ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

Review Article

Research Article

Cover caption
A road in South Africa
(photo CC: Steve Cornish, flickr). In an
article on page 63, Verster and Fourie analyse
the costs and causes of the many fatal road accidents in
South Africa.

And out into the world they go…

Thirty years ago, South Africa had 36 higher education institutions (including technikons, which later became universities of technology). Following the mergers instituted by the then Minister of Education, Kader Asmal, the country had 23 universities – now 26 after the establishment of new institutions. The specific numbers themselves are not especially significant, other than to tell us that this country has a small higher education sector: Britain has a population (this year) of about 66.5 million and over a 100 universities; South Africa's population is just over 56 million – with 26 universities.

Innumerable highly respected South African researchers have, or have had, full-time or joint appointments in universities beyond Africa's borders – too many to name. Although leading researchers such as Quarraisha and Salim Abdool Karim (Harvard and Columbia), Achille Mbembe (Columbia, California, Yale), Mala Singh (Open University), Teboho Moja (NYU) and Winston Hide (Harvard, MIT and Sheffield) come to mind (not to mention Nobel laureates), there are many more amongst them.

It is interesting, though, that from the rather small base of South African higher education institutions, nine senior university leaders (vice chancellors and deputy vice chancellors) have been appointed to senior academic or similar positions beyond, rather than in, Africa over the past 18 years.

Most recently, the Vice Chancellors of two internationally well-rated South African universities have decided to take up positions elsewhere. Professor Cheryl de la Rey – Vice Chancellor and Principal of the 110-year-old University of Pretoria, and former Chief Executive Officer of the Council of Higher Education and Deputy Vice Chancellor for Research at the University of Cape Town – has been appointed as Vice Chancellor of the University of Canterbury in New Zealand (QS 2018 ranking: 231). She will take up her new position early in 2019.

Professor Albert van Jaarsveld – Vice Chancellor and Principal of the University of KwaZulu-Natal, former Chief Executive Officer of the National Research Foundation, and before that full professor at the Universities of Pretoria and Stellenbosch – will leave South African higher education in September this year to head up the International Institute for Applied Systems Analysis, based in Austria, as its Director General and Chief Executive.

The first vice chancellor of a South African university to be appointed to a senior position outside the country was Dr Mamphela Ramphele. At the end of a term as Vice Chancellor of the University of Cape Town (where she had previously been a Deputy Vice Chancellor) she was appointed in 2000 as one of the four Managing Directors of the World Bank. Her responsibilities included overseeing the strategic positioning and operations of the World Bank Institute and she held the Vice-Presidency of External Affairs.

In early 2001, Professor Colin Bundy left the University of the Witwatersrand, after overseeing a challenging period of transformation, to become Director and Principal of the School of Oriental and African Studies at the University of London. During part of his time as Director, he also served as a Deputy Vice Chancellor of the University, before becoming the first Principal of Green Templeton College at the University of Oxford. He remains an Honorary Fellow of the School of Oriental and African Studies.

In September 2001, Professor Brenda Gourley – Vice Chancellor and Principal of the University of Natal – was appointed as the fourth Vice Chancellor and Principal of the Open University in Britain, a position she held until her retirement in 2009. During that time she became a Founder Member of the Talloires Network of universities and educational institutions – an organisation aimed at strengthening the civic roles of universities and higher education. She also served on the Board of the International Association of Universities for 8 years and as Chair, for two terms, of the Association of Commonwealth Universities.

Just over a year later, in November 2002, the University of Natal lost its Senior Deputy Vice Chancellor and Acting Vice Chancellor, David Maughan Brown, when he was appointed Deputy Vice Chancellor of York St John University in England. Amongst other duties, he was responsible for the operational management of the University, and subsequently for the University's academic activities, including learning and teaching, research and quality enhancement – while deputising, when necessary, for the Vice Chancellor and Principal.

Parenthetically, not long after, the University of Natal's other Deputy Vice Chancellor, Professor Ahmed Bawa, spent a sojourn as Professor of Physics at Hunter College, City University of New York, so the University of Natal lost three of its most senior leaders in fairly quick succession. Professor Bawa returned to South Africa in 2010 as the Vice Chancellor and Principal of the Durban University of Technology, and is now the CEO of Universities South Africa.

Mathematician and Vice Chancellor and Principal of Stellenbosch University, Professor Chris Brink, resigned in order to take up the position of Vice Chancellor and Principal of the University of Newcastle in England in 2007. He served in that position until his retirement 9 years later, at the end of 2016. Professor Brink was a founding member of the Academy of Science of South Africa and remains respected for his contributions to formal aspects of computer science with emphasis on programme semantics.

The year 2009 was marked by two departures (as is the case for 2018) – Professor Martin Hall, Deputy Vice Chancellor of the University of Cape Town, in April, and Professor Calie Pistorius, Vice Chancellor and Rector of the University of Pretoria, in September. Professor Hall became Vice Chancellor and Principal of Salford University, in England; Professor Pistorius held those positions in the University of Hull – also in England. British universities gained six of the nine departing South African higher education leaders.

In this context, the size of the sector becomes important, particularly in view of the serious challenges to be faced in providing post-school education, through to postgraduate level, in well-managed institutions, to as many young people as possible. In aiming to achieve this goal, universities cannot afford to lose highly skilled and experienced leaders – from disciplines as varied as accountancy and archaeology to mathematics and medicine – not even nine in 18 years.

Check for updates

Pint of Science: Bringing science to the public and highlighting African research

AUTHOR:
Clare Garrard[1] (iD)

AFFILIATION:
[1]Computational Biology, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Clare Garrard

EMAIL:
grrcla001@myuct.ac.za

A physicist, a biologist and a neuroscientist walk into a bar. What sounds like the beginning of a joke is in fact how the Pint of Science Festival works. This year, from 14 to 16 May, 26 scientists from varying backgrounds walked into bars at three venues across Cape Town to bring us three nights of TED-styled science talks.

Stemming from an idea to 'Meet the Researchers' in which UK scientists Dr Michael Motskin and Dr Praveen Paul brought the public to the lab, Pint of Science was born in 2013 when they decided instead to take the lab to the public, in the most quintessential of British meeting places – the pub. Now in its sixth year, Pint of Science is held annually in May across the globe, with more countries joining the festival each year.

South Africa joined in with its first Pint of Science in 2016, and remains the only African country to be a part of the festival. When the Founding Director of Pint of Science South Africa, Eugenie Marais, faced a personal challenge in 2017, Chantal Louw stepped up to take it forward, and the second Pint of Science South Africa was hosted in 2018.

> Two years ago the first Pint of Science South Africa was launched by Eugenie Marais and I joined the team as a volunteer. I loved it! I am passionate about bringing the conversations into public spaces, and hope that scientists all over South Africa will use this as an opportunity to have a drink and chat to their community about their work, says Louw.[1]

Pint of Science events are arranged in three themed evenings at each bar or pub venue. This year's Pint of Science covered four of the six Pint of Science themes: (1) Beautiful Mind, which covers neuroscience, psychiatry and psychology; (2) Atoms to Galaxies, which encompasses physics, chemistry and maths; (3) Our Body, which spans human biology, health and medicine; and (4) Planet Earth, which covers earth sciences, zoology and plant science. At each event, two or three scientists speak about science related to their research, with an opportunity for the audience to ask questions after each talk.

> We often find that the public doesn't really know what scientists do and that this gap in information contributes to a sense of mistrust, explain Motskin and Paul[2]. The Q&A session is an opportunity for attendees to ask questions directly to the people doing the research. It allows scientists to clarify confusion, to put their work in perspective, and to be seen as important members of the community or, in some cases, partners in the same fight.

According to the 2018 Edelman Trust Barometer, people are seeking credible sources in this era of fake news, and turning to experts more and more.[3] This makes science communication events like Pint of Science, during which the public can interact directly with scientists, all the more important to create trust in and enthusiasm for the scientific research that is driving the world forward.

Of the 26 talks that were held at Pint of Science South Africa 2018, nearly half were about research that related to human health. TB and HIV dominated the stage with two talks each – unsurprising given the prevalence of these diseases in South Africa.

'If you're worried you have an STD, go get tested,' emphasised HIV postdoctoral researcher, Dr Smritee Dabee. 'Talk about your STDs, people. The more you know about an STD, the more you can do about it.' Dabee also highlighted the problem of asymptomatic issues, such as inflammation, which can increase the risk of HIV transmission.

PhD student Hayley Tomes spoke on the 'black box' of the brain, mentioning a number of technologies that are used to study the brain. 'It maybe is not as useful as it is pretty, but it is definitely still quite useful', joked Tomes as she showed a microscopy image of a rainbow patterned brain known as a brainbow. This technique, which uses fluorescent chemicals to create the colours, is useful for studying connections in the brain.

Dr Jonathan Shock, a string theory researcher at the University of Cape Town, attracted great interest with his talk on the legacy of Stephen Hawking. Shock spoke about how Hawking's theories on black holes have informed the string theory of the universe, and how he repeatedly lost bets about scientific predictions, despite being one of the luminary scientists of our time.

> Stephen Hawking showed by looking at black holes, these weird objects which are the end points of massive stars, that something funny must happen. And that strange thing, [...] – one of the bets that in fact he lost – turned out to be the most important piece of the puzzle in trying to understand the nature of space and time.

Pint of Science speakers ranged from well-established mid-career scientists to postgraduate students doing their master's degrees. The inclusion of graduate students in the line-up of Pint of Science speakers plays an important role in giving young scientists a voice in the public sphere.

There is a growing belief that all PhD students should be required to present their research to a mixed audience of experts and enthusiasts as a way of encouraging scientists early on to engage with the public and share their science effectively with a wider audience. Pint of Science is another way of facilitating this dialogue.

In addition to giving a voice to young scientists, the role of Pint of Science in giving a platform to underrepresented population groups in science has been highlighted.[4] Notably, the Association of South African Women in Science

and Engineering, which organised the events at the northern suburbs venue, placed a particular emphasis on choosing female speakers as a way of highlighting women in science.

Along with the benefits of giving a platform to underrepresented population groups and encouraging young scientists to engage in science communication, joining with global organisations like Pint of Science also helps to showcase African research. For Africa to be able to highlight its own science is of utmost importance if it is to raise its profile and play a more active part in the scientific debates of the world. This will be especially important as genetic research expands, given that the richest genetic information in the world is found in Africa.

'South Africa is the first country in Africa to start Pint of Science and I hope it won't be the last', says International & France Director of Pint of Science, Elodie Chabrol.

Pint of Science is organised by a team of volunteers from varying walks of life, who share a passion for science. While Pint of Science South Africa is currently only held in Cape Town, as a volunteer-run festival it is an open field for anyone to pick it up and run with it in other cities in South Africa.

There is also room for expansion into the other two themes which were not covered this year: Tech Me Out, which covers technology, engineering and computing, and Our Society, which focuses on sociology, law, history and policy.

'We aim to have events all over South Africa and would love your help!' says Louw. 'We are looking for speakers, volunteers, partners and chapter heads. Please have a look at our new website and chat to us.'

'It is a wonderful opportunity to bridge the gap, making science more accessible to people who would not usually engage in science', enthuses Dr Taime Sylvester, a postdoctoral researcher at Stellenbosch University.

Pint of Science South Africa will be returning in May 2019. So next time a physicist, a biologist and a neuroscientist walk into a bar, join them for a chat.

References

1. UCT makes science accessible. Cape Times. 2018 May 14. Available from: https://www.pressreader.com/south-africa/cape-times/20180514/281728385161821

2. Paul P, Motskin M. Engaging the public with your research. Trends Immunol. 2016;37(4):268–271. https://doi.org/10.1016/j.it.2016.02.007

3. 2018 Edelman Trust Barometer [homepage on the Internet]. No date [cited 2018 Jun 16]. Available from: https://www.edelman.com//trust-barometer

4. Joubert M. White men's voices still dominate public science. Here's how to change this. The Conversation. 2018 May 27 [cited 2018 Jun 15]. Available from: https://theconversation.com/white-mens-voices-still-dominate-public-science-heres-how-to-change-this-96955

AUTHOR:
Jimmy Volmink[1] (iD)

AFFILIATION:
[1]Faculty of Medicine and
Health Sciences, Stellenbosch
University, Stellenbosch,
South Africa

CORRESPONDENCE TO:
Jimmy Volmink

EMAIL:
jvolmink@sun.ac.za

Reconceptualising health professions education in South Africa

Background

In June 2018, the South African Minister of Health announced the gazetting of the National Health Insurance Bill – an important step towards achieving the government's aim of realising universal health care. The Academy of Science of South Africa's (ASSAf's) consensus study report[1] on reconceptualising education and training of the health workforce, released in the same month, is therefore timely.

In recent years there have been several calls for changes in the way health professionals are trained around the globe. Most notable among these calls is the 2010 Lancet Commission report entitled: 'Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world.'[2] The Lancet report highlighted a number of failures in health professional education and recommended specific instructional and institutional reforms that are required to ensure that future healthcare professionals (HCPs) are 'fit-for-purpose'. Subsequently, the World Health Organization produced guidelines for transforming and scaling up HCP education and training, drawing on the findings of the Lancet Commission.[3]

Training institutions, educational groups and regulatory bodies in South Africa, have actively engaged with these global calls for reform. In 2012, ASSAf hosted a one-day workshop with key stakeholders to 'contextualise' the Lancet Commission's recommendations. The workshop culminated in ASSAf being asked to take the deliberations further, and the ASSAf Council responded by appointing a 10-person multi-professional panel to conduct a consensus study on the transformation of health professions education (HPE) in South Africa. The panel's brief was to consider a set of questions (produced by a task team) and to make evidence-based recommendations on how the transformational challenges, reflected in the pre-specified questions, could be addressed through interventions that are appropriate and feasible in the South African context.

Conceptual framework

As a point of departure, the ASSAf panel considered the goal of HPE to be 'the production of knowledgeable, competent, relevant and socially accountable HCPs who are capable of confidently and collaboratively promoting health and delivering quality health care in the context of universal health coverage'.

The conceptual framework adopted as a guide to the panel's deliberations comprised the following elements:

1. the health and higher education systems are interdependent;

2. population needs are determined by both social and individual determinants of health; these needs drive health system requirements, which, in turn, impact on the health professional mix, skills, competencies and attributes required by HCPs;

3. quantity, quality and relevance of HCPs are all important dimensions of education and training; and

4. iterative monitoring and evaluation, together with leadership and governance of HPE, are needed for transformation to be realised.

Key transformational areas considered and recommendations of the panel

The panel considered eight transformational areas for HPE in South Africa and made a total of 16 recommendations. Following is a summary of these, with further details, including the role players considered most appropriate for addressing the various issues, provided in separate concise and full versions of the report.[1]

Student selection

Currently, health sciences faculties in South Africa attempt to balance the tension between selection focused on student retention and success, and selection aimed at broadening access to training in order to achieve equity and redress. While good evidence exists that previous academic performance predicts academic success in undergraduate programmes, top performers do not necessarily turn out to be the best practitioners. The role of additional factors such as empathy, resilience, work ethic, ability to communicate and leadership qualities warrant attention in the discourse around student selection.

The report recommends evaluation of a broader set of selection criteria than those currently in use. It points to the need for rigorous research to determine which academic and non-academic selection criteria and student support measures best predict student success and promote the attainment of the desired graduate competencies in the South African setting. Academic development and support programmes targeted at both students and educators, and financial aid for needy students are considered essential for translating student access into retention and success.

Scaling up the health workforce

There is a shortage of HCPs in the country, most critically in the categories of medical practitioners, pharmacists and registered nurses. The main reason is that production of HCPs has not kept up with the increasing population growth, growing burden and complexity of disease, and expanding service workload in the public health sector.

The report recommends strengthening of public sector institutions to allow for the scaling up of HCP numbers. This scale-up will require additional training facilities and infrastructure, appropriately staffed clinical training

platforms, collaboration with the private sector, and greater use of ICT to support training. Human resources for health planning will also need to be improved, ideally with the adoption of tracking and reporting of key human resources for health indicators. The report further recommends a multi-stakeholder approach to enhance health professional retention during studies, in the profession and in the country.

Urban–rural maldistribution of HCPs

The shortage of HCPs in rural and underserved areas in South Africa is a particularly critical issue. Despite 46% of the South African population residing in rural areas, only 12% of doctors and 19% of nurses practise in rural settings. Corresponding figures for medical specialists and allied health professionals are worse.

The report stresses the need for the selection and training of HCPs to be orientated towards addressing inequity and meeting the needs of the most underserved. It recommends that health science faculties prioritise applicants from rural and remote areas who meet the minimum academic criteria for selection, and that they ensure greater exposure to rural settings during training, with a stipulated minimum of clinical time spent in rural areas. The report additionally recommends 'community-oriented primary care' (a model in which primary health care is provided to a defined community based on assessed need and integration of primary care and public health practice) as a strategy to support service and learning. It also recommends that a monitoring system to track progress on practice in rural and underserved areas be designed and implemented.

Team work

Professional 'tribalism' is recognised as an important barrier to delivering effective health care. The Lancet commission identified 'interprofessional education and collaborative practice' (IPEPC) as an important catalyst for change in health care. IPEPC promotes the idea that transformative collaborative learning at undergraduate level will promote team work across professional boundaries in subsequent practice.

The ASSAf report supports the adoption of IPEPC and recommends that a multi-stakeholder, national working group be formed to develop and guide the implementation of a strategic plan for IPECP in order to ensure that it becomes sustainably embedded in South Africa. This plan should include a detailed stakeholder analysis and guidance for stakeholder engagement and buy-in. It should also delineate an IPECP competency framework and curriculum.

Core competencies for South African HCPs

Currently, the various professional and regulatory bodies each develop their own profession-specific competencies. There is a need to establish a set of core competencies for all professions focusing on IPEPC. The report recommends a hybrid competency-based education model that emphasises the process of learning and the achievement of learning outcomes. An Inter-Professional Regulatory Council working group would help to build consensus around a set of generic competencies for all HCPs.

Faculty development

Educators influence both the type and the quality of graduates produced, yet this aspect of the educational experience receives very little attention. There is a need for educators to be supported to become 'change agents'. The report highlights the importance of health professions educators being more responsive to both their internal learning community and the community beyond the institution. It suggests that this could be progressively realised through building a supportive institutional climate which values teaching, recruitment and integration of traditional and non-traditional faculty; developing competency for change agency; implementing transformative educational strategies; promoting adaptive education communities; and encouraging scholarship and reflection.

Internship and community service

Internship is currently a component of basic professional training for medicine (2 years), pharmacy (1 year) and psychology (1 year). While internship is legally part of professional training in South Africa it does not typically involve higher education institutions. Community service was first introduced for medical doctors in 1998 and has since been rolled out to include a number of HCP groups. Community service is intended to provide health services to rural and underserved areas and to allow young HCPs to further develop their skills and competencies. In practice, the focus has been largely on development of technical competencies, rather than attitudes, team work, community outreach and management skills. The current experience of internship and community service in South Africa could therefore be undermining the vision and intention of undergraduate health professions education.

The report proposes a shift from the mindset of interns and community service officers as the lowest levels of medical workers to one of active development of young professionals who are better equipped for public service. It recommends that universities take responsibility for education and professional development throughout the continuum from the undergraduate years through to internship (renamed postgraduate years 1 and 2) and community service (renamed postgraduate year 3). As the current 9 years is a long time for the production of a generalist doctor, the report recommends earlier differentiation with support for specialty training commencing in the community service year.

Financing of HPE in South Africa

It is the unanimous view of the panel that the success of the proposed transformation efforts will require proper financing of HPE. Longstanding shortcomings in the financing, planning and organising of HPE in South Africa largely arise from poor alignment of the interests of key stakeholders, such as the Department of Health, the Department of Higher Education and Training, and National Treasury. Poor coordination between governance at national and provincial levels is a further complicating factor.

The report recommends urgent action to improve the governance of health sciences funding, including through strengthening the capacity and accelerating the momentum of the Joint Health Sciences Education Committee. It considers the strengthening of governance structures and building a joint vision as critical initial steps, without which detailed costing and planning activities might not produce their full intended benefit. Improving human resources for health planning, resource allocation and budgeting should also receive priority. The report furthermore recommends the introduction of joint governance structures at institutional level, with strong consideration being given to reviving the concept of the academic health sciences complex. This model, if properly supported, has the potential to improve skills development, institutional governance and quality, and could form a strong basis for strengthening alignment of interests and more effective service delivery, as well as health sciences education. The model also has relevance in the context of greater institutional decentralisation required for the purchaser–provider split under National Health Insurance.

Conclusion

The ASSAf study on HPE provides a consensus view on how health professions education and training in South Africa could be reconceptualised and offers a roadmap for the way ahead. It is hoped that the panel's recommendations will help to consolidate current, and spur new efforts, to address the quantitative and qualitative shortfall in the health workforce, resulting in long-term benefits to the health of the nation.

References

1. Academy of Science of South Africa (ASSAf). Reconceptualising health professions education in South Africa. Pretoria: ASSAf; 2018. https://doi.org/10.17159/assaf.2018/0021

2. Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, et al. Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world. Lancet. 2010;376:1923–1958. https://doi.org/10.1016/S0140-6736(10)61854-5

3. World Health Organization (WHO). Transforming and scaling up health professionals' education and training: World Health Organization Guidelines 2013. Geneva: WHO; 2013.
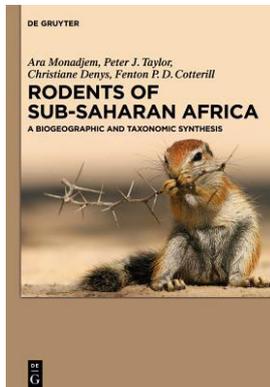
A new documentation of African rodent diversity

REVIEWER:
Brian W. van Wilgen

AFFILIATION:
Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Stellenbosch, South Africa

EMAIL:
bvanwilgen@sun.ac.za

Rodents (order Rodentia) are by far the most diverse group of mammals on earth, accounting for over one third (~2200 species) of the world's estimated 6500 species of mammals. The number of rodents known to science is constantly changing, both because new species are regularly discovered and because taxonomic revisions can either combine separate species into one, or split a single species into two or more new species. In this book, the authors have set out to provide an accessible, accurate and updated reference work that covers the identification, distribution and taxonomic relationships of the 463 known rodent species in sub-Saharan Africa. They have chosen sub-Saharan Africa, rather than providing a continent-wide coverage, to keep the material down to manageable proportions, and also because the rodents of North Africa are more affiliated to those in Europe than to the rest of Africa. Among the authors are those with extensive experience as curators in African natural history museums who are well connected with museums across both anglophone and francophone Africa.

Although there have been other recent reviews (for example Happold's[1] account of African rodents published in 2013), there is a need to update such accounts as new information comes to light. This book should be seen as complementary to works like Happold's publication. It does not include the wealth of information on the ecology, behaviour and reproductive biology covered by Happold and others, but rather focuses on taxonomic and biogeographic insights that build on an 'extraordinary renaissance' in species discovery, as well as on the fact that molecular techniques are becoming less costly and easier to use. These modern techniques often confirm the delineations put forward in early classical studies that were subsequently 'obscured by lumpers and pheneticists of the middle twentieth century'. To illustrate this, the authors point to their accounts of 31 species of Otomyinae (vlei rats and whistling rats), compared with only 17 species covered by Happold's recent account, as an example.

Besides being of biogeographic and taxonomic interest, rodents are also of importance to conservation. While none of the 463 species recognised is known to have become extinct, there are 44 species listed in the IUCN Red List as Critically Endangered (3), Endangered (21), Vulnerable (15) and Near-threatened (5). However, this picture is not complete, as there was insufficient information to determine the conservation status of 52 species, and a further 100 species have not been evaluated in terms of their conservation status. In addition, rodents are important as transmitters of diseases and as pests of crops. For example, more than 60 zoonotic diseases (i.e. those transmitted from animals to humans) have been attributed to rodents. The authors therefore identify their target audience as, inter alia, mammalian biologists, conservationists, public health workers, epidemiologists and agriculturalists.

The bulk of the book (988 out of 1092 pages) is dedicated to species accounts, based on examination of tens of thousands of museum specimens, literature records and online museum database records. There is detailed information at a genus level, including defining characters, cranial and dental features, and habitat and biogeography. The species accounts that follow have a focus on distribution and habitats, rather than on biology and ecology, which is covered in other publications. At the genus level, there are images of skulls and dentition (molars), while at the species level there are distribution maps and, in some cases, colour photographs. The distribution maps show point locations for specimens rather than a distribution range, and although the photographs include some of live specimens, many are of mounted museum specimens only. Common names are not included at a species level, although collective common names are provided for each genus (for example 'whistling rats', 'sun squirrels', 'mole-rats' or 'spiny mice').

Taxonomic treatises such as this one will always go out of date as new species are discovered, and as taxonomic revisions find new ways of classifying species. That said, this is an exhaustive examination of a significant mammal group, and it is likely to remain a valuable source of information for decades. It would, as the authors suggest, be most useful if it is treated as a companion volume to other publications, as it will assist in identifying particular species that are of conservation, biogeographic, medical or agricultural importance. The book should be a valuable addition to the library of any serious scientist working in any of these fields.

Reference

1. Happold DCD. Mammals of Africa. Volume III: Rodents, hares and rabbits. London: Bloomsbury Publishing; 2013.

AUTHORS:
George Ellis and Mark Solms

REVIEWER:
Anwar S. Mall

AFFILIATION:
Division of General Surgery,
Faculty of Health Sciences,
University of Cape Town,
Cape Town, South Africa

EMAIL:
anwar.mall@uct.ac.za

A hardwired neo-cortex – What role for neuroplasticity and developmental processes?

Is the neo-cortex of the brain hardwired or not? This is the central question George Ellis and Mark Solms of the University of Cape Town set out to answer in the seven chapters and 177 pages of this book. Regardless of which side of the argument one is on, there is a guarantee of a fulfilling read, with the latest information about the modern brain and its evolution. The main point of this book is that there is no language or any other instinctual system in the neo-cortex.

Evolutionary psychologists claim that the mind is a product of evolution and natural selection.[1,2] We have adapted successfully to our environments because the mind (specifically the neo-cortex) is modular and hardwired for specific functions, which provides us with innate knowledge from birth and equips us for survival in an environment with a 'poverty of stimulus'. An example is that of Chomsky's innate module for grammar which Pinker[3] later argued was a 'language instinct' (see Rose[4]). We required such a mind in the context of an 'environment of evolutionary adaptedness' to the African savannah but would it help us survive modern life?

The gene took central position from about the mid-20th century – the period of the new synthesis of Darwin's theory (in 1859) with modern genetics – with a convergence of ideas of heredity from several sources[5-8]: the ('selfish') gene was the unit for natural selection[8], the meme its cultural counterpart and gene-culture coevolution the new idea[9].

Textbooks traditionally focus on the physiology and anatomy of the brain, underlined by the principle of structure related to function. But the brain also produces our subjective mental life and yet the neurosciences, until quite recently, avoided studying feelings and emotions on the grounds that an objective science could not study the subjective experience, despite the implications of the injury and associated changes to the personality of Phineas Gage in 1848.[10,11] Francis Crick's 'astonishing' hypothesis and scientific study of the 'soul'[12], was possibly a step forward in 1994?

This book by Ellis and Solms represents both a recent paradigm shift from gene-centred thinking and a refutation of the claim of neo-cortical innate modularity, providing an integrated view which suggests that the structure of the neo-cortex is a 'soft-wired' arrangement of inter-connected neurons, genetically designed but sculpted by developmental processes that involve interaction between the brain and environmental stimuli under the guiding influence of the limbic system (affect): 'These specific connections develop through our life experiences and social interactions, which continually alter synaptic strengths by modifying gene expression' (p. 106). Whilst genes play a role in brain development, the brain's plasticity in response to the environment is crucial, and allows humans with roots in the African savannah to happily adapt to urban life today. This is the 'evo-devo' tradition which unites evolutionary theory with genetics and with developmental studies with less emphasis on the gene[13], and more of an understanding that evolution, development, structure, function and environment are all crucial to our survival and have to be understood in a contextual way. The unit for natural selection is far more than just the gene!

Innateness in the brain is in the limbic system which is the seat of human emotions, and is shared with other mammals.[14] Our primary emotions are hardwired in all mammals and selected for by natural selection, thus playing a key role in evolutionary development.[15] The developing cortex, in its exposure to environmental stimuli, responds by allowing competing neurons to form brain circuits by a process of adaptive selection and neural Darwinism[16], under the guidance of emotions through ascending neural pathways and neuromodulators, which in turn influence cognitive activity and brain plasticity.

Emotions produce a specific reaction to an inducing situation, tempered in humans by higher reason and wisdom, and they regulate the internal state of the body, ensuring that, for example, heart rates change during flight in fear.[15] It is the influence of this emotional system that underlies the intellect and shapes other aspects of human behaviour, giving meaning to what it means to be human. Emotion therefore has been selected for, has played a key role in evolutionary development and, in turn, has affected evolution.

Innate modules in the neo-cortex would be restrictive in responding to an unpredictable environment, and do not account for the plasticity of the brain or its compensatory mechanisms during injury. Also, the 'poverty of stimulus' argument disregards the richness of stimuli during the mother–neonate bonding period. To understand the function of the mind, the reader is asked to consider the hierarchical nature of its multi-level structure, which gives rise to complexity and many levels of emergence, thus allowing for a top-down causality in response to stimuli, and bottom-up causality where lower-level structures are adapted to higher-level functions.

The brain straddles two worlds – the inner and the outer – and the book beautifully explains the transition of activity from the physical to the mental; for example, when blood sugar levels fall, we *feel* hungry and act upon that feeling. Information from the sense organs is relayed to the appropriate cortical regions via the thalamus, forming our perceptions (percepts), which become concepts, the raw material of cognition, the further these move from the posterior to the anterior cortex. To survive and reproduce, the organism seeks its internal needs in the outside world. This is the biological purpose of cognition, which is guided by our hardwired emotions, without which there would be no survival. The fittest, whether individuals or communities, in the strongest social contexts, shall be the unit of natural selection.

Resolving the nature–nurture debate would have important implications for learning, reading and writing. A young brain is shaped by story-telling[17] and this book can make a positive impact on our failing educational system

and illiteracy. There is a difference between folk knowledge and higher academic learning which requires effort and motivation on an individual's part in dealing with counter-intuitive ideas. Also, importantly, there is an epistemology of the brain sciences and evolution, in general, in this book that the reader can appreciate.

References

1. Pinker S. How the mind works. London: Penguin Books; 1997.

2. Pinker S. The blank slate. London: Penguin Books; 2002.

3. Pinker S. The language instinct. London: Penguin Books; 1994.

4. Rose S. The 21st-century brain. London: Vintage Books; 2005.

5. Schrodinger E. What is life? The physical aspect of the living cell. Cambridge, UK: Cambridge University Press; 1945.

6. Judson HF. The eighth day of creation. The makers of the revolution in biology. New York: Simon and Schuster; 1979.

7. Watson J. The double helix: A personal account of the discovery of the structure of DNA. Phys Today. 1968;21(8):71. https://doi.org/10.1063/1.3035117

8. Dawkins R. The selfish gene. Oxford: Oxford University Press; 1978.

9. Wilson EO. Consilience: The unity of knowledge. London: Abacus, Little Brown and Company; 1998.

10. Solms M, Turnbull O. The brain and the inner world: An introduction to the neuroscience of subjective experience. New York: Other Press; 2002.

11. Damasio A. Descarte's error: Emotion, reason and the human brain. New York: Harper Collins; 1994.

12. Crick F. The astonishing hypothesis: The scientific search for the human soul. New York: Simon and Schuster; 1995.

13. Noble D. The music of life: Biology beyond the genome. Ontario: Oxford University Press; 2006.

14. Panksepp J. Affective neuroscience. The foundations of humans and animal emotions. London: Oxford University Press; 1998.

15. Damasio A. The feeling of what happens. Body, emotion and the making of consciousness. London: Vintage Books; 1999.

16. Ellis GFR, Toronchuk JA. Neural development: Affective and immune system influences. In: Ellis RD, Newton N, editors. Consciousness and emotion: Agency, conscious choice, and selective perception. Amsterdam: John Benjamins; 2005. p. 81–119. https://doi.org/10.1075/ceb.1.06ell

17. Hustvedt S. The shaking woman: A history of my nerves. New York: Henry Holt and Company; 2010.

A note on equity returns for South African investors

AUTHOR:
Eben Maré[1]

AFFILIATION:
[1]Department of Mathematics and Applied Mathematics, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Eben Maré

EMAIL:
Eben.mare@up.ac.za

KEYWORDS:
equities; earnings; volatility; risk; real return; risk adjusted returns

October: This is one of the particularly dangerous months to invest in stocks. Other dangerous months are July, January, September, April, November, May, March, June, December, August and February.

Mark Twain

Equity risk and returns

On a historical basis, South African equity markets have outperformed inflation significantly. Using returns from the 2016 version of the *Credit Suisse Global Investment Sourcebook*[1], I observe that over the period 1900 to 2015, the annualised real return (i.e. returns adjusted for inflation) for South African equities was 7.3% compared to 1.8% for bonds and 1.0% for shorter-term bills. I compare these returns with US annual real returns of 6.4% for equities, 2.0% for bonds, and 0.8% for bills over the same period.[1]

Investors buy equities to participate in the growth prospects of a company and the economy, but are exposed to significant short-term noise, namely so-called volatility. Shiller[2], for example, has shown that the market price of equities moves considerably more than warranted by underlying economic fundamentals. My aim was to demonstrate that investors, with a focus on long-term investment objectives, can capture longer-term growth and investment trends by applying focus on company earnings growth, therefore effectively ignoring short-term noise.

My data set comprises monthly closing levels of the FTSE/JSE All Share Index dating from January 1960 to December 2017. I also captured the dividend yield and price/earnings ratio. The source of my data is I-Net.

The analysis ensued by calculating the monthly price returns as well as the total returns (i.e. the price returns plus returns of dividends accrued). The same process was applied to earnings, i.e. I calculated the monthly change in earnings. I also calculated the change in the price/earnings ratio on a month-to-month basis. This gave me an effective decomposition of the monthly total return of the index into dividends received, change in earnings, as well as change in the price/earnings ratio. The monthly numbers are very volatile and obviously reflect the full ambit of noise in the turbulent equity market as a whole – my main aim was to demonstrate longer-term investment trends, hence I calculated the returns described earlier over a 20-year horizon. The rolling returns are depicted in Figure 1.
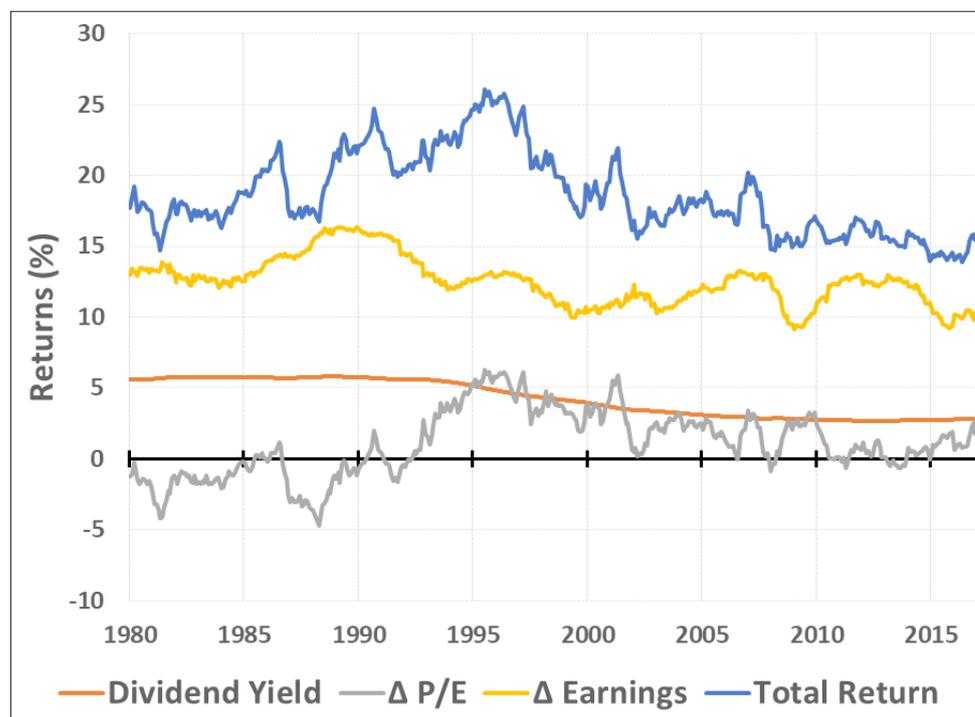
Figure 1: Decomposition of total index returns over rolling 20-year periods.

Note 1 – The main driver of total equity returns, apart from the fairly stable dividend yield, is growth in earnings, which contributes about two-thirds of the total return over time. The price/earnings ratio, which drives shorter-term valuation issues, is a lot more volatile, but, over longer-term periods, contributes marginally towards total returns.

Note 2 – Dividend yield. The dividend yield reflects a very healthy and attractive yield above 5% until the early 1990s. In 1993, a so-called Secondary Tax on Companies was introduced, which meant that companies were required to pay tax on a portion of all dividends declared to shareholders.[3] The tax was introduced to encourage

reinvestment of profits at company level – the effects of this tax are clear from 1995 onwards in a much reduced dividend yield. The Secondary Tax on Companies was eliminated in 2013 in an attempt to make South Africa a more investment friendly destination and to achieve international tax alignment. It is important to note that the dividend yield is a function of earnings growth and the company specific pay-out ratio.

Note 3 – My analysis was done on the basis of the FTSE/JSE All Share Index, which consists of multiple equities. However, the analysis is equally valid for single securities or a well-balanced diversified portfolio of equities.

It is important to note that company earnings reflect the sum total of risks and efforts experienced by a company to achieve and sustain financial profitability; understanding this fact makes us look at risks to achieve earnings growth. Understanding risks to earnings growth over time allows us to build an investment case for specific securities.

Conclusion

We are often reminded that our investments should be considered on a long-term basis. In this note I illustrate the importance of considering earnings growth as a driver of equity returns over the long term. It is important to realise that the stock market will have many short-term gyrations and exhibit significant volatility, but considering earnings growth provides us with a firm basis to look through short-term noise and assess long-term potential investment returns.

References

1. Dimson E, Marsh P, Staunton M. Credit Suisse global investment returns sourcebook. Zurich: Credit Suisse; 2016.

2. Shiller RJ. Market volatility. Cambridge, MA: MIT Press; 1989.

3. South African Revenue Services. Who must pay Secondary Tax on Companies (STC)? [webpage on the Internet]. No date [updated 2013 Feb 14; cited 2018 Apr 01]. Available from: www.sars.gov.za/FAQs/Pages/210.aspx.

AUTHORS:
Mwazvita T.B. Dalu[1]
Ashley W. Gunter[1]
Ryan J. Wasserman[2]
Tatenda Dalu[3]

AFFILIATIONS:
[1]Department of Geography, University of South Africa, Pretoria, South Africa

[2]School of Science, Monash University Malaysia, Bandar Sunway, Malaysia

[3]Department of Ecology and Resource Management, University of Venda, Thohoyandou, South Africa

CORRESPONDENCE TO:
Mwazvita Dalu

EMAIL:
smazvita@gmail.com

Owning the lake, not just the rod: The continuing challenge of 'the old boys' in knowledge production

Knowledge from the Global South, and particularly Africa, is continuously exported and repackaged, thereby transferring its ownership to those able to conform it to the paradigms of consumption in the knowledge economy of the Global North. The list of the top 40 scientific papers by country, according to Scopus, reflects a significant under-representation of publications from Africa. It is significant to note that only one African country – South Africa – features on the list, but only in the bottom five journals.[1] This exclusion may be for many reasons, not withstanding those related to funding and government support, the developmental needs of universities, a lack of ability on the part of the authors themselves to write to Western paradigms and standards, and the career aspirations and needs of African academics. Yet we propose that these issues are not the only reasons for the lack of academic voice from the African continent – there is a substantial amount of research on inequality in global knowledge production which largely focuses on income and resource inequality as the major reason for this situation.[2-4] It is, however, arguable that focusing only on the technical and economic limitations of African academics, whilst ignoring the greater cultural and political context within which the practice of academia is in itself deeply entrenched, does not sufficiently account for the challenges that they face. However, a significant cause of academic silence is the consequence of barriers resulting from practices of 'the old boys' network'. Thus, focusing on the global publication practices in academia, we present fresh arguments to bring to centre stage the consequence of barriers resulting from these networks. Here, the relative socio-political challenges of African academics are critically interwoven into the understanding and functioning of the informal old boys' network.

The idea of the 'old boys' network' stems from the British elite school system in which men of influence and means used their status to advance other men from the same school. There is much literature on the 'old boys' network' in numerous sectors. Yet this concept is still elusive and vague and few would openly acknowledge being part of any network that propagates the advancement of individuals from a similar social background. The developmental implications thereof, especially in light of the current prioritisation of the decolonisation agenda in South Africa, are also therefore discussed.

It could equally be argued that the idea of 'the old boys' may simply be a consequence of Western academia being the historical core of the university system and, as such, certain academics have mastered the paradigms, systems and language through many cycles of evolution and refinement. However, the oldest university (University of al-Qarawiyyin, Fez) is located in Morocco – on the African continent.[5] This raises the question of why similar cycles of evolution and refinement have not taken place within the African science and higher education context, leading to Morocco's influence within science and education among the ranks of world-leading scientific outputs. This may critically show that 'contemporary social, political, economic and cultural practices continue to be located within the processes of cultural domination through the imposition of imperial structures of power'[6]. The systems of academic ratings, credible outputs, journal listings and publication status is set, monitored and regulated almost exclusively by individuals based in the Global North.[7,8]

The idea of monopolies of power with regard to who actually moderates academic outputs, including journal publications, points to a centre of gatekeeping in the Global North. By consistently maintaining the reins on determining what is accepted, rejected or simply unpublishable, the old boys' network continues to hold up barriers to prominence on the global knowledge economy by African academics and institutions. The historical legacy of imperialism and colonialism is widely recognised as a barrier to development in Africa, and yet within the knowledge economy little has been done to redress the problems that this former system has left. While inferior education was forced upon Africans during the colonial era,[9] it is more apparent in some nations (e.g. South Africa) than in others (e.g. Zimbabwe). As the countries similar to the latter conformed to the 'colonial master's' standards in their general and academic culture, they seem to fair slightly better on the international academic arena, which is governed largely by the old boys' network. Similar situations likely ensue in other parts of the world where regions are currently geographically or politically fragmented as a result of the legacy of colonialism.

Although many colonial inequalities have since been redressed in most African countries to some extent, the stains of the colonial legacy still seem to manifest in the international publishing arena, as a form of mistrust by editors to African authors affiliated with African academic institutions. It appears that all knowledge is often evaluated against 'expert' knowledge based on Western scientific paradigms, before it is considered valid and useful, even to a non-Western context. This mistrust expresses itself through various means, such as intense and vigorous checks of native African writing or even the writing of authors of European descent who are affiliated with an African academic institution.[10] To date, European languages remain the languages of power, in spite of the many developments of local journals within the Global South. Thus, the colonial experience inherently continues to shape and influence research and its representation within the reproduction of knowledge in publications. As such, more native African academics come across more negative comments centred on language issues from reviewers, in comparison to their Northern counterparts. According to Sithole[11], the environment in the Global North for the African scholar is hostile, and is made insecure because, somehow, the African scholar is assumed to need the tutelage even of the most junior scholars from the Global North and it is assumed that their facts (even those originating from the more familiar local Global South context to the scholar) must be checked. This experience is also shared by academics of European descent who are based at African institutions, although to a lesser extent.[12]

One also gets a sense that patronage exists in the selection of a reviewer, as typically with some journals, the editor selects reviewers from an existing pool of members of a journal's board. Thus, very often, African authors must

face up to reviewers who are of North American or European descent, many of who (although well-meaning) harbour a bias towards Northern research interests where most journals of international accreditation are based.[13] This bias is often further exacerbated by the additional criterion of assigning reviewership and editorship to what, or rather who, constitutes an 'authoritative voice' within a certain area of research within the Global South, having been systematically determined by the old boys themselves. The latter paradox seems to have resulted in the monopoly of certain research areas to be associated with northern institutions or non-African scholars based at these institutions, in spite of an African scholar appearing to be in a more credible contextual position to conduct more comprehensive research. It has been argued that African scholarship needs to be synergistically organised and applied to subject matters and fields in which African scholars have a comparative advantage.[14] This argument, however, is gravely challenged by the old boys' network, as often the Global North tends to be responsible for conception, while the Global South is involved merely in the execution of tasks (e.g. data collection and field experiments), thus facilitating their limited inclusion in knowledge packaging and ownership of the knowledge market in publishing.[15,16] An example is that of large charismatic African wildlife species research (e.g. large mammals), which attracts much media attention and funding, yet publications thereon and grants for research continue to greatly exclude native African scholars, particularly in prominent positions within research projects. Even in cases in which African scholars appear to have prominent research positions in these areas, some of our intellectuals may be merely figure heads. These same scholars are constantly faced with the danger of alignment with whoever is paying for knowledge, as many funders are more favourable to scholars publishing in Global North journals, which often aim to fulfil Northern research interests. Thus, local scholars cease to sufficiently address the research needs and interests of the locality of origin, with detrimental implications to the development of the Global South.

Even African rating systems perpetuate the dominance of the old boys' network, by holding journals based in the Global North to a higher esteem than those based in the local Global South. For example, the system in South Africa will rank a local scientist with more publications in journals of Global North origin (which also, consequentially, usually have the higher impact factor), higher than those whose publications are mostly in local journals of Global South origin (http://www.nrf.ac.za/rating). Thus, higher knowledge exchange, and quality of thought, occurs between the Global North and South in comparison to that which occurs within the Global South exclusively. This situation has facilitated the 'lack of and/or poor development of internal scientific discussions and debates within and between our scientific communities, resulting in a general stampede of Global South scholars for individual acknowledgment by the Global North'[14]. Thus, although not intentional, these actions continue to actively facilitate the Global North in dominating policy development and implementation in the local sphere. Should scientific knowledge indeed inform policy, then the Global North continues to reinforce its power over the Global South, through dictating which or what knowledge is considered 'good', thereby presenting a significant challenge to the decolonisation agenda within policy, development and education in the Global South. Sithole[11] quotes Karioki as pointing out that 'Africa demands of its intellectuals to serve the communities from which they spring'. Hence, the Western saying 'give a man a fish, and feed him for one day; give him a rod, feed him for a lifetime', warrants an African revision: redress rightful entitlement to own the lake within which they fish, and you empower a continent.

