk's $\Lambda=0.76$, $\eta_p^2=0.24$, 95% CI [0.15, 0.31], significantly predicted one or more subcomponent of MT. Specifically, self-reflection predicted confidence, $F(1, 320)=16.85$, $p<0.001$, $\eta_p^2=0.05$, 95% CI [0.01, 0.10], and constancy, $F(1, 320)=21.72$, $p<0.001$, $\eta_p^2=0.06$, 95% CI [0.02, 0.12], but not control, $F(1, 320)=1.47$, $p=0.678$, $\eta_p^2=0.01$, 95% CI [0.00, 0.03]. Self-insight was a significant predictor of confidence, $F(1, 320)=44.01$, $p<0.001$, $\eta_p^2=0.12$, 95% CI [0.06, 0.19], constancy, $F(1, 320)=51.46$, $p<0.001$, $\eta_p^2=0.14$, 95% CI [0.08, 0.21], and control, $F(1, 320)=51.37$, $p<0.001$, $\eta_p^2=0.14$, 95% CI [0.08, 0.21].

Discussion

The purpose of the present study was to examine the relationships between MT and self-awareness dimensions among competitive tennis players. The results provided partial support for the hypotheses, as global MT was predicted by self-insight (i.e. clarity of thought, emotional and behavioural understanding) and not by self-reflection (i.e. need for and enagement in psychobehavioural self-evaluation). In addition, confidence and constancy were each predicted by self-reflection and self-insight, although control was only predicted by self-insight. Collectively, the results support superior thought, emotional and behavioural awareness among mentally tougher tennis athletes, which is consistent with qualitative research denoting the relevance of various forms of self-awareness to MT.[25,36]

The finding that self-insight was the single significant predictor of global MT suggests the phase is particularly important to athletes' MT. However, given that attaining insight requires introspection and evaluation of the self,[31] self-reflection is a necessary part of the self-awareness process. Considering the markedly larger effect size between MT and self-insight, as compared to self-reflection, mentally tougher athletes appear to be better at progressing from self-reflective activities to achieve higher levels of self-insight. With prior studies reporting that self-awareness promotes the development of MT,[37] attaining maximal MT benefits might require athletes to engage in and proceed beyond mere self-introspection toward generating a profounder level of psychobehavioural clarity and understanding.

Comparable to the global MT outcomes, the effect sizes in relation to confidence, constancy and control were larger for self-insight. Therefore, in contrast to tennis players who are largely self-reflective, those who are more self-insightful are substantially more likely to possess higher levels of confidence, constancy and control. Accordingly, the advantage of self-awareness to athletes' MT appears to depend more strongly on obtaining self-insight than on engaging in self-reflection alone.

The findings in this study indicate self-awareness is an important determinant of an athlete's confidence and belief.[51] In fact, Beaumont et al.[52] found that developing athletes' awareness of their thoughts and feelings is an effective strategy for fostering their confidence. Referring to an athlete's perceived ability to succeed[53], confidence is enhanced by reflecting on events related to the self, particularly achievements and positive experiences[4]. This would suggest that mentally tougher athletes maintain or develop confidence levels by emphasising the strengths and encouraging features of events and their responses. Further intimating the role of reflection and insight, Jones et al.[2] found that mentally tough athletes' belief is cultivated from recognising the process involved in reaching their level of achievement. The optimistic perspective associated with mentally tough athletes[54] could also account for their persistent sense of confidence[12] despite the psychobehavioural assessment and clarity that may accompany their weaknesses or negative outcomes. Whether mentally tougher athletes display a greater tendency to focus on the positive features of various situations, outcomes or consequences is an area that requires further examination.

Tennis players with greater self-reflective and insightful tendencies also exhibited higher levels of constancy. This finding suggests that athletes who think about and distinctly understand their thoughts, emotions and behaviours are more likely to remain determined and committed. Given that mentally tough athletes optimistically reflect on circumstances and events,[12] positive perceptions about the causes of and explanations for their psychobehavioural responses in such moments may enable them to remain resolute in their efforts and goal pursuits. The positive appraisal orientation that characterises MT is supported by Kaiseler et al.'s[24] study, which found that athletes with higher levels of MT rated stressors as less intense. Furthermore, Mahoney et al.[55] associate MT with striving (i.e. sustained, consistent effort) and thriving (i.e. learning and growth). Through mentally tougher athletes' heightened self-awareness, these qualities likely contribute to appraising their negative responses as opportunities to overcome their weaknesses and improve (i.e. thrive), thereby promoting commitment and dedication (i.e. striving).

The finding that self-insight predicted control supports self-awareness as a fundamental prerequisite to athletes' aptitude to maintain control.[56] Specifically, athletes' emotional and cognitive control could be promoted by clearly and extensively understanding their emotional, cognitive and behavioural experiences. Their superior self-insight might allow mentally tougher athletes to select and effectively use psychological skills[20] to control their thoughts and emotions when they arise during competition. This supposition is supported by Mahoney et al.'s[55] notion that mentally tougher athletes' awareness of their emotional experiences enables them to choose an appropriate coping strategy to maintain desired performance levels. Considering athletes utilise a range of cognitive control strategies

(e.g. word cues) during training and competitive performance contexts,[57] self-insight might be critical to athletes' employment of situationally specific psychological strategies to sustain athletic performance.

Practical suggestions

Given the role of self-awareness in the growth of MT[37], coupled with the capacity to develop MT through specific training programmes[8], self-awareness offers a potential target area for MT interventions. Recent findings have indicated reflective practices may contribute to an athlete's awareness and athletic development.[58] The propensity to self-reflect may be enhanced through the use of a logbook[52] – an exercise that also increases athletes' awareness of negative self-talk, the situations in which it occurs, and the outcomes associated with it[59]. A similar logbook, diary, or written prompt method could be used to facilitate players' self-reflection and enhance their understanding of the cognitive and emotional experiences that occur when training or competing. Through the process of identifying the antecedents and consequences of their particularly negative responses, more adaptive future responses that reflect MT (e.g. emotional control) may be engendered.

Using this type of framework, tennis players may engage in post-competition assessment of the moments in which they experienced positive and negative thoughts (e.g. 'I'm going to lose my serve') and behaviours (e.g. racquet tossing, self-degrading comments). Explanations for these thoughts (e.g. focusing on losing a service game when break point down) or behaviours (e.g. racquet tossing following a loss of serve), along with the outcome following such thoughts or behaviours (e.g. periodic performance slumps), could subsequently be examined. With the support of sport psychology professionals or coaches, this in-depth evaluation and identification of what, when and why thoughts, emotions and behaviours occurred may be used to replace maladaptive responses that emerged previously.

The process through which maladaptive responses are changed could be facilitated in several ways, particularly cognitively. Resulting from their heightened self-awareness, athletes' deleterious responses may be altered through more favourably reappraising the situations in which the responses were triggered.[60] Alternatively, athletes could develop attentional deployment skills[61] in order to focus on the positive features of situations (e.g. one's own physical and technical strengths). Another avenue to changing athletes' emotional and behavioural responses involves cognitive reframing – an approach that encourages the contestation and replacement of debilitative, negative ideas and beliefs with positive, facilitative thought processes.[62] Research has found that reframing is beneficial to generating facilitative perceptions about the influence of competitive anxiety and physiological arousal on athletes' performance.[63] Taken together, these cognitive strategies represent some of the mechanisms through which athletes' could control their thoughts and emotions, develop confidence and maintain commitment levels[52] following their self-reflective activities and attainment of insight.

Limitations and future research directions

Selected methodological limitations should be considered alongside the contributions of this study. While a purposeful decision was made to examine MT within a specific group of athletes (i.e. competitive tennis players), the generalisability of the findings to other sports is questionable. In addition, the non-experimental approach restricts conclusions of causality among the variables included in the study. This could be addressed through experimental and longitudinal MT and athletic performance level studies that target or manipulate self-awareness. Another drawback is that the data were sourced solely from the athletes, and the inclusion of additional data sources (e.g. coaches) would have provided an opportunity to cross-verify the participants' self-reports. The measurement of participants' *average* MT and self-awareness in tennis is another limitation, with the cross-situational applicability and variability of self-awareness, MT and the relationships between the two constructs indeterminable. Future research might explore athletes' self-awareness processes and MT responses following different types of stressors (e.g. inclement weather conditions) and competitive phases (e.g. ahead versus behind). The findings should also be interpreted in conjunction with the criticisms of the SMTQ, such as its brevity, partial conceptual coverage of MT[16] and logical validity concerns[42]. Although the SMTQ has been validated and has received psychometric support, additional validation studies may be required to refine the measure.

Conclusion

The findings in this study support the positive association between MT and self-awareness in competitive tennis players. Most notably, the strongest predictor of MT and its subcomponents was self-insight. Notwithstanding the necessity of self-reflection in the process toward obtaining insight, the latter appears to be particularly important when considering MT and its development among athletes. Research identifying the contextual demands and situation-based use of self-awareness among mentally tough athletes is warranted, along with whether self-reflection and insight may be used to develop MT through interventions.

References

1. Connaughton D, Hanton S, Jones G, Wadey R. Mental toughness research: Key issues in this area. Int J Sport Psychol. 2008;39:192–204.
2. Jones G, Hanton S, Connaughton D. A framework of mental toughness in the world's best performers. Sport Psychol. 2007;21:243–264. http://dx.doi.org/10.1123/tsp.21.2.243
3. Gucciardi D, Hanton S, Gordon S, Mallett C, Temby P. The concept of mental toughness: Tests of dimensionality, nomological network, and traitness. J Pers. 2015;83:26–44. http://dx.doi.org/10.1111/jopy.12079
4. Connaughton D, Hanton S, Jones G. The development and maintenance of mental toughness in the world's best performers. Sport Psychol. 2010;24:168–193. http://dx.doi.org/10.1123/tsp.24.2.168
5. Bell J, Hardy L, Beattie S. Enhancing mental toughness and performance under pressure in elite young cricketers: A 2-year longitudinal intervention. Sport Exerc Perform Psychol. 2013;2:281–297. http://dx.doi.org/10.1037/a0033129
6. St Clair-Thompson H, Bugler M, Robinson J, Clough P, McGeown S, Perry J. Mental toughness in education: Exploring relationships with attainment, attendance, behaviour and peer relationships. Educ Psychol Int J Exp Educ Psychol. 2015;35:886–907. http://dx.doi.org/10.1080/01443410.2014.895294
7. Harmison R. A social-cognitive framework for understanding and developing mental toughness in sport. In: Gucciardi D, Gordon S, editors. Mental toughness in sport: Developments in research and theory. New York: Routledge; 2011. p. 47–68.
8. Slack L, Maynard I, Butt J, Olusaga P. An evaluation of a mental toughness education and training program for early-career English Football League referees. Sport Psychol. 2015;29:237–257. http://dx.doi.org/10.1123/tsp.2014-0015
9. Hardy L, Bell J, Beattie S. A neuropsychological model of mentally tough behavior. J Pers. 2014;82:69–81. http://dx.doi.org/10.1111/jopy.12034
10. Gucciardi D, Jackson B, Hanton S, Reid M. Motivational correlates of mentally tough behaviours in tennis. J Sci Med Sport. 2015;18:67–71. http://dx.doi.org/10.1016/j.jsams.2013.11.009
11. Crust L. A review and conceptual re-examination of mental toughness: Implications for future researchers. Pers Indiv Differ. 2008;45:576–583. http://dx.doi.org/10.1016/j.paid.2008.07.005
12. Coulter T, Mallett C, Gucciardi D. Understanding mental toughness in Australian soccer: Perceptions of players, parents, and coaches. J Sports Sci. 2010;28:699–716. http://dx.doi.org/10.1080/02640411003734085
13. Thelwell R, Weston N, Greenlees I. Defining and understanding mental toughness within soccer. J Appl Sport Psychol. 2005;17:326–332. http://dx.doi.org/10.1080/10413200500313636
14. Gucciardi D, Gordon S. Development and preliminary validation of the Cricket Mental Toughness Inventory (CMTI). J Sports Sci. 2009;27:1293–1310. http://dx.doi.org/10.1080/02640410903242306
15. Gucciardi D, Gordon S, Dimmock J. Development and preliminary validation of a mental toughness inventory for Australian football. Psychol Sport Exerc. 2009;10:201–209. http://dx.doi.org/10.1016/j.psychsport.2008.07.011

16. Gucciardi D, Mallett C, Hanrahan S, Gordon S. Measuring mental toughness in sport: Current status and future directions. In: Gucciardi D, Gordon S, editors. Mental toughness in sport: Developments in theory and research. New York: Routledge; 2011. p. 108–132.

17. Sheard M, Golby J, Van Wersch A. Progress toward construct validation of the Sports Mental Toughness Questionnaire. Eur J Psychol Assess. 2009;25:186–193. http://dx.doi.org/10.1027/1015-5759.25.3.186

18. Cook C, Crust L, Littlewood M, Nesti M, Allen-Collinson J. 'What it takes': Perceptions of mental toughness and its development in an English Premier League Soccer Academy. Qual Res Sport Exerc Health. 2014;6:329–347. http://dx.doi.org/10.1080/2159676X.2013.857708

19. Gucciardi D, Jackson B, Hodge K, Anthony A, Brooke L. Implicit theories of mental toughness: Relations with cognitive, motivational, and behavioral correlates. Sport Exerc Perform Psychol. 2015;4:100–112. http://dx.doi.org/10.1037/spy0000024

20. Crust L, Azadi K. Mental toughness and athletes' use of psychological strategies. Eur J Sport Sci. 2010;10:43–51. http://dx.doi.org/10.1080/17461390903049972

21. Mattie P, Munroe-Chandler K. Examining the relationship between mental toughness and imagery use. J Appl Sport Psychol. 2012;24:144–156. http://dx.doi.org/10.1080/10413200.2011.605422

22. Nicholls A, Levy A, Polman R, Crust L. Mental toughness, coping self-efficacy, and coping effectiveness among athletes. Int J Sport Psychol. 2011;42:513–524.

23. Crust L, Swann C. The relationship between mental toughness and dispositional flow. Eur J Sport Sci. 2013;13:215–220. http://dx.doi.org/10.1080/17461391.2011.635698

24. Kaiseler M, Polman R, Nicholls A. Mental toughness, stress, stress appraisal, coping and coping effectiveness in sport. Pers Indiv Differ. 2009;47:728–733. http://dx.doi.org/10.1016/j.paid.2009.06.012

25. Bull S, Shambrook C, James W, Brooks J. Towards an understanding of mental toughness in elite English cricketers. J Appl Sport Psychol. 2005;17:209–227. http://dx.doi.org/10.1080/10413200591010085

26. Morin A. Self-awareness part 1: Definition, measures, effects, functions, and antecedents. Soc Pers Psychol Compass. 2011;5:807–823. http://dx.doi.org/10.1111/j.1751-9004.2011.00387.x

27. Pieterse A, Lee M, Ritmeester A, Collins N. Towards a model of self-awareness development for counselling and psychotherapy training. Couns Psychol Q. 2013;26:190–207. http://dx.doi.org/10.1080/09515070.2013.793451

28. Grant A. Rethinking psychological mindedness: Metacognition, self-reflection and insight. Behav Change. 2001;18:8–17. http://dx.doi.org/10.1375/bech.18.1.8

29. Grant A, Franklin J, Langford P. The self-reflection and insight scale: A new measure of private self-consciousness. Soc Behav Pers. 2002;30:821–836. http://dx.doi.org/10.2224/sbp.2002.30.8.821

30. Xu X. Self-reflection, insight, and individual differences in various language tasks. Psychol Rec. 2011;61:41–58.

31. Hixon J, Swann W. When does introspection bear fruit? Self-reflection, self-insight, and interpersonal choices. J Pers Soc Psychol. 1993;64:35–43. http://dx.doi.org/10.1037/0022-3514.64.1.35

32. Carver C, Scheier M. On the self-regulation of behavior. New York: Cambridge University Press; 1998. http://dx.doi.org/10.1017/CBO9781139174794

33. Zizzi S, Deaner H, Hirschhorn D. The relationship between emotional intelligence and performance among college baseball players. J Appl Sport Psychol. 2003;15:262–269. http://dx.doi.org/10.1080/10413200305390

34. Krane V, Williams J. Psychological characteristics of peak performance. In: Williams J, editor. Applied sport psychology: Personal growth to peak performance. New York: McGraw-Hill; 2006. p. 207–227.

35. Gucciardi D, Gordon S, Dimmock J. Advancing mental toughness research and theory using personal construct psychology. Int Rev Sport Exerc Psychol. 2009;2:54–72. http://dx.doi.org/10.1080/17509840802705938

36. Slack L, Butt J, Maynard I, Olusoga P. Understanding mental toughness in elite football officiating: Perceptions of English Premier League referees. Sport Exerc Psychol Rev. 2014;10:4–24.

37. Gucciardi D, Gordon S, Dimmock J. Evaluation of a mental toughness training program for youth-aged Australian footballers: I. A qualitative analysis. J Appl Sport Psychol. 2009;21:324–339. http://dx.doi.org/10.1080/10413200903026074

38. Loehr J. The new mental toughness training for sport. New York: Penguin; 1995.

39. Sheard M. Mental toughness: The mindset behind sporting achievement. 2nd ed. New York: Routledge; 2013.

40. Chen M, Cheesman D. Mental toughness of mixed martial arts athletes at different levels of competition. Percept Mot Skills. 2013;116:905–917. http://dx.doi.org/10.2466/29.30.PMS.116.3.905-917

41. Meggs J, Ditzfeld C, Golby J. Self-concept organisation and mental toughness in sport. J Sports Sci. 2014;32:101–109. http://dx.doi.org/10.1080/02640414.2013.812230

42. Crust L, Swann C. Comparing two measures of mental toughness. Pers Indiv Differ. 2011;50:217–221. http://dx.doi.org/10.1016/j.paid.2010.09.032

43. Petrie T, Deiters J, Harmison R. Mental toughness, social support, and athletic identity: Moderators of the life stress–injury relationship in collegiate football players. Sport Exerc Perform Psychol. 2014;3:13–27. http://dx.doi.org/10.1037/a0032698

44. Pearson R. Statistical persuasion: How to collect, analyze, and present data accurately, honestly, and persuasively. Thousand Oaks, CA: Sage Publications; 2010. http://dx.doi.org/10.4135/9781452230122

45. Briggs S, Cheek J. The role of factor analysis in the development and evaluation of personality scales. J Pers. 1986;54:106–148. http://dx.doi.org/10.1111/j.1467-6494.1986.tb00391.x

46. DaSilveira A, DeCastro T, Gomes W. Escala de Autorreflexão e Insight: Nova medida de autoconsciência adaptada e validada para adultos Brasileiros [Self-Reflection and Insight Scale: New self-consciousness measure adapted and validated to Brazilian adults]. Psico-PUCRS. 2012;43:155–162. Portugese.