Acknowledgements

References

1. Organisation for Economic Co-operation and Development (OECD). OECD science, technology and industry scoreboard 2017: The digital transformation. Paris: OECD Publishing; 2017.

2. Kapungu RS. The pursuit of higher education in Zimbabwe: A futile effort?' A paper prepared for the Center for International Private Enterprise (CIPE) 2007 international essay competition on 'Educational reform and employment opportunities' [document on the Internet]. c2007 [cited 2017 Oct 25]. Available from: http://archive.kubatana.net/docs/chiyou/kapungu_cipe_essay_higher_education_071002.pdf

3. Aarts H, Greijn H. Globalization, knowledge and learning: Developing the capacities of higher education institutes. In: Teferra D, Greijn H, editors. Higher education and globalisation – Challenges, threats and opportunities for Africa. Maastricht: Maastricht University Centre for International Cooperation in Academic Development (MUNDO); 2010. p. 9–18.

4. Hwami M. 'The Nativist Turn' and the crisis in university education. Int J Educ. 2013;5(4):124–144. https://doi.org/10.5296/ije.v5i4.4480

5. Verger J. Patterns. In: De Ridder-Symoens H, editor. A history of the university in Europe volume I: Universities in the Middle Ages. Cambridge, UK: Cambridge University Press; 2003.

6. Madge C, Raghuram P, Noxolo P. Engaged pedagogy and responsibility: A postcolonial analysis of international students. Geoforum. 2009;40:34–45. https://doi.org/10.1016/j.geoforum.2008.01.008

7. Jöns H, Hoyler M. Global geographies of higher education: The perspective of world university rankings. Geoforum. 2013;46:45–59. https://doi.org/10.1016/j.geoforum.2012.12.014

8. Liu NC. The story of academic ranking of world universities. Int Higher Educ. 2015;54:2–3. https://doi.org/10.6017/ihe.2009.54.8409

9. Maringe F, Ojo E. Sustainable transformation in a rapidly globalizing and decolonising world. In: Maringe F, Ojo E, editors. Sustainable transformation in African higher education. Rotterdam: Sense Publishers; 2017. p. 25–39. https://doi.org/10.1007/978-94-6300-902-7_2

10. Morita N. Negotiating participation and identity in second language academic communities. Tesol Q. 2004;38:573–603. https://doi.org/10.2307/3588281

11. Sithole MP. Unequal peers: The politics of discourse management in the social sciences. Pretoria: Africa Institute of South Africa; 2009.

12. Edmondson R, Luhtakallio E. Writing sociology, publishing sociology. Euro J Cultur Polit Sociol. 2016;3:387–391. https://doi.org/10.1080/23254823.2016.1245244

13. Czerniewicz L. Inequitable power dynamics of global knowledge production and exchange must be confronted head on. LSE Blog. 2013 April 29 [cited 2017 Jun 28]. Available from: http://blogs.lse.ac.uk/impactofsocialsciences/2013/04/29/redrawing-the-map-from-access-to-participation/

14. Sithole B. My knowledge, your knowledge, whose knowledge is it? Reflections from a researcher's journey through universities in North and South. In: Adriansen HK, Madsen LM, Jenses S, editors. Higher education and capacity building in Africa: The geography and power of knowledge under changing conditions. Oxfordshire: Routledge; 2016.

15. Adriansen HK, Madsen LM, Jenses S, editors. Higher education and capacity building in Africa: The geography and power of knowledge under changing conditions. Oxfordshire: Routledge; 2016.

16. Muller J, Cloete N, Van Schalkwyk F, editors. Castells in Africa: Universities and development. Cape Town: African Minds; 2017.

Check for updates
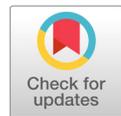
Possible predator avoidance behaviour of hominins in South Africa

AUTHOR:
Shaw Badenhorst[1]

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

CORRESPONDENCE TO:
Shaw Badenhorst

EMAIL:
shaw.badenhorst@wits.ac.za

There are many factors that contribute to the evolutionary success of species[1], for example, species often develop elaborate predator avoidance behaviour[2]. Such behaviours can be studied by ethologists in extant animals[3], but supporting empirical evidence is often lacking for such behaviours in extinct organisms[4]. Using living animals as proxies for extinct species usually allows for behavioural analogies, such as the use of baboon behaviour to infer the behaviour of our human ancestors and relatives.[5] South Africa is inhabited by a large primate, the chacma baboon (*Papio ursinus*). Baboons are comparable in size and weight to australopithecines like *Paranthropus robustus* and *Australopithecus africanus* and their human-like behaviour has long been recognised.[5,7-11]

Our understanding of the evolution of hominins is largely based on direct evidence from skeletal remains (gait, brain size, mobility, diet, etc.) and material culture (bone and stone tools), and very little is known about their behaviour. *Paranthropus robustus* was a robust australopithecine that lived between 2 and 1.2 million years ago in South Africa and remains of this species have been found at sites like Kromdraai, Swartkrans, Drimolen, Gondolin and Cooper's Cave in the Cradle of Humankind.[12] In the main, *Paranthropus robustus* was small statured but bipedal[13] and followed a patrilocal form of residence[14]. They were dietary generalists.[15] Apart from these main lines of evidence, very little is known about the behaviour of *Paranthropus robustus* during the Early Pleistocene.

From an evolutionary ecological perspective, prey species develop mechanisms to avoid being caught by predators. These anti-predatory devices are varied in animals and behavioural and anatomical adaptations may include a combination of sound, smell, colour, pattern, form, posture and/or movement devices.[2] Natural selection promotes the evolution in prey animals of features that reduce the probability of success for their predators. Animals with such anti-predatory devices tend to have a higher probability of escaping predation than animals without them.[16] Today, lions (*Panthera leo*) and leopards (*Panthera pardus*) are the main predators of baboons. Lions stalk baboons during daytime when they feed on the ground, and leopards attack them during the night.[17] Baboons have evolved numerous predatory avoidance behaviours to counter these attacks to ensure survival. The main behaviours include living in large troops, moving in a patterned way on the landscape[5], using sentries (elderly male individuals to sound alarms and defend the troop), sleeping in inaccessible places at night[18] and stone throwing[19].

Baboons also display another form of predatory avoidance behaviour: they are often found, during daytime, with other animals including buffalo (*Syncerus caffer*), elephant (*Loxodonta africana*)[20], bushbuck (*Tragelaphus scriptus*), rock hyraxes (*Procavia johnstoni*), tree hyraxes (*Heterohyrax brucei*)[21], guinea fowls (*Numida meleagris*), blue duiker (*Philantomba monticola*)[9], and, most often, with impala (*Aepyceros melampus*). The visual powers of baboons complement the acute sense of smell of impala, allowing for their mutual protection from predators.[14,17,22-27] Washburn and De Vore[11] found that olive baboons (*Papio anubis*) are closely associated with impala in open country, while in forested areas, they associate with bushbuck. The bark of these two antelope species will set a troop of baboons to flight, and a mixed herd of baboons and impalas are almost impossible to take by surprise.[11] Impala have the ability to distinguish between the alarm and contest calls of chacma baboons.[28] Both species are found in similar environments, and both are predated by lions and leopards.[29]

It is likely that hominins employed similar behaviour by associating themselves with medium and large ungulates during daytime, especially before the controlled use of fire, which would have enabled them to successfully survive for several hundreds of thousands of years. However, testing this proposed association is challenging. For example, using a simple presence–absence approach with respect to medium to large Bovidae and Equidae faunas associated with *Paranthropus robustus*[30-33] from the Cradle of Humankind does not reveal any significant absence of species (Figure 1). An absence or lower occurrence of species may indicate successful anti-predator behaviour, although sample size and taphonomic history[34,35] are also important considerations.
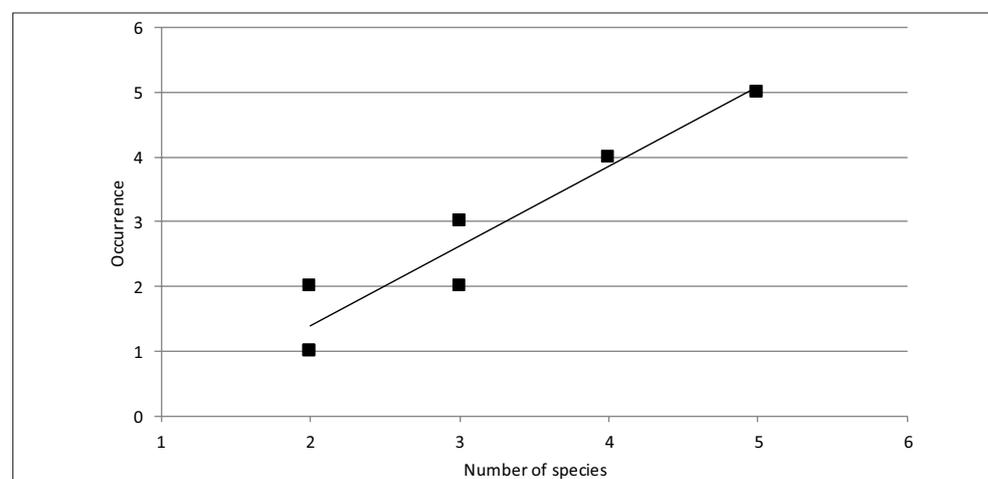
Figure 1: The number of medium and large Bovidae and Equidae species and occurrences at Kromdraai B, Sterkfontein Member 5, Swartkrans Members 1, 2 and 3, Drimolen and Cooper's D Cave. Trend line indicated.

Impalas are the perfect ungulates for primates such as the chacma baboon to associate themselves with during daytime. In addition to their acute ability to sense predators, impalas prefer woodland savanna, rarely wonder more than 2 km from permanent water, and are sedentary, in that they move less than 3 km per day within their home range.[6] During Plio-Pleistocene times, it is possible that hominins such as *Paranthropus robustus* and *Australopithecus africanus* as well as other primates associated themselves with ungulates with similar characteristics. The extinct *Gazella helmoedi* is thought to be an ancestral local form of the extant impala *Aepyceros melampus*.[36,37]

Novel approaches are required to investigate these likely associations between hominins and ungulates to reveal more about the evolution and behaviour of our human ancestors and their relatives. For example, a consideration of the predators, feeding and water requirements, habitat preference, distribution, herd size, home ranges, and anti-predator behaviour of ungulates can potentially reveal behavioural aspects of hominins like *Paranthropus robustus* and *Australopithecus africanus* to complement research on skeletal elements.

References

1. Jezkova T, Wiens JJ. What explains patterns of diversification and richness among animal phyla? Am Nat. 2017;189(3):201–212. https://doi.org/10.1086/690194

2. Pianka ER. Evolutionary ecology. New York: Addison Wesley Longman Inc.; 2000.

3. Lehner ON. The handbook of ethological methods. Cambridge, UK: Cambridge University Press; 1996.

4. Benton MJ. Studying function and behavior in the fossil record. PLoS Biol. 2010;8(3), e1000321, 5 pages. https://doi.org/10.1371/journal.pbio.1000321

5. De Vore I, Washburn SL. Baboon ecology and human evolution. In: Howell FC, Bourlière F, editors. African ecology and human evolution. Chicago, IL: Aldine Publishing Company; 1966. p. 335–367.

6. Rautenbach IL. Mammals of the Transvaal. Ecoplan Monograph No. 1. Pretoria: Ecoplan; 1982.

7. Marais EN. Burgers van die berge [Citizens of the mountains]. Pretoria: JL van Schaik; 1938. Afrikaans.

8. Marais EN. The soul of the ape. Pretoria: Human & Rousseau; 1969.

9. Bolwig N. A study of the behaviour of the chacma baboon, *Papio ursinus*. Behaviour. 1959;14(1–2):136–163. https://doi.org/10.1163/156853959X00054

10. De Vore I. The social behavior and organisation of baboon troops [PhD thesis]. Chicago, IL: University of Chicago; 1962.

11. Washburn SL, De Vore I. The social life of baboons. Sci Am. 1961;204:62–71. https://doi.org/10.1038/scientificamerican0661-62

12. Herries AIR, Curnoe F, Adams JW. A multi-disciplinary seriation of early *Homo* and *Paranthropus* bearing palaeocaves in southern Africa. Quat Int. 2009;202:14–28. https://doi.org/10.1016/j.quaint.2008.05.017

13. Klein RG. The human career: Human biological and cultural origins. Chicago, IL: University of Chicago Press; 1999.

14. Copeland SR, Sponheimer M, De Ruiter DJ, Lee-Thorp JA, Codron D, Le Roux P, et al. Strontium isotope evidence for landscape use by early hominins. Nature. 2011;474:76–78. https://doi.org/10.1038/nature10149

15. Wood B, Strait D. Patterns of resource use in early *Homo* and *Paranthropus*. J Hum Evol. 2004;46(2):119–162. https://doi.org/10.1016/j.jhevol.2003.11.004

16. Greenwood JJD. The evolutionary ecology of predation. In: Shorrocks B, editor. Evolutionary ecology. Oxford: Blackwell Scientific Publications; 1984. p. 233–273.

17. Busse C. Leopard and lion predation upon chacma baboons living in the Moremi Wildlife Reserve. Botsw Notes Rec. 1980;12:15–21.

18. Stevenson-Hamilton J. Wild life in South Africa. London: Cassell and Company; 1947.

19. Raven-Hart R. Cape Good Hope 1652-1702. The first fifty years of Dutch colonisation as seen by callers. Cape Town: A. A. Balkema; 1971.

20. Skinner JD, Chimimba CT. The mammals of the southern African subregion. Cambridge, UK: Cambridge University Press; 2005. https://doi.org/10.1017/CBO9781107340992

21. Davis BR, Ebersole JJ. Impala (*Aepyceros melampus*) associate with olive baboons (*Papio anubis*) for feeding and security in Tarangire National Park, Tanzania. Afr J Ecol. 2015;54:238–241. https://doi.org/10.1111/aje.12263

22. Pfeiffer JE. The emergence of man. London: Harper & Row Publishers; 1969.

23. Stevenson-Hamilton J. The Low-veld: Its wild life and its people. London: Cassell and Company; 1934.

24. Bigalke R. What animal is it? Pretoria: The Trustees of the South African Bird Book Fund and The Mammals of South Africa Book Fund; 1958.

25. Altmann SA, Altmann J. Baboon ecology. Chicago, IL: University of Chicago Press; 1970.

26. Morgan-Davies AM. The association between impala and olive baboon. J East Afr Nat Hist Soc. 1960;23:297–298.

27. Baenninger R, Estes RD, Baldwin S. Anti-predator behaviour of baboons and impalas toward a cheetah. E Afr Wildl J. 1977;15:327–329. https://doi.org/10.1111/j.1365-2028.1977.tb00414.x

28. Kitchen DM, Bergman TJ, Cheney DL, Nicholson JR, Seyfarth RM. Comparing responses of four ungulate species to playbacks of baboon alarm calls. Anim Cogn. 2010;13:861–870. https://doi.org/10.1007/s10071-010-0334-9

29. Cowlishaw G. Vulnerability to predation in baboon populations. Behaviour. 1994;131:293–304. https://doi.org/10.1163/156853994X00488

30. Brain CK. The hunters or the hunted? An introduction to African cave taphonomy. Chicago, IL: University of Chicago Press; 1981.

31. De Ruiter DJ. Revised faunal lists for Members 1-3 of Swartkrans, South Africa. Ann Transv Mus. 2003;40:29–41.

32. Adams JW, Rovinsky DS, Herries AIR, Menter CG. Macromammalian faunas, biochronology and palaeoecology of the early Pleistocene Main Quarry hominin-bearing deposits of the Drimolen Palaeocave System, South Africa. PeerJ. 2016;4, e1941, 46 pages. https://doi.org/10.7717/peerj.1941

33. De Ruiter DJ, Pickering R, Steininger CM, Kramers JD, Hancox PJ, Churchill SE, et al. New *Australopithecus robustus* fossils and associated U-Pb dates from Cooper's Cave (Gauteng, South Africa). J Hum Evol. 2009;56(5):497–513. https://doi.org/10.1016/j.jhevol.2009.01.009

34. Driver JC. Identification, classification and zooarchaeology. Circaea. 1991;9(1):35–47.

35. Grayson DK. Quantitative zooarchaeology: Topics in the analysis of archaeological fauna. London: Academic Press; 1984.

36. Brink JS. The evolution of the black wildebeest, *Connochaetes gnou*, and modern large mammal faunas in central southern Africa [PhD thesis]. Stellenbosch: Stellenbosch University; 2005.

37. Brink JS, Herries AIR, Moggi-Cecchi J, Gowlett JAJ, Bousman CB, Hancox JP, et al. First hominine remains from a ~1.0 million year old bone bed at Cornelia-Uitzhoek, Free State Province, South Africa. J Hum Evol. 2012;63:527–535. https://doi.org/10.1016/j.jhevol.2012.06.004

Check for updates

Possible importance of *Cannabis sativa* L. in regulation of insulin and IL-6R/MAO-A in cancer cell progression and migration of breast cancer patients with diabetes

AUTHORS:
Asis Bala[1] iD
Motlalepula G. Matsabisa[1]

AFFILIATION:
[1]Department of Pharmacology, University of the Free State, Bloemfontein, South Africa

CORRESPONDENCE TO:
Asis Bala

EMAIL:
BalaA@ufs.ac.za

According to the World Health Organization report[1], breast cancer is the most prevalent cancer in women worldwide and its incidence is significantly increasing, particularly in developing countries where the majority of cases are diagnosed in the later stages of cancer. Survivors of breast cancer have a higher chance of developing both type 1 and type 2 diabetes, specifically in the post-menopausal stage of life.[2] Conversely, female individuals with diabetes have a 20% greater chance of developing post-menopausal breast cancer than those without.[2,3] The breast cancer incidence is rising in South Africa, where some women are diagnosed late and, as a result, have poor outcomes.[3] The association of type 2 diabetes mellitus and breast cancer among African women has already been established.[4] Therefore, the relationship between insulin and breast cancer cell progression and migration should be further explored.

The South African government has yet to legalise medical cannabis; however, much debate on the subject has taken place in the last few years. The requirement for adequate research on *Cannabis sativa* L. – the variety of cannabis or marijuana found in South Africa – is periodically mentioned in local scientific publications such as the *South African Medical Journal*.[5-7] Interestingly, *C. sativa* has been found to be equally effective in inhibiting breast cancer cell proliferation and regulating insulin levels in patients with diabetes.[8,9]

The current global market for cannabis products is USD3 billion; this figure is expected to rise to USD56 billion within 10 years, as more countries and US states join the legalisation trend.[10] South Africa is one of the world's largest producers and exporters of quality cannabis, and is the main regional hub for cannabis cultivation and trading. All African nations are capable of cultivating medicinal marijuana – of the 'hemp' and 'dagga' variety – to exploit the growing international acceptance of cannabis-based medicines.[11] However, there is an apparent misidentification and confusion regarding the quality of 'hemp' and 'dagga', as *C. sativa* is known in southern Africa.

Here we summarise the regulatory role of insulin in MAO-A and IL-6/IL-6R-regulated metastasis and angiogenesis in breast cancer in patients with diabetes and the possible importance of the South African variety of *C. sativa* is outlined as a possible beneficial research option in this context. The pharmacological efficiencies of *C. sativa* in inhibiting breast cancer cell progression and migration will support the medicinal application and will act as an initial step towards its industrialisation, thereby contributing to the South African bio-economy through its R&D.

Regulatory role of insulin in breast cancer in patients with diabetes

Targeting the tumour vasculature, rather than directly targeting tumour cells, has been the more recent approach to inhibiting tumour growth in all invasive and metastatic cancers, including breast cancer.[12] The growth of breast cancer cells is under the direct regulation of hormones, growth factors, chemokines and their receptors.[13] Both hyperinsulinemia and cancer are extremely prevalent pathophysiological conditions associated with major morbidity and mortality worldwide. The implication of insulin in the pathogenesis of cancer has been in English medical literature since the early 1970s[14]; however, the possible molecular mechanisms of insulin's effect on cancer development and/or progression remain enigmatic.

Insulin acts as a major mitogen in normal mammary gland tissue as well as in breast cancer cells.[15] Many human breast cancers overexpress the insulin receptor, and overexpression of these receptors can result in a transformed phenotype in human mammary epithelial cells.[14] A recent study has demonstrated that women with insulin-receptor-positive breast cancer have worse prognoses in term of tumour progression than women with insulin-receptor-negative tumours. Recent findings indicate that fasting insulin levels predict disease-free survival and overall survival of women with breast cancer who are receiving adjuvant therapy.[14,15] In most human oestrogen-receptor- and progesterone-receptor-positive breast cancer cell lines, like MCF-7, insulin stimulates cell growth via its own receptor; however, insulin, at concentrations as high as 100 nM, has no effect on cell growth in triple negative breast cancer cells line like MDA-MB231, meriting further studies in this context.[16]

In contrast, insulin enhances the metabolic capacity of cancer cells by dual regulation of the glycolytic enzyme pyruvate kinase M2.[17] Insulin and insulin-like growth factor 1 also induce vascular endothelial growth factor mRNA expression via different signalling pathways including PI3-K and MAPK.[18,19] IL-6 is a cytokine involved in a number of cellular inflammatory disease conditions.[19] Growing evidence suggests that IL-6 is also involved in cancer progression, metastasis, chemo-resistance and angiogenesis through the activation of IL-6R.[20]

Monoamine oxidase (MAO) has two isoforms: MAO-A and MAO-B, which are generally present in the outer mitochondrial membrane. MAO-A levels are significantly lower in multiple invasive and metastatic cancer types including breast cancer. A recent study highlighted the differential expression of IL-6/IL-6R and MAO-A in the regulation of invasion/angiogenesis in breast cancer.[19,20] In addition, this IL-6R regulation of MAO-A occurs through cellular MAP kinase and Akt pathways.[20]

Other findings demonstrate that regulation of monoamine levels by MAO activity in β cells is pivotal for physiological insulin in type 2 diabetes.[21] In contrast, MAO inhibitors are capable of both potentiating and inhibiting insulin release

in vitro, depending on their concentrations. It is further concluded that the stimulation of glucose-mediated insulin secretion may be related to the MAO inhibitory effects of the drugs.[21] MAO inhibition has also been reported in a patient with type 1 diabetes and depression.[22] Reduced MAO activity in blood from insulin-dependent subjects with diabetes increased the possibility of tumour cell progression and invasion in breast cancer. Although hypoxia is a common clinical condition in the tumour environment in which it promotes cell invasion, metastasis and angiogenesis, it inhibits insulin-signalling pathways.[21,23] Therefore, the regulatory role of insulin in breast cancer cell progression in persons with and those without diabetes needs to be delineated; development of the possible therapy outlined here will provide new insight into the immunotherapy of breast cancer.

Importance of South African *Cannabis sativa* L.

Medical cannabis is a drug that comes from two species of cannabis, *C. sativa* and *C. indica*. *C. sativa* is known commonly as hemp in the northern hemisphere and as dagga or marijuana in South Africa.[10] The main active chemicals in cannabis are cannabidiol (CBD) and delta-9 tetrahydrocannabinol (THC). However, flavonoids and terpenes have been found to be equally effective in acting synergistically with CBD and THC in biological systems.[24] South African dagga (*C. sativa*) contains higher amounts of psychoactive THC than of CBD whereas hemp contains higher amounts of non-psychoactive CBD than of THC.[11,25]

Cannabinoids could be the first non-toxic agent to show promise in treating metastatic forms of breast cancer.[26] CBD inhibits the survival of both oestrogen-receptor-positive and oestrogen-receptor-negative breast cancer cell lines and induces apoptosis in a concentration-dependent manner.[26,27] Many cannabinoids mediate their effects by binding to endocannabinoid receptors like CB1, CB2 or the vanilloid receptor; however, it is indicated that the effect of CBDs is independent of these receptors.[26] Therefore, scientists are always trying to identify the receptor(s) through which CBD mediates its anticancer effects. The recently established immune biological aspect of IL-6R-regulated and MAO-A-mediated breast cancer cell progression and migration will be the immediate future target of investigations into the effect of cannabis in this context.

For people with obesity and type 2 diabetes, endocannabinoid imbalance is a key regulator.[8] THCs are more agonistic than are CBDs to endocannabinoid receptors and can control blood insulin levels and energy homeostasis. Over-regulation of the endocannabinoid system leads to excessive visceral fat accumulation in the stomach and reduces adiponectin release from the tissues, thus leading to insulin insensitivity and increases in the risk of developing type 2 diabetes.[27] Unlike THC, CBD can help to suppress appetite and help individuals to rebalance the endocannabinoid system by increasing fat breakdown, increasing mitochondrial activities and promoting metabolism, and decreasing fat storage, and is thus beneficial for diabetes management.[8,28]

Conclusion

Breast cancer cell progression and migration is directly influenced by insulin levels in patients with diabetes and IL-6/IL-6R mediates its regulatory role in MAO-A-controlled cell metastasis and angiogenesis. Importantly, the IKS Research Group of the Department of Pharmacology at the University of the Free State (Bloemfontein, South Africa) is presently working on this immunobiology and validating this dual role of cannabis in the South African context.

The CBD-enriched variety of *C. sativa* found in South Africa may offer new hope for the regulation of insulin action on MAO-A- and IL-6/IL6R-regulated metastasis and angiogenesis in breast cancer of patients with diabetes, as shown in Figure 1. The pharmacological efficiencies of *C. sativa* in inhibiting breast cancer cell progression and migration will further support its medicinal application and will lead to further industrialisation, thereby contributing to the South African bio-economy in the future.
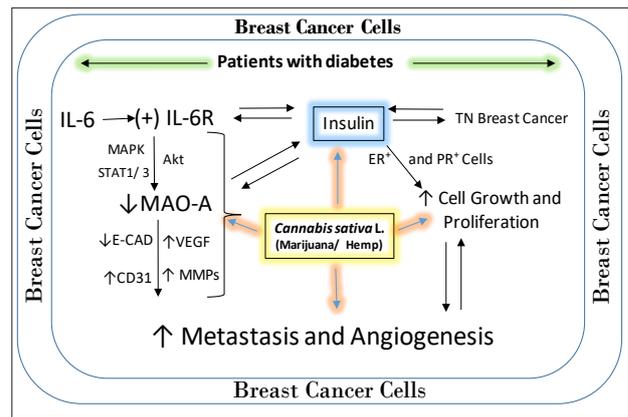
Figure 1: Regulatory role of insulin in cancer cell progression and migration in breast cancer patients with diabetes and the possible targets of *Cannabis sativa* L.

References

1. World Health Organization. GLOBOCAN 2008 update: Cancer prevalence estimates are now available [document on the Internet]. c2011 [cited 2018 Jul 05]. Available from: https://www.iarc.fr/en/media-centre/iarcnews/2011/globocan2008-prev.php

2. Autier P, Koechlin A, Robertson C, Bota M, Boniol M, Leroith D, et al. Breast cancer and diabetes. J Clin Oncol. 2013;31(15 suppl), e12507.

3. Joffe M, Ayeni O, Norris SA, McCormack VA, Ruff P, Das I. Barriers to early presentation of breast cancer among women in Soweto, South Africa. PLoS ONE. 2018;13(2), e0192071, 16 pages. https://doi.org/10.1371/journal.pone.0192071

4. Samson ME, Adams SA, Orekoya O, Hebert JR. Understanding the association of type 2 diabetes mellitus in breast cancer among African American and European American populations in South Carolina. J Racial Ethn Health Disparities. 2016;3(3):546–554. https://doi.org/10.1007/s40615-015-0173-0

5. Farham B. The cannabis debate: Let's do the research. S Afr Med J. 2016;106(9):833. https://doi.org/10.7196/SAMJ.2016.v106i9.11354

6. Parry CDH, Myers BJ. Legalizing medical use of cannabis in South Africa: Is the empirical evidence sufficient to support policy shifts in this direction? S Afr Med J. 2014;104(6):399–400. https://doi.org/10.7196/SAMJ.8135

7. Van Niekerk JP. Medical marijuana and beyond. S Afr Med J. 2014;104(6):387. https://doi.org/10.7196/SAMJ.8335

8. Penner EA, Buettner H, Mittleman MA. The impact of marijuana use on glucose, insulin, and insulin resistance among US adults. Am J Med. 2013;126(7):583–589. https://doi.org/10.1016/j.amjmed.2013.03.002

9. Velasco G, Sanchez C, Guzman M. Towards the use of cannabinoids as antitumor agents. Nat Rev Cancer. 2012;12:436–444. https://doi.org/10.1038/nrc3247

10. Hall J. Medical marijuana a huge opportunity for Africa. IOL News. 2017 October 11. Available from: https://www.iol.co.za/news/opinion/medical-marijuana-a-huge-opportunity-for-africa-11550668

11. Herbst MC. Fact sheet and position statement on cannabis. Pretoria: Cancer Association of South Africa (CANSA); 2012.

12. Shi W, Yin Y, Wang Y, Zhang B, Tan P, Jiang T, et al. A tissue factor-cascade-targeted strategy to tumor vasculature: A combination of EGFP-EGF1 conjugation nanoparticles with photodynamic therapy. Oncotarget. 2017;8(19):32212–32227. http://dx.doi.org/10.18632/oncotarget.12922

13. Papa V, Pezzino V, Costantino A, Belfiore A, Giuffrida D, Frittitta L, et al. Elevated insulin receptor content in human breast cancer. J Clin Invest. 1990;86(5):1503–1510.

14. Chappell J, Leitner JW, Solomon S, Golovchenko I, Goalstone ML, Draznin B. Effect of insulin on cell cycle progression in MCF-7 breast cancer cells. J Biol Chem. 2001;276(41):38023–38028. http://dx.doi.org/10.1074/jbc.M104416200

15. Gliozzo B, Sung CK, Scalia P, Papa V, Frasca F, Sciacca L, et al. Insulin-stimulated cell growth in insulin receptor substrate-1–deficient ZR/75/1 cells is mediated by a phosphatidylinositol/3-kinase-independent pathway. J Cell Biochem. 1998;70:268–280.

16. Costantino A, Milazzo G, Giorgino F, Russo P, Goldfine ID, Vigneri R, et al. Insulin-resistant MDA-MB231 human breast cancer cells contain a tyrosine kinase inhibiting activity. Mol Endocrinol. 1993;7(12):1667–1676.

17. Iqbal MA, Siddiqui FA, Gupta V, Chattopadhyay S, Gopinath P, Kumar B, et al. Insulin enhances metabolic capacities of cancer cells by dual regulation of glycolytic enzyme pyruvate kinase M2. Mol Cancer. 2013;12, Art. #72, 12 pages. https://doi.org/10.1186/1476-4598-12-72

18. Miele C, Rochford JJ, Filippa N, Giorgetti-Peraldi S, Van Obberghen E. Insulin and insulin-like growth factor-i induce vascular endothelial growth factor mRNA expression via different signaling pathways. J Biol Chem. 2000;275:21695–21702. https://doi.org/10.1074/jbc.M000805200

19. Scheller J, Chalaris A, Schmidt-Arras D, Rose-John S. The pro- and anti-inflammatory properties of the cytokine interleukin-6. Mol Cell Res. 2011;1813:878–888.

20. Bharti R, Dey G, Das AK, Mandal M. Differential expression of IL-6/IL-6R and MAO-A regulates invasion/angiogenesis in breast cancer. Brit J Cancer. 2018;118:1442–1452. https://doi.org/10.1038/s41416-018-0078-x

21. Aleyassine H, Gardiner RJ. Dual action of antidepressant drugs (MAO inhibitors) on insulin release. Endocrinology. 1975;96(3):702–710. https://doi.org/10.1210/endo-96-3-702

22. Hamlin E, Neptune M. Monoamine oxidase inhibition in a patient with type 1 diabetes and depression. J Diabetes Sci Technol. 2016;10(5):1203–1204. https://doi.org/10.1177/1932296816638106

23. Regazzetti C, Peraldi P, Grémeaux T, Najem-Lendom R, Ben-Sahra I, Cormont M, et al. Hypoxia decreases insulin signaling pathways in adipocytes. Diabetes. 2009;58(1):95–103. https://doi.org/10.2337/db08-0457

24. Alexander A, Smith PF, Rosengren R.J. Cannabinoids in the treatment of cancer. Cancer Lett. 2009;285:6–12. https://doi.org/10.1016/j.canlet.2009.04.005

25. Christelle MA, Jean-Francois H, Gea G. *Cannabis sativa*: The plant of the thousand and one molecules. Front Plant Sci. 2016;7–19.

26. Bandana C, Ravi J, Ganju RK. Cannabinoids as therapeutic agents in cancer: Current status and future implications. Oncotarget. 2014;5(15):5852–5872.

27. Shrivastava A, Kuzontkoski PM, Groopman JE, Prasad A. Cannabidiol induces programmed cell death in breast cancer cells by coordinating the cross-talk between apoptosis and autophagy. 2011;10(7):1161–1172. https://doi.org/10.1158/1535-7163.MCT-10-1100

28. Lianne. Diabetes and marijuana: A possible treatment? [webpage on the Internet]. c2017 [updated 2018 Mar 28; cited 2018 June 28]. Available from: https://www.thediabetescouncil.com/diabetes-and-marijuana-a-possible-treatment/

AUTHORS:
Jolandie Myburgh[1]
Maryna Steyn[2]
Maciej Henneberg[3]

AFFILIATIONS:
[1]Department of Anatomy, University of Pretoria, Pretoria, South Africa

[2]School of Anatomical Sciences, University of the Witwatersrand, Johannesburg, South Africa

[3]Biological Anthropology and Comparative Anatomy Research Unit, University of Adelaide, Adelaide, Australia

CORRESPONDENCE TO:
Jolandie Myburgh

EMAIL:
jolandie.myburgh@up.ac.za

Comments on Mpeta et al. (2018): Black living standards in South Africa before democracy

Mpeta et al.[1] recently published a paper in which they used secular trends in stature of South Africans to make inferences about living standards. The use of documentation of living standards and the fluctuations thereof as a result of economic growth to explain the secular changes observed in height is a viable approach. The authors gave an excellent review of the history of South Africa and the differences in living conditions that existed between the population groups. However, we believe the biological data were incorrectly used, in order to support inferences regarding the socio-economic situation. The average stature of a population group indeed has a direct relationship with the living conditions and the per capita income of the individuals. For example, decreased stature may be an indicator of nutritional deprivation as a result of a lower income.[2] However, it is important to note that there are many other factors that can influence stature. The size and shape of human bodies vary considerably among population groups across the world and body size is not fixed. Changes in height are continually taking place.[3-5] In this study the authors repeatedly imply that the living standards can be explained by differences and changes in height, e.g. 'by analysing the mean height…, we shed light on the standard of living' (p. 1); 'black living standards as measured by height' (p. 1). Height should not be used simplistically to indicate the standard of living or socio-economic status as it is influenced by too many other variables such as climate, diet, genetics/gene flow, family size and urbanisation.[3,4,6-9] Rather, standard of living should be used, as one of many possible factors, to *explain* the observed differences in height.

Furthermore, the authors did not take into account the normal biological variation and adaptation in height. It is rightfully stated that 'genetic characteristics explain about 80% of stature variation' because heritability is responsible for 75–90% of stature variation[10-13]; however, this factor is not taken into account in this paper even though it explains the majority of the differences in height. The authors state that 'black men are 7 cm shorter than their white counterparts' and that 'this is one of the highest within country differences in the world, but we do not know when it emerged' (p. 1). They further confuse the possible explanations for this difference by asking if genetic inheritance perhaps contributes to the patterns, but then state that the question cannot be answered as gene-level understanding is incomplete. The authors lack an understanding of the genetic variation in South African population groups and the influence of genetic variation on the heights of different population groups. They attempted to answer this question by using societies with similar genetic inheritance, that is, North and South Koreans. North and South Koreans share a similar original gene pool and the differences in their height could possibly largely be explained by differences in living standards. Although this assumption is plausible, they then continue by comparing the height gap in the Koreans to those seen in the South African population groups. Because black and white South Africans do not share a common ancestral gene pool, this comparison is redundant. The genetic origins of South African population groups are well documented. White South Africans are largely descended from colonial immigrants such as Dutch, French, German and other European groups with low frequencies of alleles typically found in Khoesan and Bantu-speaking individuals.[14-16] Population groups from the northwest of Europe (e.g. the Netherlands) are said to be the tallest individuals in the world[2,3] and their genes within white South Africans account for the taller statures seen in this group. The black South African population mainly arose from Bantu-speaking individuals from the Nigerian/Cameroonian highlands with considerable gene flow from Khoesan groups.[17,18] The Khoesan are among the shortest population groups in the world, which may have contributed to the shorter statures observed in black South African groups.[2,17,18] Therefore, the 'gap' between the stature of various South Africans is not solely an indicator of differences in standards of living but rather genetic differences. Similarly, this comparison is made between Māori and white New Zealanders who do not share a common ancestral gene pool. The convergence in stature is mainly explained by the authors as the result of improvement of living conditions and implementation of social policies without the possible effect of adaptation to similar climates or gene flow being considered. Numerous studies[19-24] have observed changes in height with a decrease or increase in socio-economic status. However, it is important to ensure that the different genetic origins of the population group are considered when comparisons are made.

The use of different types of data sets (as was the case in the Mpeta et al.[1] paper) to make comparisons needs to be approached with caution and all biases, limitations and all other possible influences need to be considered. Although the patterns in height may be accurately represented, direct comparisons in height using data sets obtained in different ways should not be made because of the inherent errors associated with each type of data set. The authors state that the World War II data set possibly had a 'preference for men of a certain height' but that 'the use of average height is unaffected' if the 'truncation is not skewed' (p. 2). Making use of a data set that pre-selects individuals of a certain height renders all comparisons with other data sets that do not select for this variable impossible. The World War II data are therefore biased towards a certain height which is not necessarily represented by other data sets. Similar bias arises when the Cape Mounted Police and South African Constabulary data are used. The use of cadaver heights or lengths from the Raymond A. Dart Collection (University of the Witwatersrand) is particularly concerning. These heights are notoriously inaccurate and are often nothing more than estimates by the morgue staff – as the authors themselves state, one finds many unlikely statures (e.g. a stature of 139 cm) among these records. Although the authors mention skeletons (p. 3) when they refer to the Dart Collection data, we assume they have used lengths recorded in the cadaver books and not statures reconstructed from skeletons. The number of bodies donated to the Dart Collection significantly increased from 1958 onwards, especially among white groups. Whereas the black South Africans in the collection were mostly unclaimed bodies, and thus presumably of lower socio-economic status, the white South African bodies were often donated and of higher socio-economic status. Additionally, the well-known measurement difference of approximately 2.5 cm between cadavers and living stature[25] is not mentioned as a

possible misrepresentation of height. Lastly, the sources of the data from the Demographic and Health Survey and National Income and Dynamic Study are based on surveys and the height values need to be approached with caution as a result of known errors in the accuracy and reliability of using self-reported height data.[26]

It is disappointing to note that the authors did not refer to any of the numerous studies on stature and secular trends that have previously been done in South Africa. Different secular patterns in height were reported during the late 19th and early 20th century in black and white South African groups. Negative and null secular trends have been observed by Kark[27], Tobias[28-31], Tobias and Netscher[32], Price et al.[33], and Louw and Henneberg[34], while limited positive trends were observed by Tobias[35], Henneberg and van den Berg[36], Steyn and Smith[37] and Myburgh[38]. For these studies, different data sets with various birth cohorts were used in order to show the different patterns in secular trends. The results from these papers explain the differences in stature between the population groups and provide theories on the possible reasons for the different patterns of change in height. Mpeta et al.[1] state that the observed decline in the stature of white men and the increase seen in black men may be a result of insufficient sample sizes. However, previous studies using larger sample sizes have observed a similar trend.[34,36,38,39] Henneberg[40] also observed instances in which the secular trend did not follow the socio-economic change in the country. He suggested that other factors, for example causative agents which affect the relevant stature-determining genes, may also be responsible for secular changes in height.[40,41] Therefore, it would appear that the major determinant of the direction and magnitude of secular changes may not only be because of the general improvement of living conditions, but rather distinct factors which are specific to the population group.[40,41] This indicates that other factors – for example urbanisation and the associated change in diet, levels of physical labour and access to health care – may also have played a role in both white and black South African groups.

We would also like to comment on terminology; the terms 'whites', 'blacks' and 'white stature' are commonly used throughout the Mpeta et al.[1] paper. These are sensitive and emotionally charged terms and are not scientifically correct. The correct terminology would be, for example, white South Africans, black South Africans or white/black males/females. As M.S. is associated with the School of Anatomical Sciences at which the Raymond A. Dart Collection is housed, we would like to express our disappointment that the use of the collection was not mentioned in their Acknowledgements.

In summary, we believe that the Mpeta et al.[1] paper would have benefited from a better understanding of biology and human variation. While their attempts to make inferences on past living standards are to be lauded, their conclusions are simplistic and tend to follow popular beliefs rather than scientific facts. Secular trends within the same group may be indicative of economic change, but trends between groups need first a genetic explanation before other variables can be considered. The reality is far more complex, as was also demonstrated by Myburgh[38] and Myburgh et al.[39], and needs in-depth analysis.

References

1. Mpeta B, Fourie J, Inwood K. Black living standards in South Africa before democracy: New evidence from height. S Afr J Sci. 2018;114(1/2), Art. #2017-0052, 8 pages. https://doi.org/10.17159/sajs.2018/20170052

2. Steckel RH. Stature and the standard of living. J Econ Lit. 1995:1903–1940.

3. Eveleth PB, Tanner JM. Worldwide variation in human growth. New York: Cambridge University Press; 1976.

4. Bogin B. Patterns of human growth. 2nd ed. Cambridge, UK: Cambridge University Press; 1999.

5. Ruff CB. Variation in human body size and shape. Ann Rev Anthropol. 2002;31: 211–232. https://doi.org/10.1146/annurev.anthro.31.040402.085407

6. Mascie-Taylor CGN, Boldsen JL. Regional and social analysis of height variation in a contemporary British sample. Ann Hum Biol. 1985;12(4):315–324. https://doi.org/10.1080/03014468500007841

7. Panek S, Piasecki E. Nowa Huta: Integration of the population in the light of anthropological data. Materialy i Prace Antropologiczne. 1971;80:1–249.

8. Bielicki T, Welon Z. Growth data as indicators of social inequalities: The case of Poland. Am J Phys Anthropol. 1982;25(S3):153–167. https://doi.org/10.1002/ajpa.1330250509

9. Matsumoto K. Secular acceleration of growth in height in Japanese and its social background. Ann Hum Biol. 1982;9(5):399–410. https://doi.org/10.1080/03014468200005941

10. Phillips K, Matheny AP. Quantitative genetic analysis of longitudinal trends in height: Preliminary results from the Louisville Twin Study. Acta Genet Med Gemello. 1990;38(2):143–163. https://doi.org/10.1017/S0001566000005389

11. Carmichael M, McGue M. A cross-sectional examination of height, weight, and body mass index in adult twins. J Gerontol A Biol Sci Med Sci. 1995;50(4):B237–B244. https://doi.org/10.1093/gerona/50A.4.B237

12. Preece MA. The genetic contribution to stature. Horm Res Paediatr. 1996;45(2):56–58. https://doi.org/10.1159/000184849

13. Silventoinen K, Kaprio J, Lahelma E, Koskenvuo M. Relative effect of genetic and environmental factors on body height: Differences across birth cohorts among Finnish men and women. Am J Public Health. 2000;90(4):627–630. https://doi.org/10.2105/AJPH.90.4.627

14. Steyn M, İşcan MY. Sexual dimorphism in the crania and mandibles of South African whites. Forensic Sci Int. 1998;98(1):9–16. https://doi.org/10.1016/S0379-0738(98)00120-0

15. Greeff JM. Deconstructing Jaco: Genetic heritage of an Afrikaner. Ann Hum Genet. 2007;71:674–688. https://doi.org/10.1111/j.1469-1809.2007.00363.x

16. L'Abbé EN, Van Rooyen C, Nawrocki SP, Becker PJ. An evaluation of non-metric cranial traits used to estimate ancestry in a South African sample. Forensic Sci Int. 2011;209(1–3):195.e1–195.e7. https://doi.org/10.1016/j.forsciint.2011.04.002

17. Herbert RK. The sociohistory of clicks in Southern Bantu. Anthropol Linguistics. 1990:295–315.

18. Stynder DD. Craniometric evidence for South African Later Stone Age herders and hunter gatherers being a single biological population. J Archaeol Sci. 2009;36(3):798–806. https://doi.org/10.1016/j.jas.2008.11.001

19. Bogin B, MacVean RB. Growth status of non-agrarian, semi-urban living Indians in Guatemala. Hum Biol. 1984;56:527–538.

20. Komlos J. Stature and nutrition in the Habsburg monarchy: The standard of living and economic development in the eighteenth century. Am Hist Rev. 1985;90(5):1149–1161. https://doi.org/10.2307/1859662

21. Floud R, Gregory A, Wachter K. Height, health and history: Nutritional status in the United Kingdom, 1750–1980. Cambridge, UK: Cambridge University Press; 1990. https://doi.org/10.1017/CBO9780511983245

22. Hauspie RC, Vercauteren M, Susanne C. Secular changes in growth. Horm Res. 1996;45(2):8–17. https://doi.org/10.1159/000184841

23. Frisancho AR, Guilding N, Tanner S. Growth of leg length is reflected in socioeconomic differences. Acta Med Auxol. 2001;33:47–50.