47. DaSilveira A, DeSouza M, Gomes W. Self-consciousness concept and assessment in self-report measures. Front Psychol. 2015;6, Art. #930, 11 pages. http://dx.doi.org/10.3389/fpsyg.2015.00930

48. Roberts L, Heritage B, Gasson N. The measurement of psychological literacy: A first approximation. Front Psychol. 2015;6, Art. #105, 12 pages. http://dx.doi.org/10.3389/fpsyg.2015.00105

49. Chou C, Bentler P. Estimates and tests in structural equation modeling. In: Hoyle R, editor. Structural equation modelling: Concepts, issues, and applications. Thousand Oaks, CA: Sage Publications; 1995. p. 37–55.

50. Cohen J. A power primer. Psychol Bull. 1992;112:155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

51. Hays K, Maynard I, Thomas O, Bawden M. Sources and types of confidence identified by world class sport performers. J Appl Sport Psychol. 2007;19:434–456. http://dx.doi.org/10.1080/10413200701599173

52. Beaumont C, Maynard I, Butt J. Effective ways to develop and maintain robust sport-confidence: Strategies advocated by sport psychology consultants. J Appl Sport Psychol. 2015;27:301–318. http://dx.doi.org/10.1080/10413200.2014.996302

53. Vealey R. Understanding and enhancing self-confidence in athletes. In: Singer R, Hausenblas H, Janelle C, editors. Handbook of sport psychology. 2nd ed. New York: Wiley; 2001. p. 550–565.

54. Guillén F, Laborde S. Higher-order structure of mental toughness and the analysis of latent mean differences between athletes from 34 disciplines and non-athletes. Pers Indiv Differ. 2014;60:30–35. http://dx.doi.org/10.1016/j.paid.2013.11.019

55. Mahoney J, Gucciardi D, Ntoumanis N, Mallett C. The motivational antecedents of the development of mental toughness: A self-determination theory perspective. Int Rev Sport Exerc Psychol. 2014;7:184–197. http://dx.doi.org/10.1080/1750984X.2014.925951

56. Brunelle J, Janelle C, Tennant L. Controlling competitive anger among male soccer players. J Appl Sport Psychol. 1999;11:283–297. http://dx.doi.org/10.1080/10413209908404205

57. Toner J, Montero B, Moran A. Considering the role of cognitive control in expert performance. Phenomenol Cogn Sci. 2014;14:1127–1144. http://dx.doi.org/10.1007/s11097-014-9407-6

58. Faull A, Cropley B. Reflective learning in sport: A case study of a senior level triathlete. Refl Pract. 2009;10:325–339. http://dx.doi.org/10.1080/14623940903034655

59. Hardy J, Roberts R, Hardy L. Awareness and motivation to change negative self-talk. Sport Psychol. 2009;23:435–450. http://dx.doi.org/10.1123/tsp.23.4.435

60. Lane A, Beedie C, Jones M, Uphill M, Devonport T. The BASES Expert Statement on emotion regulation in sport. J Sports Sci. 2012;30:1189–1195. http://dx.doi.org/10.1080/02640414.2012.693621

61. Webb T, Miles E, Sheeran P. Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. Psychol Bull. 2012;138:775–808. http://dx.doi.org/10.1037/a0027600

62. Mennis D, Ellard K, Fresco D, Gross J. United we stand: Emphasizing commonalities across cognitive-behavioral therapies. Behav Ther. 2013;44:234–248. http://dx.doi.org/10.1016/j.beth.2013.02.004

63. Moore L, Vine S, Wilson M, Freeman P. Reappraising threat: How to optimize performance under pressure. J Sport Exerc Psychol. 2015;37:339–343. http://dx.doi.org/10.1123/jsep.2014-0186

AUTHORS:
Edward J. Odes[1]
Alexander H. Parkinson[2]
Patrick S. Randolph-Quinney[1,2,3]
Bernhard Zipfel[2,4]
Kudakwashe Jakata[2]
Heather Bonney[5]
Lee R. Berger[2]

AFFILIATIONS:
[1]School of Anatomical Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa
[3]School of Forensic and Applied Sciences, University of Central Lancashire, Preston, Lancashire, United Kingdom
[4]School of Geosciences, University of the Witwatersrand, Johannesburg, South Africa
[5]Department of Earth Sciences, Natural History Museum, London, United Kingdom

CORRESPONDENCE TO:
Edward J. Odes

EMAIL:
eddieodes@gmail.com

DATES:
Received: 17 May 2016
Revised: 07 Sep. 2016
Accepted: 08 Sep. 2016

KEYWORDS:
Sterkfontein; micro computed tomography; spinal degenerative joint disease; palaeopathology; taphonomy

HOW TO CITE:
Odes EJ, Parkinson AH, Randolph-Quinney PS, Zipfel B, Jakata K, Bonney H, et al. Osteopathology and insect traces in the *Australopithecus africanus* skeleton StW 431. S Afr J Sci. 2017;113(1/2), Art. #2016-0143, 7 pages. http://dx.doi.org/10.17159/sajs.2017/20160143

ARTICLE INCLUDES:
✓ Supplementary material
× Data set

FUNDING:
DST/NRF Centre of Excellence in Palaeosciences; National Research Foundation (South Africa); University of the Witwatersrand

Osteopathology and insect traces in the *Australopithecus africanus* skeleton StW 431

We present the first application of high-resolution micro computed tomography in an analysis of both the internal and external morphology of the lumbar region of StW 431 – a hominin skeleton recovered from Member 4 infill of the Sterkfontein Caves (South Africa) in 1987. The lumbar vertebrae of the individual present a number of proliferative and erosive bony processes, which were investigated in this study. Investigations suggest a complex history of taphonomic alteration to pre-existing spinal degenerative joint disease (SDJD) as well as post-mortem modification by an unknown insect. This study is in agreement with previous pathological diagnoses of SDJD which affected StW 431 and is the first time insect traces on this hominin are described. The results of this analysis attest to the complex series of post-mortem processes affecting the Sterkfontein site and its fossil assemblages.

Significance:

- First application of high-resolution micro computed tomography of the lumbar region of StW 431, a partial skeleton of *Australopithecus africanus,* attests to pre-existing degenerative joint disease and identifies post-mortem modification by an unknown insect.
- The co-occurrence of degenerative pathology and insect modification may not be unique to StW 431. A combination of traditional morphoscopic analysis and non-invasive high-resolution tomography is recommended.

Introduction

The StW 431 hominin skeleton was discovered by excavation teams from the University of the Witwatersrand during February and March 1987,[1] at the karstic cave site of Sterkfontein, Cradle of Humankind, South Africa. This site has yielded the largest sample of the taxon *Australopithecus africanus,* fossil members of the genus *Homo* and archaeological evidence of Oldowan and more recent lithic technologies.[2-7]

The StW 431 specimen comprises a partial skeleton of *Australopithecus africanus,* consisting of 48 fragments reconstructed into 18 partial elements.[1] The skeletal remains (Figure 1) consist of portions of the right scapula and clavicle, right humerus, radius and ulna, a right rib, five thoracic vertebrae, five lumbar vertebrae, the first three sacral segments and os coxae, and part of the right acetabulum. The individual is skeletally adult (based on sacral vertebral fusion) and has been previously assigned as male on the basis of a number of morphological characteristics. These characteristics include overall robusticity and muscular markers, the proportions between the body of the first sacral segment and the sacral base, and a relatively large estimated body mass (41.1–42.5 kg) based on reduced major axis regression consistent with the male range of body size for *A. africanus.*[1] The postcranial remains were associated to a single individual, based partly on position, refit, similar colour and physical condition, state of preservation and morphology.[1]

Stratigraphic understanding of the site at the time of excavation attributed the remains to Bed B of Member 4 within the Sterkfontein Formation.[1] While the exact chronological age of this stratigraphic bed is unknown, age estimates range between 1.5 Ma and 2.8 Ma for the member as a whole.[2-7] It is widely acknowledged that accurate dating of South African cave sites has been historically problematical.[8] Age ranges of between 2.4 mya and 2.8 mya of Sterkfontein Member 4 have been established by faunal analysis and archaeology[9-12], where absolute dating methods have proved problematic[13]. Electron spin resonance methods have also been used to date South African Plio-Pleistocene sites, and dates between 1.6 mya and 2.87 mya for Member 4 Sterkfontein have been suggested, with an average electron spin resonance estimated age of 2.1±0.5 Ma.[14] Using U-Pb dating methods, a new absolute age range of between 2.65±0.30 Ma and 2.01±0.05 Ma has been assigned to fossiliferous deposits of Sterkfontein Member 4.[15] Electron spin resonance, isotopic and palaeomagnetic studies carried out on speleothem and siltstone material from Member 4 Sterkfontein Cave suggests the date of the deposits and *A. africanus* fossils at between 2.58 Ma and 2.16 Ma. Thus largely based on arguments of parsimony, provenance and morphology, most researchers studying the skeleton have accepted a taxonomic assignment to *A. africanus.*[16,17]

StW 431 is additionally important as the skeleton has been cited with regard to ongoing debate concerning the presence of skeletal pathology in the specimen; researchers have previously identified two conditions – brucellosis and spondylosis deformans – affecting this specimen.[18,19] Staps[18] diagnosed spondylosis deformans with osteophytic formation, and osteoarthritis of the facet joints from L4 to S1. In contradistinction, D'Anastasio and colleagues[19] diagnosed a case of possible brucellosis. In this paper, we attempt to resolve this debate and clarify the nature of ante- versus post-mortem processes affecting the specimen.

Scale = 50 mm

Figure 1: StW 431 – a partial skeleton of *Australopithecus africanus* discovered at Sterkfontein Caves in 1987. Stw 431 represented only the third partial skeleton attributed at the time to *A. africanus*, and represents the only probable male skeleton attributed to this taxon to date.

Materials and methods

The fourth and fifth lumbar vertebrae of StW 431 were studied macro- and microscopically to record surface morphology. Internal bone structure was imaged using micro computed tomography (micro-CT). Comparative skeletal material – including healthy modern human and pathological vertebrae from the Bone Teaching Collection and Raymond A. Dart Collection of Human Skeletons housed in the School of Anatomical Sciences at the University of the Witwatersrand – was also imaged. Comparative material with known and purported brucellar pathology from radiographical and palaeopathological literature was also studied (see Supplementary table 1 for a list of all comparative materials used in this study).

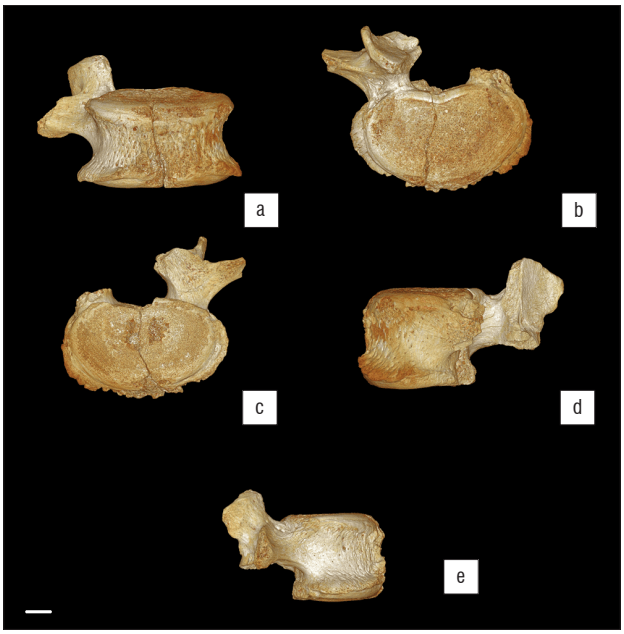
Gross surface morphology of the vertebral specimens was studied microscopically at magnifications of 7–25 times under reflected light using an Olympus SZX 16 multifocus microscope fitted with a digital camera. Micrographic imaging of the anterior and lateral bodies, antero-superior margins and endplates of the two lumbar vertebrae was carried out by applying Analysis 5.0, which includes a Z-stacking function. This function operates as an automated smoothing process whereby multiple sub-images are transformed into a single high-quality, high-resolution image with greater depth of field than a single micrograph.

In order to investigate internal (as well as external) morphology, imaging of the StW 431 vertebral specimens was carried out using micro-CT undertaken with a Nikon Metrology XTH 225/320 LC dual source industrial CT system housed in the Evolutionary Studies Institute of the University of the Witwatersrand. Both specimens were scanned using a potential difference of 85 kV and a current of 75 μA at a resolution of 33 μm; a TIFF format image stack was generated in VG Studio Max following volume reconstruction. Further reconstruction was undertaken using Avizo Amira 5.4 to generate both two-dimensional orthoslice and three-dimensional surface rendered views based on the volume data; multiplanar mode was used to allow the recovery of homologous orthoslices (superior, inferior, coronal and sagittal) through both specimens for comparative purposes.

Results

Macroscopic analysis

The fourth lumbar vertebra (L4) is largely complete (Figure 2) and presents as the bulk of the vertebral centrum, the right pedicle, with right superior articular and transverse processes. Some degree of post-mortem damage has occurred, resulting in the loss of the left pedicle, lamina and superior, inferior and transverse processes. The remaining superior and transverse processes display some degree of post-mortem damage, with apices of both processes truncated, leading to exposure of internal trabeculae. The vertebral body further displays a fracture which runs slightly antero-laterally from the margin of the vertebral foramen to the anterior border of the body, just to the right of the midline. This fracture has led to the loss of a wedge of cortical bone at the antero-inferior margin of the centrum, with loss of cortex and exposure of underlying trabecular bone either side of the plane of the fracture. This erosion is contiguous with a zone of cortical erosion affecting the anterior surface of the body, with greatest removal of cortex on the left side of the body. Additionally, there is a small area of post-mortem erosion at the antero-superior rim of the centrum, on either side of the fracture, which has shaved off part of the labrum of the body over approximately 5–7 mm of the margin.



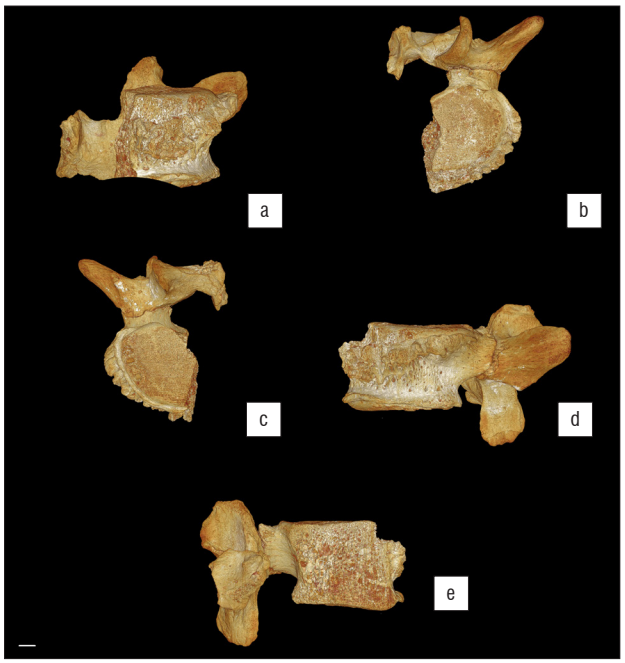
Scale = 10 mm

Figure 2: Three-dimensional volume rendered micro-CT views of the L4 vertebra of StW 431: (a) anterior, (b) superior, (c) inferior, (d) left lateral and (e) right lateral.

Slight osteophytic lipping occurs around the antero-superior region of L4, which creates a rolled appearance to the margin of the labrum, which is further disturbed by the erosion of this surface (Figure 2a). Extensive osteophytosis is expressed around the circumference of the inferior border of the vertebral body, present as a crenulated skirt of bone running from the anterior roots of the pedicles (the pedicle on the left, remaining as a short stump of the original process) around the margin of the body (Figure 2b,c). This skirt projects anteriorly a maximum of 5 mm beyond the inferior margin, and is most pronounced at the lateral

borders of the body. The superior endplate exhibits a generalised pattern of mild porosity. The surface of the inferior endplate exhibits more extensive erosion, with two major areas of cortical loss on either side of the midline of the vertebra – these are approximately 4 mm by 5 mm in diameter, the left of which is transected by the fracture which runs through the body.

The fifth lumbar vertebra (Figure 3) is less complete than L4 (Figure 2), displaying extensive post-mortem fracturing and damage. L5 presents as the left half of the vertebral body, with the pedicle, superior and inferior articular processes, and transverse process largely intact. The left lamina is present, together with a small portion of right lamina still attached, thus preserving a significant portion of the spinous process. In keeping with L4, L5 exhibits extensive osteophytic deposition, but both superior and inferior endplate margins, and the anterior surface of the body in the midline, which forms a series of cranio-caudally orientated buttresses, are more pronounced on the lateral side of the body. The osteophytic lipping extends from the left pedicle antero-laterally and continues anteriorly until it reaches the most anterior point of the vertebral body before it is interrupted by the fracture, which slices through the approximate midpoint of the vertebral body. The anterior projection of the superior osteophytic formation is interrupted by a large zone of removal of osseous tissue, which presents as a scooped, hollowed-out region that has removed both the osteophytic rim and a portion of the anterior body (Figure 3): this zone of removal has previously been described elsewhere as a region of osteolysis, possibly related to the presence of infectious disease.[19]

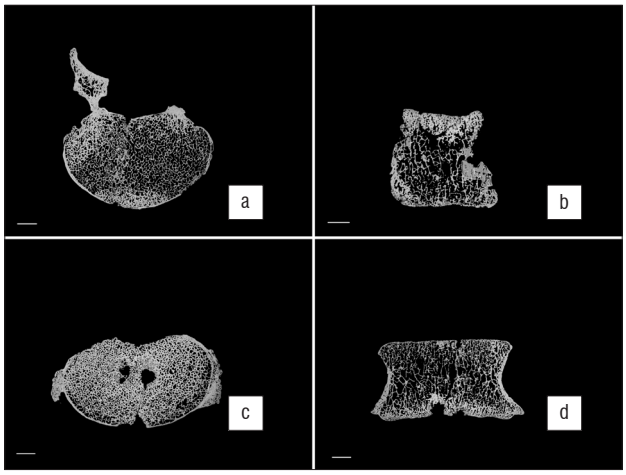


Scale = 10 mm

Figure 3: Three-dimensional volume rendered micro-CT views of the L5 vertebra of StW 431: (a) anterior, (b) superior, (c) inferior, (d) left lateral and (e) right lateral. Note the osteophytic formation on the superior and inferior endplate margins, as well as the anterior surface of the body in midline.

Microscopic analysis

Micro-tomographic imaging of the internal morphology of L4 indicates areas of both bone deposition and removal. The superior body displays very slight osteophytic lipping around the antero-superior margin, which overall presents a slightly rolled appearance in cross section. This is clearly seen in Figure 4a,b, with a slight degree of remodelling and an increase in trabecular thickness and concomitant reduction in trabecular spacing in the antero-superior margin of the centrum. The greatest expression of osteophytosis occurs in the inferior body, where the original cortical bone making up the surface of the body can be seen (Figure 4c). Secondly extensive areas of new bone formation of varying densities (from open woven to sclerotic) can be seen in transverse and coronal sections (Figure 4c,d). In addition to the marginal osteophytes, the two zones of endplate erosion are clearly evidenced. These appear as punched out cavities within the inferior body, where struts of individual trabeculae are truncated (and subsequently infilled with breccia in areas) with no evidence of sclerosis or reactive bone formation around the cavity margins (Figure 4c).

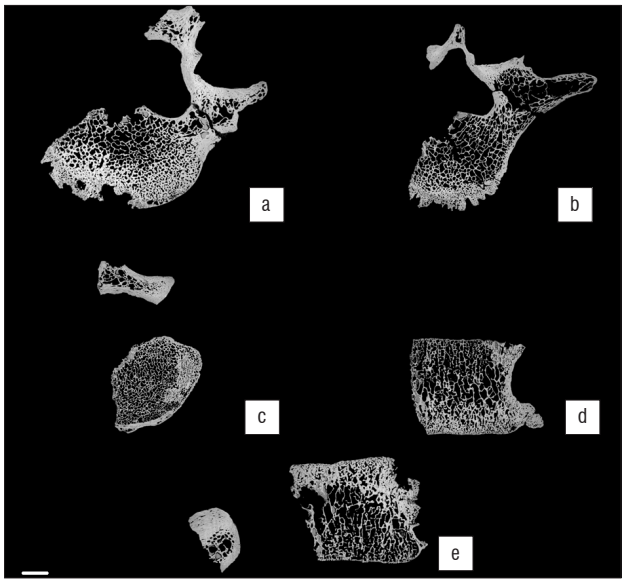


Scale = 10 mm

Figure 4: Micro-CT orthoslice views of the L4 vertebra of StW 431: (a) transverse superior, (b) sagittal midline, (c) transverse inferior and (d) anterior. Note the bilateral osteophytic formation and two zones of erosion evident on the inferior endplate surface of the L4 in (c). There is no evidence of sclerosis or new bone formation around the margins of the cavities (c). Also note areas of new bone formation ranging from open woven (c) to sclerotic (d).

In keeping with gross morphological assessment, micro-tomographic imaging of the internal morphology of L5 indicates extensive areas of both bone deposition and removal, with pronounced osteophytosis affecting the superior, anterior and inferior margins of the vertebral body. Figure 5a shows a transverse cross section through the superior region of the endplate (just inferior to the labrum), which demonstrates the extensive projection of osteophytes at the antero-superior margin. Unlike those affecting L4, these osteophytes express as direct remodelling of the cortical bone, appearing as both a thickened sclerotic margin and as crenulated buttresses of porous bone devoid of internal trabeculae; such a pattern is further reflected in the mid-transverse region of the body (Figure 5b). The inferior endplate, on the other hand, only presents osteophytosis as a thin ordered sclerotic rim extending around the inferior circumference, without buttressing (Figure 5c).

The pattern of erosion affecting L4 was identified by us (AHP) as being of potential post-mortem origin, and specifically surface modification caused by insects. In order to clarify this modification, potential insect damage was further imaged at magnifications between 7x and 115x using an Olympus SZX 16 multifocus microscope fitted with a digital camera. Terminologies used to describe the morphology of traces observed is based on a recent summary of general morphologies of bioerosional traces in bone.[20] Length and width measurements were taken using Stream Essentials© image processing software linked to the Olympus SZX multifocus microscope. Depth measurements were obtained using digital callipers. Two comparative collections of insect damage to bone were utilised in this study: these experimental collections comprise bones exposed to the following agents under control conditions; a southern African termite (*Trinervitermes trinervoides*)[21] and a dermestid beetle (*Dermestes maculatus*)[22]. Furthermore, insect damage was compared to available data gleaned from the literature.[21-24]



Scale = 10 mm

Figure 5: Micro-CT orthoslice views of the L5 vertebra of StW 431: (a,b) transverse to the midline (close to surface), (c) transverse to inferior, (d) coronal midline, (e) sagittal superior to bottom. Note new bone formation on the anterior wall (e) and major osteophytic formation at the antero-superior margin indicating remodelling of the cortex, revealing a sclerotic margin, and as crenulated buttresses of porous bone devoid of internal trabeculae (a,b). There is no evidence of sclerosis or reactive bone formation around the cavity margins (a). The inferior endplate (c) exhibits an osteophytic formation as a thin ordered sclerotic rim around inferior circumference with no presence of buttressing. Notice the channel interpreted as invertebrate damage in (e).

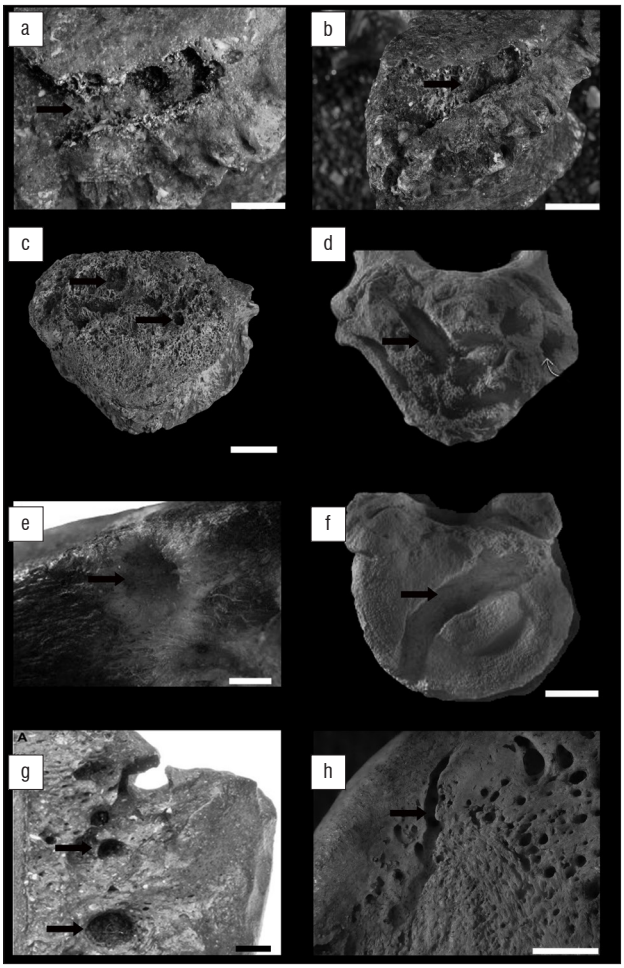
Insect traces

A channel-like structure is present on the antero-superior margin of the L5 vertebral body (Figure 6a,b). The channel comprises a series of conjoined excavations and cavities and extends from approximately the midpoint of the antero-superior rim of the body (where the vertebra has been fractured post-mortem) until the mid-lateral antero-superior margin of the upper endplate. Externally, there is no evidence of sclerosis or reactive bone formation around the cavity margins. The channel narrows from 8 mm to 3 mm in diameter and penetrates to a maximum depth of 4 mm. Embedded in this channel are distinctive circular holes which are 3–4 mm in diameter and 3–4 mm deep (Figure 5e demonstrates this pattern in cross section). No discernible mandible marks were found in association with the insect traces. The micro-CT images demonstrate an absence of bone remodelling related to either the channel or holes.

Discussion

Pathology

Gross observation of surface morphology and micro-tomographic imaging of the internal cortical and trabecular morphology of the StW 431 hominin vertebrae have indicated the presence of osseous proliferation (osteophytosis) consistent with degenerative spinal joint disease. Degenerative joint disease (DJD) is one of the most common pathological ailments observed in archaeo-skeletal assemblages, and is seen in many cases as a natural internal response of the body to 'wear and tear'.[25-29] Skeletal involvement in DJD usually consists of the covarying processes of bone formation and bone destruction, including: (1) degeneration of articular cartilage with exposure of the bone surface, leading to progressive erosive porosity in the subchondral bone; (2) bone remodelling which produces stabilising focal nodules (osteophytes) of new bone formation at joint margins or ligament/tendon insertions (entheses); (3) subchondral cysts or lytic cavities in cartilage-depleted areas; and (4) eburnation produced by direct bone-on-bone abrasion, often leaving polished wear facets on the affected areas.[30-32]



Scale = 10 mm

Figure 6: Insect traces excavated into the L5 vertebrae on StW 431 are indicated by arrows (a and b); (c) circular boring on a cercopithecus vertebra from Cooper's D; (d, f) furrows attributed to *Dermestes maculatus* from the Jurassic period[34]; (e) hole produced by the termite *Trinervitermes trinervoides*[21]; (g) holes produced by dermestids under experimental conditions[39]; and (h) furrow produced by the termite *T. trinervoides*[21].

DJD is a chronic disease process affecting joints, particularly large weight-bearing joints and is common in older individuals, but can occur in younger individuals either through a genetic mechanism or, more commonly, because of previous joint trauma.[30-32] Affected individuals may express reduced mobility, reduced flexibility and chronic pain. In modern clinical practice, DJD is equally prevalent in men and women in their mid-forties to mid-fifties, although the disease has been shown to affect much younger age groups in forensic, historical and archaeological populations in individuals with physically stressful lifestyles or occupations.[33]

The presence of osteophytosis and other associated degenerative traits elsewhere in the lumbar and sacral region of StW 431 has been proposed by Staps[18] as symptomatic of spondylosis deformans in this specimen, and our results are broadly consistent with this diagnosis, although the degenerative changes observed could also be attributed to normal age-related degenerative joint changes. However, the causal mechanism behind the destructive erosion seen on L4 and L5 has yet to be unequivocally addressed from a pathological perspective and here we attribute it to post-mortem alteration by insects.

Pirrone et al.[20] proposed that traces in bone produced by insects can be classified into one of eight general morphological categories. Subsequently, Parkinson[23] synonymised grooves and striae, as well as furrows and channels, thus reducing the number of general morphological categories to only six. These categories now are: pits, holes, chambers, tubes, furrows and grooves. The insect traces identified during the course

of this study can be classified into two of these morphological categories, namely furrows and holes. Holes (Figure 6c,e,g) are considered vertical excavations in bone which display a circular morphology in plan view and a bowl-shaped morphology in cross section. Holes are primarily found embedded in the outer cortical surface or as shallow excavations which penetrate the cortical surface and terminate in the trabecular bone.[20,23] Furrows (Figure 6d,f,h) are horizontal excavations which present as a meandering trail across the surface of a bone. Furrows are distinguishable from chambers as they lack the characteristic ellipsoidal morphology in plan view. Furrows are constructed to variable depths, and thus either record the presence or absence of vertical walls, and in cross section either display a bowl-shaped or shallow, rounded profile (see Parkinson[23(Fig.2)] and Pirrone et al.[20(Fig.3)] for a summary of plan and cross-sectional views of these common morphologies). Thus, the primary basis for identifying the insect traces on StW 431 is morphological similarity with traces widely attributed to insects in the literature (Figure 6).

The holes identified on StW 431 have a diameter range of 3–4 mm which are within the range of holes recently reported from Cooper's D (1–8.6 mm)[23] (Figure 6c) and those produced by the southern African termite *T. trinervoides* under experimental conditions (1.54–3.63 mm)[21] (Figure 6e). However, it is necessary to compare size data to data across all morphological categories because of the transitional nature of these traces. In that the morphological categories proposed by Pirrone et al.[20] and Parkinson[23] are by no means independent of one another – it is widely accepted that they represent transitional morphotypes which relate to the orientation of the excavation relative to the bone surface.[34-37] For example, if an insect excavates perpendicular to the bone surface and thus penetrates vertically in the bone, the trace would transition through a number of morphologies: initially it would be a pit, transition into a hole, and culminate in a tube. Alternatively, if the excavation is orientated parallel to the bone surface, the initial excavations would potentially take the form of a chamber and culminate in a meandering furrow. The transitional nature of these morphologies is an important consideration when one compares measurement data available in the literature. Thus the holes identified on StW 431 also fall within the range of tubes reported from Swartkrans (2–5 mm)[38], as well as those produced by *T. trinervoides* (3.41–4.2 mm)[21]. The holes on StW 431 fall outside of the range of tubes produced by *D. maculatus* (0.21–0.68 mm) and that of pits (1.5–2.5 mm) produced by dermestids[39] under experimental conditions (Figure 6g). Actualistic experimental results of bones exposed to *D. maculatus* suggests that dermestids rarely produce tubes or holes in bones, even after extended periods of exposure.[22] Lastly, the StW 431 holes also fall outside of the range of holes or tubes reported from Makapansgat which were attributed to dermestids by Kitching[40].

Morphological and metrological variables are key in the identification of insect traces on bone but they have little application in identifying a specific causal agent responsible for their creation.[23,35-37] Traces produced by insects are morphologically consistent despite geographical and/or temporal distribution; for example, furrows as described in this study from the Plio-Pleistocene are consistent with traces recently reported from the Jurassic of China.[36] Shallow circular holes have been reported throughout the Mesozoic and well into the Late Cenozoic; the most commonly inferred causal agent of traces during the Mesozoic are dermestids[34,41-44], but this shifts during the Cenozoic to termites being the most commonly inferred agent[21,24,45-47]. This temporal and geographical consistency of trace gross morphology relates to the similarity in the associated behaviour of insects whilst producing traces in bone. This unfortunate reality suggests that gross morphological categories should not be attributed to a specific agent, or should be attributed only with a high degree of caution. The trace that best illustrates this is a star-shaped pit mark.[24,45] Star-shaped pit marks have been widely reported from the Cenozoic of Africa and have been attributed to termites. Backwell and colleagues[21] sought to test this hypothesis and found that *T. trinervoides* do in fact produce star-shaped pit marks comparable to those reported in the literature.[21] Subsequent to this research, Parkinson[22] experimentally tested the impact of *D. maculatus* on bone under controlled conditions. The results of this later study suggest that dermestids also produce traces comparable to star-shaped pit marks attributed to termites. Thus inferring termites as a causal agent of star-shaped pit marks has lost a degree of credibility because various agents can produce similar morphological traces as a result of commonality in the behaviour attributed to the trace production; this then becomes an issue of equifinality which cannot be addressed directly in the palaeorecord.[22,48,49]

Despite the limitations of comparing gross morphological variables across both geologically and temporally disparate traces, one solution may be to describe traces more comprehensively within an ichnotaxonomic framework.[20,23,35] Such a description would require establishing to what degree the traces identified on bones are distinguishable from other trace taxa described in the literature, and whether the trace is substantially different to motivate the establishment of a new ichnotaxa. For example, on a gross morphological level many traces are described as chambers, but various authors have gone one step further and formally diagnosed ichnotaxa belonging to the ichnogenus *Cubiculum*. *Cubiculum ornatus* was described by Robert and colleagues[44] to include traces in bone which display a chamber-like morphology but which are characterised by the additional presence of gnawing marks/grooves. Subsequently, *Cubiculum levis* was described from Argentina as a chamber which is characterised by a bowl-shaped morphology in cross section.[35] However, in the past 18 months *Cubiculum* has expanded to include a further two taxa: *C. inornatus,* which is similar in morphology and size to *C. ornatus* but lacks the characteristic gnawing marks/grooves[36], and, most recently, *C. cooperi*, which has been described from Cooper's D in the Cradle of Humankind[23]. *C. cooperi* is distinguishable from all other *Cubiculum* taxa because of the absence of gnawing marks/grooves, as well as a uniquely consistent length to width ratio of 2:1.

This brief history of *Cubiculum* illustrates that at a gross morphological level, traces are broadly similar, but in fact the minor morphological variables between traces can be used as a basis for comparison and differentiation. The establishment of ichnotaxa in bone produced by insects also links morphological characteristics to the behavioural tendency of the trace makers. For example, *Cubiculum* ichnospecies are believed to represent the behaviour of producing pupation chambers in bone.[23,35-37,44] Interestingly, the majority of these authors who have described ichnotaxa in bone have avoided attributing a specific causal agent to the traces for various reasons.[23,35-37] These authors recognise that the trace is a reflection of behaviour, and that behaviour could easily be mimicked by numerous agents belonging to a diversity of insect groups. Simply put, more experimental research is required to begin to better understand and document the minor morphological variables between traces in bone produced by a substantially wider diversity of potential agents. Expanding research to address concerns of equifinality would be a fundamental step towards establishing practical criteria to enable causal agent determination/differentiation in the palaeorecord. In the case of the traces identified on StW 431, it is clear that these traces can be attributed to an insect based on gross morphological similarity to traces reported in the literature. High-resolution micro-CT provided clear evidence that the traces were produced post-mortem because there was no bone remodelling. However, because of the limited sample size and morphological and metrological similarity with traces produced by both termites and dermestids, we avoid attributing a specific agent. Lastly, the limited number of traces and lack of distinctive morphology do not warrant the diagnosis of an independent ichnotaxa, nor do the traces on StW 431 bear similarity to the existing ichnotaxa diagnosed in the region during the Plio-Pleistocene, namely *Munitusichnus pascens* and *C. cooperi.*[23] Lastly, drawing behavioural conclusions from a limited sample of traces which are not particularly well preserved would be futile in our opinion.

Conclusion

Our results have demonstrated that – with the exception of the presence of osteopathological lesions in the form of osteophytic lipping, demonstrating minor degenerative joint disease – the pre-mortem bone remodelling described by previous researchers[14] as evidence for infectious disease is shown here to have been caused by post-mortem processes. Our results suggest that the areas described to be as a result of infection are in fact post-mortem modifications made by insects or regions exhibiting degenerative joint disease. We draw these conclusions

based upon new imaging modalities applied to the fossil specimens. This is contrast to the methods used by D'Anastasio and colleagues[19] which comprised observation of surface morphology using light and scanning electron microscopy. In relation to the present study, in addition to macroscopic and microscopic methods, micro-CT was used in order to achieve a much higher resolution than in previous studies with full three-dimensional capacities to model and render previously identified lesions. No previous studies have used micro-CT to evaluate the surface bone of StW 431. Until now, insect traces have not been reported on StW 431.