24. Bogin B, Scheffler C, Hermanussen M. Global effects of income and income inequality on adult height and sexual dimorphism in height. Am J Hum Biol. 2017;29(2), e22980. https://doi.org/10.1002/ajhb.22980

25. Trotter M, Gleser G. Estimation of stature from long bones of American whites and Negroes. Am J Phys Anthropol. 1952;10:469–514. https://doi.org/10.1002/ajpa.1330100407

26. Rowland ML. Self-reported weight and height. Am J Clin Nutrition. 1990;52(6):1125–1133. https://doi.org/10.1093/ajcn/52.6.1125

27. Kark SL. Patterns of health and nutrition in South African Bantu [MD thesis]. Johannesburg: University of the Witwatersrand; 1954.

28. Tobias PV. Anthropometry among disadvantaged peoples: Studies in Southern Africa. In: Watts ES, Johnston FE, Lasker GW, editors. Biosocial interrelations in population adaptation. The Hague: Mouton; 1975. p. 287–305.

29. Tobias PV. The negative secular trend. J Hum Evol. 1985;14(4):347–356. https://doi.org/10.1016/S0047-2484(85)80041-5

30. Tobias PV. Adult stature in southern African Negroes – further evidence on the absence of a positive secular trend. S Afr Med J. 1990;78:97–101.

31. Tobias PV. New evidence on the absence of a positive secular trend towards increased adult stature in South African Negroes. In: Bellwood PS, editor. Man and his culture – A resurgence. New Delhi: Books and Books; 1992.

32. Tobias PV, Netscher D. Reversal of the secular trend, as evidenced by South African Negro crania and femora. Hum Biol. 1977;59:467–475.

33. Price B, Cameron N, Tobias PV. A further search for a secular trend of adult body size in South African blacks: Evidence from the femur and tibia. Hum Biol. 1987;59:467–475.

34. Louw GJ, Henneberg M. Lack of secular trend in adult stature in white South African males born between 1954 and 1975. HOMO J Comp Hum Biol. 1997;48(1):54–61.

35. Tobias PV. Stature and secular trend among South African Negroes and San (Bushmen). S Afr J Med Sci. 1975;40:145–164.

36. Henneberg M, Van den Berg ER. Test of socioeconomic causation of secular trend: Stature changes among favored and oppressed South Africans are parallel. Am J Phys Anthropol. 1990;83(4):459–465. https://doi.org/10.1002/ajpa.1330830407

37. Steyn M, Smith JR. Interpretation of ante-mortem stature estimates in South Africans. Forensic Sci Int. 2007;71(2–3):97–102. https://doi.org/10.1016/j.forsciint.2006.10.006

38. Myburgh J. Limb proportions in South Africans: Secular changes population differences and implications for stature estimation [dissertation]. Pretoria: University of Pretoria; 2016.

39. Myburgh J, Staub K, Rühli FJ, Smith JR, Steyn M. Secular trends in stature of late 20th century white South Africans and two European populations. HOMO J Comp Hum Biol. 2017;68(6):433–439. https://doi.org/10.1016/j.jchb.2017.10.001

40. Henneberg M. Secular trends in body height – Indicator of general improvement. In: Living conditions or of a change in specific factors? In: Dasgupta P, Hauspie R, editors. Perspectives in human growth, development and maturation. Dordrecht: Springer; 2001. p. 159–167. https://doi.org/10.1007/978-94-015-9801-9_14

41. Henneberg M. Possible causes of secular trends in body size – lessons from the southern hemisphere. In: Schultz M, Atzwanger K, Braüer G, Christiansen K, Forster J, Greil H, et al, editors. HOMO – unsere Herkunft und Zukunft. Göttingen: Cuvillier Verlag; 2001. p. 234–237.

AUTHOR:
Christo Boshoff[1]

AFFILIATION:
[1]Department of Business
Management, Stellenbosch
University, Stellenbosch,
South Africa

CORRESPONDENCE TO:
Christo Boshoff

EMAIL:
cboshoff@sun.ac.za

Chris Callaghan's criticism of the National Research Foundation's rating methodology: A rebuttal

No scientific endeavour – its methodologies, processes, procedures or anything related to it – should *ever* be above reproach, critical evaluation or reconsideration, including the methodology of South Africa's National Research Foundation (NRF) for rating those researchers who apply for a rating. Chris Callaghan's[1] views on the NRF rating methodology must thus be welcomed.

However, any such criticism must be fair, balanced, objective, properly justified and uncontaminated by personal grievances.

Callaghan's critical review of the NRF rating methodology falls short on a number of these grounds. Firstly, as a management *scientist*, he should have done a far better job of understanding the methodology before embarking on a critical review. Secondly, his recommendations for improvement unfortunately fall foul of the same criticism that he levels at the current NRF methodology. Thirdly, the entire rating system cannot reasonably be completely wrongheaded, with no positive consequences at all. Fourthly, he ought to have been more forthright with his readers, the reviewers of his paper and the editor of the *South African Journal of Science* by declaring his personal experience with the rating methodology. My rebuttal of Callaghan's criticism will be structured around these points, but limited to the domain of Management Sciences.

Key features of the NRF rating process

It must be noted at the outset that the NRF rating methodology is a peer review methodology. The entire process is thus constructed on this key feature and must be understood in that context.

A more careful review of the NRF rating process would have revealed that each applicant submits six names of potential reviewers to the NRF. The members of the Specialist Committee or expert panel (the members are appointed for 4-year terms) then select three of these nominated reviewers. To this list, a further three independent reviewers not nominated by the applicant are added to review the application (Callaghan's view that 'the NRF system works through reviewers chosen by the person being rated' is clearly not entirely correct). At least six reviewers are thus asked to review each applicant.

All reviewers are, of course, expected to be objective and fair in their evaluation, regardless of who nominated them. The nominated reviewers are carefully selected to ensure that they have the required knowledge and expertise to conduct the evaluation. In order to eliminate potential bias, the application form explicitly asks the applicants about their relationship with the nominated reviewer, and this explanation is screened to exclude cases in which the relationship is considered to be too close (supervisor, PhD student, research team colleague or 'life-long friends' in Callaghan's parlance). In a further attempt to preclude bias, applicants may ask the panel not to send their application to reviewers who they feel may not be objective in their assessment.

Once the peer reviewers' reports have been returned to the NRF, they are screened for suitability by the Specialist Committee, an Assessor and a Chairperson. Excessively negative and excessively positive reviewer reports are discarded as potentially marked by bias. The reasons for the rejection are formalised for auditing purposes and to counteract the 'gatekeeping' phenomenon.

The members of the Specialist Committee then consider the peer reviewers' reports and attempt to reach consensus on what an applicant's rating should be. Once the evaluations and suggested ratings of the Specialist Committee have been concluded (all members must read all the reviewers' reports), an independent Assessor, who is uninformed about what the Specialist Committee's recommendations are, enters the meeting. The Assessor will also have been tasked to review all the applications *independently* before the meeting. The Specialist Committee and the Assessor then reach consensus on the appropriate rating for each applicant (including, of course, a 'rating unsuccessful' decision, if deemed appropriate). Once this consultation has been completed, an independent Chairperson, who has also *independently* evaluated the applications of all the applicants, enters the meeting. The Chairperson's suggested ratings are then compared with those of the Specialist Committee and the Assessor; and in most cases agreement is reached on an applicant's appropriate rating. If not, the applications are referred to an Executive Evaluation Committee for review. The members of the Executive Evaluation Committee are the six Chairs of the different evaluations panels, two Convenors of the Specialist Committees, and three NRF executives, including the Deputy CEO of the NRF (who also chairs the meeting).

Clearly, despite the potential failings of human judgement, significant effort is built into the NRF rating methodology to minimise biased evaluations. With the considered involvement of six reviewers, four to eight Specialist Committee members, one Assessor and one Chairperson, no individual can manipulate the assessment in order, in Callaghan's words, to 'settle scores'. A rating outcome is thus not the decision of 'a small group of evaluators' only, as Callaghan contends. The 'power abuses' he refers to are simply not possible. In addition, if the rating outcome is considered inappropriate, the aggrieved applicant has the right to appeal.

Typically, an NRF rating application is thus reviewed by at least 12 different evaluators. It is just not credible that all 12 would be consistently biased against any individual applicant, even though the evaluations are not anonymous. It is thus difficult to accept Callaghan's suggestion that the process is prone to excessive subjectivity bias, especially if one considers that he regards the journal peer-review process – which typically consists of only

two reviewers and an editor – as superior because it is 'systematic'. I want to argue that the current NRF methodology is just as systematic, step-by-step, as the journal peer-review process.

In stating that a rating decision depends on a 'handful of reviewers' only, Callaghan demonstrates his deficient grasp of the NRF methodology. This fundamental misapprehension casts serious doubt on the validity of his entire assessment.

Callaghan's contention that the NRF's rating methodology is discriminatory also calls for scrutiny, as we must distinguish between discrimination and *unfair* discrimination. The sports scoreboard *discriminates* between the winner and the loser. The editor *discriminates* between the manuscript that will be published and the one that will not be published. The Nobel Prize committee *discriminates* between the recipient of the Nobel Prize and the nominees who do not win it. The football referee has *power* to decide on a penalty or not. The editor has *power* to decide on the suitability of a manuscript. The Nobel Prize committee has *power* to decide who receives the Nobel Prize. 'Discrimination', power and the associated hierarchies of distinction are facts of life. Why should academia be any different? Forms of 'discrimination' will thus be true of any rating system, regardless of the methodology used.

Callaghan's specific claim that the NRF rating methodology unfairly discriminates against non-white researchers (for which he offers no supporting evidence) is contradicted by the fact that 52% of scientific papers published in 2013/2014 were published by non-white authors[2] (the figure today may be even higher).

His contention that the NRF methodology discriminates against those who are unable to form 'personal relationships' borders on the laughable. The rating methodology does not require any applicant to have a 'relationship' with any nominated reviewer. There is no requirement of social skills – it only requires applicants to nominate experts who work in their field of endeavour and who are knowledgeable enough to judge the quality and impact of their research. It is not a matter of 'having connections'.

The argument that the NRF methodology is 'elitist' calls for scrutiny. If scholars enter a system of evaluation, are subject to the same system of evaluation and do not end up being evaluated as equal, is that a sign of elitism? Do *all* athletes participating in the Olympic Games expect to make good on their desire to win? Do *all* football teams competing in the FIFA World Cup expect to share the trophy? Do all students expect to receive the same marks for an examination? Is it elitist that a student receives a degree *cum laude*? Clearly not. Any sort of evaluation, by definition, implies differentiation; but does differentiation necessarily imply elitism? Callaghan's argument that the differentiation induced by the NRF rating methodology is elitist is simply not credible.

Callaghan's proposals and suggestions

Callaghan's proposals and suggestions for an alternative rating mechanism are, by their very nature, contradictory. Whatever he proposes will still be flawed (his term) by the use of power, 'discrimination' between those who do better than others and the inevitable creation of hierarchies. Whatever the rating system in place, some applicants will do better than others. No matter what he suggests, his preferred system would be as susceptible to the same criticism as the current system. Furthermore, from a practical execution point of view, some of his suggestions are simply unworkable.

To suggest that the so-called bibliometrics (h-indices in particular) are more reliable and valid because he believes that they are more objective must be questioned. Ultimately, bibliometrics also rely on peer review and human judgement. For instance, it cannot be denied that some journal reviewers explicitly favour or reject certain methodologies. It is not uncommon to find reviewers who openly state that they will not recommend papers for publication that are based on, say, Bayesian statistics, or panel data or PLS analyses. Callaghan at least acknowledges the impact of 'paradigm beliefs'; so, by implication, he acknowledges that perfect objectivity is not possible.

In any case, to get the 'wide stakeholder consensus' among scholars that he calls for in the Management Sciences on the ideal parameters seems implausible. Questions proliferate: Under whose auspices would this 'wide stakeholder consensus' be sought? Who are the stakeholders? Who will represent the stakeholders? Which indices should be included in the assessment – all of them? Should different weights be assigned to different h-indices? Should the same combination of h-indices be used for all sub-disciplines? Any attempt to introduce bibliometrics as the basis for an alternative rating system would falter and become bogged down during the first roundtable meeting, not to mention reaching consensus across disciplines. So the 'wide stakeholder consensus' he suggests is but a pipe dream. Another problem with the bibliometrics suggestion is that a sizable h-index is built up over many years. Its use will thus favour older, well-established researchers and prejudice the younger, emerging researchers on whose behalf Callaghan claims he speaks. Having said that, there is nothing in the current NRF methodology that prevents applicants from including their h-indices in an application.

Callaghan's blanket assumption that the volume of citations provides an indication of a scholar's impact on a research domain is based on the supposition that the impact was necessarily 'positive' and that this scholarship has always made a significant contribution to the chosen domain. This is simply not true. The work of many scholars is cited for 'negative' reasons: methodologies that are flawed, results that are questionable, interpretations that are poor or not justified, and the like. In other words, some authors are widely cited because their scholarship is dubious; so they rise in the citation indices for the wrong reasons.

The idea that the review process of a scholarly journal is superior to the NRF methodology because it is 'objective' or 'unbiased' because it is anonymous is also not beyond reproach. In a small academic community such as ours in South Africa, and with very few journals in which to publish, complete anonymity cannot be guaranteed. Some authors, research units and departments specialise in certain academic domains and have developed a market reputation for their specialisation. Some universities and research institutions may have specialised facilities no other institution has and only they can publish papers on particular topics. A quick Google search is not the exclusive domain of NRF reviewers. Concealing such associations may be impossible under every circumstance.

Furthermore, it is not uncommon even for international journals to ask authors to nominate potential reviewers for their papers, which also produces a degree of subjectivity bias or collegial patronage. So for Callaghan to argue that subjectivity bias is exclusive to the NRF rating methodology is simply not valid.

Callaghan contends that 'harm' has been done by the NRF methodology without explaining what that 'harm' involved or who was harmed. More importantly, he provides no evidence whatsoever of the 'harm' caused. In the absence of any scientific evidence being provided, one must assume that the criticism is no more than his own unsubstantiated opinion.

Are bibliometrics what they are cracked up to be?

The use of bibliometrics to evaluate the research impact of individual researchers and as performance indicators of research institutions is not without its critics. Hicks et al.[3] believe that the use of research metrics has become too widespread to ignore its negative consequences. This criticism can be divided into two broad categories. Firstly is that, by using bibliometrics, scholars cede their right to peer review to data, and more particularly to the data of private sector vendors who do not have the capabilities to use it appropriately.[3,4] In the words of Laloë and Mosseri[5]: '...the implementation often seems to arise from a loss of critical and rational mind'.

From a methodological point of view, Weingart[4], in assessing the introduction of bibliometrics in the UK to assess the research performance of research institutions, concluded: 'Bibliometric measures, although quantitative and therefore seemingly objective, appeared to be theoretically unfounded, empirically crude, and dependent on data that were known to be imprecise.' The Web of Science database, on which many evaluations have been based, was not initially constructed as a data source for

research performance evaluation but '...as a literature databank designed to identify uses of knowledge and networks of researchers...'[4]. Since the commercialisation of the Web of Science database, the previous costly measures to clean the data have been stopped, leading Weingart to caution against '...the uncritical embrace of bibliometric measures'[4].

Callaghan's argument that chasing a favourable NRF rating (he refers to 'gaming' – that is, 'research is conducted for the express purpose of meeting the goals of a system') leads to the publication of poor research that is hardly read by anyone does not consider that the same reactive behaviour may be produced by bibliometrics.[5] Referring to impact factors, Lawrence points out: 'It has evolved to become an end in itself – the driving force for scientists to improve their reputation…'.[6]

The validity of bibliometrics such as h-indices depends heavily on capturing the correct data, cleaning the data, skilled employees who prepare the data, proper data-processing and proper categorisation.[7,8] The use of h-indices has been found wanting on these criteria by several authors. For instance, bibliometrics have been criticised for data processing errors[4,7,8], that they collect data from only certain journals[4], that they ignore research published in books[5] and that no distinction is drawn between the credit received by the author of a sole-authored paper and the tenth author on a multi-authored paper[5].

Ironically, a further criticism made by Weingart of the validity of biblio-metric measures is the problem of the accurate categorisation of the collected data: 'In particular, *interdisciplinary* fields present a problem to proper categorization'[4]. Clearly, poor delineation between disciplines leads to mistakes in citation counts.[4]

It is clear that, in the eyes of many, the relationship between bibliometrics indices such as an h-index on the one hand, and research quality on the other hand, is tenuous to say the very least, and even unscientific.[5] More specifically, such indices are clearly not the 'strictly objective evaluation' Callaghan seems to believe they are.

Multidisciplinary and transdisciplinary research

Callaghan argues that NRF rating applicants who are involved in multi-disciplinary and transdisciplinary work are penalised ('discriminated against') by the current methodology. Again, no evidence is provided to support the contention. In the case of multidisciplinary and/or transdisciplinary applications, an applicant has several options. One is to clearly position the application by focusing on the areas in which the applicant believes the most significant impact was evident. The application form has a field where the candidate can indicate whether the research is of a multidisciplinary nature, and can identify the focus of the research during the last 8 years. The second consideration is to choose the reviewers carefully to ensure that enough expertise is available among them to encompass the ambit of the researcher's research.

Callaghan tries to make the point that researchers who have 'changed trajectory' during a later phase of their research careers are prejudiced by the NRF rating methodology. This accusation is difficult to accept if one considers that the review period is the last 8 years. The narrative section of the application form should be used to point out this change of direction to the reviewers, and what impact it has had on their research portfolio. Again, the work is reviewed by peers for its significance and impact, and not for the volume of output, as is commonly believed.

In any event, the narrative part of the rating application form offers ample opportunity for the applicant to explain the potential uniqueness of the research to reviewers.

Also disputable is Callaghan's contention that monodisciplinary research is of little value and is hardly read, while multidisciplinary and trans-disciplinary work (such as his own) is more valuable in the service of solving real-world problems. This is obviously a heresy.

Callaghan's contention that 'this rating system [is] acting as a catalyst to create a culture of competition which differentiates *publicly* between "winners" and "losers"' (p. 2) is somewhat misleading. Although the names of those who receive a rating are published on the NRF website, only the broad rating category (e.g. B) appears on the NRF website.

It is not clear how the names of those whose rating applications were unsuccessful are publicly identified.

A more balanced assessment of the NRF rating methodology

Without question, the outcomes of the NRF rating methodology are based on the perceptions, attitudes and subsequent judgements of the people involved: the Specialist Committee, the Assessor and the Chairperson. These evaluators are involved in every single rating decision. The same applies to the Appeal Committee, which handles disputes and appeals.

The NRF methodology is:

- not anonymous – it is based on past research output

- voluntary

- a peer-review process

- qualitative and subjective in nature

- as valid a measure of any researcher's peer-reviewed assessment of past research output as is humanly possible

Neither the developers nor the NRF has ever tried to pretend that it is anything else.

Its purpose is to evaluate the significance and impact of a scholar's sustained work over an extended period of time. Unlike a single paper, it evaluates the quality and impact of a 'body' of output and its significance. And, unfortunately, it is impossible to evaluate past (published) work without identifying the author. As Tijsen et al.[9] point out, excellence is by definition a matter of the *ex ante* assessment or the *ex post* evaluation of research performance.

If universities use the NRF rating as a measure of research progress and the recognition of academic standing across disciplines, it means that the system is accepted as reasonably reliable and valid. Fedderke's empirical results confirm the validity of the NRF rating methodology by stating: 'Scholars with higher NRF ratings record higher performance on average against the objective measures of absolute output and the impact of their research, than scholars at lower rating'[10]. These results obviously confirm the broad validity of the rating methodology, and seem to contradict Callaghan's point about inconsistency 'across individuals'.

Callaghan's assessment of the NRF methodology is insistently negative. A more balanced assessment would have uncovered some positives as well. Against this background, one must note that the per capita output at South African universities has increased from 0.51 in 2006 to 0.88 in 2015.[2] I argue that the NRF rating system and its associated incentives have played a significant role in this substantial improvement. The number of research publications produced by South African universities has grown from 5540 in 1994 to 15 542 in 2014 – a threefold increase in a decade (information supplied by the Department of Higher Education and Training). Again, I argue that the NRF rating system and its associated incentives have played a significant role in this substantial improvement.

Callaghan's own interaction with the NRF

To reach a balanced assessment of Callaghan's criticism of the NRF rating methodology, his readers would need to know whether he himself has applied for a rating, given his strong views of how flawed the system is. If he *has* applied, one would reasonably like to know why he applied, given his abhorrence of an unfair, discriminatory system that does no good and only causes harm.

If he did *not* apply, presumably for the reasons cited in his paper, this would add credibility to his critique. Either way, for someone who claims the moral high ground by 'having the courage to speak to power', he ought to have declared his own interaction (or otherwise) with the NRF rating methodology to all readers of his paper. This information is essential in helping readers to judge Callaghan's contribution to the debate on the validity of the NRF rating system.

In summary

Callaghan challenges the validity of the NRF rating methodology, but his own criticism lacks the validity he calls for as a result of his poor understanding of the entire rating process. A more thoughtful analysis of the rating methodology would have acknowledged its purpose and *raison d'être*, and would have revealed the considered checks-and-balances that are in place to minimise – and even avoid – the subjectivity bias to which Callaghan attributes it.

His 'analysis' is flawed because of his poor understanding of the process; he makes sweeping statements (about harm done, subjectivity, gamification, research that is not read, unfair discrimination) without offering any evidence to support these claims.

His recommendations, which suggest that the volume of citations and the peer reviews of journals will lead to bibliometric indices that are unbiased, valid measures of researchers' impact on a research domain, are not without their own limitations. Furthermore, his recommendations will not overcome his own criticism of elitism, the supposed favouring of the 'Ivy League' universities and his view that the process leads to unfair discrimination.

Callaghan contradicts himself by suggesting an alternative rating system. He advocates alternative, better measures and procedures to 'rate' researchers; but any methodology is 'elitist' by definition, and leads to a hierarchy founded on forms of differentiation. If his suggestions were to be implemented, the productive 'elites' would simply get to the top via a different route. What he regards as undue power will simply shift from 12 NRF-commissioned evaluators to (usually) two reviewers and an editor and finally to the questionable database of a commercial vendor. The evaluative hierarchy would not disappear: it might even be constrained and consolidated into an even tighter cluster of authority.

If one argues that fair discrimination between different levels of academic performance is unacceptable, then the evaluative judgement of scholarly performance is, by its very nature, the unethical use of 'power', and that the resultant hierarchies are unacceptable (because we should all be equal), then Callaghan's proposals and suggestions are not the answer. In fact, there is clearly only one answer. Extending Callaghan's arguments on elitism, power (ab)use and hierarchies, *any* method or form of 'rating' would be immoral.

I want to argue that the NRF rating system has been a huge success. It has been widely accepted by the academic community. The number of applications received by the NRF since 1984 is 6744; currently 3889 researchers hold a valid rating, of whom 196 are rated researchers in the domains of economics, management, accountancy and public administration (information supplied by the NRF). From anecdotal evidence – and many personal conversations – I know that an NRF rating has proved aspirational to many and that the prospect of a (higher) rating has motivated many researchers, both at universities and at other research institutions.

While we should all strive to find a valid and reliable measure of the standing of a researcher, it is beyond question that such evaluations should be fair, balanced and uncontaminated by personal considerations and emotions.

Finally, a rating should not be construed as a barrier: it should be seen as an objective means to demonstrate your progress as a researcher; and I want to encourage all scholars, especially young academics and researchers, to take up the challenge rather than to wallow in a pit of complaining and blaming others for 'injustices'.

Future debate

Callaghan calls for 'further research and discussions' on this topic. Some of his criticism of the NRF methodology clearly raises a number of further questions:

1. Can a voluntary system violate academic freedom?

2. Can only multidisciplinary and transdisciplinary research be to the benefit of societal stakeholders?

3. Is monodisciplinary research by definition inferior, leading to 'wasteful publications'?

4. Why are journal reviewers 'knowledgeable peers' but NRF reviewers are not?

5. Will bibliometrics demonstrate the benefit of research to societal research?

Addressing these contentious arguments in Callaghan's review could form the basis for future debate.

References

1. Callaghan C. A review of South Africa's National Research Foundation's rating methodology from a social science perspective. S Afr J Sci. 2018;114(3/4), Art. #2017-0344, 7 pages. https://doi.org/10.17159/sajs.2018/20170344

2. Centre for Research on Evaluation, Science and Technology, Stellenbosch University. SA Knowledgebase [database]. No date [cited 2018 Jul 02].

3. Hicks D, Wouters P, Waltman L, De Rijcke S, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. Nature. 2015;520(7548):429–431. http://dx.doi.org/10.1038/520429a

4. Weingart P. Impact of bibliometrics upon the science system: Inadvertent consequences? Scientometrics. 2005;62(1):117–131. https://doi.org/10.1007/s11192-005-0007-7

5. Laloë F, Mosseri R. Bibliometric evaluation of individual researchers: Not even right… Not even wrong. Europhys News. 2009;40(5):26–29. https://doi.org/10.1051/epn/2009704

6. Lawrence PA. Rank injustice. Nature. 2003;415:835–836. https://doi.org/10.1038/415835a

7. Braun T, Glanzel W, Schubert A. How balanced is the Science Citation Index's journal coverage? A preliminary overview of macrolevel statistical data. In: Cronin B, Atkins HB, editors. WoK: A Festchrift in Honor of Eugene Garfield. Medford, NJ: Information Today Inc. & The American Society for Information Science; 2000. p. 251–277.

8. Zitt M, Ramanana-Rahary S, Bassecoulard E. Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. Scientometrics. 2003;56(2):259–282. https://doi.org/10.1023/A:1021923329277

9. Tijssen RJW, Visser MS, Van Leeuwen TN. Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference. Scientometrics. 2002;54(3):381–397. https://doi.org/10.1023/A:1016082432660

10. Fedderke J. The objectivity of National Research Foundation peer review based ratings in South Africa. ERSA Working Paper 300 [document on the Internet]. c2012 [cited 2018 Jul 02]. Available from: https://econrsa.org/system/files/publications/working_papers/wp300.pdf

Review of innovations in the South African collection industry

AUTHORS:
Vinzent Blaschczok[1] iD
Tanja Verster[1] iD
Alan Broderick[2]

AFFILIATIONS:
[1]Centre for BMI, North West University, Potchefstroom, South Africa

[2]Shapiro Shaik Defries & Associates, Johannesburg, South Africa

CORRESPONDENCE TO:
Tanja Verster

EMAIL:
Tanja.Verster@nwu.ac.za

The objective of this review was to provide an overview of new developments and innovations within the collections industry that could possible enhance the performance of collection agencies, specifically in South Africa. A literature study was conducted to determine current practices in the collections industry, as well as possible future innovations. A significant trend identified throughout the literature study was the increasing prioritisation of automated digital communication in several aspects of debt collection. It is reasonable to assume that this trend will continue to become the industry standard. Four recommendations are made based on the findings of the literature study. Firstly, South African collection agencies should investigate the feasibility of developing an app-based solution to performing collections. Secondly, collection agencies should supplement traditional modelling techniques with other tools, such as those developed in the field of machine learning. Thirdly, collection agencies could consider using speech analytics to obtain insights into call centre agents' performance and adherence to business rules. Lastly, the usage of social media data in collections as well as credit risk modelling in general is recommended as a topic for future study.

Significance:

- A review of the various techniques currently employed in the field of debt collections may serve as useful reference for both academics and those working in debt collections.

- Recommendations are provided to assist businesses in aligning the operational models of their debt collection units to industry best practice.

- Topics for future research in this crucial sector of the economy, which brings together such fields as risk governance, predictive modelling, human psychology, debt management, legal compliance and business analysis, are provided.

Introduction

Collections agencies normally act as external debt collecting agencies for a diverse selection of institutions that may, during the normal course of their operations, encounter defaulters. These defaulters have breached their legal requirements and failed to fully repay the debt extended to them. This scenario may have several adverse effects on the company – the two most direct being a financial loss and the opportunity cost of the capital extended to the debtor. Debt collection is the process of trying to minimise these losses by recouping some of the debt. Companies may choose to make use of internal or external debt collection strategies, or a combination of the two. It is usual that external debt collectors receive those debtors who have proven unresponsive to internal collection efforts. Debtors referred to external collectors usually differ significantly in their repayment characteristics from those who respond to internal efforts. It has been shown that the difference can be as extreme as 10% non-payers for internal collections versus 83% for external collections, and 30% full repayments versus 0.7%, respectively.[1] Thus, with such a low probability of collecting on outstanding debt, it is critical that the collections process be as efficient as possible for external debt collectors to operate profitably.

As the purpose of this paper was to explore current best practices and possible future developments in the collections industry, as per the academic and business literature, the main research topics addressed are:

- traditional and innovative methods of creating debt collections queues;

- innovative strategies employed by leading debt collections agencies to optimise their operational profitability;

- methods for assessing and comparing the performance of different collections strategies;

- data sources not currently used by collections agencies in South Africa, which might have significant predictive power; and

- machine-learning techniques capable of modelling and improving the collections process.

We aimed to provide South African collection agencies with innovative thinking and strategies. From a brief review of the collection process, we identified four areas that became the focus of our research: (1) innovative uses of data sources in collections; (2) innovative uses of machine learning (and other statistical modelling techniques) in collections; (3) innovative ways to enhance collection strategies; and (4) innovative ways to evaluate collection strategies.

The collection process

The collections process is described briefly. Debtors purchase products or services on credit. Those who default on their repayment and are not collected on by internal collections are amalgamated into debtor lists. These lists are then provided to collection agencies which combine the information contained therein with additional data

sources to create collections queues. Collections agents are assigned to these queues with the aim of recovering as much of the debt as possible. Collection agencies typically levy a commission on the recovered debt and pay over the remainder to their clients.

The collection process is demonstrated visually in Figure 1. Our research will focus on the four main operational phases in this collection process.
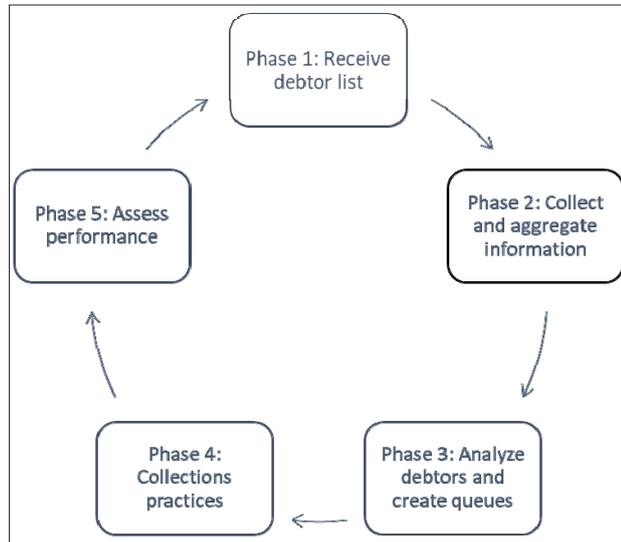

Figure 1: Operational phases of a debt collections agency.

Phase 1: Receiving debtor list

The first of these, receiving the debtor lists, is not elaborated upon specifically, but research pertaining to the other phases was conducted.

Phase 2: Collecting and aggregating information

This critical operational phase pertains to the collection and aggregation of information on those individuals on the debtor list. Several data sources can be used during this phase. Traditional data sources include data from the clients and credit bureaus. Additional data sources are summarised below.

Phase 3: Analysing debtors and creating collection queues

This phase involves the analysis of debtors and then creation of collection queues. Based on a specific collection strategy, a call agent can be assigned to a specific group of debtors, that is, a queue. This queue may also include a prioritisation of debtors.

In this phase, the data collected during the previous stage are used as inputs to a chosen model. Typically this model should return scores of individual debtors. This score should, at a minimum, provide a collection agency with estimates of the likelihood of obtaining a promise-to-pay, the conversion thereof to recovery-rate and expected amount collected, the likelihood of zero payment as well as expected collections cost and the expected collections period. It is important to note that the above aspects may be modelled separately or jointly, and that several different models can be used, and each has its own respective strengths and weaknesses. Consequently, traditional methods can be complemented by techniques from other fields such as machine learning. The collection queues are then created using the scores as generated by the model. Logically, higher scores should be associated with better repayment behaviour and thus higher operational profit.

Phase 4: Collections practices at the operational level

During this phase, the output from the previous phase is used to assign collection queues to agents, who attempt to recover as much of the debt as possible. The collections strategy becomes effective during this phase, as it determines the communications method used and provides guidelines for agent interaction with debtors. A holistic collections strategy might also contain recommendations regarding write-off of debtors, as well as the method whereby agents are assigned to queues.

Phase 5: Evaluating collections strategies and queue building models

This phase occurs concurrently with all the other phases and involves the monitoring of each phase whilst identifying possible irregularities and inefficient strategies. The performance of strategies and models should be assessed regularly and corrective measures taken as and when necessary. Collecting data on strategy performance will allow management to make data-driven business decisions and optimise operational efficiency on many fronts. It is critical that the evaluation methods be established up front and be applied consistently.

As our aim was to identify emerging practices, our research was not restricted specifically to the collections industry. Several possible new strategies and ideas emerged in the broader financial sector. It is reasonable to assume that companies might also be hesitant to publish extensive technical details pertaining to their internal systems, as considerable resources are spent developing these systems and companies will want to protect their investments. Business and academic publications were used, as they tend to have different focuses: academic publications contain more theoretical and technical details whilst business publications focus on operational implementations and business results. Case studies were preferred to highly theoretical publications as these address 'real-life' scenarios and report proven results. Note also that these case studies – specifically those from vendors – are perhaps less reliable, as they are not subject to peer review and exist primarily to serve as marketing tools; the results reported can reasonably be assumed to be biased and selectively chosen to reflect positively on the respective business.

In each of these four operational phases, specific research topics were identified as shown in Table 1. The rest of the paper will be presented around the research topics listed in Table 1.

Table 1: Overview of research topic per operational phase

Operational phase	Research topic
Phase 2: Collecting and aggregating information	Innovative uses of data sources in collections
Phase 3: Analysing debtors and creating collection queues	Innovative uses of machine learning (and other statistical modelling techniques) in collections
Phase 4: Collections practices at the operational level	Innovative ways to enhance collection strategies
Phase 5: Evaluating collections strategies and queue-building models	Innovative ways of evaluating collection strategies (in such a way that comparison between different strategies can be objectively made)

Innovative uses of data sources in collections

During the research process, several different data sources were identified that could be used in the collections industry. The main data sources identified are listed in Table 2. It is important to consider that not all these data might have significant predictive power, and it is recommended that the predictive power be investigated before inclusion in the queue-building models to prevent unnecessary inflation of the data's dimensionality. The usability of these data in terms of current legislation should also be considered – there are rich sources of data which are precluded from use in collections strategies in terms of existing legislation.

Table 2:   Innovative uses of data sources in collections

Data source	Possible application in collections
Social media data	Not used extensively in South Africa yet, but many case studies in the USA have shown possible extensions of standard models to include these data in collections[2]
Geo-location data	Can be used to identify areas with significant numbers of debtors; to identify possible relationships between location and repayment behaviour; and to identify which areas are connecting with a specific collection agency's social media
Device data	A collections strategy could be linked to device used, i.e. those who prefer mobile devices might be more responsive to telephonic or phone text communication, whilst desktop users might prefer email correspondence
Data brokers	Many case studies in the USA[2] have shown possible extensions of standard models to include the extensive types of data that data brokers can provide into predictive modelling (e.g. collections)

Social media data vendors, like Datasift and GNIP, provide services such as aggregating data across multiple social platforms (Facebook, LinkedIn, Twitter, etc.). Communication with Datasift suggests that privacy laws restrict the use of individuals' data. Only publicly available data may be used under these laws and the ethical implications thereof should be considered. Datasift[3] offers the provision of data starting at costs of USD15 000 per month. GNIP[4], a daughter company of Twitter, provides packages of one million tweets, with prices starting at USD1250. A possible application is to link a collections strategy to debtor interaction with social media, i.e. the debtors might be more receptive after interaction with agent correspondence.

Step-by-step guides[5] have been published which describe methods to search and extract the social media data based on the criteria specified by individuals on popular social networks like LinkedIn, Facebook and Twitter. Ultimately, the goal is still to link customer behaviour with propensity to pay and selection of collections strategy.[5] Web structure mining, web content mining and web usage mining are three methods discussed in Pottharst[6] to extract data from any site in general.

Another innovative use of data in collections is the utilisation of *geo-location data* (inferred from cell or mobile phone data, or from Facebook etc.). Possible applications of geo-location analysis to collection agencies include: identifying areas with significant numbers of debtors; identifying possible relationships between location and repayment behaviour; and identifying which areas align with a specific collection agency's social media footprint.

The possible use of *device data* should also be investigated as inputs to the queue-building models. A collections strategy could be linked to preferred device, i.e. those who prefer mobile devices might be more responsive to app-based communication whilst desktop users might prefer more traditional methods such as email – although this is rapidly changing as emails are equally accessible on modern mobile devices. Another possible application of social media data and device data, is to investigate the relationship between the time of day when communication occurs and likelihood of obtaining a promise-to-pay.

Data brokers, also known as data aggregators, collect data on individuals from several domains. A comprehensive overview of data brokers was compiled by Clearinghouse[7], and includes the following major data brokers based in the USA: Acxiom, Corelogic, Datalogix, eBureau, ID Analytics, Intelius, PeekYou, Rapleaf and Recorded Future. Similar companies operating in South Africa are Background Check International, Dracore and Blue Label Data Solution. Table 3 gives an overview of data provided by data brokers as well as the possible application to debt collection. Table 3 also includes some examples of the data in each data domain. The use of data brokers is already an existing practice in South African collection agencies (with compliance being an integral consideration), but the use of some data categories, specifically online behavioural data, is limited or undocumented. Many case studies in the

USA have shown possible extensions of standard models to include this type of data, e.g. an overview of data-mining techniques applied to analysis of social media data is provided by Injadat et al.[2]

In sum, collection agencies could investigate augmenting their existing data with additional data sources, which include social media data and other online behaviour, geo-location data, device data, and data provided by data brokers. These data sources can be used to augment the scoring of debtors and the determination of appropriate collections strategies.

Innovative uses of machine learning in collections

Scorecards, such as those used for credit card applications, have traditionally employed the statistical method known as logistic regression. An introduction to the methodology commonly used during scorecard building is provided by Siddiqi[8]. Scorecards have some similarity to collection queues as both involve estimates of an individual's ability to pay. Logistic regression does, however, have very specific weaknesses (e.g. assumption of linear relationship between the logit of the target variable and other explanatory variables).

Machine-learning techniques are frequently used to overcome the limitations of logistic regression. Challenger models, based on machine-learning techniques, might outperform traditional techniques when the assumptions of those traditional techniques are not appropriate.

Currently, most collection agencies in South Africa use either expert knowledge or logistic regression or a combination to model collection aspects. For example, in an industry directed research project that was done for MBD in 2007, a collections scoring model was developed using logistic regression.[9] Another industry directed research project that was done for Absa Bank in 2012 also used logistic regression to develop a collection scorecard.[10] An additional example is a project done for Shapiro Shaik Defries & Associates in 2017 which also revealed (via the personal observation of A.B.) that this perception exists in South Africa. The same observation was made by DMC Debt Management (Myburgh S, 2018, email communication, February 1), that is, that most of the collection agencies in South Africa use either expert knowledge or logistic regression (or a combination).

The field of machine learning has seen a significant growth in industry interest over the past years.[2] Machine-learning techniques are applied to social media data in the literature and it is reported that the unstructured nature of the data, as well as some applications which require unsupervised learning, were significant contributors to the choice of using machine learning.[2] The most frequently employed techniques were support vector machines, classification trees and Bayesian networks. An important result to note here is that no single machine-learning technique will consistently outperform all others in all contexts. A summary of innovative uses of machine learning and other statistical modelling techniques is provided in Table 4.

It has been argued[23] that the use of machine learning has many possible advantages for debt collectors. Several considerations when building collection queues are raised by Davey[24] which could be easily solved by using machine-learning techniques. For example, a typical trend experienced by collectors is that debtors increasingly prefer to repay their debt over long periods of time rather than once-off or over short periods. Such a repayment structure requires the modelling of payment streams, in both amounts and timing. Consequently, the models used to score debtors need to take the collections period into account, as it is a significant factor affecting operational profitability (increased costs and time value of money). One of the abilities of machine-learning techniques is to find deep patterns in the data, which humans cannot reasonably be expected to identify. A critical point raised is the need to recalculate debtors' scores as regularly as possible[24], as new data might become available which are not accounted for by older scores. Real-world results reported by Davey[24] for companies using First Data's PREDIGY® which is based on machine-learning techniques include a financial institution which reduced their losses on delinquent accounts by USD2.1 million and a management agency which increased their collection agents' performance by 18%. Note that these specific results do not convey any details on the specific machine-learning technique used, but generically refer to it as 'strategic predictive analytics'.

Table 3: Overview of data provided by data brokers

Data domain	Application to debt collection	Examples
Sensitive identity data	Possible, but ethical and legal implications	Social security number Driver's licence number Birth date Birth dates of children in household Birth dates of family members in household
Identifying data	Currently used in collections	Name Previously used names Address Address history Longitude and latitude Phone numbers Email address
Social media and technology data	Still a developing field and not used extensively in South Africa yet, but many case studies in the USA have shown possible extensions of standard models to include these data	Electronics purchases Friend connections Internet connection type Internet provider Level of usage Heavy Facebook user Heavy Twitter user (Twitter user with 250+ friends) Is a member of over five social networks Online influence estimation Operating system Software purchases Type of media posted (text, video, multimedia, etc.) Uploaded pictures Use of long-distance calling services Desktop computer access Use of mobile devices Social media and Internet accounts
Demographic data	Can be used as input to behavioural analytics models or be used to determine communications strategy	Gender Age Marital status
Court and public record data	Can be used as input to behavioural analytics models	Court orders Planned court appearances Criminal record
General interest data	Behavioural analytics and debtor profiling	Online activity Community participation
Financial data	Determine ability to pay	Employment status Length of current employment Income
Vehicle data	No case studies were found which utilised these data, but abstract underlying patterns might exist which could be used as part of the models	Vehicle ownership, including number and type of vehicles Purchase of new or second-hand vehicles
Travel data	Not used in collections currently, but applications should be investigated (e.g. perhaps people who travel extensively should rather be contacted via phone text/email and not via telephone calls)	Holiday club membership Airline loyalty member
Purchase behaviour data	Models to predict wallet prioritisation and models to profile debtors	Month-to-month variance in expenditure Fixed expenditure Blacklist status
Health data	Might determine ability to pay (fitness for work) and wallet prioritisation (e.g. medical bills)	Medical aid membership Age Parental status

Table 4: Innovative uses of machine learning and other statistical modelling techniques in collections

Machine learning / statistical modelling technique	Case study
Artificial neural network	Artificial neural networks were used to assess payment probability of customers[11], and were used for classifying social media data[12].
Bayesian network	Bayesian networks were used[2] in the field of social media data.
Markov chain	Markov chains were used for a 'payment sequence model'[13] which can be used to model repayment behaviour. A Markov decision process was also applied to improve the collections of the New York State Department of Tax and Finance.[14]
Support vector machine	Support vector machines were used to assess payment probability of customers[11], and were used for classifying social media data[12].
Text mining and speech analytics	The usage of text mining by collections agencies to assess future viability of collections is noted by De Almeida Filho et al.[15] The possibility of using speech analytics to assess the vast quantity of data generated by call centres which could not be feasibly done by human labour has been discussed by Ullrich et al.[16]
Classification tree	Classification trees were used to assess payment probability of customers[11], and for classifying social media data[12].
Random forest	Random forests were used to assess the payment probability of customers.[11]
Self-organising map	Self-organising maps combined with Cox-proportional hazard models were used to model repayment patterns of credit card debt.[17]
Multivariate kernel regression	Multivariate kernel regression was used to predict the probability of repayment.[18]
Memory-based reasoning	Memory-based reasoning was used in collections.[19]
Recency, frequency and monetary (RFM) analysis	Collection agencies can investigate applications of RFM analysis to their collections.[20] Recency might be 'How long ago was the last repayment?' Frequency might be 'How frequently does the debtor repay?' Monetary value might be 'Average repayment amount or proportion of debt repaid per repayment'. RFM analysis can then be used to segment debtors and obtain a better understanding of the factors influencing repayment behaviour, as well as as an input to a queue-building model.[21]
Multivariate regression	Collections were modelled with a multivariate regression model.[22]

Additional studies[25] used machine learning and other statistical techniques to increase debt collection. It might be of interest to collection companies to know which variables are typically used. Examples of the variables include financial ratios (e.g. age of account, size of account, size of repayment, liquidity of debtor); qualitative factors (e.g. account's business nature, market outlook); and behavioural factors (e.g. account transactions, balances, wallet priority).

In many cases it might be possible to quantify the monetary implication of a superior model. One such example is an interesting case study in which three machine-learning techniques (artificial neural networks, classification trees and memory-based reasoning) were compared with logistic regression.[19] These techniques were applied to a list of debtors whom the firm had declined to collect on. Their results showed that approximately 35% of these accounts were classified by the best model as 'good', indicating that the firm had probably misclassified a significant proportion of 'good debt' as 'bad debt' in their records. The unrecovered debt associated with these misclassified accounts totalled USD423 000. In this case, the logistic regression outperformed these three machine-learning techniques.

As mentioned in Table 4, multivariate kernel regression has also been used to predict the probability of repayment and infer the strengths of the relationships between the predictor variables and the response.[18] Some of the results included that homeowner status and self-employment status were not significant while bank account holding status was significant. The most interesting result of this study is that intuition might not necessarily be supported by the data. As an example, whilst it might be intuitive to assume that homeowners have better repayment behaviour than non-homeowners, this assumption was not supported in the analysed data. Therefore, it is strongly recommended that South African collection agencies evaluate the predictive performance of

the data used to create collection queues and ensure that expert-based systems are also supported by data.

The specific case study referred to in Table 4 that used a multivariate regression included a thorough analysis of the variables that can predict collections.[22] Data obtained from the German market, relating to more than 150 000 accounts from 78 creditors, were used to build a multivariate regression model. Again, it might be of interest for collection companies to know which variables were used, as a common challenge in modelling is identifying and acquiring appropriate input. In the German study,[22] the main variables that played a role in predicting collections were: exposure at default; age of account (fewer collections on older accounts); previous experience in collections; macroeconomic factors (unemployment rate for country and GDP growth); legal nature of debtor (corporations tend to have better repayment than individuals); intensity of customer complaint (an increase in intensity associated with lower collection rate); repayment method (instalments versus once-off) and bankruptcy (collections on accounts by debtors formally in bankruptcy were significantly lower).