Macroscopic and microscopic examinations are important prerequisites to determine the difference between post-mortem taphonomic and pre-mortem pathological processes as these two processes can look extremely similar, and are difficult to distinguish. This distinction can mean the difference between an interpretation of ante-mortem pathology or post-mortem pseudopathology. The modifications observed on StW 431 cannot be considered unique, as a similar pattern of bone modification is recorded on a fossil cercopithecoid vertebra from nearby Cooper's D (Figure 6c). Skeletal material at both sites (Cooper's D and Sterkfontein) experienced similar taphonomic sequences of events. Quadrupedal and bipedal primates apparently suffered from similar joint disease conditions[50], and after they died their bones appear to have been modified by an unknown insect.

This study is in agreement with previous pathological diagnoses of spinal DJD which affected StW 431, whilst shedding additional light on the complex series of post-mortem processes affecting the Sterkfontein site and its fossil assemblages. The co-occurrence of degenerative pathological processes *and* subsequent insect modifications of the same regions – the former being an ante-mortem in-vivo response and the latter a post-mortem or post-fossilisation modification – presents an interesting taphonomic case that may not be specific to StW 431, with a cercopithecoid vertebra from Cooper's D showing a similar taphonomic sequence of events. We suggest that it is only with a combination of traditional morphoscopic analyses of external gross morphology, coupled with non-invasive high-resolution tomographic methods (i.e. micro-CT, nano-CT or synchrotron tomography), that diagnoses can be rationalised. This comparative approach has been successfully demonstrated elsewhere in the South African fossil record, with the identification of the earliest evidence for neoplastic disease in the hominin record,[51,52] which would have been impossible using conventional diagnostic methods. Such caution is further supported by Mays[53] who suggests that, in future, secure diagnosis of infective disease processes in bone should primarily be undertaken using biomolecular means. Ancient bacterial DNA may, of course, not survive in such deep time scales, lending greater emphasis to the use of comparative non-invasive methods.[51,52]

Acknowledgements

We acknowledge the assistance and help of the following people in the production of this research: Bonita de Klerk, Wilma Lawrence and Jennifer Randolph-Quinney, and the Department of Geology (Wits). We thank the Evolutionary Studies Institute Fossil Access Advisory Panel for permission to study the StW 431 skeleton and comparative fossil specimens. The research was funded by grants to L.R.B. by the National Geographic Society, the National Research Foundation of South Africa, the South African Centre of Excellence in Palaeosciences, and the Lyda Hill Foundation. Additional direct support for E.J.O. was received from the National Research Foundation of South Africa and the South African Centre of Excellence in Palaeosciences. Direct research support for P.S.R-Q. was provided by the School of Forensic and Applied Sciences, University of Central Lancashire, UK.

Authors' contributions

E.J.O., L.R.B. and P.S.R-Q. wrote the original draft of the manuscript, incorporating additional information and data on Stw 431 from A.H.P., B.Z. and L.R.B., with additional palaeopathological commentary from H.B. K.J. undertook the micro-CT scanning of the hominin specimen and primary reconstruction. E.J.O. and P.S.R-Q. undertook secondary reconstruction from orthoslice and three-dimensional volume data; all authors contributed equally to analysis and editing.

References

1. Toussaint M, Macho GA, Tobias PV, Partridge TC, Hughes AR. The third partial skeleton of a late Pliocene hominin (Stw 431) from Sterkfontein, South Africa. S Afr J Sci. 2003;99:215–223.
2. Clarke RJ. Early Acheulean with *Homo habilis* at Sterkfontein. In: Tobias PV, editor. Hominid evolution: Past, present, and future. New York: Alan R. Liss Inc; 1985. p. 287–298.
3. Clarke RJ. On some new interpretations of Sterkfontein stratigraphy. S Afr J Sci. 1994;90:211–214.
4. Clarke RJ. The hominid species of Sterkfontein through time. Am J Phys Anthropol. 2002;suppl 34:54.
5. Wilkinson MJ. Lower-lying and possibly older fossiliferous deposits at Sterkfontein. In: Tobias PV, editor. Hominid evolution: Past, present, and future. New York: Alan R. Liss Inc; 1985. p. 165–170.
6. Brain CK. Cultural and taphonomic comparisons of hominids from Swartkrans and Sterkfontein. In: Delson E, editor. Ancestors: The hard evidence. New York: Alan R. Liss Inc; 1985. p. 72–75.
7. Stratford D, Grab S, Pickering TR. The stratigraphy and formation history of fossil- and artefact-bearing sediments in the Milner Hall, Sterkfontein Cave, South Africa: New interpretations and implications for palaeoanthropology and archaeology. J Afr Earth Sci. 2014;96:155–167. http://dx.doi.org/10.1016/j.jafrearsci.2014.04.002
8. Granger DE, Gibbon RJ, Kuman K, Clarke RJ, Bruxelles L, Caffee MW. New cosmogenic burial ages for Sterkfontein Member 2 *Australopithecus* and Member 5 Oldowan. Nature. 2015;522(7554):85–88. http://dx.doi.org/10.1038/nature14268
9. Clarke R. On the unrealistic 'revised age estimates' for Sterkfontein. S Afr J Sci. 2002;98(9–10):415–419.
10. Vrba ES. Early hominids in southern Africa: Updated observations on chronological and ecological background. In: Tobias PV, editor. Hominid evolution: Past, present, and future. New York: Alan R. Liss Inc; 1985. p. 195–200.
11. Delson E. Chronology of South African australopith site units. In: Grine FE, editor. Evolutionary history of the robust australopithecines. New York: Aldine de Gruyter; 1988. p. 317–325.
12. McKee JK, Thackeray JF, Berger LR. Faunal assemblage seriation of southern African Pliocene and Pleistocene fossil deposits. Am J Phys Anthropol. 1995;96(3):235–250. http://dx.doi.org/10.1002/ajpa.1330960303
13. Partridge T. Dating of the Sterkfontein hominids: Progress and possibilities. Trans Roy Soc S Afr. 2005;60(2):107–109. http://dx.doi.org/10.1080/00359190509520486
14. Schwarcz HP, Grün R, Tobias PV. ESR dating studies of the australopithecine site of Sterkfontein, South Africa. J Hum Evol. 1994;26(3):175–181. http://dx.doi.org/10.1006/jhev.1994.1010
15. Pickering R, Kramers JD. Re-appraisal of the stratigraphy and determination of new U-Pb dates for the Sterkfontein hominin site, South Africa. J Hum Evol. 2010;59(1):70–86. http://dx.doi.org/10.1016/j.jhevol.2010.03.014
16. McHenry HM, Berger LR. Body proportions in *Australopithecus afarensis* and *A. africanus* and the origin of the genus *Homo*. J Hum Evol. 1998;35:1–22. http://dx.doi.org/10.1006/jhev.1997.0197
17. Tobias PV. 21st Annual report of PARU and its precursors. Johannesburg: Department of Anatomy, University of the Witwatersrand; 1987.
18. Staps D. The first documented occurrence of spondylosis deformans in an early hominin. Am J Phys Anthropol. 2002;suppl 34:146.
19. D'Anastasio M, Zipfel B, Moggi-Cecchi J, Stanyon R, Capasso L. Possible brucellosis in an early hominin skeleton from Sterkfontein, South Africa. PLoS One. 2009;4(7), Art. e6439, 5 pages. http://dx.doi.org/10.1371/journal.pone.0006439
20. Pirrone CA, Buatois LA, Bromley RG. Ichnotaxobases for bioerosion trace fossils in bones. J Paleontol. 2014;88(1):195–203. http://dx.doi.org/10.1666/11-058
21. Backwell LR, Parkinson AH, Roberts EM, d'Errico F, Huchet JB. Criteria for identifying bone modification by termites in the fossil record. Palaeogeogr Palaeoclimatol Palaeoecol. 2012;337–338:72–87. http://dx.doi.org/10.1016/j.palaeo.2012.03.032

22. Parkinson AH. *Dermestes maculatus* and *Periplaneta americana*: Bone modification criteria and establishing their potential as climatic indicators [MSc dissertation]. Johannesburg: University of the Witwatersrand; 2013.

23. Parkinson AH. Traces of insect activity at Cooper's D fossil site (Cradle of Humankind, South Africa). Ichnos. 2016;23(3–4):322–339. http://dx.doi.org/10.1080/10420940.2016.1202685

24. Kaiser TM. Proposed fossil insect modification to fossil mammalian bone from Plio-Pleistocene hominid-bearing deposits of Laetoli (Northern Tanzania). Ann Entomol Soc Am. 2000;93(4):693–700. http://dx.doi.org/10.1603/0013-8746(2000)093[0693:PFIMTF]2.0.CO;2

25. Agarwall SC, Grynpas MD. Measuring and interpreting age-related loss of vertebral bone mineral density in a medieval population. Am J Phys Anthropol. 2009;139(2):244–252. http://dx.doi.org/10.1002/ajpa.20977

26. Roberts C, Cox M. Health and disease in Britain from prehistory to the present day. Stroud: Sutton Publishing; 2003.

27. Roberts CA, Manchester K. The archaeology of disease. 2nd ed. Bradford: Bradford University Press; 1996.

28. Aufderheide AC, Rodríguez-Martin C. The Cambridge encylopedia of human paleopathology. Cambridge: Cambridge University Press; 1998.

29. Ortner DJ, Putschar WGJ. Identification of pathological conditions in human skeletal remains. Smithsonian contributions to anthropology 28. Washington DC: Smithsonian Institution Press; 1981.

30. Resnick D. Orthopedics: Diagnosis of bone and joint disorders: Volume 4. 3rd ed. Philadelphia, PA: W.B. Saunders; 1994.

31. Resnick D. Diagnosis of bone and joint disorders. 3rd ed. Philadelphia, PA: W.B. Saunders Company; 1995.

32. Resnick D, Niwayama G. Diagnosis of bone and joint disorders. 1st ed. Philadelphia, PA: Saunders; 1988.

33. Rogers J. The palaeopathology of joint disease. In: Cox M, Mays S, editors. Human osteology in archaeology and forensic science. Cambridge: Cambridge University Press; 2000. p. 163–182.

34. Britt BB, Scheetz RD, Dangerfield A. A suite of dermestid beetle traces on dinosaur bone from the Upper Jurassic Morrison Formation, Wyoming, USA. Ichnos. 2008;15(2):59–71. http://dx.doi.org/10.1080/10420940701193284

35. Pirrone CA, Buatois LA, González Riga B. A new ichnospecies of *Cubiculum* from Upper Cretaceous dinosaur bones in Western Argentina. Ichnos. 2014;21(4):251–260. http://dx.doi.org/10.1080/10420940.2014.958225

36. Xing L, Parkinson AH, Ran H, Pirrone CA, Roberts EM, Zhang J, et al. The earliest fossil evidence of bone boring by terrestrial invertebrates, examples from China and South Africa. Hist Biol. 2015:1–10. http://dx.doi.org/10.1080/08912963.2015.1111884

37. Neto VDP, Parkinson AH, Pretto FA, Soares MB, Schwanke C, Schultz CL, et al. Oldest evidence of osteophagic behavior by insects from the Triassic of Brazil. Palaeogeogr Palaeoclimatol Palaeoecol. 2016;453:30–41. http://dx.doi.org/10.1016/j.palaeo.2016.03.026

38. Newman R. The incidence of damage marks on Swartkrans fossil bones from the 1979–1986 excavations. Swartkrans: A cave's chronicle of early man: Transvaal Museum Monograph. Pretoria: Transvaal Museum; 1993. p. 217–228.

39. Holden AR, Harris JM, Timm RM. Paleoecological and taphonomic implications of insect-damaged pleistocene vertebrate remains from Rancho La Brea, southern California. PLoS One. 2013;8(7):e67119. http://dx.doi.org/10.1371/journal.pone.0067119

40. Kitching J. On some fossil arthropoda from the limeworks Makapansgat, Potgietersrus. Palaeontol Afr. 1980;23(6):63–68.

41. Bader KS, Hasiotis ST, Martin LD. Application of forensic science techniques to trace fossils on dinosaur bones from a quarry in the Upper Jurassic Morrison Formation, Northeastern Wyoming. PALAIOS. 2009;24(3):140–158. http://dx.doi.org/10.2110/palo.2008.p08-058r

42. Dangerfield A, Britt B, Scheetz R, Pickard M. Jurassic dinosaurs and insects: The paleoecological role of termites as carrion feeders. Paper presented at: Salt Lake City Annual Meeting; 2005 October 16–19; Salt Lake City, UT, USA.

43. Paik IS. Bone chip-filled burrows associated with bored dinosaur bone in floodplain paleosols of the Cretaceous Hasandong Formation, Korea. Palaeogeogr Palaeoclimatol Palaeoecol. 2000;157(3):213–225. http://dx.doi.org/10.1016/S0031-0182(99)00166-2

44. Roberts EM, Rogers RR, Foreman BZ. Continental insect borings in dinosaur bone: Examples from the late Cretaceous of Madagascar and Utah. J Paleo. 2007;81(1):201–208. http://dx.doi.org/10.1666/0022-3360(2007)81[201:CIBIDB]2.0.CO;2

45. Fejfar O, Kaiser TM. Insect bone-modifications and palaeoecology of Oligocene mammal-bearing sites in the Doupov Mountains, Northwestern Bohemia. Palaeontol Electron. 2005;8(1), 11 pages. Available from: http://palaeo-electronica.org/2005_1/fejfar8/fejfar8.pdf

46. Hill A, Leakey M, Harris J. Damage to some fossil bones from Laetoli. Laetoli: A Pliocene site in Northern Tanzania. Oxford: Clarendon Press; 1987. p. 543–545.

47. Pomi LH, Tonni EP. Termite traces on bones from the Late Pleistocene of Argentina. Ichnos. 2011;18(3):166–171. http://dx.doi.org/10.1080/10420940.2011.601374

48. Lyman RL. The concept of equifinality in taphonomy. J Taphonomy. 2004;2(1):15–26.

49. Bristow J, Simms Z, Randolph-Quinney PS. Taphonomy. In: Ferguson E, editor. Forensic anthropology 2000–2010. Boca Raton, FL: CRC Press; 2011. p. 279–318. http://dx.doi.org/10.1201/b10727-10

50. Jurmain R. Degenerative joint disease in African great apes: An evolutionary perspective. J Hum Evol. 2000;39(2):185–203. http://dx.doi.org/10.1006/jhev.2000.0413

51. Odes EJ, Randolph-Quinney PS, Steyn M, Throckmorton Z, Smilg JS, Zipfel B, et al. Earliest hominin cancer: 1.7-million-year-old osteosarcoma from Swartkrans Cave, South Africa. S Afr J Sci. 2016;112(7–8), Art. #2015-0471, 5 pages. http://dx.doi.org/10.17159/sajs.2016/20150471

52. Randolph-Quinney PS, Williams SA, Steyn M, Meyer MR, Smilg JS, Churchill SE, et al. Osteogenic tumour in *Australopithecus sediba*: Earliest hominin evidence for neoplastic disease. S Afr J Sci. 2016;112(7–8), Art. #2015-0470, 7 pages. http://dx.doi.org/10.17159/sajs.2016/20150470

53. Mays SA. Lysis at the anterior vertebral body margin: Evidence for brucellar spondylitis? Int J Osteoarchaeol. 2007;17:107–118. http://dx.doi.org/10.1002/oa.903

Attenuation of pollution arising from acid mine drainage by a natural wetland on the Witwatersrand

AUTHORS:
Marc S. Humphries[1]
Terrence S. McCarthy[2]
Letitia Pillay[1]

AFFILIATIONS:
[1]Molecular Sciences Institute, School of Chemistry, University of the Witwatersrand, Johannesburg, South Africa
[2]School of Geosciences, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Marc Humphries

EMAIL:
marchump@gmail.com

DATES:
Received: 06 Aug. 2016
Revised: 06 Oct. 2016
Accepted: 08 Oct. 2016

KEYWORDS:
metal sequestration; Klip River wetland; water quality; remediation; gold mining

HOW TO CITE:
Humphries MS, McCarthy TS, Pillay L. Attenuation of pollution arising from acid mine drainage by a natural wetland on the Witwatersrand. S Afr J Sci. 2017;113(1/2), Art. #2016-0237, 9 pages. http://dx.doi.org/10.17159/sajs.2017/20160237

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
National Research Foundation (South Africa); University of the Witwatersrand

Wetlands are well known to be efficient at sequestering pollutants from contaminated water. We investigated metal accumulation in the peats of the Klip River, a natural wetland that has received contaminated water from gold mining operations in Johannesburg for over 130 years. Previous work conducted in the downstream portion identified the wetland as an important system for sequestering metals. We focused on the upstream section of the wetland, more proximal to the source of acid mine drainage, to provide a better understanding of the pollutant sources and the role of the wetland in pollutant attenuation. Geochemical and mineralogical analyses of peat cores revealed considerable metal enrichments in the peat ash, particularly in Co, Ni, Zn, Pb, Cu and U. Metal concentrations are typically between 4 to 8 times higher than those previously reported for the downstream, more distal portion of the wetland. The distribution of metal accumulation within the peat profiles suggests that contamination arises from a combination of sources and processes. Elevated concentrations in the shallow peat are attributed to the input of contaminated surface water via tributaries that drain the Central Rand Goldfield, whereas enrichments in the deeper peat suggest significant sub-surface inflow of contaminated water through the underlying dolomitic rocks. Metal immobilisation occurs through a combination of mechanisms, which include the precipitation of gypsum, metal sulfides, Fe-Mn oxyhydroxides and phosphates. Our study highlights the environmental and economic importance of natural wetland systems which have the ability to accumulate large quantities of metals and thus remediate polluted waters.

Significance:

- Considerable levels of metal accumulation are observed within the Klip River wetland peats.
- The wetland is effective in remediating highly polluted water emanating from the Witwatersrand Basin.
- The Klip River system is important for the region's future water supply.

Introduction

South Africa is renowned for the diversity and richness of its mineral wealth, the exploitation of which has sustained the economic growth of the country since the mid-1800s. However, this minerals-driven growth has been associated with a number of undesirable environmental impacts. For example, the severe health and consequent social problems arising from mine workers' exposure to asbestos-bearing dust have been well documented.[1] Similar problems were experienced in the gold mining industry as a result of exposure to silica-bearing dust, particularly in the early years of mining.[2] But perhaps the most severe and insidious adverse legacy of the mining industry is acid mine drainage (AMD), especially that arising from coal and gold mining.[3] This problem has resulted in a decline in water quality in the economic heartland of the country and could result in water shortages in the future.

AMD arises from the oxidation of pyrite which is a common gangue mineral in a variety of ore types, and is fairly well understood. AMD arising from gold and base metal mining is invariably enriched in a wide variety of metals, many of which are highly toxic. The gold ores of the Witwatersrand Basin (Figure 1, inset) are no exception and contain elevated concentrations of lead (Pb), copper (Cu), nickel (Ni), cobalt (Co), zinc (Zn) and uranium (U), amongst others. The low pH associated with AMD enhances the mobilisation of these metals, resulting in their dispersion into the ground and surface water. The major source of AMD on Witwatersrand gold mines has been seepage from tailings storage facilities (TSFs)[4-8], although discharge of water from flooded, abandoned gold mines as well as discharge of partially treated water pumped from producing mines has also contributed to the problem in recent years[9-11]. Adverse effects on water quality and the health of natural ecosystems is of growing concern[3,12], with AMD recognised as one of the most significant environmental challenges facing conservation managers and the mining industry in South Africa[13].

Numerous methods for treating AMD have been developed. The most desirable of these are passive treatment methods, two of which – wetlands and carbonate rock drains – are fairly widely used to remediate AMD. Carbonate rock drains serve to raise the pH and thus limit metal solubility in water affected by AMD. Metals continue to be transported in these drains but as insoluble, suspended particulates. However, carbonate rock drains have little influence on the sulfate content of the water. In contrast, wetlands reduce the impact of AMD by creating reducing conditions which cause reduction of sulfate ions to sulfide. The presence of sulfide results in the precipitation of many metals as sulfide minerals. The sulfate content in water is also reduced, partly because of metal sulfide precipitation but also as a result of the release of sulfur in the form of hydrogen sulfide gas. Removal may also occur through the formation of insoluble metal oxides in aerobic zones of the wetland, metal complexation with organic matter, sorption onto particles, and incorporation into the plant biomass.[14] As such, wetlands are particularly efficient in sequestering metals from inflowing waters and highly effective in remediating polluted water. The effectiveness of wetlands in removing metal contaminants from water has resulted in the use of constructed wetlands as low-cost biogeochemical systems for the treatment of AMD.[15,16]

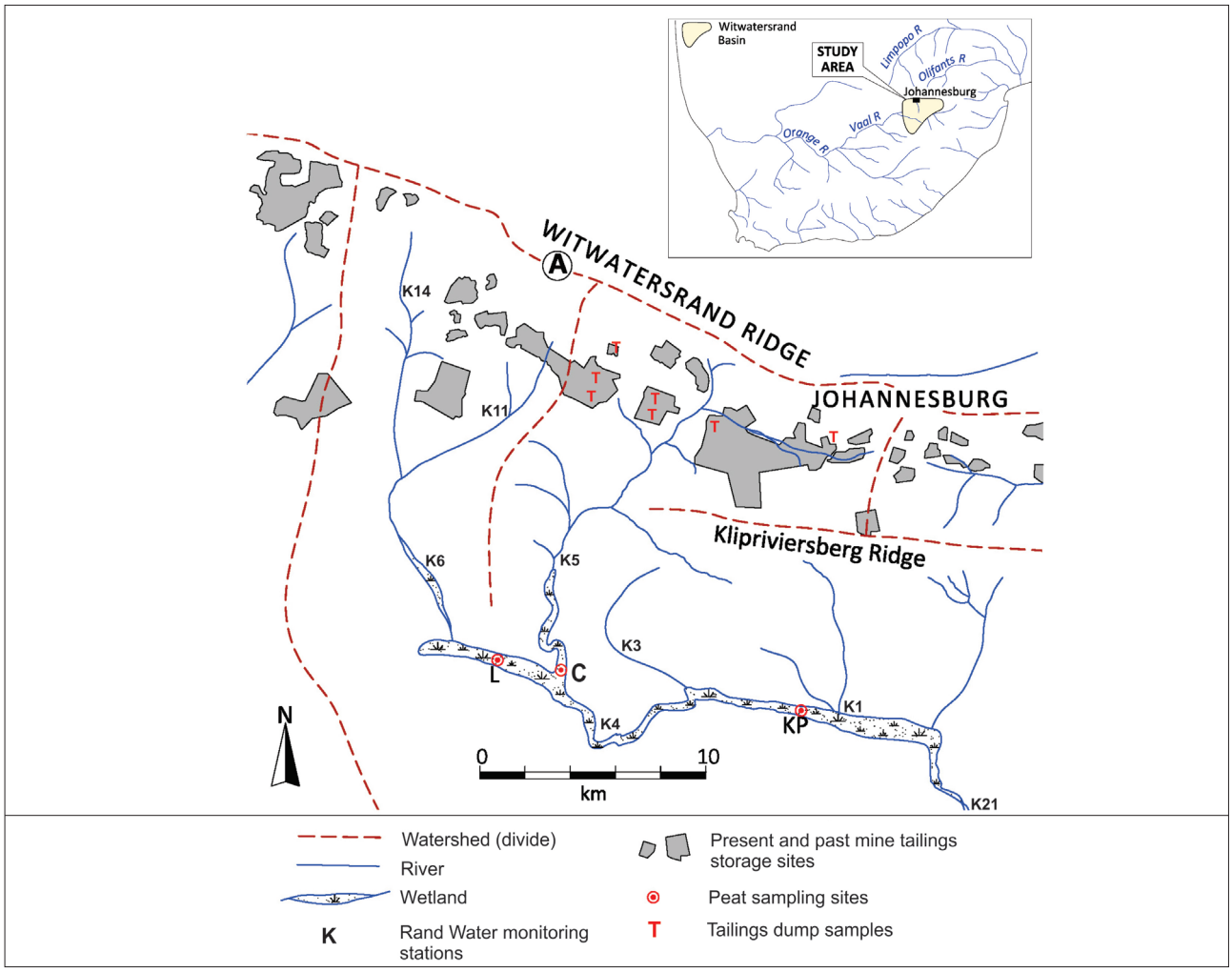


Figure 1: Map of the study area and location of the sampling sites in relation to mining activities in the Johannesburg area.

Groundwater and run-off from the Central Rand Goldfield discharges into extensive wetlands that have developed along the Klip River, the principal drainage of the southern portion of the Johannesburg conurbation. A previous study conducted in the distal, downstream portion of the wetland (Figure 1, Site KP) revealed that peat in the wetland contained elevated concentrations of certain metals as a result of sequestration from polluted water.[17] Contamination was typically confined to the uppermost section (top ~1 m) of the peat, where elevated levels of Co (460 ppm), Ni (1500 ppm), Cu (170 ppm), Pb (60 ppm) and Zn (2300 ppm) were observed in ashed material. These levels represent 15–30 times the natural background and likely arose from a combination of sources including mining, industrial effluent and sewage discharge. The previous study highlighted the importance of this wetland system in sequestering contaminants that would otherwise enter the Vaal River system and cause widespread pollution. Given the valuable function this wetland performs, a more detailed understanding of the pollutant sources, processes leading to pollutant attenuation, and the potential downstream impacts through remobilisation of toxicants is needed. The present study was undertaken in the upstream section of the wetland, more proximal to the source of acid mine drainage, in order to investigate metal sequestration in more detail and to define the sources of metals more clearly.

Study area

The study area is located southwest of Johannesburg in the Klip River valley (Figure 1).

Geology

The gold mines in the Johannesburg area (Central Rand Goldfield) commenced operations in 1886. Over the life of the goldfield, some 7700 tonnes of gold (~246 million oz.) have been extracted from the treatment of approximately 940 million tonnes of rock. The gold was contained in quartz pebble conglomerates, which are hosted in a sequence of siliceous quartzites of the Witwatersrand Supergroup. The deposits are now largely worked out and most mining has ceased. Reworking of TSFs to extract remaining gold as well as small-scale surface and underground operations continue.

The bedrock geology of the region is summarised in Figure 2a. The gold-bearing strata of the Precambrian Witwatersrand Supergroup form an east-west striking belt across the centre of the region and lie unconformably on granitoid basement rocks. Strata dip to the south at between 15° and 30°. Gold-bearing conglomerates are confined to the upper portion of the stratigraphy, which consists largely of quartzites and conglomerates. The bulk of the gold was produced from the Main Reef group of conglomerates. The TSFs are located in close proximity to the outcrop of these conglomerates. The Witwatersrand strata are conformably overlain by basaltic volcanic rocks of the Ventersdorp Supergroup. Both the Witwatersrand and Ventersdorp rocks are in turn unconformably overlain by dolomitic rocks of the Transvaal Supergroup which dip at about 10° to the south. It is these rocks that underlie and host the Klip River wetlands.

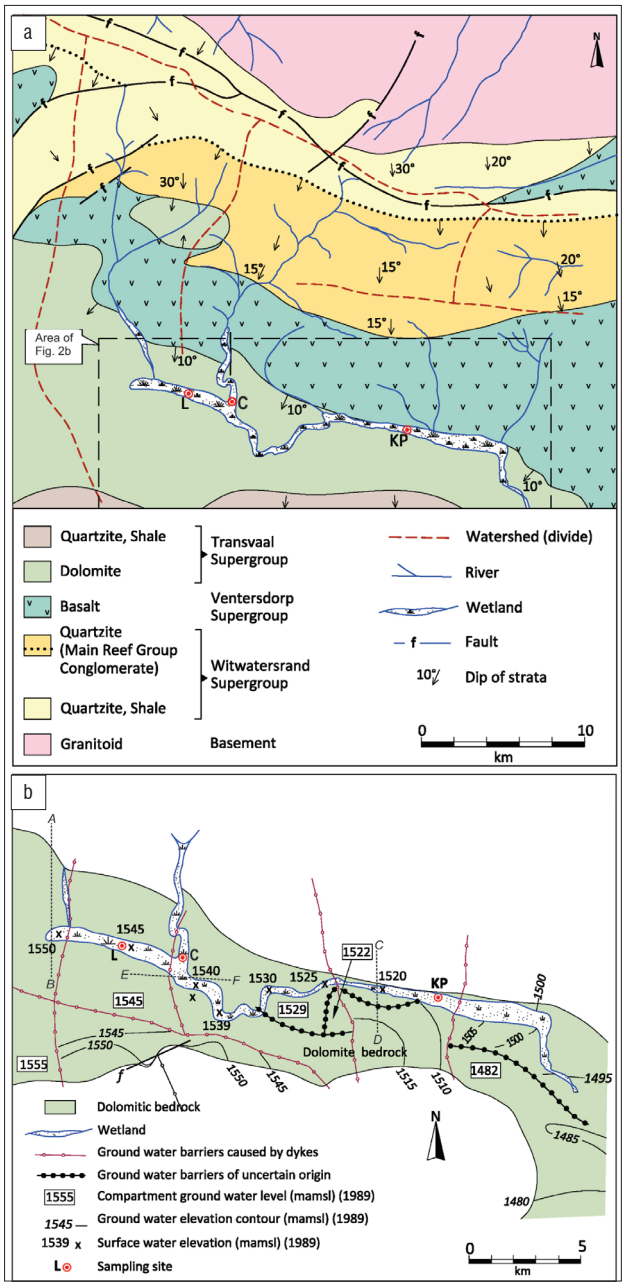


Figure 2: Maps showing (a) the simplified geology in the study area, (b) groundwater elevation and major barriers to groundwater movement in the upper Klip River Basin.[19]

Geomorphology

The primary watershed in the area (A in Figure 1) forms part of the continental divide separating the Vaal and Limpopo drainage basins. The majority of TSFs in the area lie south of this watershed in the Vaal River catchment. Minor watersheds compartmentalise run-off from the dump region (Figure 1), with the run-off arising to the west of Johannesburg entering the upper Klip River reach where study Sites L and C are situated. Tributaries discharging into the wetland further downstream, where the previously studied Kibler Park site is situated, are isolated from the dump area by the watershed associated with the low hills of the Klipriversberg Ridge.

Geohydrology

The dolomite of the Transvaal Supergroup is karstic but is typically divided up into numerous separate compartments by igneous intrusions (dykes), which restrict sub-surface flow from one compartment to the next.[18] Within each compartment, much of the surface water infiltrates into the sub-surface to fill solution cavities in the dolomite. Water typically emerges at springs located on dykes that form the downstream boundaries of compartments and cascades over these into the next lower compartment. These spring sites are host to wetlands. Sinkholes and dolines along water courses across the dolomitic areas similarly host extensive, interlinked wetland systems.

A detailed geohydrological study of the dolomitic rocks in the region of the study revealed that the dolomitic aquifer is divided into several compartments by dykes with slightly different groundwater levels in each compartment[19] (Figure 2b). According to early reports[20], springs abounded along this section of the river flanked by extensive wetlands with dense stands of *Phragmites* reeds growing atop several metres of peat. The age of the deeper peat in the wetland is in excess of 3500 years.[17] Kafri and Foster[19] found that the flow of water from one compartment to the next was largely by surface discharge across bounding barriers and there was evidently no sub-surface connection. The most important dykes that compartmentalise the dolomite in the region belong to the north-northeast-striking Pilanesberg Dyke Swarm[18], which divide the dolomite aquifer into north-south trending compartments. Compartments in the upper catchment are likely subject to ingress of polluted water from the TSFs and surface streams (Figure 3a), whereas groundwater further downstream is uncontaminated (Figure 3b) as this compartment contains no TSFs. The only source of pollution to the downstream portion of the wetland system is via overspill from the upstream compartments and run-off of polluted surface water (Figure 3c).

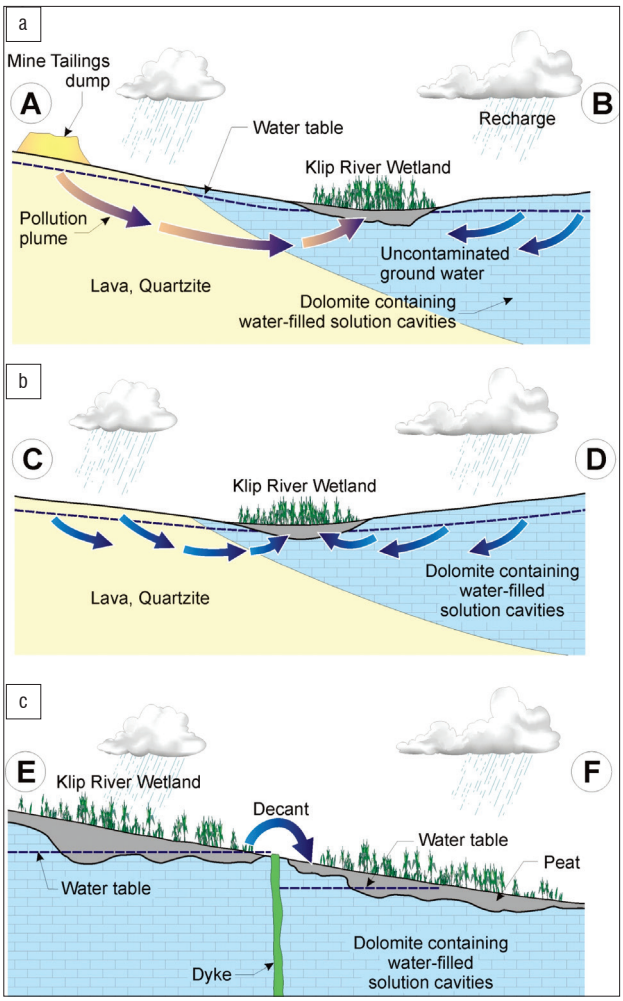


Figure 3: Schematic diagrams showing (a) the ingress of polluted groundwater into the karst-hosted Klip River wetland, (b) the recharge of groundwater in the karst-hosted Klip River wetland and (c) the flow of water from one dolomitic compartment to another across a bounding dyke. The down-channel gradient has been greatly exaggerated in this sketch. See Figure 2b for approximate profile locations.

The wetlands are sustained predominantly by groundwater, which discharges into the peat from the underlying dolomite (Figure 2a,b). In summer, this discharge is augmented by surface run-off from rainfall, with the Klip River and Klipspruit being the dominant fluvial inputs. The Klip River also receives run-off from a number of informal settlements that flank the banks of the Klip River and Klipspruit, as well as ~500 ML/day in discharge from three sewerage treatment plants, much of which enters the river a short distance downstream of Site C (Figure 1).

Little is known about the hydrochemistry of the Klip River watershed. The Rand Water Board has a number of monitoring points within the Klip River catchment (Figure 1), although data from these sites are sparse. Surface water analyses indicate that major ion concentrations are typically low, but highly variable, with pH varying between 6.0 and 8.1 (Table 1). The low metal concentrations at monitoring sites indicate that surface run-off undergoes rapid dilution upon entering the watercourse. In contrast, Naicker et al.[6] and Tutu et al.[8] found groundwater within the Central Rand mining district to be acidified (pH <4) and contaminated by high sulfate, Fe, Zn, Co and Ni concentrations. The percolation of rain water through mine dumps, creating polluted groundwater plumes, was considered to be the main contribution to contamination.

Table 1: Surface water chemistry data for the Klip River catchment for the period 2010 to 2015 (n=52)

	pH	SO_4 (mg/L)	Mg (mg/L)	Fe (mg/L)	Mn (mg/L)
Tributaries					
K14	6.6 ± 0.4	32 ± 22	3.9 ± 1.5	0.47 ± 1.0	0.26 ± 0.4
K11	6.5 ± 0.7	97 ± 40	9.6 ± 1.6	0.11 ± 0.1	0.33 ± 0.3
K5	7.5 ± 0.3	164 ± 72	15.1 ± 2.5	0.31 ± 0.4	1.9 ± 0.8
K3	7.7 ± 0.2	57 ± 8	9.5 ± 1.6	0.06 ± 0.06	0.07 ± 0.06
Klip River					
K6	7.8 ± 0.3	110 ± 20	18.6 ± 2.3	0.08 ± 0.09	0.74 ± 0.66
K4	7.8 ± 0.3	104 ± 30	13.3 ± 1.4	0.08 ± 0.1	0.83 ± 0.6
K1	7.7 ± 0.2	96 ± 22	13.4 ± 1.7	0.15 ± 0.23	1.2 ± 2.5
K21	7.9 ± 0.2	97 ± 17	13.2 ± 1.3	0.10 ± 0.08	0.36 ± 0.18

Source: Rand Water Board

Gold extraction and the formation of pollution sources

The cyanidation process used to process the Witwatersrand ore selectively removed gold and silver. Pyrite and other sulfide minerals together with the gangue minerals (primarily quartz with minor phyllosilicate minerals) reported to the TSFs. During cyanidation, the pH of the pulverised rock was kept above 9 to prevent the formation of hydrogen cyanide. However, the tailings material has a very low buffering capacity because of its high quartz content, and after disposal, oxidation of pyrite in the TSFs leads to the rapid acidification of the material. The elevated water table within the dumps[7] caused by infiltration of rain water has led to the formation of acid plumes which currently discharge from the dumps. The contaminated groundwater locally discharges into streams draining the area around the dumps.[6,8]

Methods

Sampling

Two sediment cores (L1 and L2), ~100 m apart, were collected from the upstream, western section of the Klip River wetland near the town of Lenasia (Site L, Figure 1) and one from further downstream (Site C, Figure 1) . The wetland here is approximately 500 m wide and consists of a reed-covered swamp dominated by *Phragmites australis.* These sites were selected based on examination of historical aerial photography and maps, which indicated that this region of the wetland had remained relatively undisturbed for the last ~80 years[21,22] and thus probably best resembles the hydrological functioning of the system prior to the development of the Witwatersrand mining and industrial complex. Samples were collected using a Russian peat corer, coring until the underlying clay layer was intercepted. Sub-samples for analysis were taken at 20-cm intervals. The pH of the peat was measured by inserting a pH electrode directly into collected samples. Material from several mine TSFs in the upper catchment was also sampled (Figure 1).