To summarise, we found that different modelling techniques have been successfully used in collections. Some of these techniques include artificial neural networks, support vector machines, classification trees, random forests, Markov chains, text mining, self-organising maps, multivariate kernel regression, 'recency, frequency and monetary' analysis, and multivariate regression models.

Innovative ways to enhance collection strategies

After analysing debtors and creating collection queues, the queues are assigned to agents who then communicate with debtors and attempt to collect as much as possible on the accounts in arrears. A summary of innovative ways to enhance collection strategies is provided in Table 5.

Table 5:   Overview of data provided by data brokers

Innovative ways to enhance collection strategies	Example
Machine learning	Machine-learning models are used to create queues which are assigned to agents using a method dubbed the 'snake-draft method'.[11] A combination of optimisation theory and data modelling has been used[26] to develop two possible approaches to agent scheduling that incorporate random arrival and abandonment rates.
Optimised scheduling	Traditional methods of agent scheduling can be optimised using mixed-integer formulations, e.g. a spreadsheet-based solution[27] was implemented which takes into consideration: call volume during a time period; customer abandonment of the call; business rules determining acceptable waiting times; and staff availability, preference and seniority. This solution produced 15–20% lower aggregate personnel requirements when compared to previous scheduling methods used by the centre at no significant cost to the call centre's performance.
App-based solutions (Avatars)	The stigma surrounding bad debt and delinquent accounts affects the approachability of debtors by collections agents. Avatars have been proposed as a method for circumventing this stigma. Note that the industry leader for avatar-based collections is the AvaPay system developed by BeGuided[28].
Interactive websites (online debt collection)	TrueAccord, a US software company, has developed a collections system that is reported to use behavioural analytics, machine learning and a humanistic approach to automate the collections process. An increase in collections by 20%, as well as a tenfold increase in agent productivity, has been reported.[29]
Real-life stories	It is recommended that real-life stories by debtors who have improved their lives by repaying bad debt should be included in collections strategies. This engagement might ease the communications process with debtors and promote the value of good repayment behaviour.
Ring-less voicemail drops	The DirectDrop Voicemail product[30] allows collections agents to send voicemails to debtors, without having to call. The two main benefits of such a system are that regulations restricting telephonic communication do not apply and debtors are able to access the messages in their own time, without intrusion into their personal or business lives. This approach might improve debtors' perception of collection agencies.

According to Melo[31], a representative of the analytics software company FICO, the collection strategies of the future will make use of digital, automated communication. Melo's results showed that debtors prefer digital, automated interaction to agent-based interaction. Taking this result into consideration, and the fact that South African collection agencies typically perform the majority of their communication through call centres, for which personnel salaries account for up to 70% of collection costs[27], innovative ways to enhance call centres are required (and possible ways are provided in Table 5).

Innovative ways to enhance collection strategies include the use of machine learning, optimised scheduling, app-based solutions, avatars, interactive websites (online debt collection), real-life stories and ringless voicemail drops. When considering these developing trends, it is reasonable to assume that the collection strategies of the future will mostly utilise digital, automated communication.

The impact of automated communication on the call centre industry is recommended as a crucial research topic, as estimates of the number of people employed in call centres in South Africa[32] range from 54 000[33] to 210 000[34]. Studies in the Philippines, home of the world's largest call centre industry, are already reporting reduced growth rates and increasing job complexity for call centre agents as artificial intelligence takes over simple jobs.[35]

Innovative ways of evaluating collection strategies

A collection strategy encompasses the communications strategy, write-off strategy, queue assignment and more. The integration of the collection strategy with debtor scores and, consequently, the dependence on the data sources and models used to generate these scores greatly increase the challenge in fair and practical evaluations. The importance of accurately assessing collection strategies can be summarised by the phrase 'you can't fix what you don't know is broken'. Objective evaluation and comparison of multiple collection strategies is a common challenge faced in the collections industry. Such evaluation requires an assessment of the model's performance and an unbiased comparison of the efficiency of multiple collection strategies. Causes of bias might include the application of different strategies to different debtor demographics or during periods of different macroeconomic circumstances.

There are two factors to consider when evaluating collection strategies. Firstly, before comparing the performance, it is important that macroeconomic factors be taken into account when the collection strategies under comparison are implemented at different times, as macroeconomic factors affect collections.[22] Secondly, it is also crucial that the performance evaluation criteria for a strategy be specified in advance, before the strategy is implemented.

Some innovative ways to evaluate strategies are given next. The first innovative way to evaluate such strategies is the use of speech analytics software[36] to analyse the large volume of telephonic conversations and associated data which are generated by call centres[16,37]. Because it is impractical for managers to listen to every phone call, a digital solution capable of identifying the quality of a phone call is a practical alternative. Two companies which offer commercial products capable of speech analytics are Nexidia's software analytics and CallMiner. We recommend that South African collection agencies investigate the usage of speech analytics to analyse debtor–agent interaction and identify key phrases associated with good and bad repayment behaviour. Different approaches to speech analytics are available.[38] As significant research has been done in text analytics, the usage of speech-to-text software is proposed as a method to obtain the necessary data for analysing automated collection strategies. Secondly, a specific best practice when evaluating a collection strategy is the use of a champion–challenger strategy.

Based on the literature, it is clear that separate evaluation of the critical components mentioned earlier is possible, as described in Table 6.

Table 6: Methods for separate evaluation

Methods for separate evaluation	Proposals and comments
Evaluation of the predictive power of data through the use of variable selection techniques	Automatic variable selection techniques are not recommended – variable selection should include considerations of both predictive power and the cost of obtaining the relevant data and consequently expert interpretation is required. Note that this evaluation should form part of the 'model development' phase.
Evaluate model performance: 'comparing apples with apples'	Apply 'through-the-cycle' versus 'point-in-time' model estimates to compensate for transitory macroeconomic conditions. Establish reference periods for comparing models, i.e. model comparison should be based on the same data.
Challenger models can be used to evaluate the accuracy of the current score-generating model	It is critical that these challenger models be developed on the same data sets.
Evaluating model accuracy through back-testing	The scores generated by the model should be reflective of the profit a debt collector can expect to make by targeting a specific debtor for collections. Back-testing, which is the practice of comparing model predictions to actual values, is a suitable technique for evaluating the model's accuracy. Note that this back-testing should occur at both the portfolio level and debtor level, with consideration of the individual debt amounts. Similar to account-level vs segment-level evaluation and exposure weighted vs number weighted evaluation.
Evaluating agent assignment and communication strategies	It is proposed that the existing collections strategy, specifically the method whereby agents are assigned to queues, and the communications strategy employed can be evaluated analogously to the champion–challenger methodology. The details are as follows: • Identify different segments in an existing collections portfolio. • Randomly divide each segment into two uneven populations (i.e. use an 80–20 split) and apply the current strategy to the large population and the challenger strategy to the small population. • Statistical inference can be used to evaluate whether the challenger strategy results in significant improvements over the current strategy.

Recommendations for South African collection agencies

Innovative uses of data sources in collections

The following is a list of our recommendations with respect to data sources:

• Collection agencies should consider broadening their communications strategies to include social media and digital platforms (such as applications). User interaction with such platforms generates data, which could also be used to profile debtors.

• When debtors are business owners, web-mining techniques can be used to extract data directly from individual websites. Such cases, however, might be sparse, and consequently a cost–benefit analysis of such an undertaking is advised.

• We emphasise that the predictive power and legality of any data need to be carefully assessed before data are included in any model, especially data sources from data brokers who collect data from several different domains, such as corporate (employment history, qualifications, etc.) and civil (dependants, marital status, etc.) records.

• Several social media data vendors are active in the developed world, most notably in the USA and UK. These firms provide data streams of social media data, aggregated across multiple platforms and filtered to an extent. Communication with one of these firms revealed that privacy legislation might impede the usage of such data sources to profile individual debtors. Because of these restrictions, we cannot advocate for the use of social media data, as provided by these vendors, at the current stage. Social media data, as they are currently available, may only provide a general overview of debtor demographics and high-level segmentation (such as analysing debtor repayment behaviour for rural versus urban debtors).

A notable exception is mobile phone data: despite the prevalence of mobile phones, no studies could be found which utilised mobile phone data for collections. Despite this fact, we are of the opinion that South African collection agencies should investigate the use of mobile phone data in this context.

Innovative uses of machine learning in collections

The recommendations regarding debtor analysis and queue building are:

• As traditional techniques can be implemented relatively quickly and at low cost, it is recommended that these be used as a baseline.

• A critical examination of the validity of the assumptions underlying these models is advised. Should these assumptions be violated, remedial measures need to be taken. Should remedial measures not suffice, it is suggested that alternative techniques be investigated.

• When highly non-linear or abstract patterns exist in the data, machine-learning techniques such as artificial neural networks have been shown to outperform traditional techniques.

• No single modelling technique will outperform all others in all contexts: consequently, the usage of champion–challenger strategies is advised.

• Models can be built from the outputs of more than one technique; for example, a classification technique could be used to identify segments of similar debtors, thereafter a logistic regression model could be used to estimate individual debtors' probability of payment. The collections strategy can then be aligned to the debtors' segment, whilst an agent can be allocated according to the estimated probability of payment.

• Existing models should be used, not only for collections, but also for statistical inference so that data-driven understandings of operations can be obtained.

• Collection agencies should investigate offering longer pay-back periods to debtors, in conjunction with the agency's clients.

Innovative ways to enhance collection strategies

The recommendations for collections practices at operational level are as follows:

- Mobile applications have grown in popularity, and studies have shown that an increasing number of consumers use these apps for business. Collections-specific applications already exist and have been well received by debtors; it is reported that these applications also improve collection rates and agent productivity. It is recommended that collection agencies investigate the development of a mobile application for usage by both debtors and agents, as it can provide an easily accessible, innovative and standardised platform for collections and communication.

- Studies have shown that the performance of call centres can be accurately assessed using speech analytics. It is recommended that a cost–benefit analysis be performed to evaluate whether existing software should be licenced for this purpose.

- The usage of avatars for communication purposes has been shown to be effective in reducing the stigma associated with bad debt, thus easing the communications process. It is advised that the usage of avatars for guiding debtors through the collections be investigated experimentally.

- Several different assignment strategies of agents to collections queues are possible, and recommendations in this regard are provided in the section titled 'Innovative ways to enhance collection strategies'.

- Linear programming – a commonly used method in decision support – can be utilised to optimise agent scheduling for call centres.

- Collections strategies of the future will most probably pivot towards digital, automated communication.

Innovative ways to evaluate collection strategies

Our recommendations regarding performance strategies are:

- It is critical that the collections strategy (queue-building model, business rules, etc.) used at any specific time be well documented to allow for fair comparisons of the performance of different collections strategies.

- The metric used to assess the collections strategy's performance – such as collections rate, collections period or collections cost – needs to be specified upfront.

- When assessing the performance of two different collections strategies, it is necessary to make sure that a valid comparison is made. For example, investigate the debtor profiles of the two different collection strategies. Also, the specific time period could be compared to determine a valid comparison (different macro-economic scenarios can make two strategies incompatible). Champion–challenger strategies are commonly used to evaluate and improve model performance.

The implementation of these recommendations will be critical to ensuring that South African collections agencies operate profitably, ethically and in line with international standards.

Conclusions

Research on cutting-edge innovations for collection agencies was summarised. Key phases of the collections cycle were identified, and critical aspects of those phases were researched to determine the latest developments. A comprehensive overview of the research findings was provided and theoretical studies, as well as practical applications through case studies, were included. The understanding of the literature and the collections industry in general was used to identify possible applications of these findings to South African collection agencies and recommendations were provided.

It is clear that the field of collections is a complex one, but several slight improvements in the various aspects of a collection agency's operations can, in aggregate, produce significant improvements to the efficiency and profitability thereof. The collections strategies of the future will most likely mature into digital, automated interaction. The recommendations will most probably find their way into future best practices of the collections industry.

Recommendations for future research

We have identified the following topics as viable future research prospects:

- The development of an operation evaluation strategy, comprising methods for the evaluation of all relevant aspects affecting a collections agency's performance, i.e. evaluation of data sources, score generating models, agent assignment strategies, communication methods, etc.

- The development of industry specific back-testing techniques.

- Investigation of the state of South African debt collections agencies with respect to the recommendations made in this paper.

Authors' contributions

V.B.: conceptualisation; methodology; data collection; data analysis; validation; data curation; writing revisions; project leadership. T.V.: conceptualisation; sample analysis; data analysis; writing the initial draft; writing revisions; student supervision; project management. A.B.: conceptualisation; methodology; project leadership.

References

1. Thomas L, Matuszyk A, Moore A. Comparing debt characteristics and LGD models for different collections policies. Int J Forecast. 2012;1(28):196–203. https://doi.org/10.1016/j.ijforecast.2010.11.004

2. Injadat M, Salo F, Nassif AB. Data mining techniques in social media: A survey. Neurocomputing. 2016;214:654–670. http://dx.doi.org/10.1016/j.neucom.2016.06.045

3. DataSift. Human data intelligence [webpage on the Internet]. c2017 [cited 2017 Sep 12]. Available from: www.datasift.com

4. GNIP. Release the power of social media data [webpage on the Internet]. c2017 [cited 2017 Sep 12]. Available from: www.gnip.com

5. Informatica Marketplace. Extract data from LinkedIn, Facebook, and Twitter [webpage on the Internet]. c2016 [cited 2017 Apr 12]. Available from: https://marketplace.informatica.com/solutions/extract_data_from_linkedin_facebook_and_twitter

6. Pottharst A. Webmining: Examples and applications [webpage on the Internet]. c2017 [cited 2017 Apr 12]. Available from: http://dewiar.staff.gunadarma.ac.id/Downloads/files/11103/Datamining-Webmining_Arne.pdf

7. Clearinghouse PR. Data brokers [webpage on the Internet]. c2013 [cited 2017 Apr 10]. Available from: https://www.privacyrights.org/data-brokers

8. Siddiqi N. Credit risk scorecards: Developing and implementing intelligent credit scoring. Hoboken, NJ: John Wiley & Sons; 2006.

9. De Vries L. Client specific collections scoring model (BMS9105/2008) [mini-dissertation]. Potchefstroom: North-West University; 2008. Available from: http://library.nwu.ac.za/

10. Mdlalana S. Development of a model for predicting customers who accept and stay in arrangement collections (BMS9226/2012) [mini-dissertation]. Potchefstroom: North-West University; 2012. Available from: http://library.nwu.ac.za/

11. Jongmyoung K, Pilsung K. Late payment prediction models for fair allocation of customer contact lists to call center agents. Decis Support Syst. 2016;85:84–101. Available from: http://nemo.yonsei.ac.kr/wp-content/uploads/2016/11/20161107_Late-payment-prediction-models-for-fair-allocation-of-customer-contact-lists-to-call-center-agents.pdf

12. Al-Saggaf Y, Islam MZ. Data mining and privacy of social network sites' users: Implications of the data mining problem. Sci Eng Ethics. 2014;21:941–966. https://doi.org/10.1007/s11948-014-9564-6

13. Thomas LC, Matuszyk A, So M, Mues C, Moore A. Modelling repayment patterns in the collections process for unsecured consumer debt: A case study. Eur J Oper Res. 2016;249:476–486. https://doi.org/10.1016/j.ejor.2015.09.013

14. Abe N, Melville P, Pendus C, Jensen DL, Thomas VP, Bennett JJ, et al. Optimizing debt collections using constrained reinforcement learning. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2010 July 25–28; Washington DC, USA. New York: ACM; 2010. p. 75–84. https://doi.org/10.1145/1835804.1835817

15. De Almeida Filho AT, Mues C, Thomas LC. Optimizing the collections process in consumer debt. Prod Oper Manag. 2010;6:698–708.

16. Ullrich M, McCracken D, Craft M. The delicate balancing act: Using speech analytics to optimize collections and mitigate risk. Waltham, MA: Callminer Inc.; 2010. Available from: https://callminer.com/wp-content/whitepapers/Using-Speech-Analytics-to-Optimize-Collections-and-Mitigate-Risk.pdf

17. Ho Ha S, Krishnan R. Predicting repayment of the credit card debt. Comput Oper Res. 2012;39:765–773. https://doi.org/10.1016/j.cor.2010.10.032

18. Grant C, Madula M. Using bounds to investigate household debt repayment behaviour. Res Econ. 2013;67:336–354. https://doi.org/10.1016/j.rie.2013.09.007

19. Zurada J, Lonial S. Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. J Appl Bus Res. 2005;21(2):37–54. Available from: https://pdfs.semanticscholar.org/c4d5/a5934769a6c9e851004514c01bc95a84becb.pdf

20. Ivanovic S, Mikinac K, Perman L. CRM development in hospitality companies for the purpose of increasing the competitiveness in the tourist market. UTMS J Econ. 2011;2:59–68. Available from: http://utmsjoe.mk/files/Vol.%202%20No.%201/0-2-1-6-IvanovicS-Mikinac.pdf

21. Dursun A, Caber M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism Manage Perspect. 2016;18:153–160. https://doi.org/10.1016/j.tmp.2016.03.001

22. Beck T, Grunert J, Neus W, Walter A. What determines collection rates of debt collections agencies? Finan Rev. 2017;(52):259–279. https://doi.org/10.1111/fire.12131

23. De Jager P. How does data analytics help debt collection [webpage on the Internet]. c2017 [cited 2017 Apr 16]. Available from: http://insights.principa.co.za/how-does-data-analytics-help-debt-collection

24. Davey D. Collect more for less: Strategic predictive analytics is key to growing collections and reducing costs [document on the Internet]. c2009 [cited 2017 Apr 17] Available from: https://www.firstdata.com/downloads/marketing-fs/fd_collectmoreforless_whitepaper.pdf

25. Nadolny M. The use of predictive modelling to boost debt collection [document on the Internet]. c2017 [cited 2017 Apr 17]. Available from: https://www.business-school.ed.ac.uk/waf/crc_archive/2013/116.pdf

26. Bertsimas D, Doan XV. Robust and data-driven approaches to call centres. Eur J Oper Res. 2010;207:1072–1085. https://doi.org/10.1016/j.ejor.2010.05.040

27. Dietz DC. Practical scheduling for call centre operations. Omega. 2011;39:550–557. https://doi.org/10.1016/j.omega.2010.12.001

28. BeGuided Inc. AvaPay: Payment services with a smile [webpage on the Internet]. c2017 [cited 2017 May 31]. Available from: https://www.beguided.com/#avapay

29. TrueAccord. Debt collection, reimagined [webpage on the Internet]. c2017 [cited 2017 May 31]. Available from: https://www.trueaccord.com

30. VoApps Inc. VoApps DirectDrop voicemail home [webpage on the Internet]. c2017 [cited 2017 May 31]. Available from: http://www.voapps.com/

31. Melo D. How mobile data technologies are changing debt collection and transforming businesses [webpage on the Internet]. c2013 [cited 2017 Jun 01]. Available from: http://www.itproportal.com/2013/12/17/how-mobile-technologies-are-changing-debt-collection-and-transforming-businesses/

32. Callrica – Call centre innovation [homepage on the Internet]. c2016 [cited 2017 Sep 29]. Available from: http://www.callrica.co.za/

33. Keeble J. SA Commercial [webpage on the Internet]. c2014 [cited 2017 Sep 27]. Available from: http://www.sa-commercial.co.za/sa-commercial-direct-the-growing-call-centre-sector-in-south-africa/

34. Cohen M. South Africa ideal for call centres, accent 'sounds British'. Mail & Guardian. 2013 December 02; Business [cited 2017 Sep 27]. Available from: https://mg.co.za/article/2013-12-02-south-africa-ripe-for-call-centre-expansion

35. Cusi VS. Mid- to high-level BPO jobs to be in-demand in the Philippines [document on the Internet]. Contact Center World. 2017 September 08 [cited 2017 Sep 2017]. Available from: https://www.contactcenterworld.com/view/contact-center-news/mid-to-high-level-bpo-jobs-to-be-in-demand-in-the-philippines.aspx

36. Kendrick S. The beginner's guide to speech analytics technology [webpage on the Internet]. c2016 [cited 2017 Jun 01]. Available from: https://callminer.com/blog/the-beginners-guide-to-speech-analytics-technology/

37. Davey N, Hill-Wilson M. The speech analytics guide [document on the Internet]. c2017 [cited 2017 Jun 05]. Available from: http://www.nexidia.com/media/1318/nexidia-the-speech-analytics-guide.pdf

38. Mason D, Turner P. Beginner's guide to speech analytics [webpage on the Internet]. c2013 [cited 2017 Jun 01]. Available from: https://www.callcentrehelper.com/beginners-guide-to-speech-analytics-57123.htm

AUTHORS:
Neo Molotja[1] iD
Gerard Ralphs[1] iD

AFFILIATION:
[1]Centre for Science, Technology and Innovation Indicators, Human Sciences Research Council, Cape Town, South Africa

CORRESPONDENCE TO:
Gerard Ralphs

EMAIL:
gralphs@hsrc.ac.za

A critical review of social sciences and humanities R&D expenditure in South Africa, 2005–2014

Expenditure on research and experimental development in the social sciences and humanities (SSH) in South Africa has almost doubled over the past decade. However, fine-grained analysis of patterns of R&D expenditure in SSH research fields over the period 2005/2006–2014/2015 reveals a number of critical issues for both institutional planning and national policymaking. We demonstrate that most SSH R&D expenditure in the 10-year reference period was targeted predominantly within just a few research fields: finance, economics, education, accounting and political science and public policy. By contrast, investment in SSH research fields such as architecture and habitat, media and communication studies, psychology, and transportation studies was strikingly low in the same period, with some research fields, such as dance or tourism, appearing to be at risk of decline. Using these R&D data as a proxy, we argue, principally, that institutional R&D planners and national policymakers need to find a greater balance between current priorities and future needs, if SSH R&D is to be 'leveraged' for larger socio-economic impacts, as is being envisaged in a new draft White Paper on Science, Technology and Innovation.

Significance:

- R&D expenditure in the social sciences and humanities between 2005 and 2014 was concentrated in just a few research fields, such as finance, economics and education. By contrast, R&D expenditure was comparatively low in research fields such as media and communication studies, technology management, architecture and habitat, and dance.

- In an era of rapid global technological change, but also deepening local societal challenges, South Africa's national and institutional policymakers face strategic R&D choices. This article contributes to national debate about the status and perceived role(s) of the social sciences and humanities in this context.

Introduction

In countries across the income spectrum, investment in research and experimental development (R&D) is predominantly targeted within the natural and health sciences, engineering, agriculture, and technology fields. By contrast, investment in the social sciences and humanities (SSH) is typically a small fraction of gross domestic expenditure on R&D (GERD). This scenario is also true for South Africa: SSH expenditure as a percentage of total GERD reached about 20% in 2014. However, our understanding of where specifically investment within the SSH is being targeted is limited. To redress this gap, we reviewed data from South Africa's annual R&D Survey to critically investigate three key questions: What are the notable patterns of R&D expenditure in the SSH over the period 2005–2014? Based on the evidence, what are the apparent drivers of R&D performance in SSH related fields in this period? In what ways might future national and institutional policy be geared toward the mission of advancing and coordinating SSH R&D in South Africa?

Policy context

The roots of the social sciences and humanities in South Africa can be traced through the colonial and apartheid periods, and are mirrored in the development and advancement of knowledge transfer and knowledge production through the first universities and government-funded research institutions.[1-4] After the democratic transition in 1994, the national policy framework on science and technology recognised that the SSH would have a significant place in post-apartheid society. In particular, the 1996 White Paper[5] argued that:

> Human and social scientists play a vital role in providing critical analyses of national goals, choices about development policies and strategies, and other national issues pertaining to the transformation of South African society. Their involvement is crucial to a deeper understanding of social issues and to stimulating public debate that could lead to a reconsideration of chosen paths. Equally important to any society that seeks to be innovative in its response to the demands of global change is social research that identifies and explains global trends and their implications in political and economic life, communications and lifestyle changes. Research in the social sciences is therefore of fundamental importance, particularly at this point in our history. (p.12)

This passage is instructive because it contains three key normative assertions about how the SSH should be contributing to development in South Africa. They have a role as a source of critical analysis, deepening human understanding, and insight and explanation of social, political and economic phenomena. In a 2017 discussion document entitled 'Towards the 2017 White Paper on Science, Technology and Innovation: Inclusive development through science and innovation'[6], there is an additional normative assertion about the SSH as extending beyond simply observation and commentary, to action: 'Purposeful inclusion of the SSH will be prioritised, not only in the role of observer and commentator, but also in the *conceptualisation, planning, and execution* of innovation initiatives' (p. 67; emphasis added). What this additional role points to is an active place for the SSH in a wider science, technology and innovation (STI) policy agenda; an agenda that is increasingly predicated on addressing

the continued key societal and economic challenges faced by the country – namely, poverty, inequality, unemployment and weak growth – through the promotion of innovation for inclusive development. While the contextual realities South Africa faced in 1996 have perhaps become more complex in 2017, as a result of rapid technological change, both the 1996 and the present policy outlooks express strong support for the place and value of the SSH in the country.

At ground level, debate about the status of the SSH in South Africa came to a head in 2011, with scholars and commentators disputing whether the SSH had fallen into crisis or not.[7-9] At around the same time, the National Institute for the Humanities and Social Sciences was established in 2013 through the *Higher Education Act* (*Act No. 101 of 1997*).[10] Perhaps less well known in this context is that the UK Department for International Development and the Global Development Network commissioned two separate country studies concerning the social sciences in relation to the South African research environment[11], on the one hand, and the political economy of the performance of social sciences research, on the other hand[4,12].

What has emerged from this heightened attention on the status of the SSH in South Africa in recent years is a particularly useful, if partial, body of critical analysis of different sources of evidence that is helping to develop a more robust picture of SSH research performance. These sources of evidence include not only bibliometric[11,12] and enrolment data[13], but also institutional profiling[12], qualitative data from interviews with SSH scholars[11], and political economy analysis[4,7]. In addition to this evidence and analysis, a critical review of R&D expenditure patterns in the SSH in South Africa over a meaningful period of time can help to enrich our understanding of the challenges and opportunities for the SSH going forward, particularly insofar as the allocation of financial resources to particular research fields is concerned as well as the socio-economic objectives to which R&D is directed. For institutional leaders, funders of R&D, as well as for government policymakers, the assessment documented in this article is aimed at supporting strategic planning processes and research agenda setting, as well as at contributing to the debate about the role of SSH in South Africa.

Methodology

Definitions

There are numerous classifications of what constitutes research and experimental development (R&D) within the social sciences and humanities.[14] The definition of the *social sciences and humanities* used for this research was drawn from the terms of the research fields that are used in the collection of data in the South African National Survey on Research and Experimental Development (hereafter R&D Survey). These research fields are given in Table 1.

R&D performance is defined in this article as R&D performed within five sectors in the R&D Survey: business, government, higher education, not-for-profit organisations, and science councils (Table 2). Within the R&D Survey, performance reflects only input data, including nominal expenditure and personnel involved in R&D. In this review, we do not cover data related to funding sources for R&D in the SSH, as the R&D Survey data on sources of funding is not disaggregated by research field.

According to the *Frascati Manual*[15], which was used in the collection of data referred to in this article:

> *R&D comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications.* (p.30)

The definition has since been updated by the OECD.[16] Used interchangeably with R&D performance, *R&D expenditure* is defined, in this article, as current and capital expenditure reported by respondents in the R&D Survey, covering the period 2005–2014.

Table 1: Social sciences and humanities research fields according to the South African R&D Survey

Social sciences	Humanities
Accounting	Arts and culture
Anthropology	Dances
Archaeology and history	Historical and civilisation studies
Architecture and habitat	Languages and literature
Economics	Music
Education	Other humanities not elsewhere classified
Emerging issues	Philosophy
Finance	Religious studies
Geography	
Law	
Management studies	
Media and communication studies	
Other social sciences not elsewhere classified	
Political sciences and public policy	
Population studies	
Psychology	
Sociology	
Technology management	
Tourism	
Transportation studies	

Data sources

Data for this paper were sourced from the Centre for Science, Technology and Innovation Indicators at the Human Sciences Research Council, which performs the South African R&D Survey on behalf of the Department of Science and Technology (DST). Curated, aggregate-level data are accessible to the general public on www.hsrc.ac.za. International comparative data were sourced from the UNESCO Institute for Statistics (www.uis.statdata.uis.unesco.org).

Data anomaly

In the R&D Survey, respondents classify R&D expenditure according to research field. If the research field is not listed in the code booklet, then respondents tend to report expenditure as 'Other social sciences not elsewhere classified' and 'Other humanities not elsewhere classified'. The impact is that when reviewing the data, there is often significant expenditure reported under these headings – in all sectors. This situation makes it impossible to unpack the expenditure in terms of the specific research field in which it has been invested. For this reason, definitive conclusions about the allocation of resources by research field cannot be drawn; however, it is possible to reflect on the overall quantum invested in the SSH.

Table 2 illustrates that the higher education sector has by far the largest amount of unclassified R&D expenditure (when compared with other sectors). This difference is mostly explained by the diversity of research performed at higher education institutions, on the one hand, and the scope of the research field nomenclature on which the R&D

Survey rests, on the other hand. This issue is discussed further in our recommendations. In terms of the share of unclassified R&D expenditure to total R&D expenditure in the SSH, all sectors, except the business sector, report between 11% and 19% unclassified R&D expenditure – an issue which is also addressed in the recommendations section.

Trend analysis

International trends

When compared to expenditure on R&D in the natural sciences and engineering over time, there has been an increase in the proportion of SSH R&D expenditure in South Africa – from 12.3% of GERD in 2005 (Figure 1) to 19.2% in 2014 (Figure 2).[17] Argentina, South Africa and Malaysia reported the largest proportional increases in SSH R&D in 2014 (as compared to 2005), while Chile, Turkey and Poland showed negative growth in SSH R&D expenditure in 2014 as compared to 2005.

Compared to countries for which equivalent 2014 data were available for analysis, the ratio of GERD that is dedicated to SSH in South Africa is substantially higher than, for example, the Republic of Korea (3.6%), Russian Federation (4.1%), Malaysia (8.2%), Poland (9.0%), Chile (9.8%), Turkey (15.3%), Paraguay (16.2%) and Uruguay (17.1%). The only country to surpass South Africa in this regard in 2014 was Namibia, which dedicated 24.5% of GERD to SSH R&D.

National trends

At the national level, the reported rand value of social sciences R&D expenditure over the 10-year period was ZAR27.8 billion. Of a total of 19 social science research fields, 5 received 57.4% of this expenditure over the period 2005–2014 (Figure 3), with expenditure being highest in the finance research field. The next 4 research fields made up 18.1% of the expenditure, with the remaining 10 research fields making up 12.4%. Expenditure not classified by research field amounted to 12.1%.

Table 2: R&D expenditure in the social sciences and humanities research fields, by sector, including unclassified expenditure (2005–2014)

Sector	Total R&D expenditure, social sciences and humanities (ZAR'000)	Total unclassified expenditure (ZAR'000)	Proportion of unclassified expenditure (%total)	Unclassified expenditure, social sciences (ZAR'000)	Unclassified expenditure, humanities (ZAR'000)
Business	8 580 122	52 178	0.61	50 371	1808
Not-for-profit	1 455 977	262 777	18.05	224 534	38 243
Government	3 063 845	349 382	11.40	327 398	21 984
Science councils	2 075 952	351 523	16.93	174 937	176 586
Higher education	17 598 571	3 794 335	21.56	2 574 087	1 220 248
Total	32 774 466	4 810 196	14.68	3 351 328	1 458 868
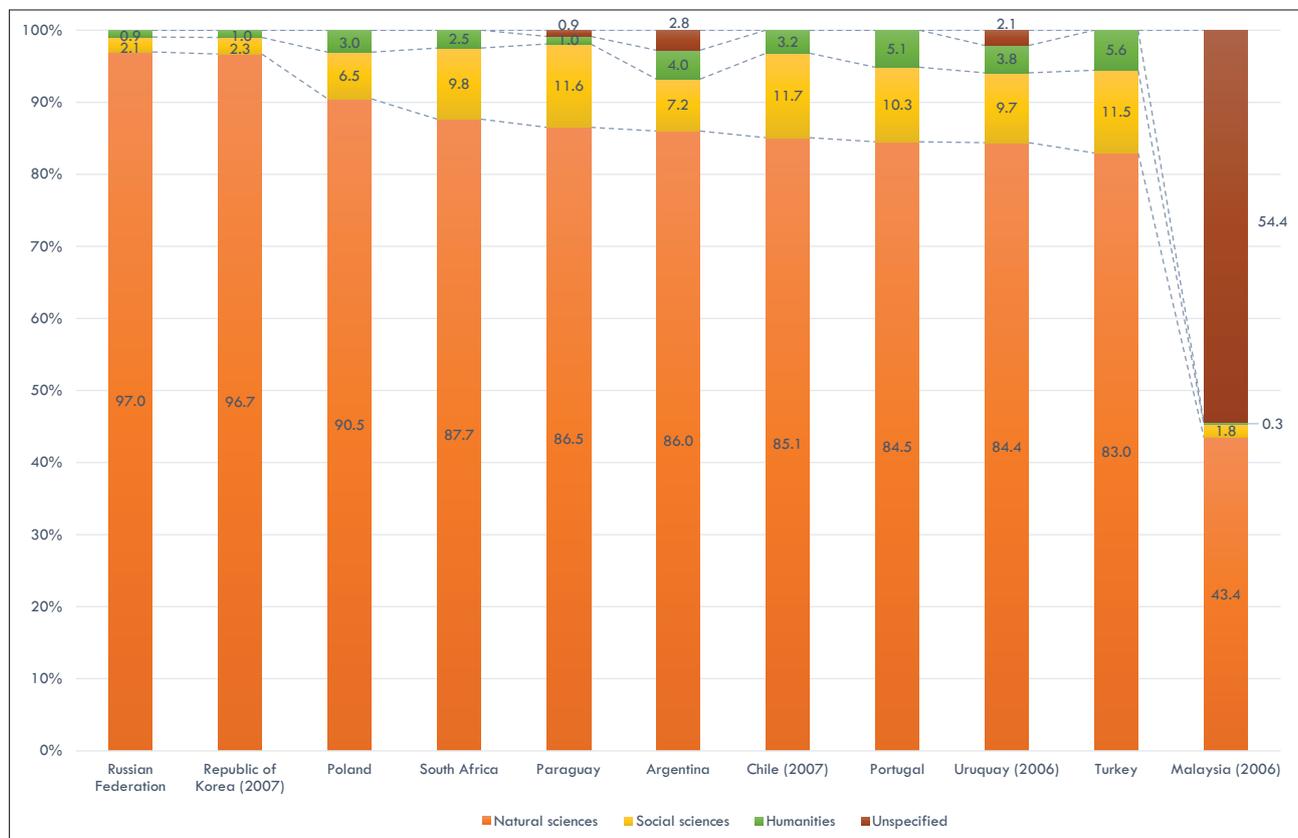
Data are for 2005 or the closest available year as indicated in brackets. The heading 'natural sciences' includes the following sub-categories: engineering and technology, natural sciences, medical and health sciences, and agricultural and veterinary sciences.

Figure 1: International comparison of R&D expenditure by research field as a % of GERD, 2005.
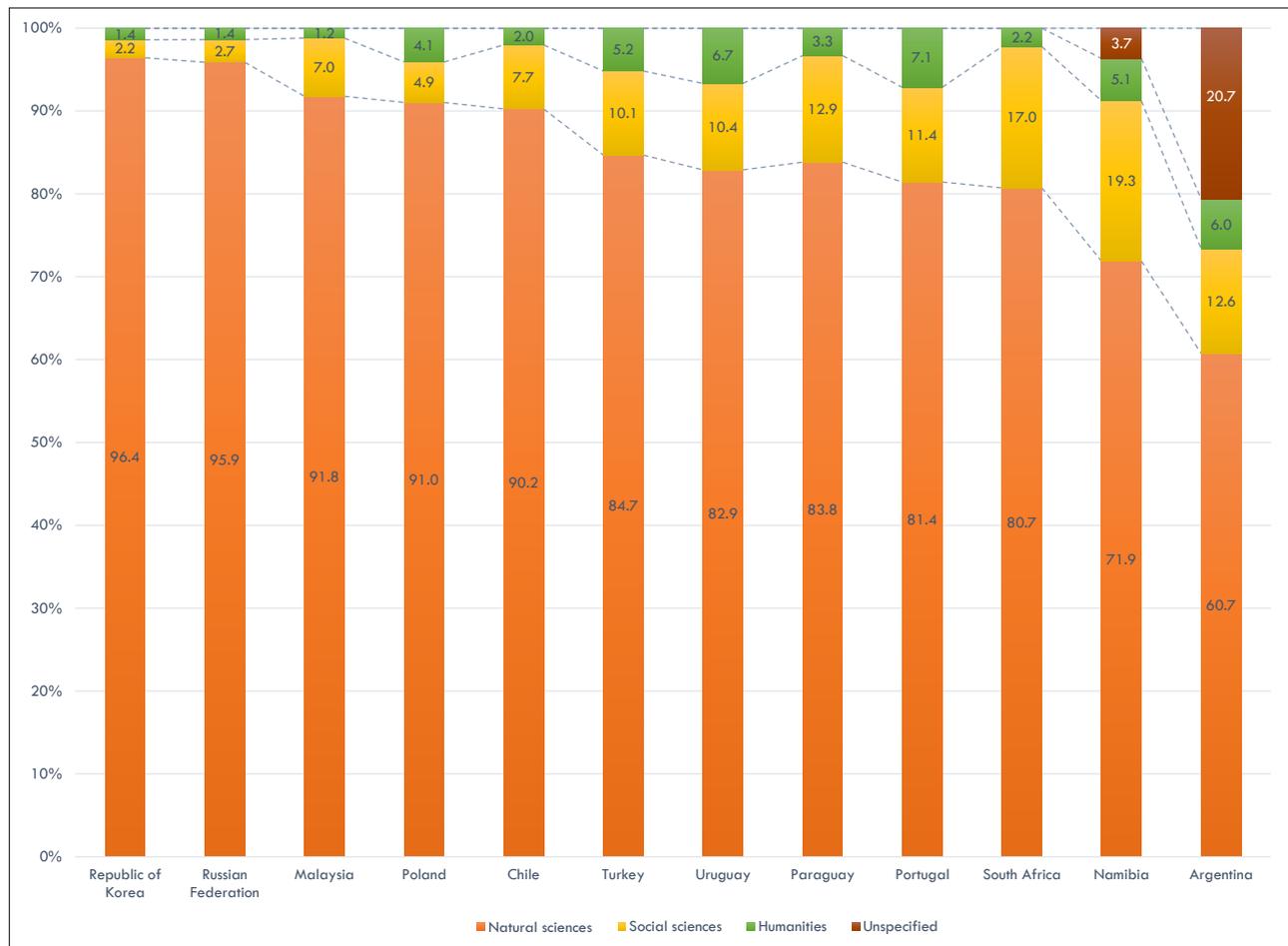
Data are for 2014 or the closest available year as indicated in brackets. The heading 'natural sciences' includes the following sub-categories: engineering and technology, natural sciences, medical and health sciences, and agricultural and veterinary sciences.

Figure 2: International comparison of R&D expenditure by research field as a % of GERD, 2014.

Social sciences R&D expenditure more than doubled between 2005 and 2014, from ZAR1.4 billion to ZAR5.0 billion.

Nationally, the reported rand value of humanities R&D expenditure over the 10-year period was ZAR5.0 billion. Excluding unclassified expenditure, five research fields of a total of seven made up 65.9% of the expenditure, with expenditure being highest in the languages and literature research field (Figure 4). Two research fields (philosophy, dances) made up the remaining 4.9% of expenditure. Expenditure not classified by research field amounted to 29.2%. Humanities R&D expenditure nearly doubled between 2005 and 2014, from ZAR350 million to ZAR657 million.

Sectoral trends

In the higher education sector, expenditure was registered in all of the SSH research fields specified in the R&D Survey. Expenditure for R&D in education (ZAR1.8 billion), law (ZAR1.7 billion), languages and literature (ZAR1.5 billion), management studies (ZAR1.3 billion) and economics (ZAR1.1 billion) was the highest over the 2005 to 2014 period. The next three research fields for which comparatively large amounts were spent were psychology (ZAR0.95 billion), accounting (R0.86 billion), and political sciences and public policy (ZAR0.80 billion). Research fields such as population studies (ZAR0.02 billon) and transportation studies (ZAR0.01 billion) only attracted a small amount of R&D funding relative to the largest cohort. Finance-related R&D grew very slowly off a low base over the 10-year period from 2005 to 2014.

In the science councils, the majority of SSH R&D expenditure (69.3%) was in a small number of research fields: political sciences and public policy (ZAR0.35 billion), education (ZAR0.35 billion), sociology (ZAR0.27 billion), population studies (ZAR0.24 billion), and economics (ZAR0.23 billion). Because of the low number of science councils that perform SSH R&D, as defined by the R&D Survey, this trend is not necessarily surprising.

In the government sector, SSH R&D was performed across a wide range of research fields, with expenditure being highest in the economics (ZAR0.66 billion), political sciences and public policy (ZAR0.64 billion), management studies (ZAR0.36 billion), law (ZAR0.16 billion), and arts and culture (ZAR0.13 billion) research fields.

In the business sector, expenditure on R&D was highest in the fields of finance (ZAR5.4 billion), accounting (ZAR1.8 billion), economics (ZAR0.59 billion), technology management (ZAR0.39 billion) and management studies (ZAR0.20 billion). Notably, in the finance research field, a sharp upward trend can be seen between 2011 and 2014. In the review of unit level data, it was noted that a single entity reported significant R&D expenditure in the finance research field in this period, which would explain the growth. Also notable is the increase in expenditure in the emerging issues research field, from ZAR0.952 million in 2013 to about ZAR17 million in 2014. There is only one humanities research field, languages and literature, which reflects expenditure between 2005 and 2010. It has to be emphasised that zero expenditure may be for other reasons, such as that the firms reporting on this expenditure might have ceased to be part of the survey or that no R&D was undertaken in this field.
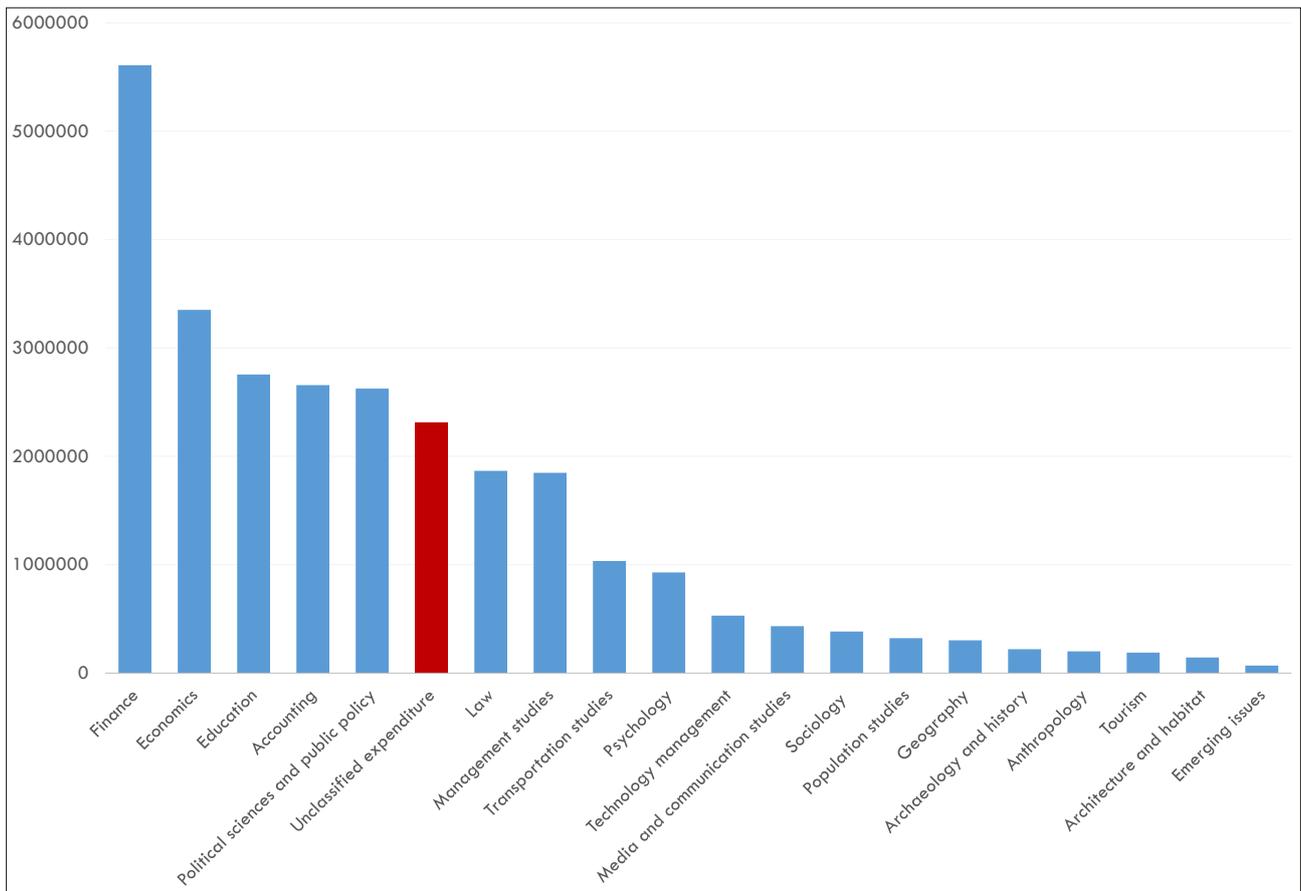
Figure 3:   Total GERD by social sciences research field over 10-year period (ZAR'000)
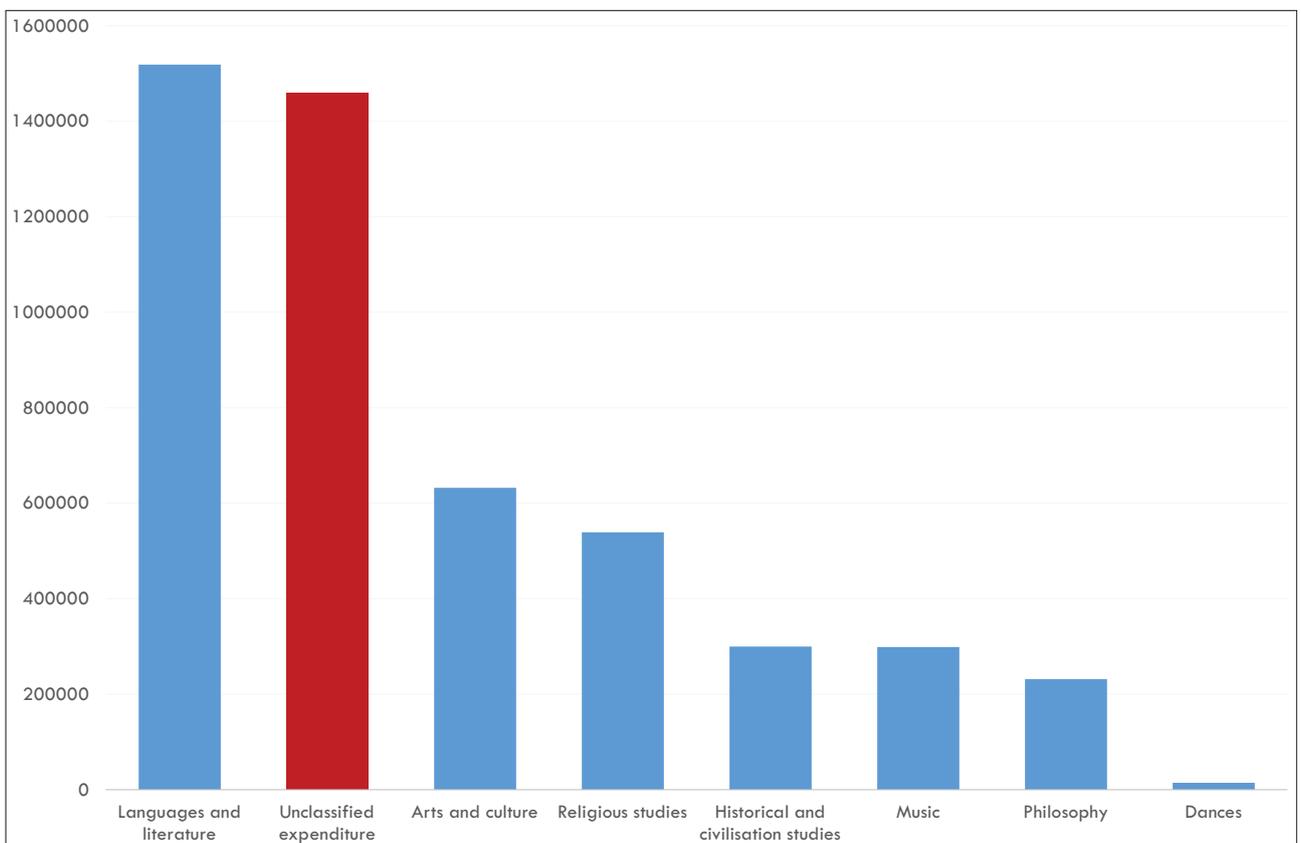

Figure 4:   Total GERD by humanities research field over 10-year period (ZAR'000)

In South Africa, R&D expenditure by not-for-profit organisations (NPOs) is typically a small fraction of total GERD (in 2014, for example, it was 2.7%; in 2005 it was 1.6%), with the bulk of funding for NPO R&D provided by foreign sources. Between 2006 and 2008, R&D expenditure in the education research field registered a sharp decline (from ZAR0.06 billion to ZAR0.02 billion), while R&D expenditure in the political sciences and public policy research field increased (from ZAR0.04 billion to ZAR0.06 billion). By 2014, however, the trend had reversed, with education R&D expenditure (ZAR0.04 billon) surpassing that in the political sciences and public policy research field (ZAR0.03 billion). Interestingly, low as their expenditure is, NPOs tend to carry out research in most research fields. Emerging issues are also researched in the NPO sector, suggesting the extent to which the nomenclature of research fields included in the R&D Survey does not capture fully the research areas in which NPOs are engaged.

Key observations

There are a series of key observations that can be drawn from these data in relation to the questions posed at the beginning of this article. The recommendations section of this paper elicits some of the implications of these observations for national and institutional planning and research agenda setting.

A first observation is that South Africa appears to invest proportionally more of its GERD in SSH than the countries for which equivalent data are available; and linked to this observation is that SSH research investment has virtually doubled over the period of a decade. At the macroscale, this observation could reinforce a view that, in fact, the SSH are not being overlooked in South Africa, but are rather being underpinned by increased investment driven by a recognition of the value of performing this type of R&D. However, it would be premature to draw that conclusion – for all SSH research – without examining and analysing the expenditure across the different research fields and sub-fields comprising the SSH, and the sectors in which the R&D is performed, because the playing fields may not be level across either. Indeed, when we performed the analysis of SSH R&D expenditure data from 2005 to 2014, as summarised in this article, what we uncovered was a picture that alarmed us as much as it served to reinforce the debate concerning the so-called 'crisis' moment for the SSH in South Africa. In particular, looked at from the perspective of research fields, the spend on SSH is very strongly geared to areas such as finance, economics, education, accounting and political science and public policy (Figures 3 and 4), in every year of the survey.

Commentators might perceive this situation to accurately reflect pre-vailing economic and social urgencies. However, the comparatively low spend in areas such as technology management and media and communication studies suggests that R&D agenda setters and decision-makers may not have paid sufficient attention to the opportunities and challenges posed by the exponential technological changes taking place globally under the heading of the Fourth Industrial Revolution.

Also concerning is the comparatively low spend in areas such as psychology, architecture and habitat, archaeology and history, socio-logy, and languages and literature. Given South Africa's complex colonial and apartheid legacies – which remain persistent and which traverse geo-spatial, subjective and linguistic experiences of citizens and communities, and which play out frequently in service delivery protests, contests over restitution, land and state-owned property, and, most recently, at our universities – there appears to be a weak alignment in research spend and these burgeoning societal trends. The same logic can be explored in the case of the relatively low levels of reported research investment in areas such as transportation studies, tourism and dance. To the extent that these areas arguably represent important potential economic and social development opportunities for South Africa, South Africans, and the globe more broadly, we would argue that SSH research decision-makers need to be alert to the apparent decline of these research areas, at worst, or the under-reporting of R&D expenditures through the Survey, at best.

A further observation is that when the 2005–2014 spending on R&D in the SSH is looked at from a sectoral perspective, the higher education sector spends by far the bulk on research in both humanities and the

social sciences. As the key 'custodians' of the SSH disciplines, therefore, decision-makers from these institutions have an especially crucial role to play in advancing SSH research in South Africa and in redressing any imbalances that might be leading to decline in certain research fields. The other major players in SSH R&D are business, government, science councils, and NPOs, in that order. Driven by their mandates, these institutions are arguably the main producers of applied SSH R&D, addressed toward specific economic, advocacy or policy objectives (as opposed to basic research objectives), and their research agendas can be driven by the availability of research funds, whether local or international. Collectively, these institutions spend marginally more on social sciences R&D than do higher education institutions, although of course comparatively less on humanities R&D. The question this split raises for decision-makers from all sectors relates to the issue of impact, or perhaps more crudely put, value for money. The question is: are R&D investments resulting in widespread impact, of the type South Africa desperately requires?

A caveat is in order at this point: what we know for sure is that a decade worth of R&D spending in the SSH has not provided a miracle cure for persistent poverty, inequality, unemployment, and weak growth. But nor has the other 80% of South Africa's R&D spend in the natural sciences, engineering and health research fields. However, to think that there is a linear relationship between R&D and the solving of complex societal challenges is to misunderstand the many and, in some cases, competing drivers that underpin R&D in different sectoral contexts, or within different research fields. Therefore, fine-grained analyses of research performance and research expenditure are required at both sectoral and research field level. Through this paper, we have tried to make a start on this agenda, although the need for further in-depth empirical work is acute.

Finally, from the perspective of the R&D Survey instrument, what emerged clearly from the trend analysis presented above is the problem of unclassified R&D expenditure in SSH. This unclassified R&D expenditure is a problem because it is impossible to analyse it at the level of research fields, not only for researchers and institutional planners, but also for funders of research (both private and public, local and international). This challenge presents a particular research design imperative for the managers of the R&D Survey, to the extent that it would be desirable to reduce the reporting of unclassified R&D expenditure as much as possible.

Conclusions

Since the dawn of democracy in South Africa, there has been intense debate about the role of the social sciences and humanities. On one level, this debate has been about the survival of individual research and teaching agendas, disciplines, and even entire institutions – essentially, a debate about funding. However, there is now a strong narrative in both institutional and national policy discourse, as reflected in the 1996 White Paper and its soon-to-be successor, that SSH research has a vital role to play in helping us to not only understand, but also address, the structural societal and economic challenges of our transition. The formation of the National Institute for Humanities and Social Sciences in 2013, and continued investment in R&D through the National Research Foundation, the universities, the South African Research Chairs Initiative, and the Human Sciences Research Council, are reflections of the state's commitment in this regard.

At another level, there is possibly a tendency in our discourse to consider the SSH as an afterthought to the natural sciences and, as a result, to limit or shortchange SSH research. Whereas, what the data in this paper show is that the SSH in fact covers a multiplicity of research agendas, which are carried out under the umbrella(s) of different sectoral or organisational interests or mandates, which may be in direct competition or contradiction with each other, or, indeed, with the vision of the SSH in the White Paper. It is therefore unhelpful to characterise the SSH as a 'singular' entity in this way. Rather, what we need to better ascertain is the nature of the disciplinary funding economies of SSH R&D, in order to assess whether inequities or skewed distribution of scarce resources (whether public, philanthropic or private) are hindering knowledge production in the public interest.

For the R&D Survey, as one of the instruments that can assist us to achieve this goal in the future, it will be essential to carefully review the nomenclature or taxonomy of research fields used to track R&D expenditure and performance, and possibly even in the future, impact.

Recommendations

The three recommendations detailed below are directed at institutional planners, funders and the South African R&D Survey managers, respectively.

For institutional planners: While the data show an alignment between R&D and the major societal trends, what is also revealed is 'neglect' of potentially future strategic R&D opportunities. We recommend therefore that institutional planners consider the South African R&D data and suite of analytical products as one key source of evidence in their forecasting and scenario building.

For funders: The data presented in this paper provide a mirror of the key R&D performance trends over a 10-year period, including those research fields that are thriving, as it were, and those that are diminishing. We recommend that funders also pay particular attention to South African R&D data and results in funding allocation decisions.

For R&D Survey managers: We have shown that there is room for improvement in the classification of research fields. To limit the reporting of unclassified expenditure, it is essential for the R&D Survey managers to ensure that appropriate space and scope for reporting expenditure are provided.

Acknowledgements

Authors' contributions

N.M. was responsible for the data collection; N.M. and G.R. were responsible for conceptualisation, methodology, data analysis, validation and the write-up.

References

1. Dubow S. Scientific racism in modern South Africa. Cambridge, UK: Cambridge University Press; 1995.

2. Dubow S. A commonwealth of knowledge: Science, sensibility and white South Africa 1820-2000. Oxford: Oxford University Press; 2006.

3. Shepherd N. State of the discipline: Science, culture and identity in South African archaeology, 1870–2003. J South Afr Stud. 2003;2(4):823–844. https://doi.org/10.1080/030.570.7032000135842

4. Centre for Research on Evaluation, Science and Technology (CREST). Mapping social sciences research in South Africa Part II: The political economy of social sciences research in South Africa. Stellenbosch / London: CREST / UK Department for International Development; 2014.

5. South African Department of Arts, Culture, Science and Technology (DACST). White paper on science and technology: Preparing for the 21st century. Pretoria: DACST; 1996.

6. South African Department of Science and Technology (DST). Towards the 2017 white paper on science, technology and innovation: Inclusive development through science and innovation. Pretoria: DST; 2017.

7. Academy of Science of South Africa (ASSAf). Consensus study on the state of the humanities in South Africa: Status, prospects and strategies. Pretoria: ASSAf; 2011. Available from: http://hdl.handle.net/20.500.11911/33

8. South African Department of Higher Education and Training (DHET). Report commissioned by the Minister of Higher Education & Training for the Charter for Humanities and Social Sciences. Final report. Pretoria: DHET; 2011.

9. Mouton J. The Humanities and Social Sciences in SA: Crisis or cause for concern. S Afr J Sci. 2011;107(11/12), Art. #961, 4 pages. http://dx.doi.org/10.4102/sajs.v107i11/12.961

10. Higher Education Act (101/1997), Republic of South Africa. Regulations for the establishment of a national institute for the humanities and social sciences. Government Gazette, 5 December 2013. Pretoria: Government Printers; 2013.

11. Potgieter C, Bob U, Sooryamoorthy R. Assessing the environment for social science research in developing countries: The case of South Africa. Final report. Durban: University of KwaZulu-Natal / Global Development Network; 2016.

12. Centre for Research on Evaluation, Science and Technology (CREST). Mapping social sciences research in South Africa Part I: Mapping of research centres. Stellenbosch London: CREST / UK Department for International Development; 2014.

13. Essop A. Report on the state of the arts, humanities and social sciences in South African universities: Prepared for the Andrew W. Mellon Foundation [document on the Internet]. c2015 [cited 2018 Feb 14]. Available from: https://mellon.org/media/filer_public/c6/38/c6388113-6eb5-4f0e-82ff-0639b733abc3/arts_humanities_social_sciences_south_african_universities.pdf

14. American Academy of Arts and Sciences (AAAS). The heart of the matter. Cambridge, MA: AAAS; 2013.

15. Organisation for Economic Co-operation and Development (OECD). Frascati manual: Proposed standard practice for surveys on research and experimental development. Paris: OECD; 2002.

16. Organisation for Economic Co-operation and Development (OECD). Frascati manual: Guidelines for collecting and reporting data on research and experimental development. Paris: OECD; 2015.

17. Human Sciences Research Council (HSRC). South African national survey on research and experimental development: Statistical report 2014/15. Cape Town: Centre for Science, Technology and Innovation Indicators, HSRC; 2017.

Nurses' knowledge of and willingness to promote female condom use in Johannesburg Health District

AUTHORS:
Magdelena S. Petkova[1]
Zuberu B. Elabor[2]
Trayana Djarova-Daniels[3]
Emilia G. Pavlova[4]
Olufemi B. Omole[5]

AFFILIATIONS:
[1]Department of Family Medicine, University of the Witwatersrand, Johannesburg, South Africa

[2]Principal Family Physician, Eastern Cape District Clinical Specialist Team, Port Elizabeth, South Africa

[3]Department of Biochemistry and Microbiology, University of Zululand, KwaDlangezwa, South Africa

[4]Department of Physiology and Biochemistry, National Sports Academy, Sofia, Bulgaria

[5]Department of Family Medicine, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Magdelena Petkova

EMAIL:
dr_meggdim@yahoo.com

The female condom is the only current method for female partners to simultaneously prevent both pregnancy and sexually transmitted infections. Studies of various contraceptive methods suggest that providers' acceptance and endorsement may be a key factor in their clients' uptake and continued use of a method. Our aim in this study was to assess the relationship between nurses' knowledge of and their willingness to promote female condom use in 18 public healthcare facilities within the Johannesburg District. The mean score of correct answers of 398 nurses completing an anonymous, self-administered, six-item questionnaire was 4 out of a maximum of 6. Data analysis included the use of descriptive statistics and a chi-square test. It was found that 79% of participants were knowledgeable but only 59% were willing to promote female condom use. There was no association between knowledge of and willingness to promote female condom use. The following significant associations with knowledge and willingness to promote female condom use were found: family planning experience, being employed by a local government authority, working in a primary care clinic and having had informal training on female condom use. Informal training of nurses within the family planning unit in the clinics has the potential to improve nurses' knowledge and willingness to promote female condom use.

Significance:

- This is the first study conducted in South Africa on the knowledge of and willingness of healthcare providers to promote female condom use.

- An effective strategy is needed to motivate healthcare workers to promote female condom use with their patients.

- Informal training of nurses within the family planning unit has the potential to improve nurses' knowledge of and willingness to promote female condom use.

Introduction

The female condom has been developed as an alternative to the male condom and remains the only female-initiated means of preventing both pregnancy and sexually transmitted infections (STI) including the human immunodeficiency virus (HIV). HIV incidence has slightly decreased in South Africa, but many people still engage in risky sexual behaviour. Findings also indicate higher condom use among young men than among women.[1] The incidence of HIV varies across South Africa's provinces, with KwaZulu-Natal having the highest incidence.[2] The Sexual and Reproductive Week campaign was launched in February 2016 under the umbrella of the national health department to promote female and male condom use. Recently other viral infections such as Ebola and Zika were found to be sexually transmitted.[3-5] Despite the fact that cases of Ebola and Zika virus have never been reported in South Africa, the fact[4] that these emerging infectious diseases could spread rapidly should be emphasised when promoting female condom use.

The female condom is considered to be more effective than the male condom because it is less likely to break or burst and it covers a larger portion of the genital area.[6] Despite these views, female condom acceptability is low, mainly because of interpersonal factors. The method's higher cost may also underpin many challenges in female condom promotion.[6-11]

A post-marketing survey in Zimbabwe in 1998 highlighted the key role of healthcare providers and lay educators in increasing women's access to the female condom.[12] In a study conducted in 1996, preliminary data showed that the female condom was an acceptable method for some Kenyan couples, but the authors noted the necessity of further research on safety, cost-effectiveness, and obstacles to acceptance.[13] The authors of a large community-based female condom intervention trial in rural Kenya reported that in spite of the intervention, healthcare service providers in the community failed to promote female condoms actively and believed them to be unsuitable for most women.[14] The interpretation of these results also indicates the need for an alternative method of prevention, one of which is the female condom. This was supported with other antiretroviral treatment as suggested by the results in a study conducted on an urban US population.[15]

Data on 2712 subjects were collected through a cross-sectional survey in Tanzania using mass media as a mode of promoting the female condom. In the above study, as well as in a South Africa based study, communication with a peer educator or healthcare provider was found to have a direct positive impact on female condom uptake.[16,17]

Research on the acceptability of the female condom carried out in 1999 in Brazil with 2453 women showed high rates of acceptability. It has been noted that women in Costa Rica, Indonesia, Mexico, Senegal, Tanzania, Zimbabwe and the USA feel safe using the female condom.[18] In spite of these findings, the authors of the first family condom study in Italy using a Lickert-scale questionnaire after post-test counselling consider that promotional strategies and training activities should be designed to increase willingness to use the female condom.[19] Research done in

different national settings demonstrates that healthcare providers often have negative views of this method and lack information on how to promote it.[20-23]

Nurses providing family planning services in southern Mozambique saw the provider–client relationship for choosing contraceptive methods as an integral part of their job as opposed to an added burden or contributing to job dissatisfaction.[24]

A study done in Botswana showed that the perception and attitude of healthcare workers on female condoms played a positive role in advocating against STIs, HIV and unwanted pregnancy in women.[25] Health worker attitudes toward clients has also been found to play a role in clients' sexual health choices.[26] Training on female condoms for healthcare workers and the availability of female condoms are essential.

Another study on clients' sexual health choice in South Africa showed that the lack of distribution of female condoms in family planning units could prevent their promotion.[27]

The study conducted in Brazil to ascertain knowledge and promotion of female condom use by healthcare providers found similar results but did not examine any possible association between knowledge of female condom use and willingness of the providers to promote their use.[18] The only available data on introduction of the female condom to the public healthcare sector are those from urban Zimbabwe, presented in a Horizons project.[12]

A nationally representative survey was done in Zimbabwe and South Africa. In South Africa, providers of family planning or HIV/STI services were more likely to counsel women on the use of male condoms and less likely to counsel women on female condom use than their counterparts in Zimbabwe. At the time, new types of female condoms were not available and the providers did not see the available type as a good female barrier method.[28]

The provision of contraceptive services during other primary healthcare consultations is an implementation strategy of the National Contraceptive Policy guidelines[29] to provide high-quality contraceptive services since 2001. The policy guidelines encourage more extensive use of barrier methods and emphasise that at least one female-initiated barrier method should be made available for contraception. As at 2003, when these service delivery guidelines were made, the female condom was already available in selected sites throughout the country.

The authors of a study which reviewed the condom programme implemented by the South African National Department of Health, concluded that although the female condom programme was well established, both cost and education were factors in the uptake of the use of the female condoms.[30] The popularity and acceptability of the female condom was dependent on its much higher cost (~18 times[30] more expensive than the male condom), which necessitated the development of new, cheaper types of female condoms. FC2, a type of female condom, was available in South Africa between 1998 and 2013. A South African national government programme piloted female condoms in 1999–2000, but FC2 was not evaluated.

A randomised South African study looking at the higher cost of female condoms introduced new types of female condoms and found similar acceptability among clients of all three types of female condoms.[31]

Through a joint parallel programme of Universal Access to Female Condoms and Maternal and Child Health, strategies, recommendations and guidelines for the new female condoms were developed to make female condoms more accessible and affordable. The parallel programming also looked at the training of the providers and the marketing of female condoms.[32] The parallel programming aimed to make the different types of female condoms available to clients to choose for themselves. Both clients and providers shared positive views on the availability of more than one type of female condom; however, the healthcare professionals expressed a need for training on each type of female condom. Some facilities did not provide all three types of female condoms and clients therefore did not have a choice.[33]

New designs and materials were introduced for female condoms in 2014, specifically the variants Cupid and Pleasure More. All three types of female condoms are made of rubber latex. The two newer types are made of natural rubber latex whilst FC2 is composed of synthetic nitrile rubber latex.[33]

The use of female condoms has doubled since 2008.[34] South Africa's female condom programme is the largest and most comprehensive in the world.[33,34]

In 2014–2016, a study by the MatCH Research Unit at the University of the Witwatersrand, Johannesburg, assessed the national programme on female condoms and identified strategies to increase acceptability.[34] An increase (from 0% to 34%) in female condom use was identified in 2015/2016 compared to male condom use in a national evaluation.[34] The literature review and the evaluation of the South African national programme showed that although there is gross availability in the public sector, distribution by the providers is not taking place.

Several studies on the knowledge, attitudes, practices and behaviours associated with the use of female condoms have been reviewed.[35] It was concluded that knowledge on female condom use was low in most of the countries among the different users. Low use was associated with the limited availability of female condoms.

In South Africa, no published studies exist that have examined the relationship between healthcare providers' knowledge of female condoms and their willingness to promote female condom use.

Methods

We undertook a cross-sectional descriptive study with an analytical component. The study was conducted in 18 public healthcare facilities located in the sub-district 'F' of the Johannesburg Health District in the Gauteng Province of South Africa. These facilities included 16 primary healthcare clinics run by the local municipal government authority, a provincial government community health centre and a provincial district hospital. Female condoms (of type FC2) were freely available in all these healthcare facilities and nurse clinicians in these facilities were active distributors of the female condoms. These facilities provide prevention, promotion, rehabilitative and curative health services. At the time of this study in 2011, the total population of nurses employed by these facilities was 465. A sample size of at least 384 participants was calculated to give 5% precision with a 95% confidence level. We, however, sought to use the entire population of 398 nurses who met the inclusion criteria and consented to participate. Exclusion criteria were the unwillingness to consent or being a nurse in training.

Data were collected using an anonymous, structured, self-administered questionnaire. No questionnaire from any previous research was used in this study. The questionnaire was designed for data collection on the participants' characteristics: knowledge of female condom use and willingness to promote the use of the female condom. The determination of knowledge was, therefore, done using six questions based on the conclusions and findings of various previous studies as discussed in the literature review.[6,18,19] 'Knowledgeable on female condom use' was determined if a participant correctly answered four or more questions of the six knowledge questions. 'Not knowledgeable' was determined if a participant answered fewer than four of the six knowledge questions correctly. The mean score of correct answers out of the six-item knowledge questionnaire was calculated and used to assess the level of knowledge. The study also categorised participants into two groups based on their willingness to promote female condom use. 'Willing to promote female condom use' was defined as a participant who sees the promotion of female condom use as part of their work and despite being overstretched at work they are willing to promote the use of the female condom. If a participant in the study did not have the two elements of the above statement, they were considered 'not willing to promote female condom use'.

In this study, formal training was defined as prescribed and organised learning in the presence of a designated teacher with the award of a qualification or a learning credit and the external specifications of

outcomes.[36] Informal training was defined as learning that takes place outside of a dedicated learning environment and which arises from the activities and interests of individuals or groups, but which may not be recognised as learning.[37]

A pilot study was undertaken on 15 nurses at Mofolo Community Health Centre (who were not included in the analysis of the final study) to identify sensitive and ambiguous questions that needed modification, and to determine the time needed to complete the questionnaire. For example, the question 'the female condom works to prevent pregnancy and STI' was answered 'yes' by everyone, and so was changed to 'female condoms do not prevent pregnancy and STI, including HIV'. The internal consistency of the knowledge questions was tested using Cronbach's alpha test (in this pilot study the coefficient was 0.636). The final six-item questionnaire was designed and prepared on the basis of the analysis of the outcomes of the pilot study.

Each facility was visited at different times. These visits were preceded by a phone call to the facility manager to arrange the time of the visit. On the day of the visit, study information leaflets and questionnaires were handed directly to nurse clinicians. It was clearly explained to each nurse clinician that completion of the questionnaire implies consent. Participants were asked to drop completed questionnaires in an identified ballot-type box in their own time.

Data were analysed using STATA version 10.0 (College Station, TX, USA). Descriptive analysis of categorical data is presented as percentages and frequencies. Continuous data are presented as means with standard deviations. The association between knowledge and willingness to promote female condom use was investigated using the multinomial logistic regression. A p-value of less than 0.05 was considered statistically significant.

Ethical clearance

The University of the Witwatersrand Human Research Ethics Committee granted ethical approval for the study (certificate number M101017). Permission was obtained from the Chief Director of the Johannesburg Health District and the Director of the City of Johannesburg Health to conduct the study in their facilities. The questionnaire was anonymous and did not collect identifying information from participants. Participants were told that completion of the questionnaire implies consent.

Results

A total of 398 nurses participated in the study; 67 nurses were unwilling to participate or unavailable at the time of the study. The participants' characteristics are presented in Table 1.

The male:female ratio was 1:10; therefore, the sample was mostly female nurses. The mean age of the nurses was 40.7 years (s.d.=10.8). Fewer than half (48%) the nurses had nursing experience of 1–9 years and even fewer (45%) had a diploma certificate in nursing. More than half (60%) of the nurses were professional nurses and 77% practised a specialty of one kind or another. The majority (81%) of nurses was employed by provincial government facilities and a minority (19%) was employed by local government facilities. A majority (62%) of the participants had experience in family planning, and more than half (53%) were working in a district hospital. A minority (27%) had formal training on female condom use and the majority (56%) had informal training.

Participants' knowledge of female condoms is shown in Table 2; most participants answered the questions correctly. At least three of the six questions used to assess nurses' knowledge were answered correctly by 90% of the nurses. The question 'female condom cannot be put in place by female 8 hours before intercourse' was correctly answered by 74% of the participants. More than half (54%) of the nurses did not know whether a female condom can be used simultaneously with a male condom and 72% did not know whether the outer ring of the female condom stays inside the vagina just around the cervix. The mean number of correct answers was 4. The majority (79%) of nurses was found to be knowledgeable on female condom use and more than half (59%) were willing to promote female condom use.

Table 1: Characteristics of the participants who completed the questionnaire

Variable	Number of participants (n=398)	Percentage (%)
Gender		
Male	35	8.79
Female	363	91.21
Age in years		
<29	66	16.58
30–39	127	31.91
40–49	107	26.88
50–59	82	20.60
60–69	16	4.02
Race		
Black	366	91.96
White	16	4.02
Coloured	14	3.52
Indian	2	0.50
Years of experience		
<10	191	47.99
10–19	79	20.00
20–29	90	22.61
30+	38	9.55
Qualification		
Diploma	179	44.97
Nursing certificate	159	39.95
Degree	39	9.80
Postgraduate degree	21	5.28
Category		
Enrolled nursing assistant	90	22.61
Enrolled nurse	71	17.84
Professional nurse	237	59.55
Family planning experience		
Yes	248	62.31
No	150	37.69
Type of employer		
Provisional government	322	80.90
Local government	76	19.10
Type of workplace		
District hospital	212	53.27
Clinic	98	24.62
Community health centre	88	22.11
Formal training on female condom use		
Yes	109	27.39
No	289	72.61
Informal training on female condom use		
Yes	223	56.03
No	175	43.97

Table 2: Knowledge questions

Question	Correct answer	Answered correctly	
		Number	Percentage
Female condoms do not prevent pregnancy and STIs	False	351	88.19
Female condom cannot be used simultaneously with a male condom	True	183	45.98
Female condom cannot be put in place by female 8 hours before intercourse	False	294	73.87
Female condoms remove sexual arousal	False	363	91.21
Female condom has flexible ring at the closed end of the pouch with a slightly larger ring at the open end	True	352	88.44
At the open end of the sheath, the ring stays inside the vagina just around the cervix	False	113	28.39

Table 3: Association between participants' characteristics and knowledge of female condom

Variable	Knowledgeable (*n*=315)	Not knowledgeable (*n*=83)	*p*-value
Gender			
Female *n*=363	291 (80.17%)	72 (19.83%)	0.107
Male *n*=35	24 (68.57%)	11 (31.43%)	
Category of nurse			
Enrolled nurse *n*=71	56 (78.87%)	15 (21.13%)	0.085
Enrolled nursing assistant *n*=90	64 (71.11%)	26 (28.89%)	
Professional nurse *n*=237	195 (82.28%)	42 (17.72%)	
Family planning experience			
Yes *n*=248	213 (85.89%)	35 (14.11%)	**0.000**
No *n*=150	102 (68%)	48 (32%)	
Type of employer			
Local government *n*=76	70 (92.11%)	6 (7.89%)	**0.002**
Provincial government *n*=322	245 (76.09%)	77 (23.91%)	
Workplace			
Community health centre *n*=88	69 (78.41%)	19 (21.59%)	**0.001**
Clinic *n*=98	90 (91.84%)	8 (8.16%)	
District hospital *n*=212	156 (73.58%)	56 (26.42%)	
Formal training			
Yes *n*=109	90 (82.57%)	19 (17.43%)	0.302
No *n*=289	225 (77.85%)	64 (22.15%)	
Informal training			
Yes *n*=223	192 (86.10%)	31 (13.90%)	**0.000**
No *n*=175	123 (70.29%)	52 (29.71%)	

p≤0.05 is significant; indicated in bold

The association between participants' characteristics and knowledge of female condoms is shown in Table 3. Of all the participants' characteristics, having family planning experience, being employed by a local municipal government authority or working in a clinic and having had informal training on female condom use, were associated with being knowledgeable about female condoms.

Nurses who answered two of the knowledge questions – 'Female condoms do not prevent pregnancy and STI' and 'Female condoms remove sexual arousability' – correctly were more willing (61%) to promote female condom use.

The variables which showed significant association with the willingness of nurses to promote the use of female condoms, following a multinomial logistic regression analysis, are displayed in Table 4. Family planning experience, being employed by a local government authority, working in a clinic, having had formal or informal training or being a professional nurse were all significantly associated with willingness to promote female condom use. Although the majority (82%) of the nurses who were willing to promote female condom use were noted to be knowledgeable, there was no statistically significant association between these variables (Table 5).

Table 4: Characteristics of the nurses that were significantly associated with willingness to promote female condom use

Characteristic of nurse	*p*-value	RRR	s.e.	95% CI
Having a postgraduate degree	0.852	0.935	0.340	0.458–1.906
Having more experience as a nurse (years of practice)	0.940	0.967	0.433	0.402–2.328
Being a professional nurse	0.023*	1.739	0.475	1.018–2.971
Having family planning experience as a nurse	0.039*	1.539	0.322	1.020–2.320
Being a local government employee	0.001*	2.489	0.716	1.416–4.373
Working in a clinic	0.005*	1.999	0.624	1.084–3.685
Having formal training on female condom use	0.020*	1.729	0.409	1.087–2.749
Having informal training on female condom use	0.000*	2.089	0.433	1.392–3.137

RRR, relative risk reduction

*Significant at p≤0.05

Variable	Knowledgeable (*n*=315)	Not knowledgeable (*n*=83)
Willing to promote female condom use *n*=233	191 (81.97%)	42 (18.03%)
Not willing to promote female condom use *n*=165	124 (75.15%)	41 (24.85%)

Pearson chi-squared(1)=2.7243, p=0.099

Discussion

We found that 79% of participants had good knowledge on female condoms but only 59% of them reported willingness to promote female condom use among their patients. This worrying finding indicates that it is necessary to have effective mechanisms to increase not only knowledge on female condom use, but also motivation in healthcare providers to promote such use.

The low level of willingness has serious implications for reducing the spread of HIV and other STIs such as Ebola, Zika and hepatitis B.[3-5] While different studies have reported the role of healthcare providers' knowledge in promoting female condom use,[7,22,38] we suggest strategies to improve the motivation for healthcare workers to advocate use of the female condom to their patients. Considering that such studies are non-existent in the South African primary healthcare context, further studies are needed to explore the willingness of the healthcare providers to promote female condom use. While many international funding agencies – such as the WHO, Universal Access to Female Condom and the Swedish International Development Agency – promote female condom use, South Africa is not on the list of the Country Level Focus.[39] This signifies the need for local policies aimed at ensuring that female condom use is promoted as a priority means of contraception, especially within family planning[24,28] and HIV/Aids prevention programmes[2,4,6,7].

Our findings suggest that knowledge influences willingness to promote female condom use, although this association was not statistically significant. In a project introducing the female condom to the public health sector, Mantell et al.[17] reported that personal practice with the female condom generated a more positive attitude towards its promotion. It should be noted that 'hands-on' training was implemented in this project and therefore the relationship between knowledge and positive attitude was not very clear. The same variables that predicted knowledge were also used to predict willingness to promote the female condom. However, the relationship between variables and willingness was not as strong as those for knowledge. The demographic characteristics of the nurses in this study were not significantly associated with willingness to promote female condom use. This finding is similar to that of Oliveira da Silva et al.[18]

Some demographic characteristics were found to be associated with knowledge of female condoms. Nurses with family planning experience were three times more likely to have knowledge on female condoms than those without family planning experience. Whilst such a finding may be expected, this observation confirms the importance of ensuring that all nurses are trained in providing family planning services within the integrated primary healthcare services.

Participants employed by a local government authority were four times more likely to have knowledge on female condoms than those employed by the provincial government authority. Of the local government nurses, 75% were willing to promote female condom use, compared with 55% of provincial government nurses. Because local government clinics provide mostly prevention services, this finding may be a reflection of the nature of work expected of nurses in the different workplaces. The current salary dispensation favours nurses employed by a local government authority.[40] Nurses working in clinics were three times more likely to

have knowledge of female condoms than those working in the district hospital. Health promotion and prevention tend to be more emphasised in clinics than in the district hospital.

Nurses with informal training were three times more likely to have knowledge of female condoms than those without informal training. This result suggests that female condom providers would achieve better informal training with rotations in family planning units. Similarly, the use of a female condom through demonstration on a model and a step-by-step process of inserting could improve informal training.[35] These findings affirm those of Mantell et al.[17] The group of healthcare providers had 3 days' 'hands-on' training in the use of and the introduction to the female condom in South Africa. It should be noted that the post-training assessment was done immediately and showed increased knowledge. Our study indicates that informal training has the potential to improve nurses' knowledge of and willingness to promote female condom use in public healthcare facilities in the Johannesburg Health District.

Additionally, the possession of a university degree or postgraduate degree did not predict knowledge; however, formal training showed a statistically significant association with willingness to promote female condom. Studies previously cited have pointed out the relevance of training on female condom promotion.[8,17]

Although direct comparison is limited by different methodologies, our findings that higher degree qualification and formal training do not predict knowledge, are consistent with those in the literature.[17,18]

A majority of the participants knew that female condom use prevents pregnancy and STIs. This finding was expected, as this question relates to prevention of STIs and pregnancy and nurses are usually sent for different workshops/training on prevention of STIs and pregnancy. A significant number of nurses demonstrated adequate knowledge that the female condom 'can be placed up to 8 hours before intercourse', which was also noted in the study done by Oliveira da Silva et al.[18] The participants therefore knew the difference between the use of male and female condoms. A majority of the participants in our study had adequate knowledge on the female condom being 'a flexible ring at the closed end of the pouch with the slightly larger ring at the open end' compared with less than 40% who were familiar with female condom design in the study by Oliveira da Silva et al.[18] Similarly to the findings of Mantell et al.[6], most participants in our study had knowledge on 'Female condom removes sexual arousability'. This implies that the statement was clear and understandable. For two of the knowledge questions, either knowledge or clarity was poor. The question 'At the open end of the sheath, the ring stays inside the vagina just around the cervix' was answered correctly by only 28% of the participants. This is a technical question and can only be correctly understood by participants who had received training. More than half of the participants did not know whether 'Female condom can be used simultaneously with a male condom'. This statement was clear and understandable. The fact that most participants failed to answer it correctly may be a gap in the participants' knowledge.

The questionnaire was not comprehensive and we did not aim to cover all the important issues on female condom use. This study was focused on the association between providers' knowledge of and willingness to promote the use of the female condom. The review of the current literature shows that South Africa still lacks studies on this association as an important addition to sexual health in South Africa.

The study nevertheless had some limitations. This study had a cross-sectional design and at best only describes associations and not causality. Although family planning experience was associated with knowledge of female condom use, the duration of experience that might have an impact on knowledge and willingness was not explored. Because the questionnaire has not been validated, the questions on knowledge and willingness may not have exhaustively assessed what is required to uplift the knowledge and improve the willingness of nurses to promote female condom use. In addition, the computation of a knowledge/willingness score based on correct answers to a set of questions has not been validated. The analysis also did not incorporate differential weightings that may be placed on different questions.

Conclusion

The vast majority of nurses were knowledgeable about female condoms, but this knowledge did not translate into willingness to promote female condom use. The knowledge on simultaneous use of male and female condoms was insufficient. Clinical experience and having had specialty or postgraduate training did not predict the nurses' knowledge on female condoms. Nurses working in clinics were three times more likely to be knowledgeable and two times more likely to promote female condom use. Employment by local government was significantly associated with knowledge of and willingness to promote the use of female condoms. Professional nurses and nurses with a postgraduate qualification were more willing to promote the use of female condoms. Focusing on the informal training of nurses within the family planning unit has the potential to improve their knowledge and willingness to promote female condom use. Our results show that rotating through a family planning unit would not only improve knowledge on female condoms, but also help to change their attitudes towards female condom promotion.

Future research in a randomised controlled trial is suggested to test the relationship between knowledge of and willingness to promote the use of female condoms. Important research on the use of the female condom could be done in different post-natal facilities. Further studies could assess formal training on female condom use as an integral part of nurses' training. Motivation mechanisms need to be developed for the distribution of female condoms by healthcare service providers in the provincial government authority.

Acknowledgements

Author's contributions

This paper was based on the master's degree of M.S.P. at the University of the Witwatersrand. M.S.P and Z.B.E. conceived and designed the study; M.S.P. collected the data; M.S.P. and Z.B.E. performed and interpreted the statistical analyses; Z.B.E., O.B.O. and T.D-D. prepared the manuscript; the final drafting of Introduction, Results and Discussion was done by T.D-D., E.G.P. and M.S.P.; and T.D-D. and E.G.P. updated the literature review. E.Z. provided supervision. All co-authors contributed to the improvement of the final manuscript.

References

1. Matseke G, Peltzer K, Mchunu G, Louw J. Correlates of condom use among male and female aged 18-24 years in South Africa. Gender Behav. 2012;10(2):1–10. Available from: http://www.ajol.info/index.php/gab/article/view/83663

2. South African National AIDS Council. South Africa Global AIDS Response Progress Report (GARPR) [document on the Internet]. c2015 [cited 2017 Nov. 14]. Available from: http://sanac.org.za/wp-content/uploads/2016/06/GARPR_report-high-res-for-print-June-15-2016.pdf

3. Musso D, Roche C, Robin E, Nhan T, Teissier A, Cao-Lormeau V-M. Potential sexual transmission of Zika virus. Emerg Infect Dis. 2015;21(2):359–361. https://doi.org/10.3201/eid2102.141363

4. Oduyebo T, Petersen EE, Rasmussen SA, Mead PS, Meaney-Delman D, Renquist CM, et al. Update: Interim guidelines for health care providers caring for pregnant women and women of reproductive age with possible Zika virus exposure — United States, 2016. MMWR Morb Mortal Wkly Rep. 2016;65(5):122–127. https://doi.org/10.15585/mmwr.mm6505e2

5. Thorsen A, Formently P, Lofthouse C, Broutet N. Systematic review of the literature on viral persistence and sexual transmission from recovered Ebola survivors: Evidence and recommendations. BMJ Open. 2016;6(1), e008859, 7 pages. https://doi.org/10.1136/bmjopen-2015-008859

6. Mantell JE, West BS, Sue K, Hoffman S, Exner TM, Kelvin E, et al. Health care providers: A missing link in understanding acceptability of the female condom. AIDS Educ Prev. 2011;23(1):65–77. https://doi.org/10.1521/aeap.2011.23.1.65

7. Andrade SS, Zaccara AA, Leite KN, Brito KK, Soares MJ, Costa MM, et al. Knowledge, attitude and practice of condom use by women of an impoverished urban area. Rev Esc Enferm USP. 2015;49(3):364–372. https://doi.org/10.1590/S0080-623420150000300002

8. Hoffman S, Mantell J, Exner T, Stein Z. The future of the female condom. Perspect Sex Reprod Health. 2004;36(3):120–126. https://doi.org/10.1363/3612004

9. Holmes L, Ogungbade GO, Ward DD, Garrison O, Peters RJ, Kalichman SC, et al. Potential markers of female condom use among inner city African-American women. AIDS Care. 2008;20(4):470–477. https://doi.org/10.1080/09540120701867016

10. Holtgrave DR, Maulsby C, Kharfen M, Jia Y, Wu C, Opoku J, et al. Cost-utility analysis of a female condom promotion program in Washington, DC. AIDS Behav. 2012;16(5):1115–1120. https://doi.org/10.1007/s10461-012-0174-5

11. Maksut JL, Eaton LA. Female condoms=missed opportunities: Lessons learned from promotion-centered interventions. Womens Health Issues. 2015;25(4):366–376. https://doi.org/10.1016/j.whi.2015.03.015

12. Kerrigan D, Mobley S, Rutenberg N, Fisher A, Weiss E. The female condom: Dynamics of use in urban Zimbabwe. USAID contract no. HRN-A-00-97-00012-00, document number 153884. Washington DC: Population Council, Horizons Project; 2000.