Chemical analyses

Sediment samples were dried at 110 °C and milled into homogeneous powders. Organic content was measured by loss on ignition at 450 °C. Ash residues were analysed for major elements by X-ray fluorescence (XRF) using fused glass beads and a PANalytical PW1400. Calibration was performed using a range of international (NIST) and national (SARM) rock standards. Trace metal and rare earth element (REE) analyses were carried out using laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). The analyses were performed using an Agilent 7700 ICP-MS connected to a Resolution M-50-LR Excimer (Resonetics) laser ablation system (193 nm). Helium was used as a carrier gas with a flow of 0.3 L/min. Analyses were calibrated using a certified BCR glass reference material as an external standard. SiO_2 concentrations (from XRF analyses) were used for internal standardisation. Accuracy, as the relative difference from reference values, was typically better than 10%, and most elements plotted in the range ±5%. Total carbon and sulfur were measured on bulk peat samples using a Euro EA3000 elemental analyser. Material collected from TSFs was analysed for major and minor elements using XRF.

Sequential extractions

To examine the partitioning of metals among various phases, sequential extractions on bulk sediment samples from Core L1 were performed following the modified BCR procedure.[23] Samples were sequentially extracted in four stages to identify metal associated with different fractions, namely acid soluble (exchangeable and carbonates), reducible (Fe-manganese (Mn) oxyhydroxides), oxidisable (organic matter/sulfides) and residual (silicates). The extracts from each step were analysed for selected major and trace elements using an Agilent 7700 ICP-MS in gas and no gas modes where appropriate.

Mineralogy

Mineralogical investigations were performed using X-ray diffraction on un-orientated powder samples using a Bruker D2 Phaser with monochromated CoKα radiation (10–90° 2θ). Scanning electron microscope energy dispersive X-ray (SEM-EDX) analyses were performed on selected minerals within the peat using a FEI Nova 600 microscope. Samples were dispersed in water, mounted onto aluminium (Al) stubs and coated with 10 μm gold-palladium. Peat samples were selected based on bulk geochemical analyses which showed high metal concentrations.

Results and discussion

Core profile characteristics and major chemistry

The proportion of organic matter in the samples collected from the Klip River wetland is highly variable. Cores from Site L reveal ~3 m of peat (defined here as organic-rich sediment) underlain by cohesive grey clayey sand. The peat is very fibrous and relatively low in inorganic components, with carbon contents varying between 24% and 48% (Figure 4). Total sulfur concentrations were variable, ranging between 2% and 5.5%, while pH displayed little variation with depth, averaging 6.9. Peat accumulation at Site C is less well developed and generally confined to the upper 40 cm of the profile (Figure 4). The peat is characterised by elevated sulfur concentrations (up to 5%), with pH again showing little variation with depth.

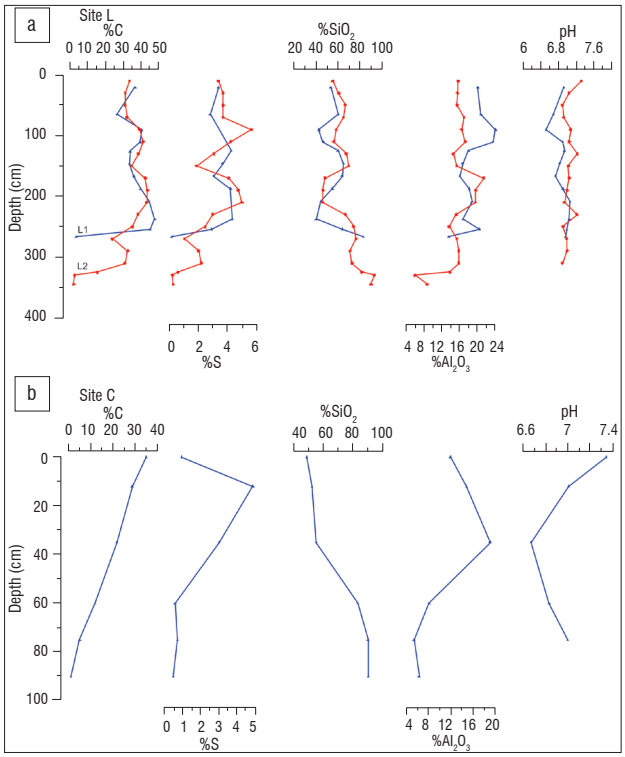


Figure 4: Variation in pH and bulk chemical composition of peat with depth at (a) Site L and (b) Site C.

The inorganic fraction of the peat is largely dominated by SiO_2 and Al_2O_3, which together comprise between 55% and 80% of the ash fraction (Figure 4). The inorganic component is derived predominantly from clastic material in run-off and airborne dust.[17] Plots of Al_2O_3-SiO_2, Fe_2O_3-Al_2O_3 and CaO-Al_2O_3 abundances in the clastic material underlying the peat at Sites L, C and KP and in mine tailings show linear trends reflecting varying proportions of quartz and an iron (Fe)-bearing aluminous clay mineral end member (Figure 5). The peat samples all show departures from the mixing lines as a result of the presence of Fe- and calcium (Ca)-bearing minerals, which are not part of the clastic load. Departure from the mixing line is most clearly observed in samples from Site L, indicating extensive in-situ precipitation and sequestration of chemical components within the peat.

Metal profiles

In order to compensate for the varying proportion of inorganic material contained within the peat and allow for comparison between sediment profiles, metal concentrations were normalised by expressing data relative to $SiO_2+Al_2O_3$, which are primarily of clastic origin. The clay underlying the wetland is characterised by very low metal ratios (Table 2) and provides a baseline level for pre-mining, uncontaminated sediment. In contrast, normalised metal values within the peat from Sites L and C reveal strikingly different downcore variations (Figure 6a,b). At sampling Site L, metals show highly elevated ratios, although enrichments occur in different sections of the peat profiles. Normalised concentration profiles for CaO, Fe_2O_3, Co and Ni show similar downcore trends, with highest enrichment found near the base of the peat sequence (200–250 cm) where Co and Ni in particular are enriched up to 700 and 200 times, respectively, relative to the underlying clay. In contrast, Pb and Cu show pronounced increases in concentration, of about four-fold, over the uppermost metre of the peat profile. Zn values are variable through the profile, although higher enrichments are generally observed in the deeper section of the profiles, particularly between 70 cm and 200 cm, where samples are enriched up to 900 times relative to the underlying clay.

Normalised metal profiles at Site C show similar trends for all metals investigated, revealing increasing enrichment toward the top of the profile (Figure 6b). Co, Ni and Zn are enriched 30–50 times in the surface peat relative to the underlying material. Surface enrichment in metals was also observed at the previously studied downstream Site KP. Here metals within the surface peat (upper 1 m) were enriched 2–10 times relative to the underlying peat material (Figure 6c).

Spatial differences in chemical accumulation

Average metal ratios at Site L are significantly higher when compared to those from Sites C and KP (Table 2). Site L is located ~15 km upstream of Site KP and is in closer proximity to potential surface and groundwater pollution sources (Figures 1 and 2). Higher pollution levels at Site L thus reflect proximity to source and the efficacy of the wetland system in removing metals from solution. Between sampling Sites L and KP, Co, Ni, Cu and Zn ratios decrease by between 75% and 80%, while average Zn and U ratios decrease by ~90% and 96%, respectively.

At Sites KP and C, contamination is confined to the near surface peats, whereas peat from deeper within the profiles (>150 cm) is unpolluted (Figure 6). This result is in contrast to both surface and sub-surface metal enrichments observed at Site L. The strong enrichment in metals in the deepest peat here strongly suggests inflow of contaminated groundwater into the wetland from below. The formation of gypsum crusts observed along the margins of the wetland near Site L indicates that groundwater seeping into the wetland is highly concentrated in Ca and SO_4. Similar groundwater seepage zones associated with surface gypsum precipitate formation were observed along reaches of the Natalspruit drainage network.[6]

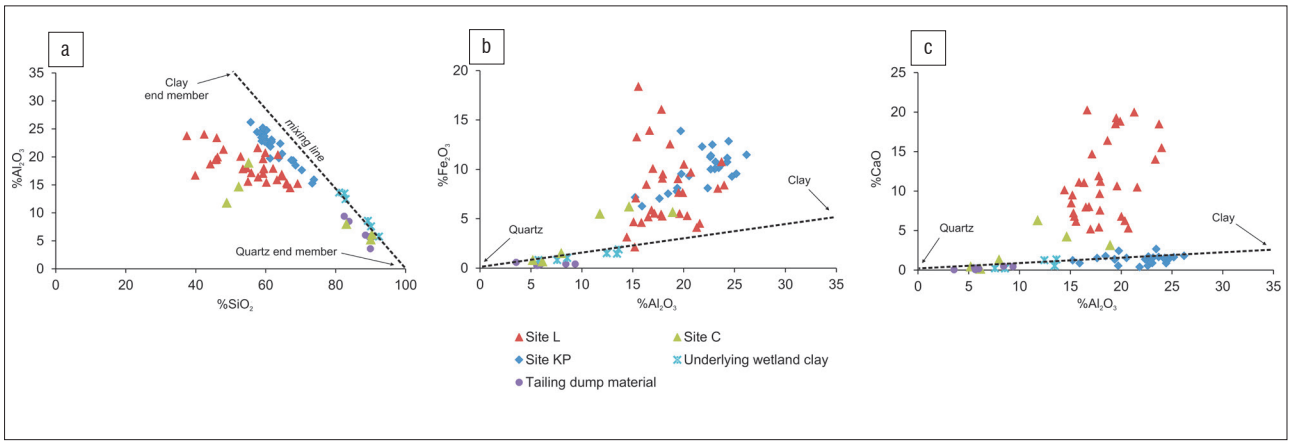


Figure 5: Relationship between (a) SiO_2 and Al_2O_3, (b) Al_2O_3 and Fe_2O_3 and (c) Al_2O_3 and CaO in the underlying wetland clay, tailings storage facilities material, and peat ash at sampling Sites L, C and KP.

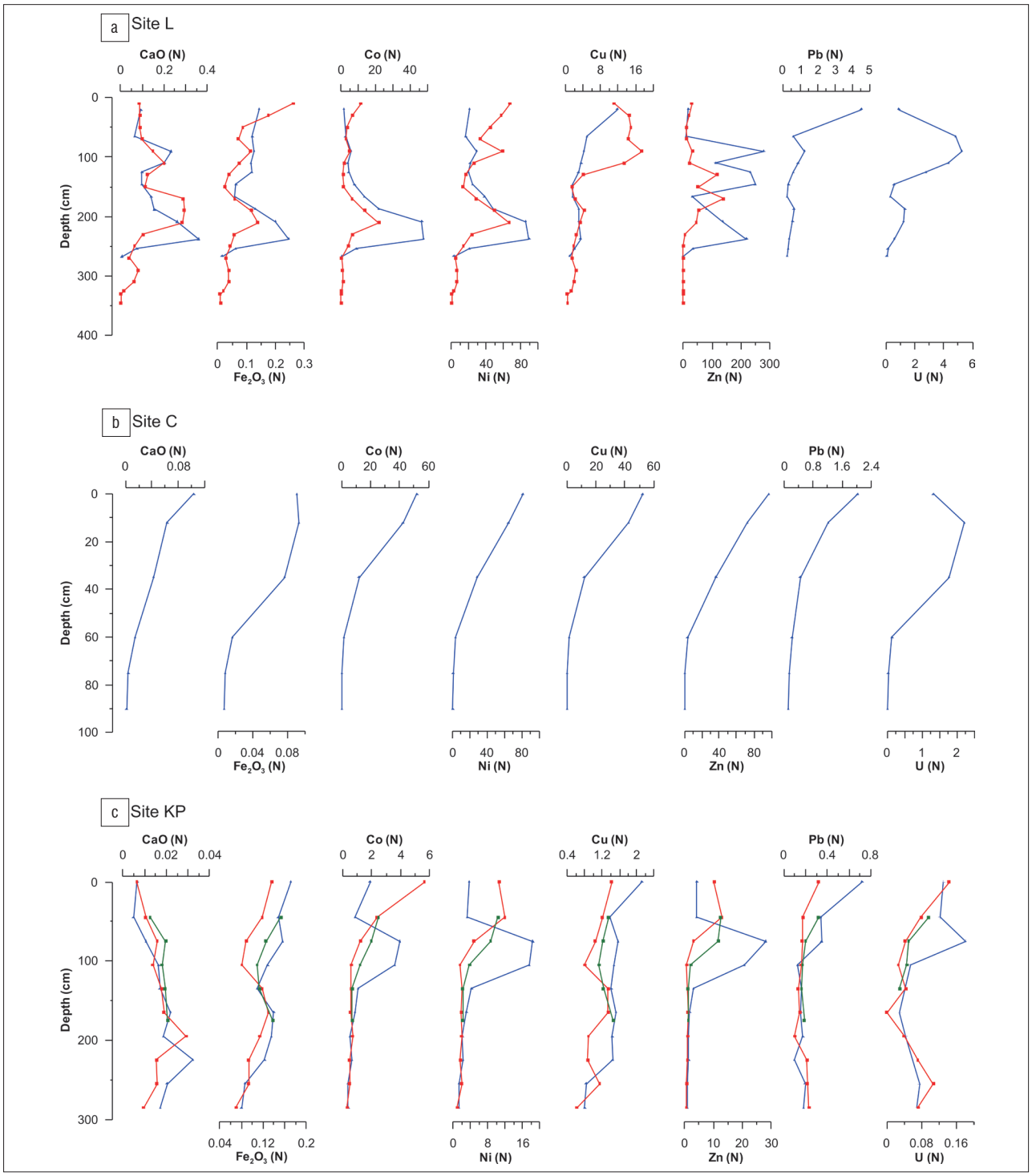


Figure 6: Normalised metal concentration profiles of peat samples from (a) Site L, (b) Site C and (c) Site KB.

Table 2: Average normalised metal ratios within tailings dump material and the inorganic fraction of peat deposits at Sites L, C (this study) and KP.[17] Normalised metal ratios in the clastic layer underlying the peat are given in parentheses.

	Fe_2O_3 / (SiO_2+Al_2O_3)	CaO / (SiO_2+Al_2O_3)	Co / (SiO_2+Al_2O_3)	Ni / (SiO_2+Al_2O_3)	Cu / (SiO_2+Al_2O_3)	Zn / (SiO_2+Al_2O_3)	Pb / (SiO_2+Al_2O_3)	U / (SiO_2+Al_2O_3)
Site L	0.11 (0.01)	0.16 (0.002)	10.8 (0.06)	38.9 (0.40)	6.3 (0.33)	84.7 (0.20)	0.9 (0.24)	2.0 (0.06)
Site C	0.05 (0.01)	0.04 (0.003)	18.0 (0.12)	29.6 (0.57)	2.5 (0.20)	35.0 (0.44)	0.7 (0.13)	0.9 (0.02)
Site KP	0.12 (0.07)	0.02 (0.01)	1.45 (0.30)	4.9 (0.94)	1.2 (0.62)	5.1 (0.63)	0.2 (0.23)	0.1 (0.07)
Tailings dump material	0.04	0.002	0.34	0.78	0.47	0.51	0.35	0.25

Table 3: Water chemistry from sampling Site L and typical chemical composition of mine void water (50th percentile) from the Central Witwatersrand Basin[24]

	pH	Electrical conductivity (mS/m)	Mg (mg/L)	Al (mg/L)	Ca (mg/L)	Fe (mg/L)	Mn (mg/L)	SO_4 (mg/L)	Co (ug/L)	Ni (ug/L)	Cu (ug/L)	Zn (ug/L)	Pb (ug/L)	U (ug/L)
Central Basin mine void water (n=12)	3.0	397	172	122	279	40	47	2831	4684	10 589	328	9122	28	606
Surface water from Site L (n=1)	6.9	62	23	0.04	44	0.01	0.2	–	3.8	5.2	2.9	7.7	0.06	–

Metal distributions within the peat at Site L appear to reflect different sources of water to the wetland, with some trace elements showing increasing abundance in the uppermost metre (e.g. Cu and Pb), whereas Co, Ni and Zn are highly enriched in the deeper peat. AMD-contaminated water from the Central Witwatersrand Basin is highly acidic (pH 3) and remarkably enriched in Co, Ni and Zn (Table 3[24]). The Central Rand is dominated by the numerous mine TSFs which form large footprint plumes within the Klip River catchment.[25] Groundwater entering the wetland thus likely carries high metal loads, which precipitate within the peat under higher pH conditions.

Fe, S and Ca also show some relative enrichment in the deeper peat at Site L, with total Ca concentrations 4–10 times higher when compared to Sites C and KP. The peat at Site L is likewise enriched in REEs. Post-Archaean Australian Shale (PAAS)-normalised REE patterns (Figure 7) indicate that MREEs in particular are enriched relative to both the light (LREEs) and heavy REEs (HREEs). The observed pattern is typical of AMD-affected water and sediments and likely indicates that REEs are fractionated during pyrite oxidation, as has been observed in other studies.[26,27]

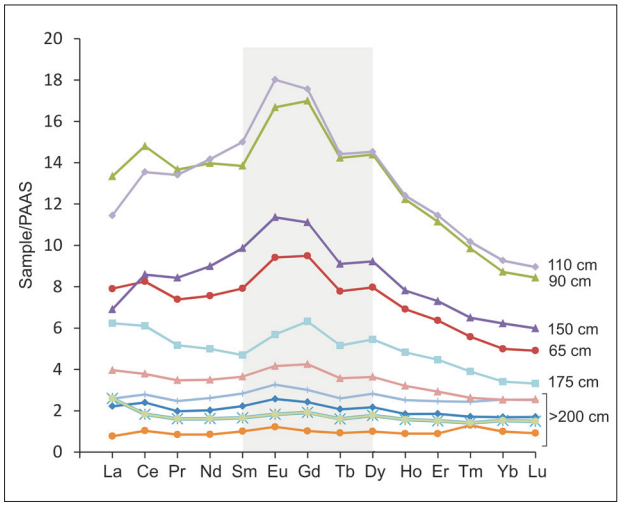


Figure 7: Post-Archaean Australian Shale (PAAS)-normalised rare earth element (REE) patterns of peat samples from Site L. Note that samples from 65 cm to 175 cm show significant medium REE enrichment.