13. Ruminjo JK, Steiner M, Joanis C, Mwathe EG, Thagana N. Preliminary comparison of the polyurethane female condom with the latex male condom in Kenya. East Afr Med J. 1996;73(2):101–106. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8756048

14. Feldblum PJ, Kuyoh MA, Bwayo JJ, Omari M, Wong EL, Tweedy KG, et al. Female condom introduction and sexually transmitted infection prevalence: Results of a community intervention trial in Kenya. AIDS. 2001;15(8):1037–1044. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11399986

15. Weeks MR, Zhan W, Li J, Hilario H, Abbott M, Medina Z. Female condom use and adoption among men and women in a general low-income urban U.S. population. AIDS Behav. 2015;19(9):1642–1654. https://doi.org/10.1007/s10461-015-1052-8

16. Agha S, Van Rossem R. Impact of mass media campaigns on intentions to use the female condom in Tanzania. Int Fam Plan Perspect. 2002;28(3):151–158. https://doi.org/10.2307/3088258

17. Mantell JE, Scheepers E, Karim QA. Introducing the female condom through the public health sector: Experience from South Africa. AIDS Care. 2000;12(5):589–601. https://doi.org/10.1080/095401200750003770

18. Oliveira da Silva N, Ferreira Moura SR, Guedes TG, Cesar de Almeida P. Conhecimento e promoção do uso do preservativo feminino por profissionais de unidades de referência para DST/HIV de Fortaleza-CE: o preservativo feminino precisa sair da vitrine [Knowledge and promotion of the female condom by professionals of reference units for STD/HIV Fortaleza-CE: The female condom needs to get out the window]. Saude Soc. 2008;17(1):107–116. Portuguese. https://doi.org/10.1590/S0104-12902008000100010

19. Spizzichino L, Pedone G, Gattari P, Luzi AM, Gallo P, Valli R, et al. The female condom: Knowledge, attitude, and willingness to use. The first Italian study. Ann Ist Super Sanità. 2007;43(4):419–424. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18209275

20. Mantell JE, Smit JA, Exner TM, Mabude Z, Hoffman S, Beksinska M, et al. Promoting female condom use among female university students in KwaZulu-Natal, South Africa: Results of a randomized behavioral trial. AIDS Behav. 2015;15(7):1129–1140. https://doi.org/10.1007/s10461-014-0860-6

21. Masvawure TB, Mantell JE, Mabude Z, Ngoloyi C, Milford C, Beksinska M, et al. "It's a different condom, let's see how it works": Young men's reactions to and experiences of female condom use during an intervention trial in South Afria. J Sex Res. 2014;51(8):841–851. https://doi.org/10.1080/00224499.2013.814043

22. Speizer IS, Hotchkiss DR, Magani RJ, Hubbard B, Nelson K. Do service providers in Tanzania unnecessarily restrict clients' access to contraceptive methods? Int Fam Plan Perspect. 2000;26(1):13–20. Available from: https://www.guttmacher.org/sites/default/files/article_files/2601300.pdf

23. Weeks MR, Coman E, Hilario H, Li J, Abbott M. Initial and sustained female condom use among low-income urban U.S. women. J Womens Health (Larchmt). 2013;22(1):26–36. https://doi.org/10.1089/jwh.2011.3430

24. Hayford SR, Agadjanian V. Providers' views concerning family planning service delivery to HIV-positive women in Mozambique. Stud Fam Plan. 2010;41(4):291–300. https://doi.org/10.1111/j.1728-4465.2010.00254.x

25. Mashanda-Tafaune B, Monareng LV. Perception and attitude of healthcare workers towards the use of a female condom in Gaborone, Botswana. Health SA Gesondheid. 2016;21:162–170. https://doi.org/10.1016/j.hsag.2015.12.002

26. Ng C, Pacqué-Margolis S, Kotellos K, Brantley S, IntraHealth International. Integrating family planning and HIV/Aids services: Health workforce considerations [document on the Internet]. c2012 [cited 2017 Nov. 14]. Available from: https://www.capacityplus.org/technical-brief-5/content/health-worker-attitudes-toward-clients

27. Ramkissoon A, Searle C, Burns C, Beksinska M. Chapter 3: Sexual and reproductive health and rights [document on the Internet]. c2010 [cited 2015 Jun 03]. Available from: http://sanacws.org.za/resources/sexual-and-reproductive-health-and-rights-in-south-africa

28. Holt K, Blanchard K, Chipato T, Nhemachena T, Blum M, Stratton L, et al. A nationally representative survey of healthcare provider counselling and provision of the female condom in South Africa and Zimbabwe. 2013;3(3):1–8. https://doi.org/10.1136/bmjopen-2012-002208

29. Bookshelf: National contraception and fertility planning policy and service delivery guidelines. Reprod Health Matter. 2014;22(43):200–203. https://doi.org/10.1016/S0968-8080(14)43764-9

30. Beksinska ME, Smit JA, Mantell JE. Progress and challenges to male and female condom use in South Africa. Sex Health. 2012;9(1):51–98. https://doi.org/10.1071/SH11011

31. Joanis C, Beksinska M, Hart C, Tweedy K, Linda J, Smit J. Three new female condoms: Which do South African women prefer? Contraception. 2011;83(3):248–254. https://doi.org/10.1016/j.contraception.2010.08.002

32. MatCH Research Unit (Maternal Adolescent and Child Health). Developing a strategy for female condom parallel programming: Proceedings of an expert meeting; 2011 Nov 23–24; Durban, South Africa. Durban: Universal Access to Female Condoms Joint Programme; 2011. Available from: http://www.hst.org.za/publications/NonHST%20Publications/Developing%20a%20Strategy%20for%20Female%20Condom_0.pdf

33. Beksinska M, Zulu B, Phungula L, Mabude Z, Nkosi P, Greener R, et al. Evaluation of the National South African Female Condom Programme: Parallel programming of female condoms [document on the Internet]. c2016 [cited 2017 Nov. 14]. Available from: http://www.femalecondoms4all.org/wp-content/uploads/2016/06/Parallel-Programming-of-Female-Condoms-Evaluation-of-the-South-African-Female-Condom-Programme.pdf

34. MatCH Research Unit, University of the Witwatersrand; HIV Center for Clinical and Behavioral Studies, New York State Psychiatric Institute and Columbia University Medical Center; USAID; and Project SOAR. Evaluation of South Africa's National Female Condom Programme. Washington DC: USAID; 2017. Available from: https://www.usaid.gov/sites/default/files/documents/1864/sa-female-condom-brief-508.pdf

35. Moore L, Beksinska M, Rumphs A, Festin M, Gollub EL. Knowledge, attitudes, practices and behaviors associated with female condoms in developing countries: A scoping review. Dovepress. 2015;2015(6):125–142. https://doi.org/10.2147/OAJC.S55041

36. Eraut M. Non-formal learning and tacit knowledge in professional work. Br J Educ Psychol. 2000;70(1):113–136. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10765570

37. Cook J, Smith M. Beyond formal learning: Informal community eLearning. Comput Educ. 2004;43(1–2):35–47. https://doi.org/10.1016/j.compedu.2003.12.003

38. Mantell JE, Hoffman S, Weiss E, Adeokun L, Delano G, Jagha T, et al. The acceptability of the female condom: Perspectives of family planning providers in New York City, South Africa, and Nigeria. J Urban Health. 2001;78(4):658–668. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11796812

39. Godhelp K. Advocating female condoms. A reflection on stakeholder commitments & trends in support for female condoms. End-line stakeholder analysis [document on the Internet]. c2015 [cited 2017 Nov. 14]. Available from: http://www.femalecondoms4all.org/Wp-content/uploads/2015/12/151210-Finalreport-end-line-stakeholder-analysis-UAFC-Rutgers.pdf

40. Bachmann MO, Makan B. Salary inequality and primary care integration in South Africa. Soc Sci Med. 1997;45(5):723–729. https://doi.org/10.1016/S0277-9536(96)00406-6

I believe I can do science: Self-efficacy and science achievement of Grade 9 students in South Africa

AUTHORS:
Andrea Juan[1] iD
Sylvia Hannan[1] iD
Catherine Namome[2] iD

AFFILIATIONS:
[1]Human Sciences Research Council, Durban, South Africa

[2]Human Sciences Research Council, Pretoria, South Africa

CORRESPONDENCE TO:
Andrea Juan

EMAIL:
ajuan@hsrc.ac.za

An important component of an individual's scientific literacy is a positive attitude towards science. However, emphasis is too often placed on achievement scores rather than attitude. While individuals' relative levels of problem-solving skills, inherent aptitudes for the subject matter and teaching practices are conveyed through achievement scores, attitudes to science convey individuals' emotional evaluation of the subject. Attitudes have a strong impact on behaviour: through either facilitating the learning process or hindering it. Furthermore, attitudes towards science reflect the culture which exists within a school, as well as the wider social context within which learning takes place. As a result, understanding attitudes is a key component of the interpretation of achievement results. We used data from 12 514 Grade 9 students in South Africa who participated in the 2015 Trends in International Mathematics and Science Study to investigate students' self-efficacy in science. Multiple linear regression analyses were used to address the following key research questions: (1) What is the relationship between self-efficacy and science achievement for Grade 9 students in South Africa? and (2) What are the contextual factors associated with the self-efficacy of Grade 9 students in South Africa? The findings reveal a positive relationship between self-efficacy and science achievement and suggest a need to also focus on non-cognitive aspects in order to improve science achievement.

Significance:
- The study contributes to understanding the determinants of science performance at school.
- The findings highlight the importance of non-cognitive dimensions in science achievement at school.
- The findings have policy implications for education programmes and teachers in relation to interventions which incorporate non-cognitive dimensions.

Introduction

Science is considered an important area of education in any country, as it contributes to increased science and technology knowledge and increased scientific development in higher education and other related fields, while scientific knowledge has an economic utility and cultural significance. It is therefore concerning that there has been a global decrease in the number of students choosing to pursue science. This shift is particularly apparent in the final years of secondary education.[1-3] South Africa is no exception to this trend. Many students perform poorly in science, and do not choose to continue with it, or may not qualify to study science at university.[4,5] A significant amount of science education research has therefore been dedicated to identifying ways to improve the quality of science education, and subsequently increase enrolments in science-related studies at the post-school level.[6] In the pursuit of remedying this situation, studies have examined the determinants of science achievement. One of the areas of study has been student attitudes. There has been an increased focus on the need to develop positive attitudes towards science, learning science and scientists.[1,3,7]

At a national level, the 2010 and 2013 South African Social Attitudes Survey results showed that generally South Africans expressed positive attitudes about science. However, the statements that measured attitudes about the benefits of science showed a general decline from 1999 when these questions were introduced to the study. An analysis of the items which measured scepticism towards science showed that the South African population seems to have become more cautious about the level of trust they place in science, and more concerned about the impacts of science.[8,9]

The survey also asked questions related to the attitudes of respondents towards science as a subject at school. The results showed that 41% of South Africans felt that the science learned at school was not useful in their daily lives, while almost a third (30%) did not see its value in their jobs. These findings may indicate that people are unaware of the applications of school science in their daily lives, or that school science is taught in a way that is disconnected from individuals' daily life experiences.[9]

Experiences of school science play a critical role in shaping peoples' attitudes towards the science they are taught at school, as well as science-related topics which they encounter outside of school.[6,10] Attitudes towards science that are formed early in life may have an impact on the relationships that individuals have as adults with science in general.[6] It is therefore important to understand the way in which attitudes are shaped and the associated behaviours in order to enhance science education in South Africa.

Positive attitudes towards science are considered to be a central component of an individual's scientific literacy, but are often overshadowed by the emphasis placed on science achievement scores.[6,11] Although achievement scores provide an understanding of an individual's aptitude and problem-solving skills, attitudes offer an emotional evaluation of science and are significant determinants of behaviour in relation to the learning process.[1,6,7] In addition, attitudes towards science are a reflection of the school climate and culture, and highlight the social context in which learning takes place.[6] As such, understanding attitudes is important in the interpretation of achievement results.

Creating positive attitudes towards science is a desirable outcome in and of itself; however, positive attitudes are also associated with science-related behaviours both at school and outside of school.[6,7]

In the 2015 Trends in International Mathematics and Science Study (TIMSS), Grade 9 students in South African schools scored an average of 358 points on a scale with a mean of 500 and a standard deviation of 100, and ranked last of the 39 participating countries, three of which participated in TIMSS at the Grade 9 level. Only 21% of the 12 514 South African students reported high levels of self-efficacy, with 45% reporting moderate levels and 35% reporting low levels of self-efficacy. Those students who reported high levels of self-efficacy scored 64 points higher in the science achievement test than those who reported low levels.[12] This finding has negative implications for the continuity of scientific endeavour, as well as for the scientific literacy of future generations, and the country's innovation economy.

The commitment and motivation to learn science can be influenced by whether students enjoy the subject, attach value to it in terms of its usefulness to both themselves and society, and by their science self-efficacy which relates to their self-confidence in their ability to accomplish science-related tasks and activities. In order to investigate students' self-efficacy in relation to science, we address two key research questions:

(1) What is the relationship between self-efficacy and science achievement for Grade 9 students in South Africa?

(2) What are the contextual factors associated with the self-efficacy of Grade 9 students in South Africa?

Literature review

Self-efficacy is a core feature of social cognitive theory, which posits that people do not merely respond to influences within their environment, but actively seek out and interpret information in order to make decisions. This human agency relies, in part, on a person's beliefs in their efficacy to exercise control over their own functioning, as well as events that have an impact on their lives.[13] The founder of the theoretical underpinnings of self-efficacy, Albert Bandura, stated that self-efficacy relates to a personal expectation of one's ability to perform in order to reach specific goals.[13] When applying this to education, the expectation, in turn, affects students' motivation, interest and performance in a subject. This relates to all subjects, including science. Efficacy beliefs therefore affect how much effort people apply to an activity, how long they will continue when encountering obstacles (persistence), and how resilient they are when confronted with difficult situations.[7,13-15] Resilience contributes to the alleviation of stress and promotes adaptation and the development of skills to cope with adversity and change.[14,16,17]

The sources of self-efficacy can be grouped into four categories:

- *Mastery experience* which results from students engaging in science tasks, interpreting the outcomes of those tasks and using this interpretation to formulate beliefs about capability and capacity to perform the task. Successful outcomes result in confidence.

- *Vicarious experience* which stems from students observing peers performing tasks, and then evaluating their own prospect of success in similar tasks.

- *Social persuasion* which refers to judgements that other people make regarding the capabilities of the student. Positive persuasions build stronger beliefs in capabilities and in the successful attainment of goals.

- *Physiological states* experienced when engaging in science tasks: anxiety, stress or excitement. The degree of confidence is gauged by the physical state that is experienced while engaging in the task.[18]

Self-efficacy emerges from what students infer from these sources, and teachers can enhance students' self-efficacy through understanding and utilising these sources.[15,19] Schunk[20], for example, highlights the impact of modelling (reproducing the strategies and behaviours of one or more models, which may be teachers or peers, and patterning one's thoughts

and beliefs after theirs), goal setting and self-evaluation on self-efficacy. Pajares highlights that teachers should assess and attempt to develop students' self-efficacy as they progress through school, as these beliefs 'can have beneficial or destructive influences'[15(p.120)].

Factors associated with self-efficacy

The strength of self-efficacy also has an impact on changes in behaviour, as those who have a higher self-efficacy are more likely to exhibit greater persistence in achieving success, no matter what the level of difficulty, and they approach these tasks as challenges to be overcome.[21,22] Conversely, those who have low self-efficacy believe that tasks are more difficult than they realistically are, and consequently experience stress and anxiety when facing these challenges.[23] Low self-efficacy beliefs have a negative impact on academic achievement, and can, over time, 'create self-fulfilling prophecies of failure and learned helplessness that can devastate psychological well-being'[19(p.219)].

Although the findings regarding the link between self-efficacy and science achievement vary, most of the research has shown that there is a significant correlation between the two constructs.[15,18,24-26] This positive relationship remains across levels of education, and is strongly associated with taking science-related majors and the pursuit of science-related careers.[10,18]

Studies have found gender differences related to self-efficacy.[15] However, this association differs internationally, and has been attributed to differences in the opportunities which are afforded to female individuals.[7,27] Britner and Pajares[18] found that gender differences found in the sources of self-efficacy were minimal. In their study, girls reported greater anxiety about their performance in school science, but were more confident in their ability to manage their studies successfully.

The influence of other attitudes to science, in terms of enjoyment and value, has also been examined. Students' enjoyment of science and the value they place on it in terms of its usefulness, can affect their commitment and motivation to learn the subject.[7]

The role of the teacher in terms of encouragement, teaching approach and the learning environment they create is also important.[15,27] Positive teacher-directed instruction involves the explicit teaching of scientific rules, principles, concepts and strategies for problem-solving. This instruction often includes modelling a variety of examples and guiding students through their review and practice. If students do not have a clear understanding of how a task will look when it is completed, their efforts to complete the task will often be ineffective. Feedback from teachers clarifies areas in which a student is doing well, but may also show how far a student must improve. This feedback is a form of social persuasion.[18] It refers to judgements that teachers make regarding the capabilities of the student. Positive persuasions build stronger beliefs in capabilities and in the successful attainment of goals. Self-efficacy and achievement may be enhanced through teaching practices that incorporate progress feedback, modelled strategies, self-evaluation and goal setting. Through the use of such methods, teachers can promote improved academic performance and motivation for continued learning.[15,20] Resilience is affected by the mindset of students, and students who have the belief, or are taught, that they can develop their intellectual abilities, tend to perform better. Students with a growth mindset therefore persevere when faced with challenges.[28,29]

When examining academic achievement in South Africa, the role of socio-economic status (SES) cannot be overlooked. The educational outcomes of students in South African schools are influenced by the socio-economic imbalances which exist in the country. Students attending schools in the richest quintile (quintile 5) significantly outperform students from schools in the other four quintiles.[30] Further important factors are parental involvement and attitudes, with students from family environments in which parents highly value science being more likely to achieve well in science.[31,32] Various forms of parental involvement have been found to be associated with self-efficacy, including involvement in school functions and extra-curricular activities, modelling persistence and teaching children how to cope with difficulties, encouraging children to try various activities and supporting their efforts,

and providing an environment that allows for mastery experiences and stimulates curiosity.[33,34] Students from high SES families are more likely to achieve better results in science[35,36], and students who attend schools where their peers come from higher SES families also outperform their peers at schools with lower concentrations of wealth.[37] Students from families with higher SES have more home resources, such as computers and books, as well as 'social capital' in terms of supportive relationships among schools and individuals which highlight societal norms and values, which lead to school success.[34,36,38]

Methodology

The data used for this article were from the TIMSS 2015 study conducted by the International Association for the Evaluation of Educational Achievement (IEA). The sampling procedure followed a two-stage stratified cluster sample design.[40] In the first stage the population of South African schools that offered Grade 9 classes in 2013 was stratified according to province, type of school (public or independent) and language of instruction.[39] Schools were then sampled using a systematic random sampling approach. In the second stage of sampling, intact classes were selected with equal probability from each participating school.[40] The 2015 sample consisted of 12 514 students in 292 schools.

Measures and variables

This section details the variables that were used and prepared for inclusion in regression analyses for the purpose of addressing the research questions. In the TIMSS contextual questionnaire for students, questions fall within three broad measures of attitudes and self-belief. These measures include students enjoying science, valuing science and self-efficacy in science. The scaled index for self-efficacy was created by the IEA from the following statements: 'I usually do well in science'; 'Science is not one of my strengths'; I learn things quickly in science'; 'Science is harder for me than any other subject'; 'I am good at working out difficult science problems' (all indicating mastery experience); 'Science is more difficult for me than for many of my classmates' (vicarious experiences); 'My teacher tells me I am good at science' (social persuasion); 'Science makes me confused' (physiological states).[12] Indices for enjoyment of and valuing science were also created in a similar fashion by the IEA. The enjoyment index was created from the following statements: 'I enjoy learning science'; 'Science is boring'; 'I like science'; 'I wish I did not have to study science'; 'I learn many interesting things in science'; 'I look forward to learning science in school'; 'Science teaches me how things in the world work'; 'I like to conduct science experiments'; and 'Science is one of my favourite subjects'.[12]

The valuing index was created from the following statements: 'I think learning science will help me in my daily life', 'I need science to learn other school subjects'; 'I need to do well in science to get into the university of my choice'; 'I would like a job that involves science'; 'I need to do well in science to get the job I want'; 'Learning science will give me more job opportunities when I am an adult'; 'It is important to learn about science to get ahead in the world'; 'My parents think that it is important that I do well in science'; and 'It is important to do well in science'.[12]

Other variables in the regression model included students' classroom and home experiences. The index for classroom experiences was created by the IEA from the following statements: 'I know what my teacher expects me to do'; 'My teacher is easy to understand'; 'I am interested in what my teacher says'; 'My teacher gives me interesting things to do'; 'My teacher has clear answers to my questions'; 'My teacher is good at explaining science'; 'My teacher does a variety of things to help us learn'; 'My teacher lets me show what I have learned'; 'My teacher tells me how to do better when I make a mistake'; and 'My teacher listens to what I have to say'. Home experiences were assessed by the following items: 'My parents ask me what I am learning in school'; 'My parents make sure that I set aside time for my homework'; and 'My parents check if I do my homework'. Responses to these statements were converted to dummy variables, where one (1) indicates that the activity was done and zero (0) indicates the absence thereof.

The values for all indices were converted to scale scores (all indices were put on a common scale to account for differences across the indices) and then converted to z scores (mean=0, standard deviation=1). Using z-scores enables the comparison of two or more scores that are from different normal distributions.

The following variables were included in this regression analysis as controlling variables: gender, age, squared age and home SES. The home SES is a composite variable with 14-item variables if a learner indicated having a fridge, television, own room, own/shared computer, Internet connection, landline telephone, motor car, water flush toilets, running tap water, electricity, dictionary, a gaming system, own cell phone and a study desk. Using factor scores from a principal component analysis as weights, a home SES variable was then constructed for each learner. The eigenvalue (variance) for each principal component indicated that the principal component accounted for 79% of the variation. In using the square of the age variable, we were testing a quadratic relationship between age, self-efficacy and science achievement. The assumption was that the age variable was non-linearly related to science achievement and self-efficacy.

The TIMSS 2015 study included a science test. Science scores were reported on a common scale with a centre point of 500 and standard deviation of 100 by using item response theory.

Data analysis

Using multiple regression analyses, we developed two models to investigate whether self-efficacy explains science performance, and whether it is associated with student contextual factors. Student measures that were statistically significant at the 95% ($p<0.05$) and 99% ($p<0.01$) confidence levels were identified. The first model explains the association between science performance and self-efficacy while controlling for attitudinal factors including enjoying science and valuing science as well as gender, age, square of age and home SES. In the second model, we investigated the association between self-efficacy and students' home and school experiences controlling for home SES, science performance, gender, age and the square of age. The data analysis was done with SPSS version 24 and the IEA International Database (IDB) Analyzer (version 3.1.17), which is a plug-in for SPSS developed by the IEA specifically to analyse data from IEA surveys. The software takes into account the complex sample design and makes use of plausible value technology. The IDB Analyzer generates SPSS syntax that takes into account the sampling design in the computation of sampling variance, and the plausible values used to compute achievement scores. The IDB Analyzer enables the calculation of descriptive statistics and conducts statistical hypothesis testing among groups in the population without the need for programming code.[41] All models estimated in our analysis rely on this procedure.

Results

Table 1 sets out the means and standard deviations of the variables and constructed indices. Table 2 represents the results of the multiple regression analysis conducted to test if self-efficacy significantly predicts overall science performance. This model accounted for 27% of the variance in science achievement (adjusted $R^2=0.27$, $p<0.00$). The gender, home SES and age variables behave as expected. On average, girls tended to score almost 6 points lower than boys ($\beta=$ -5.86, $p<0.01$) and each additional home asset was associated with a 31-point increase in science scores ($\beta=30.86$, $p<0.01$). The findings regarding age show that science achievement is positively associated with age ($\beta=90.82$, $p<0.01$) up to a point but as students get older (above the grade appropriate age), the effect of age is negative ($\beta=$ -3.72, $p<0.01$).

After taking into account other key background characteristics such as gender, age and home SES, as well as other measures of attitude, self-efficacy still had a significant positive association with science achievement. Every standard deviation increase in the self-efficacy index was associated with a 7 scale score point increase in science achievement ($\beta=7.29$, $p<0.01$). Interestingly, valuing science was associated with a decrease in science achievement. This finding

suggests that valuing science alone is not a strong enough attitude to ensure that students perform well in science. This result highlights the need to focus on encouraging the enjoyment of science, as well as building students' self-efficacy in science.

Table 3 presents results of the second model – a multiple regression analysis conducted to determine the contextual factors associated with self-efficacy. This model accounted for 31% of the variance in self-efficacy (adjusted R^2=0.31, p<0.00). The model shows that after controlling for gender, age and home SES, interaction between teachers and students (classroom experiences) had the highest effect on self-efficacy (β= 0.51, p<0.01), science achievement had the second highest effect (β= 0.17, p<0.01) and parents checking homework had a very low (β= 0.04, p<0.05) but significant effect on self-efficacy.

Students were asked how often certain activities took place at home. These activities included parents wanting to know what was learned at school and parents checking homework, and ensuring that time was set aside for homework. The responses to these questions reflected the different levels of engagement between parents and students. The act of parents checking whether homework was done was found to be positively associated with the expression of self-efficacy by students.

Students whose parents checked their homework had a 0.04 of a standard deviation score higher for self-efficacy than students whose parents did not check their homework, with all other factors in the model taken into account (p<0.05). This relationship remained, even when factors such as science achievement, home SES and gender were taken into account. This finding points to a need to raise awareness among parents and caregivers of the importance of engaging with students about their homework and reading about science. Higher levels of home SES were related to more positive attitudes about science.

Classroom experiences, as measured by classroom interactions between students and teachers, were assessed by examining instructional practices. The focus of the items that constituted the index was the degree of engagement between student and teacher. It was found that positive classroom interactions were significantly associated with increases in self-efficacy (β= 0.51, p<0.01).

The influence of gender is striking. When factors such as science achievement, home SES and classroom experiences were controlled for, girls were less likely to have confidence in their ability to learn science (β= -0.05, p<0.05).

Table 1: Descriptive statistics of predictor variables

Variable *N*=12 514	Continuous variable		Dummy variable	
	Mean	**Standard deviation**	**Percentage of students with dummy=1**	**Percentage of students with dummy=0**
Science achievement	372	100		
Self-efficacy	10.20	1.95		
Enjoy learning science	10.64	1.93		
Value science	10.85	1.90		
Age	15	1.22		
Age²	249	39.56		
Home socio-economic status	2.91	0.95		
Gender (girl)			51	49
Parents ask me what I am learning in school			63	37
Parents make sure that I set aside time for my homework			66	34
Parents check if I do my homework			57	43

Author's own calculations from the TIMSS 2015 database[42]

Table 2: Multiple regression results investigating the association between science performance and self-efficacy

Explanatory variable	Regression coefficient	Regression coefficient (s.e)	Regression coefficient (*t*-value)	Standardised coefficient	Standardised coefficient (s.e.)	Standardised coefficient (*t*-value)	*p*-value
Gender (girl)	-5.86	1.90	-3.08	-0.05	0.02	-3.03	<0.01
Age	90.82	25.49	3.56	1.03	0.29	3.61	<0.01
Age²	-3.72	0.78	-4.74	-1.37	0.28	-4.88	<0.01
Home socio-economic status	30.81	2.92	10.57	0.27	0.02	12.96	<0.01
Enjoy science	16.58	2.58	6.42	0.16	0.03	6.16	<0.01
Value science	-14.23	2.10	-6.78	-0.14	-0.02	-6.32	<0.01
Self-efficacy	7.29	2.63	2.77	0.07	0.02	2.83	<0.05
Constant	-232.98	208.98	-1.11	–	–	–	

Authors' own calculations from the TIMSS 2015 database[42]

Table 3: Multiple regression results investigating the association between self-efficacy and contextual factors

Explanatory variables	Regression coefficient	Regression coefficient (s.e)	Regression coefficient (*t*-value)	Standardised coefficient	Standardised coefficient (s.e.)	Standardised coefficient (*t*-value)	*p*-value
Gender (girl)	-0.05	0.02	-2.81	-0.04	0.02	-2.84	<0.05
Age	-0.35	0.20	-1.70	-0.41	0.23	-1.75	
Age^2	0.01	0.01	1.52	0.36	0.23	1.56	
Home socio-economic status	0.05	0.01	4.40	0.05	0.01	4.32	<0.01
Science achievement	0.00	0.00	8.59	0.17	0.02	8.29	<0.01
Parents ask what has been learnt in school	0.00	0.01	0.19	0.00	0.01	0.19	
Parents make sure that time is set aside for homework	0.01	0.01	0.91	0.01	0.01	0.91	
Parents check if homework is done	0.04	0.02	2.33	0.04	0.02	2.37	<0.05
Classroom experiences	0.51	0.02	31.01	0.51	0.01	48.79	<0.01
Constant	2.33	1.68	1.39	.	.	.	

Author's own calculations from the TIMSS 2015 database[42]

Discussion and conclusion

The aim of our study was to determine the degree to which self-efficacy beliefs are associated with science achievement of a sample of Grade 9 students in South Africa. In addition, we examined contextual factors that may shape science self-efficacy in these students. Self-efficacy was found to be positively associated with achievement. This finding is consistent with a number of studies such as Britner and Pajares[18], Singh et al.[24], Juan et al.[25] and Sabah and Hammouri[26]. Higher levels of enjoyment, which had the highest association with science achievement compared with the other two attitudinal indices (value and self-efficacy in science), and improved self-efficacy, may therefore translate into more effort being put into learning and understanding school science.

The second model examined the contextual shapers of self-efficacy. The importance of positive classroom experiences in relation to the engagement between teachers and students was highlighted. The implication is that, at the school level, it is crucial to invest in approaches that inculcate positive attitudes and learning behaviours, and to view this as a key goal of quality education. Students spend the majority of their time at school, and schools are therefore the primary settings for developing students' self-efficacy. Schools should develop policies that highlight the crucial role that teachers play in shaping attitudes towards science, and ensure that techniques are employed to foster positive attitudes in general and self-efficacy in particular.

In order to promote self-efficacy in science, classroom teaching should incorporate practices which motivate students and enhance their confidence in their abilities. Children rely on the judgement of others to make their own judgements of confidence and self-value (social persuasion). As such, teachers have a responsibility, beyond classroom instruction, to nurture the self-beliefs of their students.[15] Teachers should also aim to increase their students' confidence as they progress through school, starting as early as possible so that self-efficacy is developed as a habit.[15]

Providing students with complex, meaningful tasks that can be mastered while offering support and encouragement may help to develop students' self-efficacy (through mastery experience). Individualised classroom structures and instructions which are tailored to the academic capabilities of students may minimise social comparisons. Students may therefore be inclined to judge their progress in relation to their own standards rather than in comparison with their classmates' progress (vicarious experience).[15]

Encouraging a growth mindset, in which students are taught that they can develop their intellectual abilities over time, thereby encouraging resilience when encountering challenges, may provide a piece of the puzzle.[28,29] This requires fostering a school culture in which students are encouraged and teachers are provided with the knowledge and skills to enable them to cultivate their students' self-efficacy.

Outside of the classroom environment, parental involvement in checking homework was found to be related to students' self-efficacy. This finding is consistent with other studies which have found that various forms of parental involvement are associated with students' self-efficacy.[33,34] This finding also highlights a need to further involve parents and caregivers in their children's science learning and emphasises their role in engaging with students about their homework and reading about science. School policies should focus on encouraging the active involvement of parents in the educational process. In line with other studies[35,36], we found that higher levels of home SES were related to more positive attitudes about science. This result is a concern, as a large percentage of South African students come from households with low home SES. Encouraging positive classroom interactions and parental involvement for all students, and particularly those of lower socioeconomic status, is therefore critical for increasing students' self-efficacy.

The stereotypical view is that doing science is more consistent with a male self-image than with a female one. We examined whether a gender-related image of science impacted the self-efficacy that students expressed towards science. Unlike the findings of Britner and Pajares[18], we found that when factors such as science achievement, home SES and classroom experiences were controlled for, girls still reported lower levels of self-efficacy than did boys. This may be a result of the 'traditional' gender roles which are ingrained in society, where science has historically been seen as a career for men rather than women. This low level of self-efficacy may have future implications for subject choice and the representation of women in science, technology, engineering and mathematics (STEM) careers and the continued gender bias in STEM endeavours. Particular attention must be paid to inequalities between the genders in relation to the effectiveness with which schools, and societies as a whole, promote motivation and interest in science. In the classroom, teachers should be

gender sensitive in terms of actively engaging both genders in the lesson. Parents and teachers should encourage girls to tackle academic and social obstacles with confidence, and challenge the gender stereotypes associated with specific academic fields.[15] Furthermore, emphasis should be placed on academic success in science as being a result of sustained effort rather than an outcome of individual characteristics such as gender.[15] Programmes that use female role models may encourage girls to recognise that they are capable of high science achievement and pursuing science careers. At a broader societal level, initiatives that challenge the pervasive discourse surrounding gender appropriate subjects and careers should be embarked upon by the national government. This complex phenomenon requires further research.

Our findings point to a need to acknowledge the role of attitudes in academic science achievement by policymakers, parents and teachers. This acknowledgement must be accompanied by initiatives that actively develop a positive sense of self-efficacy in students. Additionally, the findings point to the importance of socialising children early on about the benefits of science, in order to promote interest and science self-efficacy.

Acknowledgements

Author's contributions

A.J.: conceptualisation, methodology, project leadership, overall writing. S.H.: literature review, overall writing. C.N.: data analysis, sample analysis.

References

1. Osborne J, Simon S, Collins S. Attitudes towards science: A review of the literature and its implications. Int J Sci Educ. 2003;25:1049–1079. https://doi.org/10.1080/0950069032000032199

2. Barmby P, Kind PM, Jones K. Examining changing attitudes in secondary school Science. Int J Sci Educ. 2008;30:1075–1093. https://doi.org/10.1080/09500690701344966

3. Sarjou AA, Soltani A, Kalbasi A, Mahmoudi S. A study of Iranian students' attitude towards science and technology, school science and environment, based on the ROSE Project. J Stud Educ. 2012;2(1):90–103. https://doi.org/10.5296/jse.v2i1.1438

4. Mji A, Makgato M. Factors associated with high school learners' poor performance: A spotlight on mathematics and physical science. S Afr J Educ. 2006;26:253–266.

5. Martin MO, Mullis IVS, Foy P, Stanco GM. TIMSS 2011 international results in science. Chestnut Hill, MA/Amsterdam: Boston College/IEA Secretariat; 2012.

6. Juan A, Reddy V, Zuze TL, Namome C, Hannan S. Does it matter whether students enjoy learning science? Exploring student attitudes towards science in South Africa. HSRC Policy Brief [document on the Internet]. c2016 [cited 2017 Jul 04]. Available from: http://www.timss-sa.org.za/download/Attitudes%20Policy%20Brief.pdf

7. Hassan G. Attitudes toward science among Australian tertiary and secondary school students. Res Sci Technol Educ. 2008;26:129–147. http://dx.doi.org/10.1080/02635140802034762

8. Reddy V, Gastrow M, Juan A, Roberts B. Public attitudes to science in South Africa. S Afr J Sci. 2013;109(1/2), Art. #1200, 8 pages. http://dx.doi.org/10.1590/sajs.2013/1200

9. Reddy V, Juan A, Hannan S, Arends F, Gastrow M, Roberts B. Science awareness, attitudes and astronomy: Continuity and changes [document on the Internet]. c2015 [cited 2017 Jul 12]. http://www.timss-sa.org.za/download/SASAS%20final%20report.pdf

10. Lyons T. Different countries, same science classes: Students' experiences of school science in their own words. Int J Sci Educ. 2006;28:591–613. http://dx.doi.org/10.1080/09500690500339621

11. Bybee R, McCrae B. Scientific literacy and student attitudes: Perspectives from PISA 2006 science. Int J Sci Educ. 2011;33:7–26. http://dx.doi.org/10.1080/09500693.2010.518644

12. Martin MO, Mullis IVS, Foy P, Hooper M. TIMSS 2015 international results in science [document on the Internet]. c2016 [cited 2017 Jun 22]. Available from: http://timssandpirls.bc.edu/timss2015/international-results/

13. Bandura A. The explanatory and predictive scope of self-efficacy theory. J Soc Clin Psychol. 1986;4:359–373. https://doi.org/10.1521/jscp.1986.4.3.359

14. Keye D, Pidgeon AM. An investigation of the relationship between resilience, mindfulness, and academic self-efficacy. Open J Soc Sci. 2013;1(6):1–4. http://dx.doi.org/10.4236/jss.2013.16001

15. Pajares F. Gender and perceived self-efficacy in self-regulated learning. Theor Pract. 2002;41:116–125. http://www.jstor.org/stable/1477463

16. Wang MC, Haertel GD, Walberg HJ. Educational resilience in inner cities. In: Wang MC, Gordon EW, editors. Educational resilience in inner-city America: Challenges and prospects. Hillsdale, NJ: Lawrence Erlbaum Associates; 1994.

17. Ahern NR, Kiehl EM, Sole ML, Byers J. A review of instruments measuring resilience. Issues Compr Pediatr Nurs. 2006;29:103–125. http://dx.doi.org/10.1080/01460860600677643

18. Britner SL, Pajares F. Sources of science self-efficacy beliefs of middle school students. J Res Sci Teach. 2006;43:485–499. https://doi.org/10.1002/tea.20131

19. Margolis H, McCabe PP. Improving self-efficacy and motivation: What to do, what to say. Interv Sch Clin. 2006;41:218–227. https://doi.org/10.1177/10534512060410040401

20. Schunk DH. Self-efficacy for reading and writing: Influence of modelling, goal setting and self-evaluation. Read Writ Q. 2003;19:159–172. https://doi.org/10.1080/10573560308219

21. Bandura A. Perceived self-efficacy in cognitive development and functioning. Educ Psychol. 1993;28:117–148. https://doi.org/10.1207/s15326985ep2802_3

22. Tuan H, Chin C, Shieh S. The development of a questionnaire to measure students' motivation towards science learning. Int J Sci Educ. 2005;27:639–654. http://dx.doi.org/10.1080/0950069042000323737

23. Pajares F. Self-efficacy beliefs in academic settings. Rev Educ Res. 1996;66:543–578. http://dx.doi.org/10.3102/00346543066004543

24. Singh K, Granville M, Dika S. Mathematics and science achievement: Effects of motivation, interest, and academic engagement. J Educ Res. 2002;95:323–332. https://doi.org/10.1080/00220670209596607

25. Juan AL, Reddy V, Hannan S. Attitudes to science: Part of the puzzle to improve educational achievement? Africa Growth Agenda. 2014;7(1):13–16. http://hdl.handle.net/10520/EJC163568.

26. Sabah S, Hammouri H. Does subject matter matter? Estimating the impact of instructional practices and resources on student achievement in science and mathematics: Findings from TIMSS 2007. Eval Res Educ. 2012;23:287–299. http://dx.doi.org/10.1080/09500790.2010.509782

27. George R. A crossdomain analysis of change in students' attitudes toward science and attitudes about the utility of science. Int J Sci Educ. 2006;28:571–589. http://dx.doi.org/10.1080/09500690500338755

28. Dweck C. Mindset: The new psychology of success. New York: Random House Inc.; 2006.

29. Yeager DS, Dweck C. Mindsets that promote resilience: When students believe that personal characteristics can be developed. Educ Psych. 2012;47:302–314. http://dx.doi.org/10.1080/00461520.2012.722805.

30. Van der Berg S. How effective are poor schools? Poverty and educational outcomes in South Africa. Stud Educ Eval. 2008;34:145–154. https://doi.org/10.1016/j.stueduc.2008.07.005

31. George R. Measuring change in students' attitudes toward science over time: An application of latent variable growth modeling. J Sci Educ Technol. 2000;9:213–225. https://doi.org/10.1023/A:1009491500456

32. Perera LDH, Bomhoff EJ, Lee GHY. Parents' attitudes towards science and their children's science achievement. Discussion paper 02/14. Melbourne: Monash University; 2014.

33. Fan W, Williams CM. The effects of parental involvement on students' academic selfefficacy, engagement and intrinsic motivation. Educ Psych. 2010;30:53–74. http://dx.doi.org/10.1080/01443410903353302

34. Schunk DH, Pajares F. The development of academic self-efficacy. In: Wigfield A, Eccles J, editors. A volume in the educational psychology series. Development of achievement motivation. San Diego, CA: Academic Press; 2002.

35. Özdemir E. Modelling of the factors affecting science achievement of eighth grade Turkish students based on the third international mathematics and science study – Repeat (TIMSS - R) data [dissertation]. Ankara: The Middle East Technical University; 2003.

36. Sun L, Bradley KD, Akers K. A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. Int J Sci Educ. 2012;34:2107–2125. http://dx.doi.org/10.1080/09500693.2012.708063

37. Webster BJ, Fisher DL. Accounting for variation in science and mathematics achievement: A multilevel analysis of Australian data third International Mathematics and Science Study (TIMSS). Sch Eff Sch Improv. 2000;11:339–360. http://dx.doi.org/10.1076/0924-3453(200009)11:3;1-G;FT339

38. Dika SL, Singh K. Applications of social capital in educational literature: A critical synthesis. Rev Educ Res. 2002;72:31–60. https://doi.org/10.3102/00346543072001031

39. Reddy V, Visser M, Winnaar L, Arends F, Juan A, Prinsloo CH, et al. TIMSS 2015: Highlights of mathematics and science achievement of Grade 9 South African students. Pretoria: Human Sciences Research Council; 2016.

40. LaRoche S, Joncas M, Foy P. Sample design in TIMSS 2015. In: Martin MO, Mullis IVS, Hooper M, editors. Methods and procedures in TIMSS 2015 [document on the Internet]. c2016 [cited 2017 Jul 24]. Available from: http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

41. Foy P. TIMSS 2015 user guide for the international database [document on the Internet]. c2017 [cited 2017 Jun 15]. Available from: https://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_UserGuide.pdf.

42. Human Sciences Research Council. Trends in International Mathematics and Science Study (TIMSS) 2015 Grade 9: Learner background (BSG) – South Africa as one of 39 countries [data set]. In: TIMSS 2015 learner background. Version 1.0. Pretoria: Human Sciences Research Council [producer]; 2015. Human Sciences Research Council [distributor]; 2017. http://dx.doi.org/10.14749/1499247520

Metafrontier analysis of commercial and smallholder tomato production: A South African case

AUTHORS:
Jabulile Z. Gwebu[1,2]
Nicolette Matthews[1]

AFFILIATIONS:
[1]Department of Agricultural Economics, University of the Free State, Bloemfontein, South Africa

[2]Agricultural Research Council, Pretoria, South Africa

CORRESPONDENCE TO:
Nicolette Matthews

EMAIL:
MatthewsN@ufs.ac.za

South African agriculture is a dualist agricultural system with well-developed commercial farmers and resource-poor smallholder farmers. In an effort to address the dualist nature of agriculture, the South African government has developed a strategic plan to assist smallholder farmers in entering commercial markets. The strategic plan aims to advance subsistence and smallholder farmers into commercial production through improved resource management for sustainable food security and smallholder livelihood. However, the productivity of smallholder farmers continues to be very low compared with that of commercial farmers. Our aim was to compare tomato productivity for commercial and smallholder tomato farmers in the Nkomazi area (Mpumalanga Province) using a metafrontier analysis. We used an output-oriented data envelopment analysis metafrontier approach and the Tobit model to investigate smallholder and commercial farmers' technical efficiencies and related factors which affect tomato production. Results indicate that smallholder farmers have high levels of technical efficiency compared to the group frontier (0.74), but they are less technically efficient compared to the metafrontier (0.51). The group efficiencies of the smallholder farmers also showed a large variation ranging from 3% to 100%, while commercial farmers have high levels of efficiency compared to both the group frontier (0.89) and the metafrontier (0.88). Results from the Tobit regression indicate that farmers' managerial decisions are an important determinant of their technical efficiency. We conclude that smallholder farmers first need to increase their level of technical efficiency relative to their peers before aiming to compete with commercial farmers.

Significance:
- Smallholder farmers should first improve their resource use efficiency compared to their fellow smallholder farmers before they consider comparing themselves against the commercial farmers.

Introduction

The structure of the agricultural sector in South Africa is dualistic in nature, with the sector comprising the well-developed commercial sector and the resource-poor smallholder farmers.[1] Sandrey and Vink[2], as cited by Tshuma[3], argue that the commercial sector in South Africa consists of a few very big, successful and profit-oriented farmers who mostly use advanced production technology. The smallholder-farming sector is, however, dominant in rural areas of South Africa where about 70% of the poorest households are found.[3] According to the National Department of Agriculture[4], commercial agriculture follows a more capital-intensive growth path while substantial agricultural resources lie unused or underutilised in rural areas.

In 2001, a strategic plan for South African agriculture was established[4] with the vision of a united and prosperous agricultural sector. The vision was designed to bridge the inherent dualism and to maximise the contribution of the agricultural sector for economic growth and development. Since then the Department of Agriculture, Forestry and Fisheries[5] developed a strategic plan to assist smallholder farmers to enter commercial markets. The strategic plan aims to advance subsistence farmers into commercial production through improved resource management for sustainable food security and smallholder livelihood. However, the productivity of smallholder farmers continues to be very low compared with that of commercial farmers.[6]

Others[7,8] have argued that smallholder farmers' productivity could be improved by improving the technological application of resources. They further argued that yield could be improved by increasing technical efficiency of production, without involving additional production resources and without adopting new technology.

Most studies estimate technical efficiency levels and identify the sources of inefficiency for smallholder farmers[9-17] or commercial farmers[18,19]. Studies that compare smallholder and commercial farmers' technical efficiencies within vegetable production are limited. Murthy et al.[20] estimated technical and scale efficiencies for tomato producers in Karnataka, considering different farm sizes (small, medium and large farms), with the use of a data envelopment analysis (DEA). The estimated efficiency levels were 77.7% for small farms, 82.5% for medium farms and 72.9% for large farms. The results showed that an optimal farm size would result in higher technical efficiency levels. They continued to pool the data to estimate the metafrontier (average efficiency score of 86.7%), but failed to compare the results for the various sized farms against the metafrontier.

Some of the studies conducted within South Africa include those of Mkhabela[21] and Khaile[22]. Mkhabela[21] used a stochastic production function (SFA) approach to investigate vegetable production for small- and large-scale farmers. Mkhabela identified vegetable farmers who produce on an area smaller than 1 ha as small-scale farmers, while those who produce on more than 1 ha as large-scale farmers, and found that large-scale farmers were technically more efficient in vegetable production than were small-scale producers. Khaile[22] used a DEA approach to estimate technical efficiencies of small- and large-scale raisin producers in Eksteenskuil. Khaile used

separate optimisations to estimate technical efficiencies for small- and large-scale producers and also found large-scale producers to be more efficient than small-scale producers. The main drawback of the abovementioned studies is that they failed to account for differences in production technology when comparing the production systems. Chen and Song[23] stated that differences in farming technologies could change the production frontier. Therefore, technical efficiencies from different production frontiers are not comparable. The metafrontier model[24] makes it possible to calculate comparable technical efficiencies for agricultural farms that operate under different technologies.

The main aim of this study was to compare tomato productivity of commercial and smallholder tomato farmers in the Nkomazi area, Mpumalanga, using a metafrontier analysis. Smallholder and commercial farmers in Nkomazi operate under different production environments, using different technologies to produce tomatoes. Information on commercial and smallholder farmers' production efficiency can help identify strategies to move smallholder farmers into commercial markets.

For this study, smallholder farmers were defined by their limited resource endowment compared to other farmers in the agricultural sector. These farmers were classified as farmers who produced on a very small piece of land, relied mostly on family labour, and produced primarily for home consumption. Commercial farmers produced primarily for a market, had access to productivity-improving resources, and used hired labour.

Data and procedures

The data for the research were collected from smallholder and commercial tomato farmers in the Nkomazi local municipality located in the Mpumalanga Province of South Africa. Data were collected with the use of a structured questionnaire. Personal interviews were conducted with farmers during May/June 2015 for the 2014/2015 tomato production season. The study was approved by the Ethics Committee of the Faculty of Natural and Agricultural Sciences of the University of the Free State (UFS-HSD2018/0934).

Study area

The Nkomazi region is known for its dual agriculture system which consists of commercial and smallholder farming. The area has also been described to be among the areas with the highest agricultural potential in South Africa.[25] The local municipality is situated within the eastern part of the Ehlanzeni District municipality. The Nkomazi region experiences frost-free winters with an average minimum temperature of about 8 °C, which allows production of vegetables and subtropical fruit.[26] The municipality falls within the summer rainfall region with the rainy season normally lasting from October to March. The estimated mean rainfall for the municipal area varies between approximately 500 mm and 1 600 mm per annum.[27] Summers are very hot and humid with an average maximum temperature of about 33 °C.[26] The extreme maximum temperature is 41.7 °C and the extreme minimum -2.0 °C. The climatic conditions of Nkomazi are suitable for tomato production as tomatoes are a warm season crop that is sensitive to frost. The study area was chosen because of its distinctiveness with regard to agricultural potential. Moreover, the Nkomazi area, known as Malelane, is the main area of tomato production in Mpumalanga.

Sampling procedure and data collection

A total of 102 tomato farmers were purposively selected and interviewed using a structured questionnaire. The group of sampled farmers consisted of 65 smallholder and 38 commercial farmers. After outliers were removed from the data the sample size was reduced to a total of 87 farmers of which 52 were smallholder farmers and the remaining 35 were commercial farmers. Production data collected with the questionnaire were the amount of tomatoes produced (kg/ha), fertiliser use (kg/ha), water use (number of irrigations in a season), seedlings (number of seedlings/ha), labour used (man-days/ha), and the amount of land used for tomato production (ha). The questionnaire also covered socio-economic, institutional and management factors, which were hypothesised to influence technical efficiency of farmers.

Theoretical framework

Specifying the production frontier

Others[28-30] have used both parametric and non-parametric methods to estimate technical efficiency. Both the parametric and non-parametric approaches share a common objective in that a benchmark needs to be identified to test the performance of the rest against that of the best.[31] The commonly used methods or models for measuring technical efficiency are the stochastic production function and DEA. The former uses econometric methods whereas the latter uses linear programming techniques.[28] As the stochastic production function and DEA use different methods to fit or determine the frontier, they accommodate differently for random noise and for flexibility in the structure of production technology.[32]

Although both techniques have their advantages and disadvantages, DEA was used in the current study because it does not require specification of a production function and therefore avoids model misspecification.[33] Because the aim of this study was to investigate smallholder and commercial tomato production using current input levels, an output-orientated variable-returns-to-scale model was used to estimate technical efficiency. The output-oriented model allows for the expansion of the amount of tomatoes produced without increasing resource use. The linear programming models for both smallholder and commercial farmers were solved separately using the Benchmarking Package[34] of R[35]. The DEA model used to determine farmers' technical efficiency was specified as follows:

$$Max_{\theta ik, \lambda ik} \quad \theta_{ik}$$
$$s.t. \ \theta_{ik} y_{ik} - Y_k \lambda_i \leq 0$$
$$X_k \lambda_i - x_{ik} \leq 0 \qquad\qquad \text{Equation 1}$$
$$\sum_i \lambda_i = 1$$
$$\lambda_i \geq 0$$

where θ_{ik} is the estimated technical efficiency that shows by how much the output (y) of the ith farm in the group k frontier (smallholder or commercial group) can increase using the same level of inputs (x); y_{ik} is a vector (m x 1) of tomato output produced by the ith farm in group k; x_{ik} is a vector (k x 1) of inputs used by the ith farm in group k; Y_k is the tomato output matrix (n x m) for all L_k tomato farms in group k; X_k is the tomato input matrix (n x k) for the L_k tomato farm in group k; and λ_i is a multiplier weight used to weight the input–output decisions of a farmer. The restriction $\sum_i \lambda_i = 1$ allows variable returns to scale. Because Equation 1 estimates output efficiency, the estimated measure of technical efficiency (θ_{ik}) ranges from one to infinity. From this score, the technical efficiency of each farmer in each group is estimated as:

$$TE^k = \frac{1}{\theta_{ik}} \qquad\qquad \text{Equation 2}$$

where TE^k defines a technical efficiency score that varies between zero and one. Values of the efficiency score (Equation 2) less than one indicate that output could increase through efficiency gains, without changing the levels of the inputs. The DEA model is estimated k times (once for smallholder and once for commercial tomato farmers), because the smallholder and commercial farmers use heterogeneous technologies. Smallholder and commercial tomato farmers are distinct from one another, because they face different constraints, have different resource endowments, and different opportunities for growth. As a result, the estimated technical efficiencies of heterogeneous groups cannot be compared. The performance of the smallholder and commercial tomato farmers can only be compared when taking the metafrontier as a reference.

Specifying the metafrontier to estimating technical efficiencies

Comparison of TE^k estimated in Equation 2 could be misleading as the smallholder and commercial farmers do not use the same technology. Therefore, we applied a metafrontier approach to compare technical efficiency of smallholder and commercial farmers. A convex metafrontier

was fitted using the pooled data (L = $\Sigma_k L_k$ farms) to estimate the farmers' technical efficiencies relative to a common benchmark (the metafrontier). The linear programming model for the pooled data solved using the Benchmarking Package[34] of R[35] was specified as:

$$Max_{wi,\lambda i} \quad W_i$$
$$\text{s.t. } W_i b_i - B\lambda_i \leq 0,$$
$$A\lambda_i - a_i \leq 0, \qquad \qquad \text{Equation 3}$$
$$\sum_i \lambda_i = 1$$
$$\lambda_i \geq 0$$

where W_i is the technical efficiency estimated for every tomato farmer in the pooled sample; b_i represents the M x 1 vector of output produced by the ith farm; a_i represents the N x 1 vector of inputs used by the ith farm; B represents the M x L matrix of output produced for all L farms; and A represents the N x L matrix of inputs used by all L farms. The metafrontier DEA model follows the same mathematical approach as the group frontiers, with only the size of decision-makers (L) changing. Equation 4 was used to estimate the technical efficiencies relative to the metafrontier (TE^*):

$$TE^* = \frac{1}{w_i} \qquad \qquad \text{Equation 4}$$

The metafrontier will never be below the group frontiers; hence, it captures the unrestricted technology set by enveloping the group frontiers.[36] This implies that even if farmers are technically efficient compared with their own peers in group k, they are not necessarily efficient when assessed against the metafrontier.

Specifying the inefficiency model

James Tobin first introduced the Tobit regression model in 1958.[37] The Tobit regression is a censored regression model. A censored model implies that the dependent variable is limited by a maximum or minimum value or both.[38] A Tobit model is preferred for cases in which the dependent variable is restricted in some way.[39] The estimated technical efficiency score is a censored variable with a lower limit of zero (0) and an upper limit of one (1), therefore the Tobit regression model is well suited to estimate the factors affecting farmers' sources of efficiency. The Tobit regression model has also been used in a number of studies[25,40-44] to determine the factors affecting variation in technical efficiency of farmers.

In the current study, a Tobit model was used to identify sources of variation in technical efficiency for smallholder and commercial tomato farmers. For both regression models, the selected variables were regressed against the technical efficiency scores obtained with Equation 2. The Tobit regression model used to determine the factors that improved the efficiency levels of the smallholder and commercial tomato farmers were specified as:

$$TE^k = \delta_0 + \sum_{i=1}^{10} \delta_i Z_i + \varepsilon_i \qquad \qquad \text{Equation 5}$$

where TE^k is the technical efficiency score estimated for farmer i in group k using Equation 2; δ_0 is the constant term; δ_i is the vector of unknown parameters to be estimated; Z_i represents the variable that may influence the efficiency; and ε_i is the error term. The Tobit models were estimated using Stata.[45]

Empirical application

Estimating efficiencies

The group frontiers (smallholder and commercial) and the metafrontier were estimated with the use of the production information obtained with the structured questionnaire. The output variable used in the DEA was tomato output measured in tonne/ha. While the input variables consisted of farmers' fertiliser use (measured in kg/ha), number of seedlings planted on a hectare of land (#seedlings/ha), and area under tomato production (ha). Farmers were also asked to indicate the volume of water used for irrigation during the production season.

However, because farmers did not know the actual volume of irrigation water applied (mm), they were asked to indicate the number of irrigation sessions during the production season (#irri/season). Farmers were also asked to indicate the number of labourers employed during the production season. Typically smallholder farmers rely on family labour to produce crops while commercial farmers use hired labour. In this study, both family labour and hired labour were considered for the estimation of the technical efficiencies of tomato farmers. Although the farmers who used family labour did not necessarily pay for labour, there was still an opportunity cost for the family labour (i.e. missed opportunity to increase off-farm income). Furthermore, these farmers still had non-family members who assisted with production activities. As few farmers' knew how many hours of labour were spent on tomato production, they were asked to indicate labour use in terms of man-days used to produce tomatoes (man-days/ha). The average, standard deviation, minimum and maximum amounts of tomatoes produced and inputs used by smallholder and commercial tomato farmers are presented in Table 1.

The data showed that smallholder farmers used on average 197 kg/ha of fertiliser, 26 irrigation sessions in a season, 8975 seedlings/ha, 755 man-days/ha and 0.3 ha to produce 5 tonnes of tomatoes on a hectare. Commercial farmers used 412 kg/ha of fertiliser, 21 irrigation sessions, 20 170 seedlings/ha, 336 man-days/ha to produce 96 t/ha on 5 ha of land. Data indicated that commercial farmers used more fertiliser, seedlings and land to produce more tomatoes. More importantly, the commercial farmers used fewer irrigation sessions and less labour to produce more tomatoes compared to the smallholder farmers. The reader is reminded that irrigation is measured as the number of irrigation sessions during a production season and not volume of irrigation water applied. Investigation of the farmers' irrigation practices revealed that smallholder farmers used predominantly watering cans and buckets to water tomato plants and irrigated the plants when the crop or soil appeared dry, whereas the commercial farmers used irrigation technology (e.g. drip irrigation) and a fixed irrigation schedule. The commercial farmers indicated that they irrigated tomato plants every 3–4 days. It is therefore possible that the smallholder farmers applied water every day or every other day, resulting in a higher number of irrigation sessions compared to those of the commercial farmers.

Explaining technical inefficiencies

The questionnaire included questions on socio-economic, institutional and management factors that were hypothesised to affect the farmers' level of technical efficiency. These variables were regressed against the technical efficiency scores estimated with Equation 2. Because the smallholder and commercial farmers were dealt with separately in the estimation of the efficiencies, two regressions were estimated to explain the sources of inefficiencies. The hypothesised variables, the description of the explanatory variables, and the expected relationship to technical efficiency are given in Table 2.

The socio-economic variables considered in this study included education, experience in tomato production, gender of farm manager (decision-maker), and access to off-farm income. Education and experience were included as continuous variables for which the farmer indicated their age and experience in tomato production in years. Gender of the farm manager was a binary variable, for which 1 indicated male and 0 indicated female. Farmers were asked to indicate if they had access to off-farm income, for which 1 indicated access and 0 no access. The sources of off-farm income were government grants and money sent home by family members who worked in nearby towns or cities.

The farmers were also asked to indicate if they produced tomatoes on rented (coded as a 1) or owned land (0). The expectation was that farmers who used rented land would aim to increase farm profit through better management of resources, in other words that they would aim to maximise technical efficiency. To determine farmers' management practices, farmers were asked to indicate the number of times they applied fertiliser and if they had access to sufficient irrigation water (sufficient water indicated by a 1) to fulfil the crop water requirement. Farmers were also asked to indicate if they used advanced irrigation technologies (1 for drip irrigation) that would increase water use efficiency.

Table 1: Average, standard deviation (s.d.), minimum (min) and maximum (max) tomato output and inputs used in estimation of technical efficiency

Variable	Unit	Smallholder farmers				Commercial farmers			
		Average	s.d.	Min	Max	Average	s.d.	Min	Max
Tomato output†	t/ha	5	5	0.33	29	96	17	67	128
Fertiliser†	kg/ha	197	78	0	340	412	71	250	590
Water†	#irri/season	26	9	14	44	21	4	12	30
Seedlings†	#seedlings/ha	8975	2821	5400	18 000	20 170	3042	16 000	28 000
Labour†	man-days/ha	755	300	390	1560	336	55	233	420
Land	ha	0.3	0.2	0.1	1	5	1.34	3	7

†Values were rounded to the closest whole number

Table 2: Description of the production variables and their expected signs

| Variable	Description	Expected sign
Socio variables		
Z_1 = education	Years of formal education of farmers	+
Z_2 = experience	Years of experience in tomato production	+
Z_3 = gender	Farmer's gender, 1 if male, 0 if female	+
Z_4 = off-farm income	1 if farmer received off-farm income, 0 otherwise	+
Institutional variable		
Z_5 = rental land	1 if farmer rented the farmland, 0 otherwise	+
Management variables		
Z_6 = frequent fertiliser application	Number of fertiliser applications per season	+
Z_7 = access to sufficient irrigation water	1 if the farmer had access to sufficient irrigation water, 0 otherwise	+
Z_8 = use drip irrigation	1 if the farmer used drip irrigation method, 0 otherwise	+
Z_9 = timely planting	1 if farmer planted at the beginning of the planting season, 0 otherwise	+
Z_{10} = staking	1 if the farmer practised staking, 0 otherwise	+

Most smallholder farmers used either buckets or cans and flood irrigation to irrigate tomatoes. The farmers were therefore not certain whether they provided the crop with adequate water, which could have affected the amount of tomatoes produced. The presence of a more advanced irrigation technology was therefore expected to increase technical efficiency.

Farmers were asked to indicate if they were able to plant tomatoes in a timely manner (1 indicated ability to plant in a timely manner and 0 not). Lastly, farmers were asked to indicate if they used staking to reduce the occurrence of tomato diseases and pests (1 indicated the use of staking, 0 otherwise). It was expected that the use of staking would increase tomato production and thereby increase technical efficiency.

Results and discussion

Technical efficiency of the group frontiers

The technical efficiency scores of the smallholder and commercial tomato farmers are presented in Figure 1 as a cumulative density function (CDF).[46]
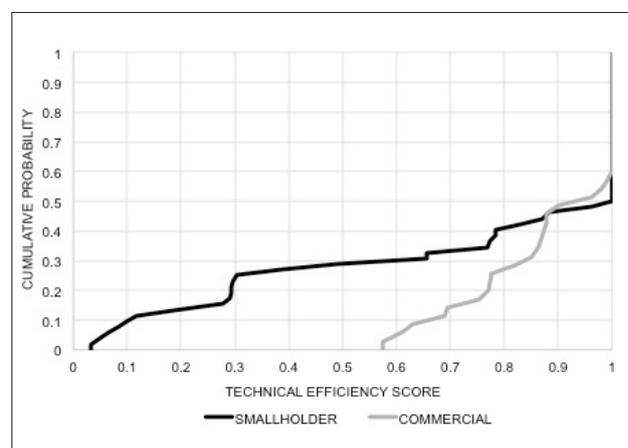
Figure 1: Cumulative probability of smallholder and commercial tomato farmers' technical efficiency scores.

The CDF indicated the technical efficiency scores as a continuous distribution that could be investigated. The CDF indicated that technical efficiencies of smallholder tomato farmers ranged from 3% to 100%. The difference between the lowest and the highest efficiency scores indicated that resource use efficiency differed greatly among smallholder tomato farmers. The estimated average technical efficiency score of smallholder tomato farmers was 74%, with a standard deviation of 35%. The average efficiency score revealed that smallholder tomato farmers had the potential to increase tomato output by 26% (1-0.74) using their current farm resources and technologies. About half (52%) of the smallholder farmers operated on the efficiency frontier (TE=1) while the remaining 48% operated under the efficiency frontier (TE<1).

Surprisingly, only 43% of the commercial farmers operated on the frontier while the majority (57%) were not operating on the efficiency frontier. However, the variation in the estimated efficiency scores for the commercial farmers was low, with an average of 89% and a standard deviation of 13%. In addition, a staggering 95% of the commercial farmers had a technical efficiency score above 60%, while only 70% of the smallholder farmers had technical efficiency scores above 60%.

The average technical efficiency scores of smallholder (74%) and commercial farmers (89%) were higher than the respective 69% and 65% reported by Enwerem and Ohajianya[47] for rice farmers in Nigeria. Mburu et al.[48] in their analysis of economic efficiency of wheat farmers in Kenya found slightly higher technical efficiency scores for smallholder (85%) and large-scale (91%) farmers. Similarly, Khaile[22] also found that commercial raisin farmers in Eksteenskuil, South Africa, were more technically efficient (85%) than the smallholder raisin farmers (81%). About half of the smallholder and commercial tomato farmers in this study were thus technically inefficient based on their estimated group frontiers, meaning that both groups could improve their resource use decisions.

Factors influencing technical efficiency in tomato production

In order to improve performance, tomato farmers must know what to change or increase in order to increase efficiency. The results for the factors that influenced technical efficiency of smallholder and commercial farmers are presented in Table 3. Results for the smallholder farmers indicated that education, off-farm income, rental land, frequency of fertiliser application, use of drip irrigation and staking of tomatoes had a significant effect on technical efficiency. Increased levels of education resulted in a significant increase ($p<0.05$) in smallholder tomato production. Results were consistent with those of Itam et al.[49] and Chepng'etich et al.[30] who found that smallholder farmers' with a higher level of education were more technically efficient than uneducated farmers. Formal education increased farmers' ability to investigate and adopt new technology, which tended to move farmers closer to the frontier. Similar to the results for commercial farmers, smallholder farmers' efficiency was increased by the availability of off-farm income ($p<0.01$) and the use of rental land ($p<0.01$). Again, farmers who had access to off-farm income were better able to obtain improved production inputs and to obtain these inputs in a timely manner. The use of rental land to produce crops indicated that the farmers were not interested in producing for home consumption alone. These farmers would therefore have been interested in increasing the returns they received from renting production land, and therefore it was expected that the renting of agricultural land was associated with more efficient farmers. An increase in the frequency of fertiliser application significantly increased ($p<0.05$) the efficiency of smallholder tomato production. This result indicates that smallholder farmers who applied fertiliser more often during tomato production were more technically efficient.

Table 3: Factors influencing technical efficiency for smallholder and commercial tomato farmers

Variable	Smallholder			Commercial		
	Coefficient	s.e.	Probability	Coefficient	s.e.	Probability
Socio variables						
Education	0.016**	0.007	0.028	-0.001	0.003	0.555
Experience	0.009	0.006	0.107	0.001*	0.000	0.060
Gender	0.047	0.047	0.271	0.030*	0.015	0.069
Off-farm income	0.417***	0.076	0.000	0.072***	0.019	0.001
Institutional variable						
Rental land	0.191***	0.058	0.002	0.039*	0.020	0.062
Management variables						
Frequency of fertiliser application	0.039**	0.016	0.019	-0.011***	0.003	0.003
Access to irrigation water	0.080	0.072	0.268	0.122***	0.025	0.000
Use drip irrigation	-0.139**	0.061	0.028	0.010	0.014	0.481
Timely planting	0.065	0.055	0.244	0.068**	0.025	0.012
Staking	-0.097**	0.044	0.032	0.004	0.016	0.775
Constant	0.230***	0.065	0.001	0.745***	0.057	0.000
/sigma	0.133	0.013		0.034	0.004	
Log likelihood	29.232324			65.189		
LR chi² (10)	100.01			93.24		
Prob> chi²	0.000***			0.000***		
Number of observations	52			35		

*10%, **5%, ***1%

The use of drip irrigation ($p < 0.05$) and staking ($p < 0.05$) decreased the level of efficiency of tomato production. The results indicated that smallholder farmers who used drip irrigation in the Nkomazi area were technically inefficient. The unexpected negative result was probably because smallholder farmers in the study area lacked the knowledge to use the irrigation technology correctly. Thus the farmers were either under- or over-irrigating the tomatoes. The negative result for staking was not consistent with the results of Gojeh et al.[50] who found that staking was beneficial in tomato cultivation. During the staking process, plants could have been handled incorrectly (and damaged), thus reducing fruit formation, or the weight of the hanging fruit resulted in injury or stress, thereby reducing the quality and size of the fruit. Alternatively, the use of stacking in combination with the use of inefficient irrigation systems (e.g. watering cans and buckets) could have resulted in water-stressed tomato plants. Staking lifts the plant from the ground, thus exposing the plant and soil to the sun and wind. As a result, plants would have required increased irrigation water as a result of increased evapotranspiration. Failure to supply the required water would result in water stress, decreased plant development, decreased fruit formation, and decreased levels of technical efficiency.

The results in Table 3 indicate that experience, gender, off-farm income, rental land, frequent fertiliser application, access to sufficient irrigation water and timely planting had a significant impact on the technical efficiency of commercial tomato farmers. Experience (0.001) and gender (0.030) were positively related and statistically significant at a 10% level. The expectation was therefore that male farmers with more experience would be more efficient in commercial tomato production. Over time, farmers are better positioned to obtain new knowledge and skills necessary for choosing new farm technologies that increase technical efficiency. Also tomato production is labour intensive, especially during planting, weeding and harvesting and, therefore, male farmers are expected to be more efficient than their female counterparts. The results are consistent with the results of Mango et al.[51]

As expected, off-farm income (0.072) for commercial farmers was positively related and statistically significant at a 1% level. It was expected that farmers who had a source of off-farm income were more technically efficient than commercial farmers with no off-farm income. Off-farm income increased the chance for farmers to easily and timeously buy important inputs such as fertiliser and pesticides. Results also indicated that the use of rental land to produce tomatoes ($p < 0.1$) increased the farmers' technical efficiency. These results were similar to that of Chimai[29] who found that using rented land had a positive influence on efficiency in field crop production because renting land is an added cost to the farmer. The farmer would therefore wish to be more efficient in their production, in order to justify the additional costs of renting land.

It was expected that frequent application of fertiliser would increase farmers' technical efficiency. However, results indicate that frequent fertiliser application reduced commercial farmers' technical efficiency at a statistically significant 1% level. There are a number of possible explanations for the negative relationship to fertiliser application. One possible explanation could be that the farmers over-utilised fertiliser trying to improve output, while the over-utilisation of fertiliser probably resulted in a decrease in tomato output. Another possible explanation could be that the fertiliser was applied at the wrong times. However, it is difficult to tell exactly what the cause of the negative coefficient was, as we did not test for over-application of fertiliser or the timing of fertiliser application.

Access to sufficient irrigation water ($p < 0.01$) and the timely planting of seedlings ($p < 0.05$) increased technical efficiency for commercial tomato production. Results indicated that access to enough water for irrigation in tomato production was important and hence increased technical efficiency. The planting of tomatoes early in the production season (timely planting of tomatoes) would have increased technical efficiency. Commercial farmers indicated that the challenge with late tomato planting in Nkomazi was the weather. During planting, rain could have resulted in blight fungal disease, because the leaves remained wet for extended periods.

Tomato farmers' technical efficiency relative to the metafrontier

The results for the group frontiers indicated that the technical efficiency levels for the smallholder and commercial farmers were different. However, the group frontiers cannot be used to draw a conclusion regarding the similarity or differences in the farmers' resource use management decisions, as the farmers faced different environmental and production conditions. As a result, a metafrontier was estimated to present a common frontier that could be used for comparison of the farmers with the common benchmark. A comparison between the distribution of technical efficiency scores for the group frontiers and relative to the metafrontier is indicated in Figure 2 as a CDF. SMALLHOLDER and COMMERCIAL are used to denote the technical efficiency of the smallholder and commercial tomato farmers, respectively. An asterisk (*) is used to indicate the technical efficiency relative to the metafrontier.
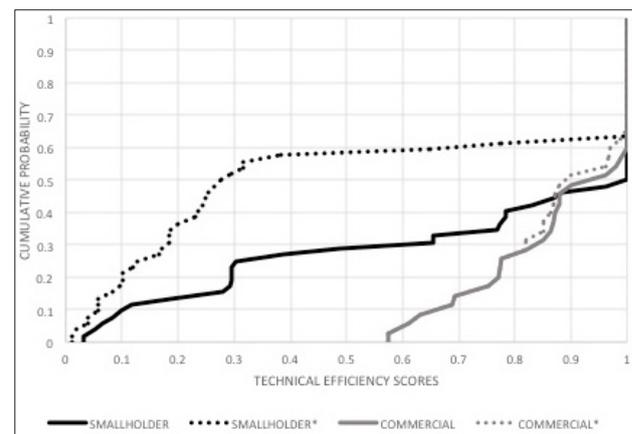

Figure 2: Cumulative probability for the group frontiers (solid lines) and technical efficiency relative to the metafrontier (broken lines).

The CDF indicates that smallholder farmers' technical efficiency relative to the metafrontier (TE*) ranged from 1% to 100%. The wide range in estimated technical efficiency scores of smallholder farmers indicates the huge variation in resource use efficiency. It is interesting to note that about 38% (1-0.62) of the smallholder farmers were technically efficient (100%) relative to the metafrontier, while nearly 50% of the farmers were efficient relative to the group frontier. A large group of the farmers (58%) had an efficiency score (TE*) below 40%; as a result the smallholder farmers' average efficiency level relative to the metafrontier was 51% compared to the 74% for the group frontier. Results for the technical efficiency of smallholder farmers indicate that the level of inefficiency estimated relative to the metafrontier was higher than that estimated for the smallholder group frontier. The implication of this result is that the average smallholder farmers found it difficult to compete with the farmers who formed the common benchmark.

The average TE* score estimated for the commercial farmers was 88% with a standard deviation of 13%. The variation in the estimated TE* for commercial farmers was much lower than that estimated for the smallholder farmers. Even though some farmers showed a slight decrease in their estimated TE* compared with the TE estimated from the group frontier, the average TE* decreased by 1% to 88%. The number of farmers who were producing on the frontier decreased from 43% for the group frontier to 37% for the metafrontier. Therefore, the results indicate that, compared to the common benchmark, the commercial farmers were performing much better than the smallholder farmers, with more commercial farmers operating closer to the common benchmark. This result is similar to that of Temoso et al.[52] who found that commercial beef producers' average technical efficiency relative to the metafrontier (74%) was higher than that of traditional farms (smallholder farmers) (71%) in Botswana.

Conclusion

Our aim was to investigate the technical efficiencies of smallholder and commercial tomato farmers in the Nkomazi area of South Africa. Identification of the differences in these farmers' technical efficiencies and the factors that influence their level of inefficiency can help identify strategies to move smallholder farmers into commercial markets.

Results for the group technical efficiencies showed that smallholder and commercial tomato farmers were fairly efficient with an average efficiency of 74% and 89%, respectively. There was, however, still some room for improvement, which was highlighted by the large variability in smallholder technical efficiency. In conclusion, both groups of farmers were fairly efficient when compared to their peers. However, some of the commercial and smallholder farmers could still improve their technical efficiency by improving their input use decisions.

The results from the Tobit model show that experience, gender, off-farm income, rental land, frequent fertiliser application, access to sufficient irrigation water and timely planting had a significant impact on technical efficiency of commercial tomato farmers in Nkomazi. Education, off-farm income, rental land, frequent fertiliser application, use of drip irrigation and staking of tomatoes were significant factors for technical efficiency in smallholder tomato production. Results for the factors influencing farmers' technical efficiency indicate that the factors which increased or decreased technical efficiency were different for smallholder and commercial farmers. Only two factors had a significantly positive effect for both groups of farmers: off-farm income and the renting of additional land. The remaining variables indicated that management decisions (frequency of fertiliser applications, use of irrigation and irrigation technology, timely planting and staking) impacted smallholder and commercial farmers' technical efficiency in tomato production. The conclusion is that assistance provided to tomato farmers should focus on production management.

The results for the metafrontier again show that the commercial farmers were fairly efficient in tomato production, with an average efficiency of 88%. The smallholder tomato farmers were much less efficient compared to the metafrontier, with an average efficiency of 51%, although there were some smallholder farmers with scores on the metafrontier. The results therefore indicate that the smallholder tomato farmers would have to make fairly substantial changes to their production practices, which include their production technologies, in order to compete with the farmers who represented the metafrontier. However, it would not make sense for the smallholder farmers to improve their production relative to the metafrontier if they were still under-performing compared with their peers, as measured with the group frontier. Therefore, smallholder farmers must first increase their technical efficiency relative to other smallholder farmers, before aiming to improve relative to the metafrontier and, in so doing, commercial farmers.

A possible means to provide assistance to farmers could be through extension services. However, the extension officers should be trained to provide guidance to farmers regarding appropriate management decisions which would improve farmers' technical efficiency. Farmers should be encouraged to utilise existing agricultural support programmes like *Masibuyele emasimini* ('going back to till the land') that aim to improve farmers' productivity. These support programmes should be tailored to provide production management guidance and not only provide inputs to beneficiaries. However, more empirical research is needed to fully understand the technical efficiency differences between smallholder and commercial farmers. A better understanding of such differences will allow for the development of better systems to assist both smallholder and commercial farmers, and could provide the necessary understanding on how to integrate smallholder farmers into commercial markets.

Acknowledgement

Authors' contributions

J.Z.G.: data collection, data analysis, writing the initial draft, funding acquisition. N.M.: student supervision, writing revisions, critically reviewing the writing.

References

1. Greyling JC. The role of the agricultural sector in the South African economy [master's dissertation]. Stellenbosch: Stellenbosch University; 2012.

2. Sandrey R, Vink N. Trade and innovation project case study 4: Deregulation, trade reform and innovation in the South African agriculture sector. OECD Trade Policy Working Papers No.76. Paris: OECD; 2008. http://dx.doi.org/10.1787/240428414841

3. Tshuma MC. Understanding the small-scale agricultural sector as precondition for promoting rural development in South Africa. Afr J Agric Res. 2014;9(31):2409–2418. https://doi.org/10.5897/ajar12.1631

4. South African Department of Agriculture. The strategic plan for South African agriculture. Pretoria: Directorate Agricultural Information Services; 2001.

5. South African Department of Agriculture, Forestry and Fisheries. Strategic plan for smallholder support 2011–2014/15. Pretoria: Directorate of Small Holder Development; 2013.

6. Pienaar CJ. Increased water productivity in irrigated tomato production in the smallholder farming community of Giyani [master's dissertation]. Stellenbosch: Stellenbosch University; 2014.

7. Nkamleu GB, Nyemeck J, Diakalia S. Metafrontier analysis of technology gap and productivity differences in African agriculture. J Agri Food Econ. 2006;1(2):111–120.

8. Obwona M. Determinants of technical efficiency differentials amongst small and medium scale farmers in Uganda: A case of tobacco growers. AERC Research paper 152. Kenya: African Economic Research Consortium; 2006

9. Najjuma E. Technical efficiency of tomato production: A comparative study between small scale green house and open field tomato farmers in Kiambu. J Agr Policy. 2017;1(1):16–36.

10. Shrestha RB, Huang WC, Pradhan U. Evaluating the technical efficiency of smallholder vegetable farms in diverse agroecological regions of Nepal. Int J Food Agric Econ. 2016;4(4):97–112.

11. Belete A, Setumo MP, Laurie SM, Senyolo MP. A stochastic frontier approach to technical efficiency and marketing of orange fleshed sweet potato (ofsp) at farm level: A case study of KwaZulu-Natal Province, South Africa. J Hum Ecol. 2016;53(3):257–265.

12. Malinga NG, Masuku MB, Raufu MO. Comparative analysis of technical efficiencies of smallholder vegetable farmers with and without credit access in Swaziland: The case of the Hhohho Region. Int J Sustain Agric Res. 2015;2(4):133–145. https://doi.org/10.18488/journal.70/2015.2.4/70.4.133.145

13. Etim NA, Udoh EJ. Identifying sources of efficiency among resource poor indigenous vegetable farmers in Uyo, Nigeria. Int J Food Agric Econ. 2014;2(1):33.

14. Donkoh SA, Tachega M, Amowine N. Estimating technical efficiency of tomato production in Northern Ghana. Am J Exp Agr. 2013;3(1):56–75. https://doi.org/10.9734/AJEA/2013/1447

15. Khan REA, Ghafar S. Technical efficiency of tomato production: A case study of District Peshawar (Pakistan). World App Sci J. 2013;28:1389–1392.

16. Ogunniyi LT, Oladejo JA. Technical efficiency of tomato production in Oyo State Nigeria. Agr Sci Res J. 2011;1(4):84–91.

17. Abdulai A. Resource use efficiency in vegetable production: The case of smallholder farmers in Kumasi metropolis [master's dissertation]. Kumasi: Kwame Nkrumah University of Science and Technology; 2006.

18. Karunarathna M. Estimating technical efficiency of vegetable farmers in Anuradhapura District in Sri Lanka. J Econ Res. 2014;2(2):55–67.

19. Sartorius Von Bach HJ, Townsend RF, Van Zyl J. Technical inefficiency of commercial maize producers In South Africa: A stochastic frontier production function approach. Agrekon. 1998;37(2):162–170. https://doi.org/10.1080/03031853.1998.9523502

20. Murthy DS, Sudha M, Hegde MR, Dakshinamoorthy V. Technical efficiency and its determinants in tomato production in Karnataka, India: Data envelopment analysis (DEA) approach. Agr Econ Res Rev. 2009;22(2):215–224.

21. Mkhabela T. Technical efficiency in a vegetable based mixed-cropping sector in Tugela Ferry, Msinga District, KwaZulu-Natal. Agrekon. 2005;44(2):187–204. https://doi.org/10.1080/03031853.2005.9523709

22. Khaile PME. Factors affecting technical efficiency of small-scale raisin producers in Eksteenskuil [master's dissertation]. Bloemfontein: University of the Free State; 2012.

23. Chen Z, Song S. Efficiency and technology gap in China's agriculture: A regional metafrontier analysis. China Econ Rev. 2008;19(2):287–296. https://doi.org/10.1016/j.chieco.2007.03.001

24. Battese GE, Rao DSP, O'Donnell C. A metafrontier production function for estimation of technical efficiencies and technology gaps for firms operating under different technologies. J Productiv Anal. 2004;21:91–103. https://doi.org/10.1023/b:prod.0000012454.06094.29

25. Thabethe LS. Estimation of technical, economic and allocative efficiencies in sugarcane production in South Africa: A case study of Mpumalanga growers [master's dissertation]. Pretoria: University of Pretoria; 2013.

26. Van Niekerk G. Farms on Mpumalanga achieving high return on investment [webpage on the Internet]. c2015 [cited 2016 Jun 10]. Available from: http://www.pamgolding.co.za/property-news/2015/3/13/farms-in-mpumalanga-achieving-high-return-on-investment

27. South African Department of Agriculture and Land Administration. Strategic plan 2007/2008-2009-2010. Nelspruit: Mpumalanga Provincial Government; 2007.

28. Tipi T, Yildiz N, Nargeleçekenler M, Çetin B. Measuring the technical efficiency and determinants of efficiency of rice (*Oryza sativa*) farms in Marmara region, Turkey. N Z J Crop Hortic Sci. 2009;37(2):121–129. https://doi.org/10.1080/01140670909510257

29. Chimai BC. Determinants of technical efficiency in smallholder sorghum farming in Zambia [master's dissertation]. Columbus, OH: Ohio State University; 2011.

30. Chepng'etich E, Nyamwaro SO, Bett EK Kizito K. Factors that influence technical efficiency of sorghum production: A case of small holder sorghum producers in lower Eastern Kenya. Advances Agr. 2015;2015, Art. #861919, 11 pages. http://dx.doi.org/10.1155/2015/861919

31. Fried HO, Lovell CK, Schmidt SS. Efficiency and productivity. The measurement of productive efficiency and productivity growth. New York: Oxford University Press; 2008. https://doi.org/10.1093/acprof:oso/9780195183528.001.0001

32. Porcelli F. Measurement of technical efficiency. A brief survey on parametric and non-parametric techniques. Coventry: University of Warwick; 2009. Available from: http://www.warwick.ac.uk/fac/soc/economics/staff/phd_students/porcelli/porcelli_dea_sfm.pdf

33. Sarafides V. An assessment of comparative efficiency measurement techniques. Occasional Paper 2. London: Europe Economics; 2002.

34. Bogetoft P, Otto L. Package benchmarking. Benchmark and frontier analysis using DEA and SFA [webpage on the Internet]. c2015 [cited 2015 Sep 16]. Available from: https://cran.r-project.org/

35. R Core Team. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/

36. O'Donnell CJ, Rao DP, Battese GE. Metafrontier frameworks for the study of firm-level efficiencies and technology ratio. Empirical Econ. 2008;34(2):231–255. https://doi.org/10.1007/s00181-007-0119-4

37. Tobin J. Estimation of relationships for limited dependent variables. Econometrica. 1958;26(1):24–36. https://doi.org/10.2307/1907382

38. Carson RT, Sun Y. The Tobit model with a non-zero threshold. Econometrics J. 2007;10:488–502. https://doi.org/10.1111/j.1368-423x.2007.00218.x

39. Bhatt MS, Bhat SA. Technical efficiency and farm size productivity-micro level evidence from Jammu and Kashmir. Int J Food Agric Econ. 2014;2(4):27–49.

40. Awotide DO, Kehinde AL, Akorede TO. Metafrontier analysis of access to credit and technical efficiency among smallholder cocoa farmers in Southwest Nigeria. Int Bus Res. 2015;8(1):132–144. https://doi.org/10.5539/ibr.v8n1p132

41. Sibiko KW, Owuor G, Birachi E, Gido EO, Ayuya OI, Mwangi JK. Analysis of determinants of productivity and technical efficiency among smallholder common bean farmers in eastern Uganda. Current Res J Econ Theory. 2013;5(3):44–55.

42. Bougna E, Crozet Y. Estimating technical and allocative efficiency using stochastic cost frontier: an application to railway. Paper presented at: The European Transport Conference 2013: Big Data; 2013 September 30 – October 2; Frankfurt, Germany.

43. Otieno DJ, Ruto E, Hubbard L. Determinants of technical efficiency in beef cattle production in Kenya. Paper presented at: The International Association of Agricultural Economists (IAAE) Triennial Conference; 2012 August 18–24; Foz do Iguacu, Brazil.

44. Chavas JP, Petrie R, Roth M. Farm household production efficiency: Evidence from the Gambia. Am J Agr Econ. 2005;87(1):160–179. https://doi.org/10.1111/j.0002-9092.2005.00709.x

45. StataCorp. Stata statistical software release 14. College Station, TX: StataCorp LP; 2015.

46. Hardaker JB, Lien G, Anderson JR, Huirne RB. Coping with risk in agriculture: Applied decision analysis. Oxfordshire: CABI Publishing; 2015. https://doi.org/10.1079/9781780645742.0000

47. Enwerem VA, Ohajianya DO. Farm size and technical efficiency of rice farmers in Imo State, Nigeria. Greener J Agri Sci. 2013;3(2):128–136.

48. Mburu S, Ackello-Ogutu C, Mulwa R. Analysis of economic efficiency and farm size: A case study of wheat farmers in Nakuru District, Kenya. Econ Res Int. 2014:1–10. http://dx.doi.org/10.1155/2014/802706

49. Itam KO, Ajah EA, Ofem UI, Abam OE. Technical efficiency analysis of small scale cassava farmers in Cross River State, Nigeria: A stochastic production frontier approach. Appl Econ Finan. 2015;2(4):10–18. https://doi.org/10.11114/aef.v2i4.1028

50. Gojeh AJ, Belew D, Mohammed A. Yield and quantity of indeterminant tomato (*Lycopetsicon esculentum* Mill.) varieties with staking in Jimma. Singapore J Sci Res. 2012;2(2):33–46. https://doi.org/10.3923/sjsres.2012.33.46

51. Mango N, Makate C, Hanyani-Mlambo B, Siziba S, Lundy M. A stochastic frontier analysis of technical efficiency in small-scale maize production in Zimbabwe: The post-fast-track land reform outlook. Cogent Econ Finance. 2015;3, 1117189, 14 pages. https://doi.org/10.1080/23322039.2015.1117189

52. Temoso O, Villano R, Hadley D. Evaluating the productivity gap between commercial and traditional beef production systems in Botswana. Agric Syst. 2016;149:30–39. https://doi.org/10.1016/j.agsy.2016.07.014

AUTHORS:
Tanja Verster[1] iD
Erika Fourie[2] iD

AFFILIATIONS:
[1]Centre for BMI, North West University, Potchefstroom, South Africa

[2]Statistics, North West University, Potchefstroom, South Africa

CORRESPONDENCE TO:
Erika Fourie

EMAIL:
erika.fourie@nwu.ac.za

The good, the bad and the ugly of South African fatal road accidents

We reflect on the good, the bad and the ugly of the fatal accidents occurring on South Africa's roads. The cost of human lives indisputably equates to 'the ugly' and the economic cost of accidents associates with 'the bad'. 'The good' relates to the reduction of both these costs that may result from the entrance of self-driving cars into the South African market as well as awareness campaigns like the Arrive Alive National Road Safety Strategy. The general contribution of this paper is to raise awareness of the effects of accidents, more specifically fatal accidents. Current trends in terms of human factors as well as road and environmental factors involved in the fatal accidents on South African roads are summarised. This paper also serves as a preliminary investigation into possible factors influencing these accidents, which ought to be of interest to a very broad readership, more specifically those focusing on risk analysis, and certainly is of interest to any citizen of South Africa.

Significance:

- Awareness is raised of the effects of fatal accidents on South African roads.

- Current trends in terms of human factors as well as road and environmental factors on road accidents are reflected upon.

- The futuristic effect of self-driving cars is explored.

Introduction

The high number of accidents on South African roads not only results in loss of human life with the associated pain, grief and suffering but also has a negative effect on the well-being of South Africans and on South Africa's socio-economic development. The total cost of South African road traffic accidents in 2015 was estimated to be ZAR142.95 billion, which equates to 3.4% of South Africa's gross domestic product (GDP). The average cost of accidents in similar countries is reported at only 2.2% of their GDP.[1] South Africa's economy is not only affected by the direct costs of accidents, but also via the death toll itself as the fatalities most often are the breadwinners of the households, which leaves the rest of the household members financially unattended.

Therefore, the fatal accidents on South African roads pose a threat to all South Africans which, in turn, indicates a need for South Africans – academics and average citizens alike – to be educated about the accidents occurring on South African roads. This need has also been emphasised by the Automobile Association Foundation of South Africa[2] and the international community (World Health Organization, World Bank and United Nations) in their Decade of Action for Safety from 2010 to 2020 declaration. South Africa joined this initiative by developing its own National Road Safety Strategy 2011–2020.[3]

Awareness campaigns alone may not be sufficient to reduce the number of accidents on South African roads.[4] One possible solution is to impose compulsory third-party insurance to all motorists. According to John Melville, head of risk services at Santam, a leading South African short-term insurance company, this mandatory third-party vehicle insurance would lower the cost of cover for South African consumers and could also reduce the number of road accidents, because it will bring down the number of damaged vehicles on our roads.[5]

Whether it is for the calculation of general insurance premiums or how the calculation for this third-party insurance premium is done, some risk analyses are required to determine which premium should be charged. Many factors are taken into account when determining the short-term insurance premium. One factor is usually a prediction of the number of accidents; other factors include the type of vehicle driven (how safe is the vehicle) and how often and how far the person drives.[6,7]

The risk posed to all South Africans via the indirect risk of the loss of income (should a fatality be the main income provider of the household) affects life insurance too. The number of fatal accidents also influences the Road Accident Fund. Note that the Road Accident Fund covers personal injury to the third party, but not damage to material possessions such as their vehicle. Third-party insurance usually covers this damage, but up to 35% of vehicles in South Africa are uninsured.[8] Again, a risk analysis is required because an increase in fatalities will lead to an increase in claims which will ultimately result in an increase in premiums and increased cost to the Road Accident Fund.

In order for these risk analyses to be conducted, factors that could possibly affect the number or severity of accidents on South African roads need to be determined. Many different industries will benefit from this information. Data sources that contain information regarding these factors also need to be investigated. We undertook an exploratory investigation for those specifically interested in these risk analyses that may be required in future for prediction purposes (such as predicting the number of accidents, the number of claims and the cost of accidents). At the same time, we report some important and interesting findings regarding South African accidents, focusing on fatal accidents, to inform and to raise awareness as a general contribution to the wider audience.

In 'the good, the bad and the ugly of South African fatal road accidents', the cost of human lives undeniably equates to 'the ugly' and the economic costs associated with these accidents to 'the bad'. 'The good' is addressed by the reduction of both these costs resulting from accident awareness campaigns (e.g. this article), and other initiatives like compulsory third-party insurance, and, in the future, the possible entrance of self-driving cars into the market.

Current situation: Factors influencing accidents

In order to be able to predict the number of vehicle accidents, one first needs to understand the factors that influence these accidents. Bayam et al.[6] explain that a vehicle crash is considered to be an interacting system of five factors: (1) driver, (2) vehicle, (3) environmental and geographical conditions, (4) roadway and (5) occupants and other road users. An illustration of Bayam et al.'s[6] five factors is provided in Figure 1. The 'driver' factor relates to personal attributes like the age and gender of the driver. The 'occupants and other road users' factor relates to information regarding the occupants (such as age, gender and other characteristics), the occupants of the other vehicle as well as pedestrians and cyclists or any other road users who had an impact on the accident. These two factors together can also be viewed as the *human factors* contributing to accidents (indicated by the purple grouping in Figure 1). The 'vehicle' component reflects vehicle type and age (year) of the vehicle. 'Environmental and geographical conditions' comprise weather condition, lighting condition, day and time of day, as well as area type. The 'roadway' factor reflects road attributes such as road condition and road surface. The latter two factors are also referred to as the *road and environmental factors* (indicated by the green grouping in Figure 1). Information regarding the accident itself, e.g. accident type, number of fatalities and severity, is provided under 'accident', that is, in the middle of the figure.

We set out to explore these factors in relation to fatal accidents. A fatal accident is defined as a crash resulting in the death of one or more persons killed during, immediately after or within 30 days of an accident in which the death was as a direct result of the crash. The persons killed may be the driver of a vehicle involved, passengers of vehicles involved, cyclists or pedestrians.[9]

The data used in this paper refer specifically to fatal accidents that occurred on South African roads during 2015, and were retrieved from a report published by South Africa's Road Traffic Management Corporation, which is the only currently available official accident data.[1] This information is supplemented with findings from the literature, for example, a project report entitled 'Roads infrastructure policy for South Africa' that was done in collaboration with the Department of Transport, ITS Engineers, Alta Swanepoel & Associates, and Urban-Econ, released in December 2014. The report notes that the Road Traffic Management Corporation should report and monitor all crashes,[10] not only fatal crashes.