It is clear that the upstream section of the wetland where Site L is located is significantly more impacted by AMD than at Sites C and KP further downstream. Although water discharging from mines on the Central Rand is highly acidic and carries high metal loads, this water is diluted and neutralised as it flows toward the wetland. The presence of dolomite facilitates the infiltration of water and further raises the pH of the water as a result of the presence of carbonate, although the redox potential may still remain low. Flow into the head of the wetland near Site L is therefore likely largely from below, except during heavy rainstorm events. The absence of a pollution signature in the deeper part of the peat at Sites C and KP suggests that groundwater entering the wetland is unpolluted, and that pollutant accumulation in these areas is largely caused by surface water flow and possible atmospheric fallout. Sites C and KP are located in dolomitic compartments in which the groundwater is isolated from pollution plumes from TSFs by dykes (Figure 2b).

Mechanisms of sequestration

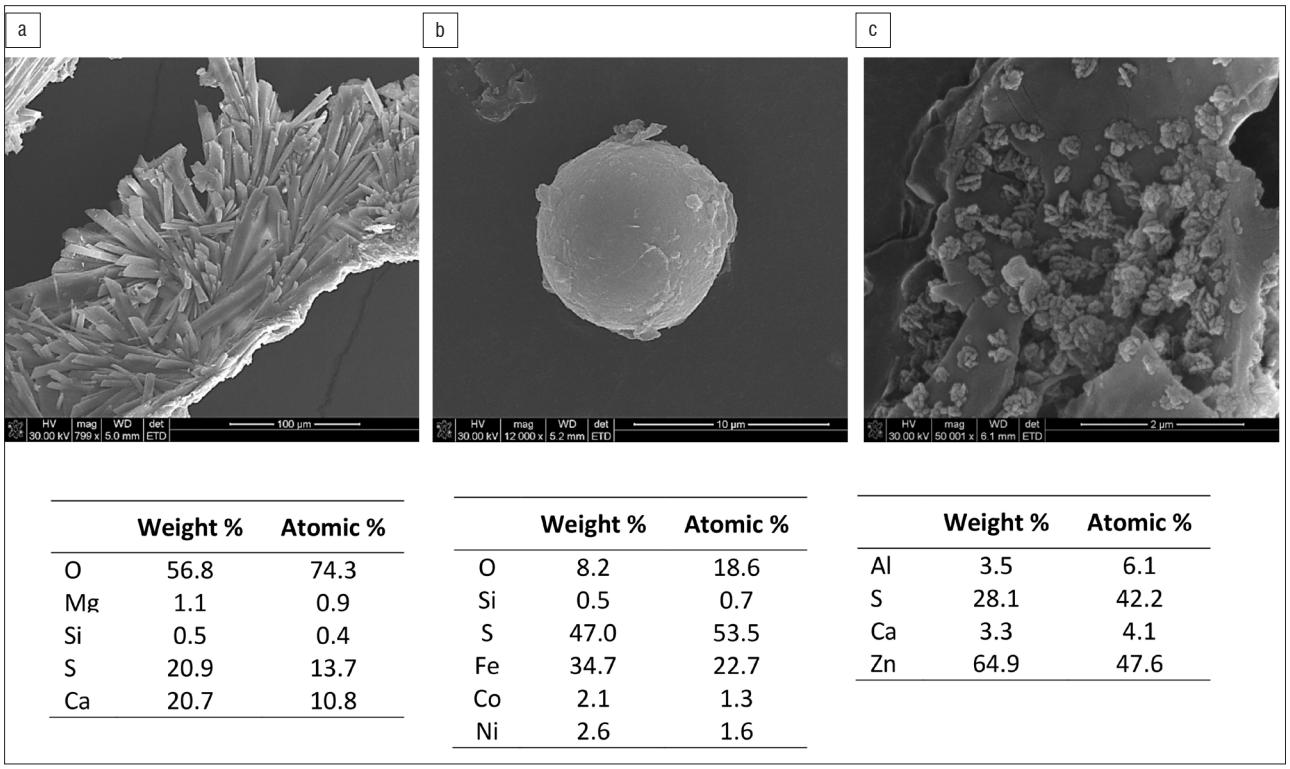
The majority of metal sequestration in the Klip River appears to occur in the upper section of the wetland near sampling Site L. Investigation of the mechanisms involved in the sequestration of metals in the peat therefore focused on core samples from this area.

The BCR sequential extraction data reveal similarities in the partitioning of Pb, Ni and Zn, with Fe-Mn oxyhydroxides the predominant phase for sequestration (Table 4). Metal associated with Fe-Mn oxyhydroxides are expected given the pH of the peat (pH 6.5–7) which favours the precipitation of Fe and Mn minerals and has also been noted in the AMD-affected Blesbokspruit wetland.[28] The oxidisable fraction becomes an increasingly important phase for Co (43%), Ni (37%) and Cu (75%) below 150 cm, indicating that the sequestration of these metals in the deeper peat is likely associated with sulfides and/or organic matter. Highly chalcophile metals such as Ni and Co are thus expected to precipitate with FeS_2 in the deeper peat where sulfur is present in the reduced S^{2-} form. Scanning electron microscopy investigations of the deeper peat reveal the presence of pyrite spheres that are enriched in Ni and Co (Figure 8), as well as Zn sulfide clusters. Cu is predominantly found within the oxidisable fraction, likely associated with sulfides. The partitioning data suggest that changes to pH or redox could result in the remobilisation of metals, particularly Pb, Ni, Zn, Co and Cu which are held predominantly in potentially available forms.

Peat samples showed evidence of extensive gypsum precipitation (Figure 8). Correlation between Ca and total S (R^2=0.62) suggests that this is likely the dominant phase for these two elements. Ca is precipitated as gypsum likely in response to evapotranspiration which leads to saturation in Ca minerals. The formation of gypsum crusts observed along the margins of the wetland near Site L indicates that groundwater seeping into the wetland is highly concentrated in Ca and SO_4. Kafri and Foster[19] also reported elevated sulfate in the groundwater. Gypsum precipitates are likely important adsorption sites for trace metals. Analysis of gypsum crusts from the Natalspruit wetland revealed high concentrations (2000–5000 ppm) of several metals, including Co, Ni, and Zn.[6] The BCR partitioning data indicate that gypsum adsorption could account for 12–18% (acid soluble fraction) of the total Co, Ni and Zn sequestered within the peat at Site L.

Implications and conclusions

Mining operations over the last ~130 years on the Witwatersrand have released high levels of acid and metals into surface and groundwater, with abandoned mines continuing to produce acid and contaminated water. The considerable enrichment in metals observed within the Klip River peats reflects exposure to contaminated water since the establishment of mining operations on the Witwatersrand and demonstrates the value of this wetland system in sequestering metals from polluted water. The wetland system therefore performs a vital ecosystem service in the environment by trapping metals that would otherwise enter the Vaal River system downstream.



	Weight %	Atomic %
O	56.8	74.3
Mg	1.1	0.9
Si	0.5	0.4
S	20.9	13.7
Ca	20.7	10.8

	Weight %	Atomic %
O	8.2	18.6
Si	0.5	0.7
S	47.0	53.5
Fe	34.7	22.7
Co	2.1	1.3
Ni	2.6	1.6

	Weight %	Atomic %
Al	3.5	6.1
S	28.1	42.2
Ca	3.3	4.1
Zn	64.9	47.6

Figure 8: Scanning electron microscopy images and elemental results indicating various forms in which pollutants are sequestered within the peat: (a) precipitation of gypsum crystals (at 100 cm), (b) precipitation of Co- and Ni-enriched pyrite spheres (at 250 cm) and (c) formation of Zn sulfide clusters (at 250 cm).

Table 4: Partitioning of metals in Core L1, defined using the BCR extraction procedure. Data are reported as averages (n=12) with the predominant fraction shown in bold.

%	Acid soluble	Reducible (Fe-Mn-oxyhydroxides)	Oxidisable (organic matter /sulfides)	Residual	Potentially available†
Al	0.9	26	8	**65**	35
Fe	0.2	8	28	**64**	36
Pb	10	**63**	25	8	75
Ni	12	**42**	25	21	74
Zn	13	**54**	7	28	64
Co	18	25	**31**	23	68
Cu	5	30	**57**	15	92

†The potentially bioavailable fraction is considered to be the sum of metal concentrations in the acid soluble, reducible and oxidisable fractions.

This occurs through a complex combination of metal sequestration mechanisms, which include mineral precipitation, co-precipitation and adsorption. Reducing wetland conditions and sustained groundwater discharge from the underlying dolomite produce biogeochemical conditions that favour metal sequestration, creating a natural passive treatment system.

The Klip River case study highlights the importance of natural wetlands as vital biogeochemical systems that have a substantial ability to accumulate large quantities of metals and thus remediate polluted waters, particularly those affected by acid mine drainage.

Although the Klip River peats are an important sink for contaminants, the accumulation of a large chemical reservoir presents a possible future source of pollutants. Pollutant metals are associated with relatively unstable phases, potentially susceptible to chemically or biologically mediated release into interstitial waters. This emphasises the importance of conserving the Klip River system as degradation, particularly in the most proximal region, would likely cause the wetland to become a source of contamination.[22] It also highlights the need for future research focused on a better understanding of metal sequestration within the peat and the potential for remobilisation. In addition, an increase in chemical loading within the system may ultimately result in a decrease in metal retention efficiency over time. The Klip River peats are therefore unlikely to act as an infinite metal sink and an increase in contaminated discharge into the system could have devastating consequences for both the wetland and the region's water supply.

Acknowledgements

The University of the Witwatersrand and the National Research Foundation of South Africa are acknowledged for providing financial support. Musarrat Safi, Sarah Pope and Fantasia Makhuvha assisted with the preparation of samples.

Authors' contributions

M.S.H. and T.S.M. collected the samples and interpreted the data. All authors contributed to writing the manuscript.

References

1. McCulloch J. Asbestos blues: Labour, capital, physicians and the state in South Africa. Oxford: James Currey; 2002.
2. Katz E. The white death: Silicosis on the Witwatersrand gold mines 1886-1910. Johannesburg: Wits University Press; 1994.
3. McCarthy TS. The impact of acid mine drainage in South Africa. S Afr J Sci. 2011;107(5/6), Art. #712, 7 pages. http://dx.doi.org/10.4102/sajs.v107i5/6.712
4. Wittmann GTW, Förstner U. Heavy metal enrichment in mine drainage II: The Witwatersrand Goldfields. S Afr J Sci. 1976;72:365–370.
5. Rosner T, Van Schalkwyk A. The environmental impact of gold mine tailings footprints in the Johannesburg region, South Africa. Bull Eng Geol Environ. 2000;59:137–148. http://dx.doi.org/10.1007/s100640000037
6. Naicker K, Cukrowska E, McCarthy T. Acid mine drainage arising from gold mine activity in Johannesburg, South Africa and environs. Environ Pollut. 2003;122:29–40. http://dx.doi.org/10.1016/S0269-7491(02)00281-6
7. Ntsume GM, McCarthy TS. A preliminary study of the relative contributions of diffuse and point sources of pollution arising from gold mining activity in a Witwatersrand goldfield. In: Loredo J, Pendás F. Mine water 2005 – Mine closure. Oviedo: University of Oviedo; 2005. p. 169–175.
8. Tutu H, Cukrowska EM, McCarthy TS, Hart R, Chimuka L. Radioactive disequilibrium and geochemical modelling as evidence of uranium leaching from gold tailings dumps in the Witwatersrand Basin. Int J Environ Anal Chem. 2009;89:687–703. http://dx.doi.org/10.1080/03067310902968749
9. Hobbs PJ, Cobbing JE. Hydrogeological assessment of acid mine drainage impacts in the West Rand Basin, Gauteng Province. Report no. CSIR/NRE/WR/ER/2007/0097/C. Pretoria: CSIR Natural Resources and the Environment; 2007.
10. Hobbs PJ. Pilot implementation of a surface water and groundwater resources monitoring programme for the Cradle of Humankind World Heritage Site. Report no. CSIR/NRE/WR/ER/0088/B. Pretoria: CSIR Natural Resources and the Environment; 2012.
11. Van der Merwe W, Lea I. Towards sustainable mine water treatment at Grootvlei mine. In: Armstrong D, De Villiers AB, Kleinmann RLP, McCarthy TS, Norton PJ, editors. Proceedings of the 8th International Mine Water Association Congress; 2003 October 19–22; Johannesburg, South Africa. IMWA; 2003. p. 25–36.
12. Durand JF. The impact of gold mining on the Witwatersrand on the rivers and karst system of Gauteng and North West Province, South Africa. Afr J Earth Sci. 2012;68:24–43. http://dx.doi.org/10.1016/j.jafrearsci.2012.03.013
13. Bobbins K. Acid mine drainage and its governance in the Gauteng City-Region. Johannesburg: Gauteng City-Region Observatory; 2015.
14. Sheoran AS, Sheoran V. Heavy metal removal mechanism of acid mine drainage in wetlands: A critical review. Minerals Eng. 2006;19:105–116. http://dx.doi.org/10.1016/j.mineng.2005.08.006
15. O'Sullivan AD, Moran BM, Otte ML. Accumulation and fate of contaminants (Zn, Pb, Fe and S) in substrates of wetlands constructed for treating mine wastewater. Water Air Soil Poll. 2004;157:345–364. http://dx.doi.org/10.1023/B:WATE.0000038882.09628.ab
16. Dean AP, Lynch S, Rowland P, Toft BD, Pittman JK, White KN. Natural wetlands are efficient at providing long-term metal remediation of freshwater systems polluted by acid mine drainage. Environ Sci Technol. 2013;47:12029–12036. http://dx.doi.org/10.1021/es4025904
17. McCarthy TS, Venter JS. Increasing pollution levels on the Witwatersrand recorded in the peat deposits of the Klip River wetland. S Afr J Sci. 2006;102:27–34.
18. Wolmarans JF. Some engineering and hydrogeological aspects of mining on the West Wits Line. In: Anhaeusser CA, Maske S. Mineral deposits of southern Africa. Johannesburg: Geological Society of South Africa; 1986. p. 791–796.
19. Kafri U, Foster MBJ. Hydrogeology of the malmani dolomite in the Klip River and Natalspruit Basins, South Africa. Environ Geol Water Sci. 1989;13:153–166. http://dx.doi.org/10.1007/BF01664700
20. Mellor ET. The geological map of the Witwatersrand goldfield (1:60 000). Union of South Africa Department of Mines and Industries geological survey; 1917.
21. Vermaak V. Geomorphological investigation of the Klip River Wetland, South of Johannesburg [MSc dissertation]. Johannesburg: University of the Witwatersrand; 2009.
22. McCarthy TS, Arnold V, Venter J, Ellery WN. The collapse of Johannesburg's Klip River wetland. S Afr J Sci. 2007;103:391–397.
23. Pueyo M, Rauret G, Lück D, Yli-Halla M, Muntau H, Quevauviller P, et al. Certification of the extractable contents of Cd, Cr, Cu, Ni, Pb and Zn in a freshwater sediment following a collaboratively tested and optimised three-step sequential extraction procedure. J Environ Monit. 2001;3:243–250. https://doi.org/10.1039/b010235k
24. South African Department of Water Affairs (DWA). Feasibility study for a long-term solution to address the acid mine drainage associated with the East, Central and West Rand underground mining basins. Assessment of the water quantity and quality of the Witwatersrand mine voids. Report no. P RSA 000/00/16512/2. Pretoria: DWA; 2012.
25. Tutu H, McCarthy TS, Cukrowska E. The chemical characteristics of acid mine drainage with particular reference to source, distribution and remediation: The Witwatersrand Basin, South Africa as a case study. Appl Geochem. 2008;23(12):3666–3684. http://dx.doi.org/10.1016/j.apgeochem.2008.09.002
26. Grawunder A, Merten D, Buchel G. Origin of middle rare earth element enrichment in acid mine drainage-impacted areas. Environ Sci Pollut Res. 2014;21:6812–6823. http://dx.doi.org/10.1007/s11356-013-2107-x
27. Sharifi R, Moore F, Keshavarzi B. Geochemical behavior and speciation modeling of rare earth elements in acid drainages at Sarcheshmeh porphyry copper deposit, Kerman Province, Iran. Chemie der Erde. 2013;73:509–517. http://dx.doi.org/10.1016/j.chemer.2013.03.001
28. Roychoudhury AN, Starke MF. Partitioning and mobility of trace metals in the Blesbokspruit: Impact assessment of dewatering of mine waters in the East Rand, South Africa. Appl Geochem. 2006;21:1044–1063. http://dx.doi.org/10.1016/j.apgeochem.2006.02.024

Stormwater harvesting: Improving water security in South Africa's urban areas

AUTHORS:
Lloyd Fisher-Jeffes[1]
Kirsty Carden[1]
Neil P. Armitage[1]
Kevin Winter[2]

AFFILIATIONS:
[1]Department of Civil Engineering, University of Cape Town, Cape Town, South Africa
[2]Department of Environmental and Geographical Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Kirsty Carden

EMAIL:
Kirsty.carden@uct.ac.za

DATES:
Received: 26 May 2016
Revised: 02 Oct. 2016
Accepted: 05 Oct. 2016

KEYWORDS:
water scarcity; alternative water resources; flood management; climate change resilience; sustainable drainage

HOW TO CITE:
Fisher-Jeffes L, Carden K, Armitage NP, Winter K. Stormwater harvesting: Improving water security in South Africa's urban areas. S Afr J Sci. 2017;113(1/2), Art. #2016-0153, 4 pages. http://dx.doi.org/10.17159/sajs.2017/20160153

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
South African Water Research Commission

The drought experienced in South Africa in 2016 – one of the worst in decades – has left many urbanised parts of the country with limited access to water, and food production has been affected. If a future water crisis is to be averted, the country needs to conserve current water supplies, reduce its reliance on conventional surface water schemes, and seek alternative sources of water supply. Within urban areas, municipalities must find ways to adapt to, and mitigate the threats from, water insecurity resulting from, inter alia, droughts, climate change and increasing water demand driven by population growth and rising standards of living. Stormwater harvesting (SWH) is one possible alternative water resource that could supplement traditional urban water supplies, as well as simultaneously offer a range of social and environmental benefits. We set out three position statements relating to how SWH can: improve water security and increase resilience to climate change in urban areas; prevent frequent flooding; and provide additional benefits to society. We also identify priority research areas for the future in order to target and support the appropriate uptake of SWH in South Africa, including testing the viability of SWH through the use of real-time control and managed aquifer recharge.

Significance:

- Addresses water scarcity through building resilience to the impacts of climate change; improving the liveability of cities; and prioritising water-sensitive urban design.

Introduction

South Africa experienced the worst drought in decades in 2016. This current drought has left many towns and cities with extremely compromised water supply systems, and food production has been limited across the country, thus placing pressure on the already fragile economy. In order to avert a future water crisis, the country needs to reduce its reliance on conventional surface water schemes based on impoundments on rivers and to seek alternative sources of water supply. Within urban areas, municipalities must find ways to adapt to, and mitigate the threats from, water insecurity resulting from, inter alia, droughts, climate change and increasing water demand driven by population growth and rising standards of living. Stormwater harvesting (SWH) is one alternative water resource that could supplement traditional urban water supplies, as well as simultaneously offer a range of benefits including the management of flooding and the provision of recreational areas. For the purposes of this paper, SWH refers to the collection and storage of run-off from an urban region and its subsequent use irrespective of location, and is usually implemented by the relevant local authority.[1] In comparison, rainwater harvesting is the collection and storage of run-off from an individual property (usually from the roofs of buildings) and its subsequent private use within that property.[1]

Based on the results of recent research in South Africa[1], as well as a review of the relevant international literature, we set out three position statements in this paper relating to how SWH can contribute to: improving water security and increasing resilience to climate change in urban areas; preventing frequent flooding; and providing additional benefits to society, such as creating amenity and preserving biodiversity. We have included priority research areas for the future in order to identify and support the appropriate uptake of SWH in South Africa, as well as recommendations regarding issues that need to be addressed to enable this research.

Position 1: Stormwater harvesting improves water security

The Atlantis Water Resource Management Scheme (AWRMS) has been in operation since 1979[2] and provides a useful South African example of SWH on a large scale. An important design aspect of this SWH system was the use of the town of Atlantis as a significant component of the catchment. The town was planned with separate residential and industrial areas, which allowed for the separation of high- and low-quality wastewater effluent. Stormwater and higher-quality treated municipal effluent are used to recharge an unconfined aquifer for later extraction and use. Low-quality water is disposed through recharge near the coast in such a way as to create a hydraulic barrier between the cleaner groundwater and the seawater.[3] The AWRMS has successfully ensured a supply of water for the town of Atlantis over the last 37 years, with approximately 30% of the groundwater supply augmented through artificial recharge. Interestingly, the establishment of the scheme was initially in response to the need to find an alternative to marine wastewater discharge[2], but after many successful years in operation, it is now seen internationally as an exemplar of a stormwater and wastewater reuse scheme[4].

Aside from the AWRMS, SWH has not been widely exploited in South Africa, and is limited to a number of small on-site systems used for irrigation at factories or distribution centres – even though the possibility of widespread use of stormwater as a resource in the country was mooted some time ago.[5] The reasons for this are not entirely clear, but may relate to issues of social perception, as well as institutional processes associated with the operation and maintenance of such schemes.[2]

Fisher-Jeffes[1] undertook one of the few detailed studies of the viability of SWH in South Africa, focusing on the residential areas of the Liesbeek River Catchment in Cape Town. Whilst it was acknowledged that there is

significant climatic variation across South Africa, he found that SWH had the potential to reduce the total current residential potable water demand of the catchment by more than 20% if the stored stormwater was used for non-potable purposes such as irrigation and toilet flushing – a significant saving for the City of Cape Town. However, in order for such reductions in water demand to be realised, the vast majority of residents and businesses would be required to make use of harvested stormwater. This requirement would likely necessitate changes in the regulations related to the supply of water in the City of Cape Town. Additionally, as Ellis et al.[6] indicated as part of their research in South Africa, significant social and institutional barriers – similar to those encountered elsewhere in the world[7-10] – may be an impediment to the adoption of SWH. This highlights the need for further research that accounts for the local context as most of the existing research into the implications of SWH has been undertaken in developed countries. International examples of large-scale SWH include:

- Singapore – which has one of the most comprehensive SWH systems that has proven itself to be a useful high-quality water resource.[11]
- USA and Australia – harvested stormwater is used for a range of end uses including irrigation, toilet flushing, commercial and industrial uses.[4]

Of significant concern to water resource planners is the uncertainty of the effects of climate change on water resources. For example, Fisher-Jeffes[1] highlighted that for a catchment in Cape Town, evaporation is expected to increase, while precipitation is expected to decrease. Using adjusted run-off data, the analysis showed that, based on the expected changes in evaporation and precipitation from 31 different climate change scenarios, it is very likely that SWH systems (as with other water resource schemes) will be negatively impacted by climate change. Losses could, however, be reduced through the use of managed aquifer recharge in place of open storage – as is the case for the AWRMS.

While local and international examples provide support for the wider adoption of SWH to address water security in South Africa, local climatic factors can influence its viability. In Cape Town, for example, the Mediterranean-type climate results in most of the harvestable stormwater being available during the wet winter months, when the reservoirs are typically filling in any case. Harvesting stormwater during this time may seem unnecessary; however, it could be utilised as a way to reduce normal demand from the city's reservoirs during the wet winter months (by increasing the rate at and level to which these reservoirs fill up) – thereby ensuring an increase in the availability of water during the dry summer months.

Position 2: Stormwater harvesting prevents flooding

SWH schemes all make use of some form of storage system. Some make use of retention ponds, while others make use of temporary detention ponds before either infiltrating or injecting water into an aquifer – also known as managed aquifer recharge. In either case – detention or retention – run-off is detained in an open storage system. The functioning of detention and retention ponds is well known[12,13]: by storing run-off volume, downstream flows are attenuated, resulting in reduced flooding. International case studies have demonstrated these benefits of SWH systems.[4]

Fisher-Jeffes[1] demonstrated the impact that such reductions in peak flows might have on flooding (and flood risks) using a two-dimensional flooding model. Figure 1 illustrates the flood hazard levels – using the City of Cape Town's definition of flood hazard (a combination of depth and velocity of water)[14] – for a storm event on 12 July 2009, with and without SWH. It is evident that SWH has the potential to significantly reduce flood risks in storm events.

A further opportunity exists for stormwater managers to actively manage SWH systems using real-time control in such a manner that, prior to a predicted storm event, the storage is partially emptied. In this way, significant attenuation could be achieved without compromising the ability to meet water demand. This option would require the development of a calibrated run-off model that could make use of predicted rainfall to estimate the run-off for a particular storm. Based on the anticipated run-off, the stormwater manager could partially empty the SWH system's storage a day or more before the rain event (depending on the availability of rainfall predictions), resulting in an increase in the pre-event flow rate in the river, but a decrease in the peak flows, which could prevent flooding.

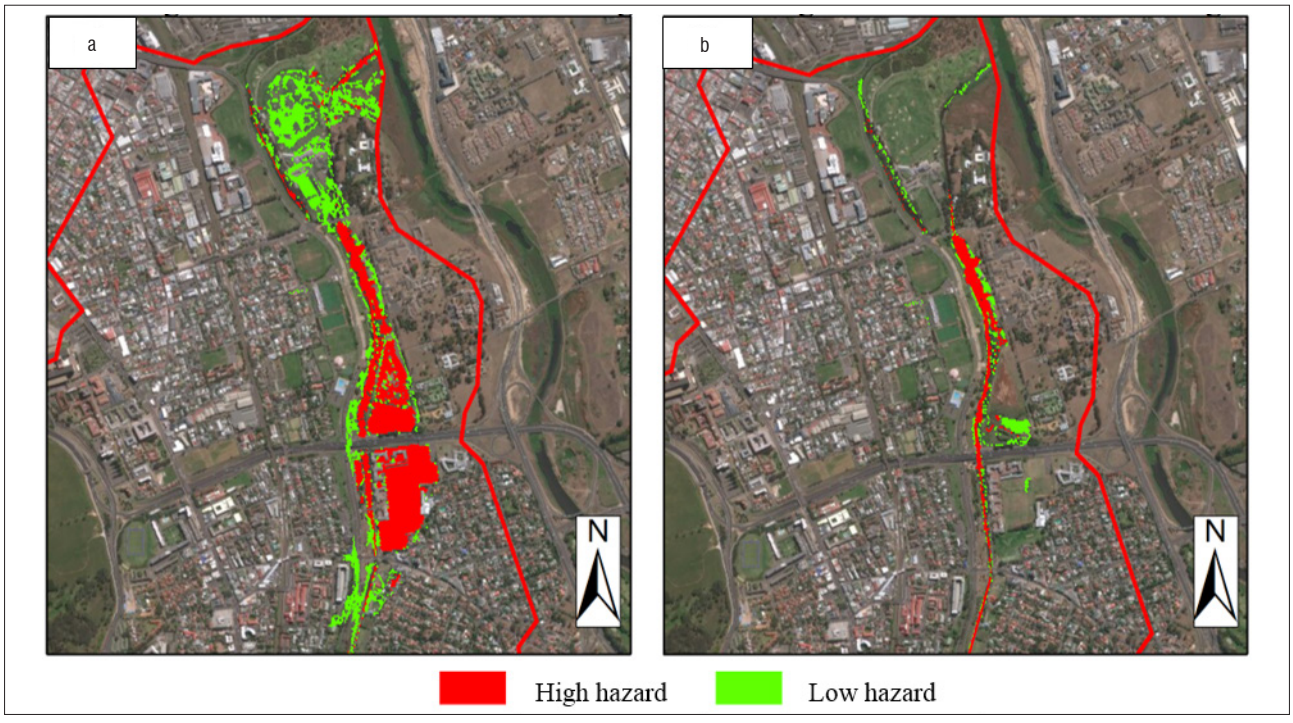


Figure 1: Flooding in the Liesbeek River Catchment on 12 July 2009 shown (a) without and (b) with stormwater harvesting.[1]

Position 3: Stormwater harvesting provides additional benefits

There is extensive literature on the value of the substantial benefits that natural assets – parks, wetlands, ponds etc. – can offer society.[15,16] In one such study, De Wit et al.[15] investigated the value of natural assets in the City of Cape Town. Through their own investigation and review of the literature, the authors monetised the value of different natural assets and ecosystem goods and services – including amenity value, biodiversity values, and water treatment capabilities. Whilst parks, wetlands and open spaces, such as those that might be created for SWH systems are typically considered to provide a positive amenity value, De Wit et al.[15] note that some can create negative amenity, particularly if they are not maintained, or even provide a risk to society. Fisher-Jeffes[1] equated the positive amenity generated by SWH to an estimated ZAR2–7.2 million per year in 2013 for the Liesbeek River Catchment – a catchment of only 2600 hectares. It is also worth recognising that by using harvested stormwater it will be possible to delay, and possibly avoid, the need for future water schemes based on impoundments on rivers. In so doing, SWH indirectly protects the ecosystem services that are naturally provided from being destroyed. Furthermore, if intentionally designed, SWH can offer significant multifunctional use and amenity benefits – such as recreational areas – and can aid in supporting biodiversity and mitigating urban 'heat island' effects, e.g. through the formation of 'blue-green' corridors with indigenous vegetation.[17]

Future research

Current research in South Africa has thus far focused on the financial, economic, technical and practical viability of SWH and has highlighted the need for further investigation, including into the social aspects of SWH – such as whether all sectors of South African society would be willing to use harvested stormwater, and if so, for which end uses they would be willing to use it? The preliminary study in the Liesbeek River Catchment[1] highlighted the potential for conducting future research into using SWH combined with real-time control to significantly reduce flooding during major storm events, as well as mitigating water scarcity. Similarly, research is required to determine whether SWH might be especially valuable as a water supply during droughts, as run-off from urbanised areas is typically greater than that from natural catchments during these events because of the extent of impervious areas. The experience and knowledge from the AWRMS should be expanded into studies on the viability of managed aquifer recharge systems to store and treat stormwater – using both confined as well as unconfined aquifers.

Most importantly, research thus far has highlighted that many issues need to be addressed to enable future studies, including the installation of basic monitoring and data logging (rainfall, flow and quality) equipment across urban areas in South Africa, to address the urgent need for basic calibration data.

Conclusions

Stormwater harvesting offers an alternative water supply source – one that is almost entirely untapped in South Africa – that could ensure improved water security for towns and cities across the country. While stormwater could be treated to potable standards – as has been done in Singapore – it may not be economically feasible or desirable and it may be preferable to use the stored water for non-potable purposes. SWH systems – especially those enhanced with real-time control – could offer additional benefits by mitigating flooding through storing run-off, thereby attenuating downstream flows. SWH can be designed to offer multifunctional use and amenity benefits, and can, through the formation of 'blue-green' corridors, aid in supporting biodiversity and mitigating urban 'heat island' effects.

In conclusion, while there is currently a need for ongoing research to quantify the additional benefits of optimally designed, built and operated SWH systems, indications are that SWH has the potential to contribute to improving water security and increase resilience to the impacts of climate change in urban areas, as well as simultaneously offer a range of social and environmental benefits.

Acknowledgements

We acknowledge the South African Water Research Commission for funding this study as part of project no. K5/2412: A feasibility study to evaluate the potential of using water-sensitive design principles to strengthen planning for water sensitive cities of the future.

Authors' contributions

L.F-J. was responsible for the background research (supervised by N.P.A.), and wrote the manuscript together with K.C., who was the project leader. L.F-J., N.P.A. and K.W. were key researchers on this project. Both N.P.A. and K.W. provided critical reviews of the paper, which was then revised into its final state by L.F-J. and K.C.

References

1. Fisher-Jeffes L. The viability of rainwater and stormwater harvesting in the residential areas of the Liesbeek River Catchment, Cape Town [PhD thesis]. Cape Town: University of Cape Town; 2015.
2. South African Department of Water Affairs and Forestry (DWAF). The Atlantis Water Resource Management Scheme: 30 years of artificial groundwater recharge. Pretoria: DWAF; 2010. p. 76.
3. Murray EC, Tredoux G. Planning water resource management: The case for managing aquifer recharge. In: Proceedings of the 2004 Water Institute of Southern Africa (WISA) Biennial Conference; 2004 May 2–6; Cape Town, South Africa. Johannesburg: Water Institute of South Africa; 2004. p. 430–437.
4. Philp M, Mcmahon J, Heyenga S, Marinoni O, Jenkins G, Maheepala S, et al. Review of stormwater harvesting practices. Technical report no. 9. Queensland: Urban Water Security Research Alliance; 2008.
5. Wright A. Urban stormwater, correctly managed, is a resource rather than a nuisance. In: Proceedings of the Water Institute of South Africa 1996 Biennial Conference & Exhibition; 1996 May 20–23; Port Elizabeth, South Africa. Johannesburg: Water Institute of South Africa; 1996. 8 pages. Available from: http://www.ewisa.co.za/literature/files/1996%20-%2032.pdf
6. Ellis D, Armitage NP, Carden K. Water sensitive design drivers and barriers. In: Proceedings of the Water Institute of South Africa (WISA) 2016 Biennial Conference; 2016 May 15–19; Durban, South Africa. Johannesburg: Water Institute of South Africa; 2016.
7. Brown R, Farrelly M, Keath N. Summary report: Perceptions of institutional drivers and barriers to sustainable urban water management in Australia. Melbourne: National Urban Water Governance Programme; 2007.
8. Nurick SD, Cattell K. An investigation into the mechanisms driving large property owning organisations to implement green building features. In: Proceedings of the South African Council for Quantity Surveyors Profession (SACQSP) 2013 Conference; 2013 June 20–21; Cape Town, South Africa. Cape Town: SACQSP; 2013. p 92–104.
9. Tjandraatmadja G, Cook S, Chacko P, Myers B, Sharma AK, Pezzaniti D. Water sensitive urban design impediments and potential: Contributions to the SA Urban Water Blueprint – Post-implementation assessment and impediments to WSUD. Adelaide: Goyder Institute for Water Research; 2014.
10. Dobbie M, Brooks K, Brown R. Risk perceptions of Australian urban water practitioners towards alternative water systems, including stormwater harvesting and quality treatment systems [document on the Internet]. c2012 [cited 2013 Aug 29]. Available from: http://waterforliveability.org.au/wp-content/uploads/Project6-SocietyInstitutions-Apr2012.pdf
11. Lim MH, Leong YH, Tiew KN, Seah H. Urban stormwater harvesting: A valuable water resource of Singapore. Water Pract Technol. 2011;6(4), Art. #wpt20110067. http://dx.doi.org/10.2166/wpt.2011.0067
12. Armitage N, Vice M, Fisher-Jeffes L, Winter K, Spiegel A, Dunstan J. The South African guidelines for sustainable drainage systems. Report TT558/13. Pretoria: Water Research Commission; 2013.
13. Woods-Ballard B, Kellagher R, Martin P, Jefferies C, Bray R, Shaffer P, et al. The SuDS manual [homepage on the Internet]. London: CIRIA; 2007. Available from: http://www.ciria.org/Resources/Free_publications/SuDS_manual_C753.aspx
14. City of Cape Town: Catchment, Stormwater and River Management. Floodplain and River Corridor Management Policy. Cape Town: Catchment, Stormwater and River Management, Directorate of Transport, Roads and Major Projects; 2009. p. 19.

15. De Wit M, Van Zyl H, Crookes D, Blignaught J, Jayiya T, Goiset V, et al. Investing in natural assets: A business case for the environment in the City of Cape Town [document on the Internet]. c2009 [cited 2010 Aug 25]. Available from: http://resource.capetown.gov.za/documentcentre/Documents/City%20research%20reports%20and%20review/EnvResEconomics-Final_Report_2009-08-18.pdf

16. Kumar P, editor. The economics of ecosystems & biodiversity: Ecological and economic foundations. London and Washington: Earth Scan; 2010.

17. Schaffler A, Christopher N, Bobbins K, Otto E, Nhlozi M, De Wit M, et al. State of green infrastructure in the Gauteng City-Region. Johannesburg: Gauteng City-Region Observatory; 2013. Available from: http://www.gcro.ac.za/media/reports/sogi_final_P6nDjP5.pdf