Another example found in the media[11] stated that the Road Traffic Management Corporation's collection and distribution of data are in chaos and the integrity of the data available is questionable. Specifically, data on non-fatal accidents were corrupted when they were migrated to a new database.[11] This issue is also raised in the 2016 Road Traffic Management Corporation report in which the importance of quality data on road traffic crashes is emphasised.[1]

Figure 2 provides the breakdown of the number of accidents as well as the cost involved per accident type for the accidents that occurred on South Africa's road network in 2015. It should be noted that fatal crashes constitute only 1.3% of the number of accidents occurring on South African roads, but 42.4% of the costs involved.[1]
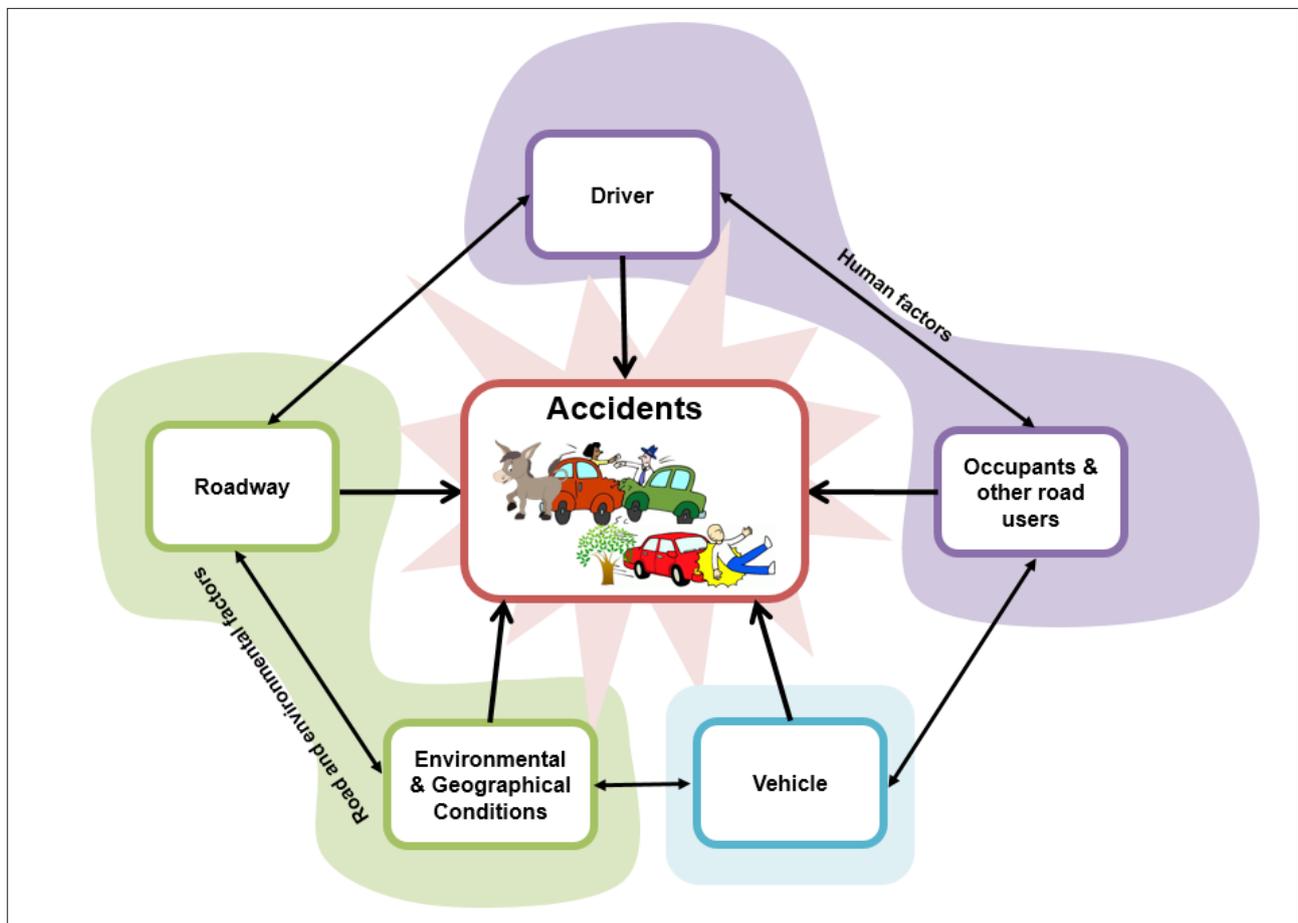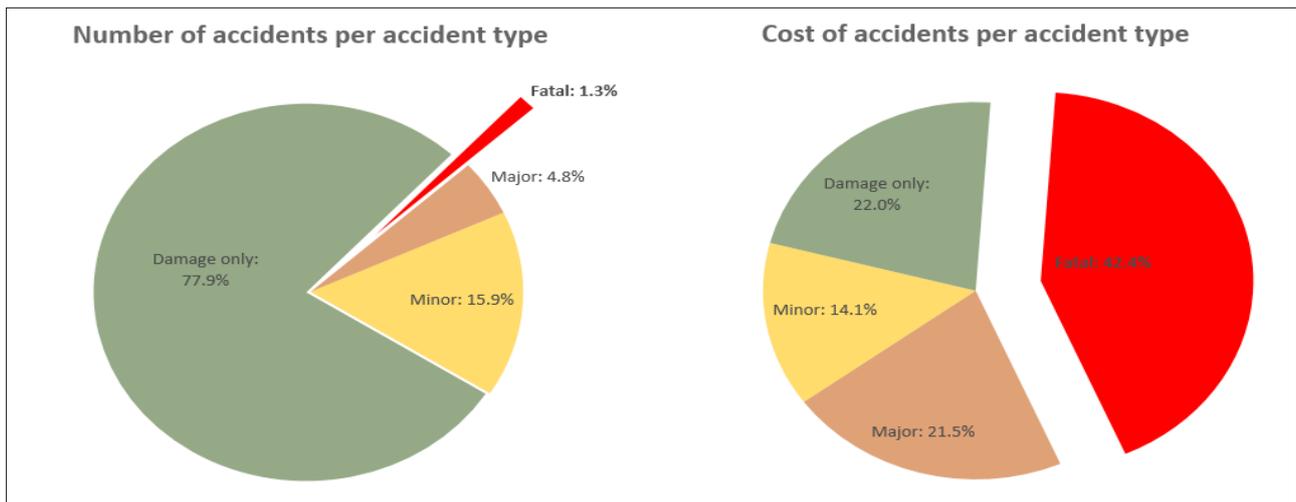

Figure 1: Five factors identified[6] to cause or interact to cause road accidents.

Figure 2: Breakdown of (a) the number of accidents and (b) costs involved per accident type, for the road accidents that occurred in South Africa during 2015.

We focused on fatal accidents because of the more detailed information available. It should be noted that it may take some time before reasonable reporting on non-fatal accidents can be expected.[1]

We first provide some accident information regarding the fatal accidents occurring during 2015 and then discuss each of the five interacting factors grouped into three general contributing factors (as indicated in Figure 1): *human factors* (i.e. drivers, occupants and other road users), *vehicle* and *road and environmental factors*.

Accident information

As mentioned before, all information is based on fatal accidents. One way to assess the information is to consider the severity of fatal crashes (i.e. the average number of fatalities per crash). Table 1 shows the number of fatal crashes, the number of fatalities and the severity per year, from 2010 to 2015. A total of 10 613 fatal accidents occurred from 1 January to 31 December 2015 and resulted in 12 944 fatalities. The severity for 2015 is therefore reported at 1.2. Severity can further be explored by considering the day of the week. The highest severity rates are reported for Mondays, Fridays, Saturdays and Sundays, with rates ranging from 1.22 to 1.25.

Table 1: Accident information for all fatal road accidents reported in 2010 to 2015

Year	Number of fatal crashes	Number of fatalities	Severity
2010	10 837	13 967	1.3
2011	11 228	13 954	1.2
2012	10 977	13 528	1.2
2013	10 170	11 844	1.2
2014	10 367	12 702	1.2
2015	10 613	12 944	1.2

Data source: South African Department of Transport[9]

Another way of assessing this type of information is by reviewing the number of fatal crashes as well as the number of fatalities per 10 000 vehicles. Figure 3 provides a breakdown of this information per province
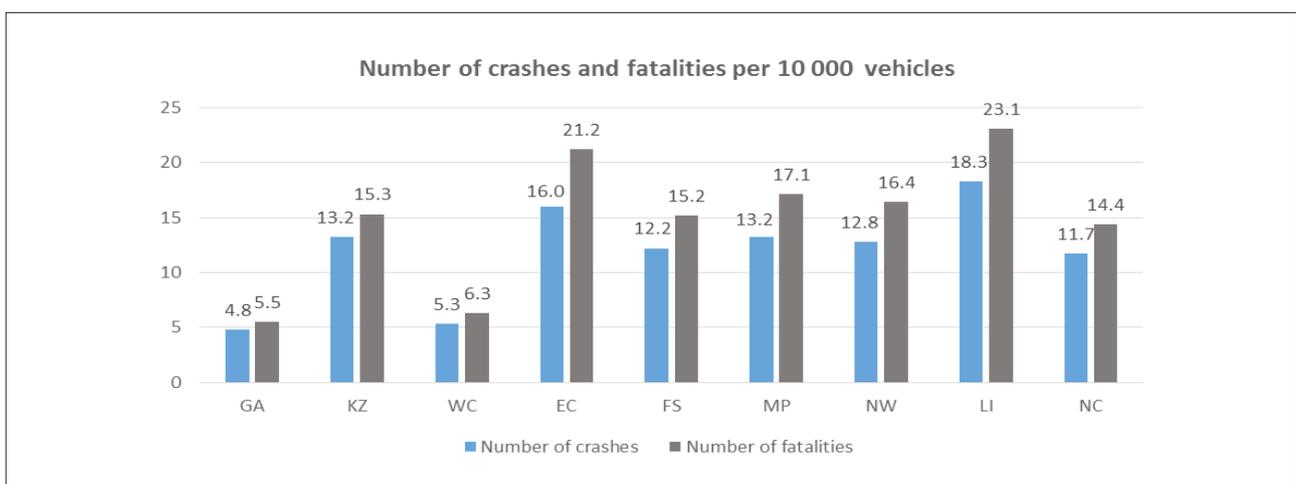
Data source: South African Department of Transport[9]

GA, Gauteng; KZ, KwaZulu-Natal; WC, Western Cape; EC, Eastern Cape; FS, Free State; MP, Mpumalanga; NW, North West; LI, Limpopo; NC, Northern Cape

Figure 3: Number of crashes and fatalities per 10 000 vehicles per province.

for 2015; the blue bars represent the number of fatal crashes and the grey bars represent the number of fatalities. The highest number of fatal crashes per 10 000 motorised vehicles was reported for Limpopo Province and the lowest for Gauteng. The same trend is evident for the number of fatalities per 10 000 vehicles. The number of crashes per 10 000 motorised vehicles for South Africa as a whole is reported at 12 and the number of fatalities at 11.

Human factors

Figure 4 gives the distributions of the fatalities per road user type, gender, race and age based on the information provided by the South African Department of Transport.[9]

Of the 12 944 fatalities, 27.0% were drivers and two thirds of these drivers (66.6%) were younger than 30 years old. This finding is supported by that of Bédard et al.[7] who also found that most fatally injured drivers are from this age group. Summala and Mikkola[12] gave the following reason for this phenomenon: younger drivers tend to avoid and delay taking breaks and also take more deliberate risks than older drivers do.

In addition to the 27.0% of drivers killed, 32.8% of the fatalities were passengers and 37.7% were pedestrians. Male individuals made up 78.0% and female individuals 22.0% of the recorded fatalities. This distribution is somewhat different to the national gender distribution of 49% males and 51% females.[13] Based on the known races, the majority (81.4%) of the fatalities were black individuals. This finding is not surprising, as black people comprise the largest population group in South Africa at 80.7%[13] and thus the majority of road users. It can also be seen from Figure 4 that approximately half of the fatalities on South African roads in 2015 were people aged 20–29 years (24.6%) and 30–39 years (24.9%).

Although pedestrians are amongst the most vulnerable road users, it is alarming that 16.3% of the pedestrians killed in vehicle accidents were young children (younger than 10 years) and 20.8% were children 14 years or younger. The Automobile Association Foundation[2] has also reported that two pedestrians are killed daily by 'vehicle unknown', which means that they have been hit by a vehicle and left to die.

Vehicle

There are a number of different types of vehicles operating on South African roads, including motor cars (and station wagons), motorcycles, bakkies (or pickups), trucks, buses (both normal and mini), caravans and trailers (both heavy and light).

Almost half (48.5%) of the vehicles involved in fatal accidents during 2015 were motor cars (and station wagons), with bakkies second highest (18.4%). This result is not surprising as these types are also the most registered vehicles on the road at 64.6% and 22.5%, respectively.

The vehicle type that is third most (9.0%) involved in fatal accidents is minibuses. It is reported that 3 of the 36 daily fatalities on South African roads are related to minibus taxis. Public opinion in South Africa is that minibus taxis are unsafe and are being operated in an offensive way. Some of the reasons listed for this phenomenon are that minibus taxis are overloaded in order to increase the number of passengers and are being operated at high speed to reduce travel time.[14] The most widely publicised, and certainly the most ambitious government intervention in the minibus taxi industry, is the Taxi Recapitalisation Programme. Through this project, the South African government seeks to challenge head-on the problem of an ageing fleet within the South African taxi transportation system. The first objective of this project is the systematic introduction of safe and comfortable vehicles for taxi commuters through a scrapping allowance which acts as an incentive for taxi operators to hand in, on a voluntary basis, their very old vehicles for decommissioning. The second objective is the economic empowerment of the taxi industry through a package of business opportunities that the Recapitalisation Project affords the taxi industry to participate in nationally through
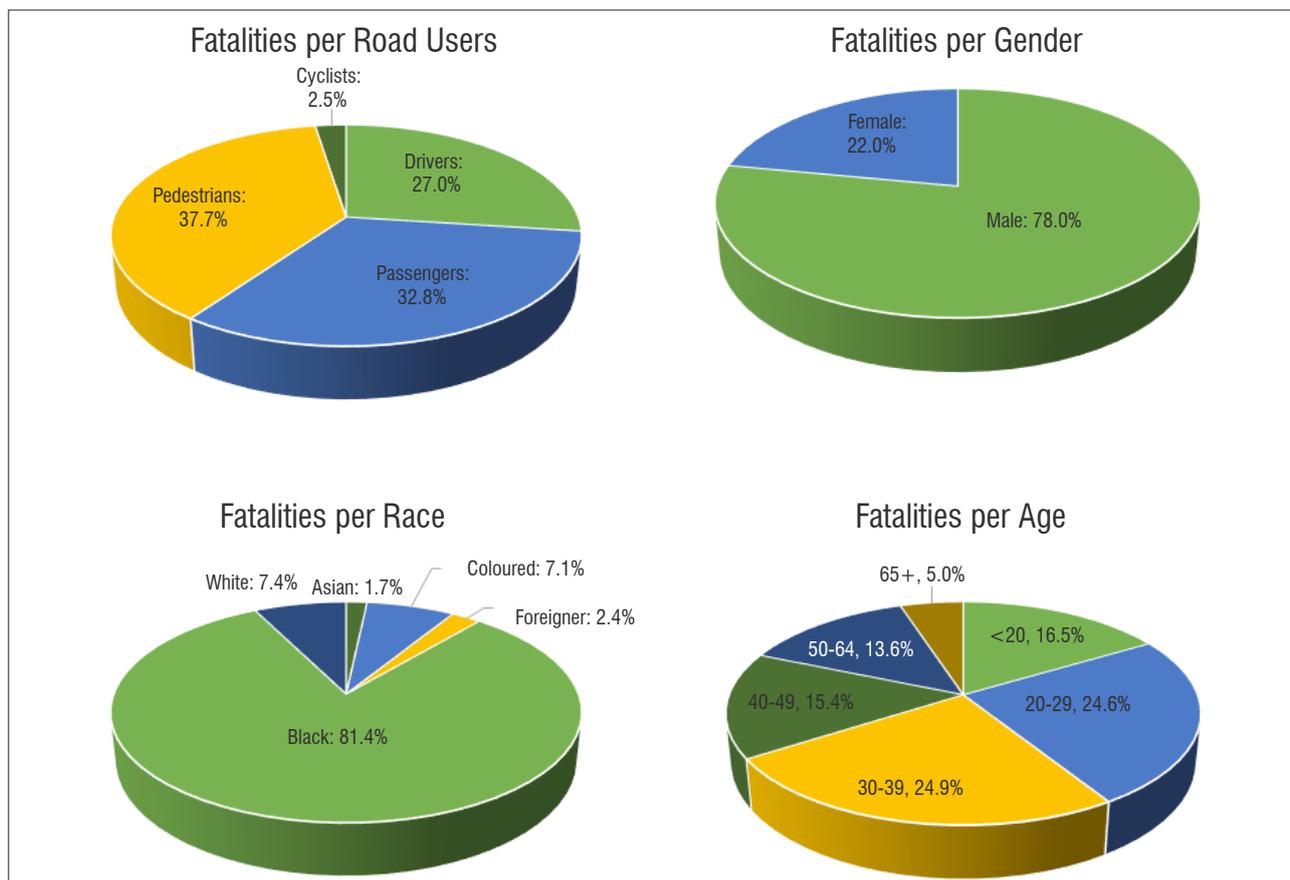
Data source: South African Department of Transport[9]

Figure 4: Distribution of road fatalities per road users, gender, race and age.

the South African Taxi Council structures as well as at the level of the provincial co-operatives.[14]

The number of unlicenced and/or unroadworthy vehicles on South African roads is alarming – the total number of vehicles falling in this specific category on 31 December 2015 was 1 013 278.

Road and environmental factors

One of the most pertinent environmental factors that must be taken into account when assessing accident information is the day of the week on which the accident occurred as well as the month in which the accident occurred. Figure 5 displays the number of fatalities per weekday and per month as reported by the South African Department of Transport.[9]

In total, 60.3% of the fatal accidents on South Africa's road network occurred on weekends (Fridays, Saturdays and Sundays) and slightly over 40% of the accidents occurred late afternoon and at night (16:00–22:00). This pattern is commented on in the National Road Safety Strategy 2011–2020 that states that these days are the most dangerous on South African roads.[3] This strategy document also reports drunk driving as one of the main reasons for the high number of accidents taking place late afternoon and at night.

The month with the highest fatality rate is December at 11.2%, but this result is to be expected as the roads are busiest in December because it is the main holiday period for South Africans. The figures reported for May, July and August are 9.2%, 9.2% and 10.1%, respectively. The lowest rates are reported for January and February, at only 6.0% and 5.0%; one could argue that people tend to stay at home after their December travels (Figure 5).

The following road and environmental conditions were listed[9] as reasons for fatal accidents (in descending order):

- Sharp bends
- Poor visibility
- Wet and slippery road surfaces
- Stray or wild animals
- Poor road surfaces
- Poor lighting
- Road works
- Poor and inadequate road markings
- Blind corners

The listed factors do not come as a surprise as various other authors (e.g. Chang and Mannering[15] and Lord et al.[16]) have investigated the effects of road and environmental conditions on road accidents. Only 21% of South Africa's road network is paved[17], which is why poor road surfaces, road works as well as poor and inadequate road markings are amongst the reasons listed for fatal accidents.

Summary

We reviewed the accident information by looking at three general contributing factors: human factors, road and environmental factors and the vehicles involved in the accident. All these factors also influence the risk analysis arena, specifically the short-term insurance and life-insurance industry. Figure 6 indicates the percentage of fatal accidents caused by each of the general contributing factors.

Human factors contributed to almost 80% of the fatal accidents on South African roads, for which 52.5% of these incidents related to jaywalking (when a pedestrian crosses a roadway where regulations do not permit doing so) and 11.6% to speeding. Road and environmental conditions contributed to 12.7% of the accidents, with sharp bends (22.0%) and poor visibility (16.5%) the main causes. The remaining 7.8% of the fatal accidents were caused by vehicle factors, with tyres bursting prior to the accident in 71.7% of the cases (Figure 7).

Should the observations discussed be tracked over time, they could be used as part of the trend analyses required for the previously discussed risk analyses. Even more useful would be trend analyses on all vehicle crashes, and not only the fatal ones. This analysis could be a valuable input in predicting the number of crashes, which in turn could be used as a factor to determine, for example, short-term insurance premiums.

Looking into the future

The historical fatality rates in South Africa as a result of road accidents paint a dire picture. Some sources even report that road accident fatalities may soon surpass the number of deaths attributed to HIV/Aids.[2] However, there may be light at the end of the proverbial tunnel as self-driving cars are expected to enter the market in the near future (see Figure 7 for an illustration of a self-driving car). It is reported that the first self-driving cars might be available from as early as 2018.[18]
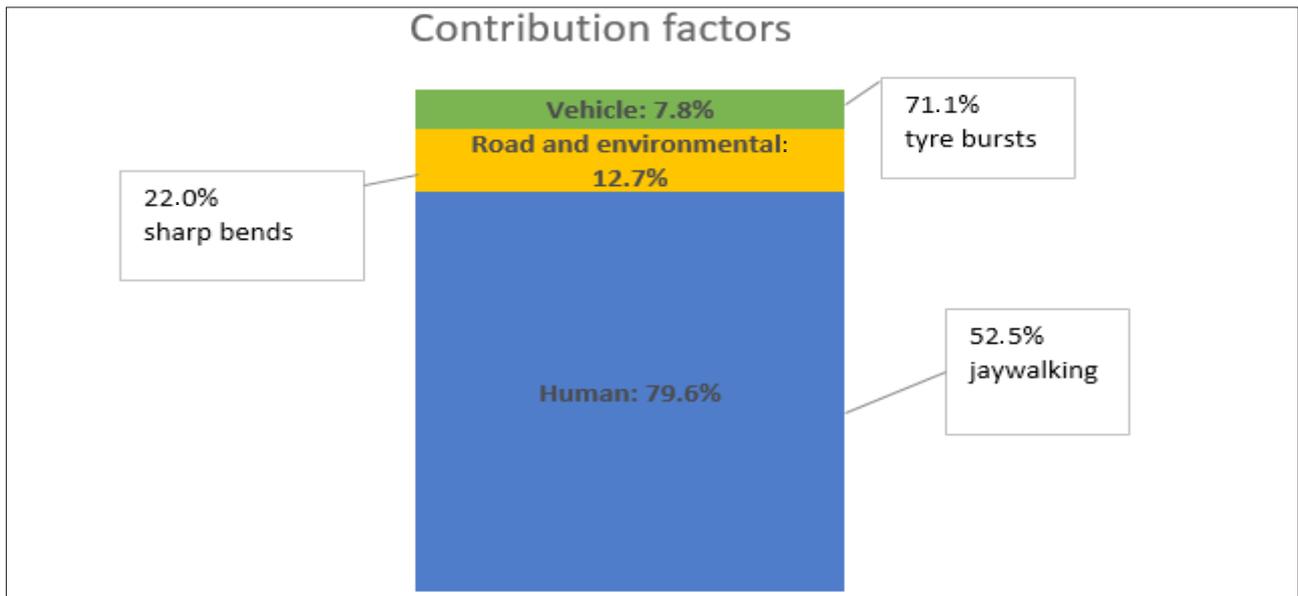
A number of benefits are anticipated to emerge from self-driving cars. Those benefits directly relating to road accidents and thus road fatalities are fewer crashes, reduced severity of crashes and improved emergency response times to crashes. Other expected benefits include less traffic congestion, shorter travel times, lower vehicle emissions, better fuel economy and lower insurance rates.[20]

Some of the concerns associated with self-driving cars are that these cars may not be driven as well as those driven by humans and that the cars might not know how to respond in unexpected situations. Other concerns relate to the safety consequences of equipment or system failures and system performance in inclement weather.[20]
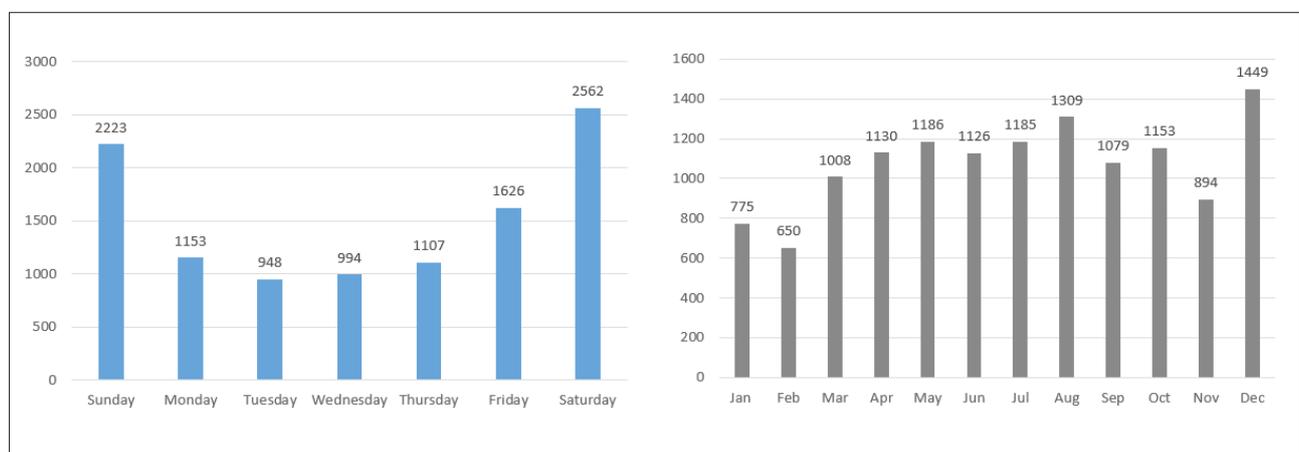
Data source: South African Department of Transport[9]

Figure 5: Number of road fatalities per (a) weekday and (b) month.

Contribution factors

Vehicle: 7.8%
Road and environmental: 12.7%
Human: 79.6%

71.1% tyre bursts

22.0% sharp bends

52.5% jaywalking

Data source: South African Department of Transport[9]

Figure 6: Graphical summary of the factors contributing towards fatal road accidents.

Image: Frank Derks, flickr, CC-BY-2.0

Figure 7: Illustration of a self-driving car.[19] Green indicates safe zones picked up by the sensors of the self-driving car and red indicates danger zones.

However, it should be noted that the timeline projections for self-driving cars entering the markets are aimed at developed countries. Developing countries like South Africa will have to wait another couple of years. Thus, although South Africans might still miss out on the benefits for another couple of years, hopefully some of the concerns will also have been addressed by the time that self-driving cars enter the South African market.

Another possible 'light' for the future is the impact of the various awareness campaigns. This article serves as one example, but a more formal campaign in South Africa is the Arrive Alive National Road Safety Strategy. The vision of Arrive Alive is 'to develop an effective road safety information portal that will enhance awareness of road safety and save lives'[3]. The World Health Organization has a Decade of Action for Road Safety 2011–2020 Strategy that aims to stabilise the growing number of road traffic crashes and to reduce the number of fatal road traffic crashes by half over a 10-year period.[3,21]

The possibility of mandatory third-party vehicle insurance might also shed some hope for the future. As mentioned, such insurance could possibly reduce the number of road accidents by reducing the number of damaged vehicles on the roads.[5]

Conclusion and ideas for future research

A number of different conclusions relating to fatal accidents on South African roads can be made based on the information reported. However, it should be noted that this report contains only information relating to fatal accidents. Although fatal accidents contributed 42.4% of the costs involved in accidents on South African roads, they equate to only 1.3% of all the accidents and have quite a unique trend. Thus, focusing on only fatal accidents (because of the lack of available data for other accident types) could result in incorrect conclusions.

Nonetheless, South African citizens should take note that more than 75 000 lives have been lost since 2010 as a result of road accidents, of which an alarming 37.7% were pedestrians. Another noteworthy observation is that slightly more than 40% of those deceased in 2015 were younger than 30 years of age. The high number of unlicenced and/or unroadworthy vehicles as well as poor road conditions are contributing to South Africa's high road network fatality rate. All of the abovementioned trends could be considered for risk analyses or future predictions, such as the number of accidents and insurance costs.

In 'the good, the bad and the ugly of South African fatal road accidents', the cost of human lives (12 994 lives in 2015) indisputably equates to 'the ugly' and the economic cost of accidents (ZAR142 951 million in 2015) associates with 'the bad'. 'The good', as set out in this paper, relates to the reduction in both of these costs that may be brought forward by the entrance of self-driving cars into the South African market as well as awareness campaigns like Arrive Alive.[3]

Both the cost of human life and the economic effect of road accidents in South Africa paint a shocking picture. This shocking picture emphasises the need to inform South Africans and to raise awareness of the effects of all types of accidents. That said, the South African Department of Transport[1] has acknowledged the need for a comprehensive road traffic incident database which should include fatal as well as non-fatal accident information. This database should ideally contain the following per accident: the severity, road users' gender and age, day and time of accident, type of road, road condition and location. However, a substantial coordinated effort across a number of sectors will be required to achieve this database.

Once a comprehensive database on all vehicle accidents has been constructed, our analyses can be repeated with a focus on all the different types of accidents, and can include investigation into other possible factors influencing the accident types. From a risk-analysis

perspective, with specific focus on short-term insurance, a database holding the information of all short-term insurance claims should also be available.

To some extent, such a database already exists and is managed by the South African Insurance Association in conjunction with their key stakeholders, including National Treasury, the Financial Services Board and the Department of Transport as well as all other industries closely related to the short-term insurance industry.[8] Managed by the South African Insurance Association, the Insurance Data System Committee has steered the collection and management of these crucial data to the point at which all contributing members have submitted their historical files and are submitting daily claims updates to the database. However, the voluntary submission of company statistics remains a challenge.[8] Alternatively, this central database could also be managed by an objective and independent institution like a university, in which case the data would also be more accessible for future research projects.

Authors' contributions

Both authors contributed equally to the conceptualisation of the study, data collection and writing of the manuscript.

References

1. South African Department of Transport. Cost of crashes in South Africa. Research and development report. Tshwane: Road Traffic Management Corporation; 2016. Available from: http://www.rtmc.co.za/images/media/Cost%20of%20Crashes%20in%20South%20Africa%20-%20RTMC%20-%20September%202016.pdf

2. AA Foundation. The impact of speed on South African road fatalities [document on the Internet]. c2003 [cited 2017 Jan 05]: Available from: http://aa-upgrade.ensight.ws/upload/documents/general/aa-foundation-speed.pdf

3. Arrive Alive. Decade of action and arrive alive. National road safety strategy 2011–2020. Bloemfontein: Arrive Alive; 2011. Available from: https://www.arrivealive.co.za/documents/Road%20Safety%20Strategy%20for%20South%20Africa%202011.pdf

4. Barry C. South Africa needs a third party insurance programme. FleetWatch. 2017 July 20 [cited 2018 Jan 20]: Available from: https://fleetwatch.co.za/south-africa-needs-a-third-party-insurance-programme/

5. Hesse M. 3rd-party cover for your car. IOL Personal Finance. 2015 July 04 [cited 2018 Jan 25]. Available from: https://www.iol.co.za/personal-finance/3rd-party-cover-for-your-car-1880315

6. Bayam E, Liebowitz J, Agresti W. Older drivers and accidents: A meta-analysis and data mining application on traffic data. Expert Syst Appl. 2005;29:598–629. https://doi.org/10.1016/j.eswa.2005.04.025

7. Bédard M, Guyatt GH, Stones MJ, Hirdes JP. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accid Anal Prev. 2002;3:717–727. https://doi.org/10.1016/S0001-4575(01)00072-0

8. South African Insurance Association. Annual review 2016. Metamorphosis. Johannesburg: South African Insurance Association; 2016. Available from: https://www.saia.co.za/annual-review-2016.pdf

9. South African Department of Transport. Road traffic report. Calendar: 1 January – 31 December 2015. Tshwane: Road Traffic Management Corporation; 2015. Available from: http://www.rtmc.co.za/rtmc1/Docs/JanDec2015.pdf

10. South African Department of Transport. Roads infrastructure policy for South Africa: Policy framework. Pretoria: ITS Engineers (Pty) Ltd; 2014. Available from: http://www.swartland.org.za/media/docs/2015/Documents/Services/Salga/Roads%20Infrastructure%20Policy%20-%20Policy%20Framework.pdf

11. Joseph R. Who are South Africa's worst drivers? The truth doesn't lie in the data. Africa Check. 2013 July 04 [cited 2018 Jan 12]. Available from: https://africacheck.org/2013/07/04/who-are-south-africas-worst-drivers-the-truth-isnt-in-the-data/

12. Summala H, Mikkola T. Fatal accidents among car and truck drivers: Effect of fatigue, age and alcohol consumption. Hum Factors. 1994;36(2):315–326. https://doi.org/10.1177/001872089403600211

13. Statistics South Africa (StatsSA). Community survey 2016. Statistical release P0301. Pretoria: StatsSA; 2016. Available from: http://cs2016.statssa.gov.za/wp-content/uploads/2016/07/NT-30-06-2016-RELEASE-for-CS-2016-_Statistical-releas_1-July-2016.pdf

14. Arrive Alive. Minibus taxis and road safety [webpage on the Internet]. c2017 [cited 2017 Jan 23]. Available from: https://www.arrivealive.co.za/Minibus-Taxis-and-Road-Safety

15. Chang LY, Mannering F. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accid Anal Prev. 1999;31(5):579–592. https://doi.org/10.1016/S0001-4575(99)00014-7

16. Lord D, Washington SP, Ivan JP. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accid Anal Prev. 2005;37(1):35–46. https://doi.org/10.1016/j.aap.2004.02.004

17. National Treasury. Local government budgets and expenditure review: 2006/07 – 2012/13. Pretoria: National Treasury; 2011. Available from: http://www.treasury.gov.za/publications/igfr/2011/lg/02.%202011%20LGBER%20-%20Final%20-%2013%20Sept%202011%20(renumbered).pdf

18. Kalogianni A. The next 10 years in car tech will make the last 30 look like just a warm-up. Digital Trends. 2016 January 12 [cited 2017 Jan 05]. Available from: http://www.digitaltrends.com/cars/the-future-of-car-tech-a-10-year-timeline/

19. Posky M. Saying what's popular: Suppliers claim automakers are overselling the future of 'mobility'. The Truth about Cars. 2017 August 03. Available from: http://www.thetruthaboutcars.com/2017/08/suppliers-claiming-automakers-oversold-future-mobility/

20. Schoettle B, Sivak M. A survey of public opinion about autonomous and self-driving vehicles in the U.S., the U.K., and Australia. Ann Arbor, MI: University of Michigan Transportation and Research Institute; 2014. Available from: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/108384/103024.pdf?sequence=1&isAllowed=y

21. World Health Organization (WHO). Global plan for the decade of road safety action 2011–2020. Geneva: WHO; 2011. Available from: http://www.who.int/roadsafety/decade_of_action/plan/plan_english.pdf

AUTHORS:
Christina M. Botai[1]
Joel O. Botai[1,2]
Abiodun M. Adeola[1]

AFFILIATION:
[1]South African Weather Service, Pretoria, South Africa

[2]Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Christina Botai

EMAIL:
christina.botai@weathersa.co.za

Spatial distribution of temporal precipitation contrasts in South Africa

The focus of the present study was to investigate the spatial-temporal variability and trends of precipitation concentration across South Africa using the Tropical Rainfall Measuring Mission (TRMM) 3B42 version 7 satellite precipitation data sets spanning 1998–2015. In the analysis, the precipitation concentration index (PCI) was used to infer the variability of temporal precipitation contrasts and the spatial distributions at annual, seasonal and supra-seasonal timescales. The results indicate that precipitation concentration across South Africa exhibits noticeable spatial-temporal variability. In terms of PCI classification criteria, the precipitation distribution ranges from relatively uniform (mainly in the central and southern interior of South Africa) to highly irregular (especially to the northeastern and western parts of South Africa) at annual timescales. At seasonal timescales, the precipitation distribution is uniform during December–February season, moderate during March–May and September–November seasons while during the June–August season, the precipitation distribution is highly irregular. Furthermore, during the 1998–2015 period, there exists a spatial and temporal pattern of PCI trends which are generally statistically insignificant. The PCI analysis results reported in this study are essential because they provide valuable information on the long-term total variability in the precipitation records across South Africa. In particular, this study contributes towards evaluating the spatial contrasts or concentration of the different accumulated amounts of the received precipitation. Results from this study have important scientific and practical applications in hydrological hazard risks (floods and droughts) and soil erosion monitoring.

Significance:
- Precipitation concentration exhibits spatial-temporal variability.
- At an annual timescale, precipitation concentration is highly irregular in most parts of the country.
- Precipitation concentration distribution varies across seasons.

Introduction

There is an increasing global consensus that climate change and variability is a reality, as manifested both in terms of the varying climatic variable mean and extremes. A number of scientific publications from diverse fields point to this eventuality.[1-4] The implication of the changing climate has elicited great concern in individuals and societies in general as well as in various sectors (e.g. agriculture, water, energy and health) directly or indirectly affected by the impacts of climate change. In South Africa, the imminent implication of climate change is for instance the strain imposed on water resources as a result of prolonged dry spells. This unfortunate situation of limited water supply reverberates across South Africa currently (e.g. 2016/2017 hydrological year), as manifested in the significant drop of dam levels, see for example Figure 1. In addition, changes in precipitation patterns are contributing to the on-going drought experienced in various provinces, with five of the provinces officially declared as drought disaster regions. This hydrological drought has serious socio-economic consequences, including threats to the nation's food security, health and economy.

Because of the dire consequences imposed by climate change, understanding processes that relate to this change can be vital for climate-related preparedness measures, which thereby enhance economic growth sustainability, including planning for agriculture, water resources management and planning, economic planning and health. As precipitation is one of the essential variables associated with climate change, the changing patterns and trends of this climatic variable require proper and systematic attention within the topic of climate change.

South Africa, which is considered a semi-arid country, is characterised by highly variable diurnal[5], intra-seasonal[6,7] as well annual timescales[8]. These variations are reported to be intensifying (i.e. becoming more variable) over time (see for example studies by Hewitson and Crane[9], Engelbrecht et al.[10], Shongwe et al.[11], Pohl et al.[12] and references therein). This inherent variability characterises the country's annual rainfall distribution patterns, concentration and intensity. Furthermore, rainfall variability in southern Africa has also been linked to variations in teleconnection patterns such as the El Niño Southern Oscillation (ENSO).[6] Understanding precipitation characteristics (e.g. patterns, variability, intensity and concentration) forms part of preparedness and mitigation measures that can be put in place to reduce the impacts of weather and climate extreme events. In particular, an in-depth analysis of precipitation at sub-regional level is important as the results can be used to quantify spatial and temporal variation of precipitation concentration and distribution patterns in the region.

The precipitation concentration index (PCI) has been proposed as an indicator to assess spatial and temporal distributions as well as concentration of precipitation.[13,14] The PCI has the ability to provide valuable information on long-term total variability in the amount of rainfall received.[15,16] Research studies on PCI include the computation of PCI to evaluate the statistical structure of precipitation rates for Pearl River Basin situated in Southern China.[14] In addition, Martin-Vide[17] calculated PCI values with the aim of assessing the spatial patterns of rainfall in Peninsular, Spain. Li et al.[18] estimated the PCI values for the Kaidu River Basin and assessed the spatial and temporal variability of the statistical structures of precipitation across Xinjiang, China. De Luis et al.[19] made an analysis of the PCI
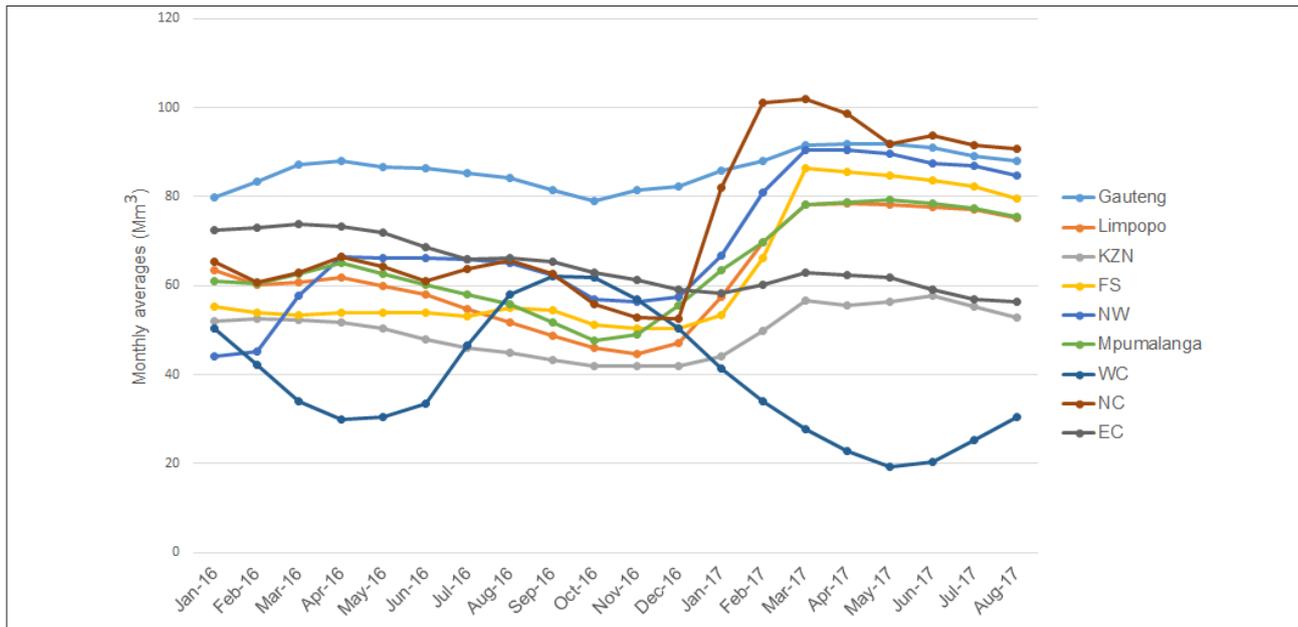
KZN, KwaZulu-Natal; FS, Free State; NW, North West; WC, Western Cape; NC, Northern Cape; EC, Eastern Cape

Figure 1:   Status of averaged dam levels from January 2016 to August 2017 at provincial level.

using a monthly precipitation database of Spain. Valli et al.[20] computed PCI values at annual and seasonal scale and analysed the patterns of rainfall in agro-climatic zones of Andhra Pradesh, India. Iskander et al.[21] calculated PCI values and Simple Daily Intensity Index (SDII) and investigated trends and variability of annual precipitation total and annual number of rainy days in 34 stations distributed over Bangladesh. Gocic et al.[22] computed PCI values using monthly precipitation data sets from 29 stations in Serbia and analysed spatial variability of monthly precipitation and long-term total variability in the precipitation series. In addition, Gocic et al.[22] used three support vector machine (SVM) models coupled with the discrete wavelet transform, the SVM firefly algorithm and the SVM radial basis function to estimate and predict PCIs in Serbia.

It is apparent from the literature that analysis studies on PCI have been undertaken in various countries and regions in the past decades. Most of the reported studies, however, were conducted in India, China, Bangladesh and Spain. In South Africa, a number of studies have been carried out on precipitation patterns (e.g. Kruger[23], New et al.[24], Kruger and Sekele[25], MacKellar et al.[26], Botai et al.[27] and references therein), yet characterising the variability of precipitation using the PCI has never been done. The aim of the present study was to analyse the spatial-temporal variability of PCI across South Africa using the Tropical Rainfall Measuring Mission (TRMM) daily 3B42 satellite derived data sets. In this regard, the PCI analysis calculated at annual, supra-annual and seasonal timescales is used to characterise the precipitation contrasts across South Africa over the period 1998–2015. This study provides valuable scientific and practical insights on the overall patterns of spatial-temporal contrasts of accumulated rainfall relevant for agricultural planning and water resources management.

Study area

South Africa is known to be a semi-arid to arid country, and is particularly characterised by a highly variable climate with constrained water resources as a result of weather extremes enforced by climate change and variability. The country's climate conditions range from the Mediterranean in the southwestern corner of the country to temperate in the interior plateau and subtropical in the northwestern region. South Africa's average annual rainfall is about 450 mm/year, which is below the world's 860 mm average per year. South Africa is conventionally characterised by four main seasons, i.e. summer (December–January–February (DJF)), autumn (March–April–May (MAM)), winter (June–July–August

(JJA)) and spring (September–October–November (SON)). Rainfall in South Africa exhibits seasonal variability, with most of the rainfall occurring mainly during summer months; see for example Figure 2. However, in the southwestern region of the country, rainfall occurs mostly in winter months. South Africa experiences rainfall that varies significantly from west to east. Annual rainfall in the northwestern region often remains below 200 mm, whereas much of the eastern Highveld receives between 500 mm and 900 mm (occasionally exceeding 2000 mm) of rainfall per annum. The central part of the country receives about 400 mm of rain per annum, with wide variations occurring closer to the coast.

Data

We utilised the Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) daily 3B42 precipitation products covering the period from 1 January 1998 to 1 November 2015. These products cover the 50°S to 50°N latitude belt at a spatial resolution of 0.25° × 0.25°.[28] The TRMM data sets are freely available from http://disc.gsfc.nasa.gov/uui/datasets. They were retrieved from the TRMM calibrated-based multiple satellite microwave and infrared measurements using the TMPA algorithm reported by Chiu et al.[29] In particular, the TMPA algorithm involves four stages and can be summarised as follows: (1) calibrating and combining the microwave precipitation estimates; (2) creating infrared precipitation estimates using the calibrated microwave precipitation; (3) combining the microwave and infrared estimates; and (4) incorporating rain gauge data. Overall, the TRMM products provide a significant opportunity to monitor precipitation over tropical and sub-tropical regions wherein the TRMM Microwave Imager becomes reliable and useful.[30,31]

Methodology

In this study, PCI was calculated and used to assess rainfall concentration in South Africa for the period 1998–2015. The PCI values were calculated on three different timescales, namely annual, seasonal and supra-seasonal. At annual timescale, the PCI values were calculated according to Equation 1 as described in Oliver[13] as well as in Zhang and Qian[32].

$$PCI_{annual} = \frac{\sum_{i=1}^{12} Pi^2}{(\sum_{i=1}^{12} Pi)^2} \times 100 \qquad \text{Equation 1}$$
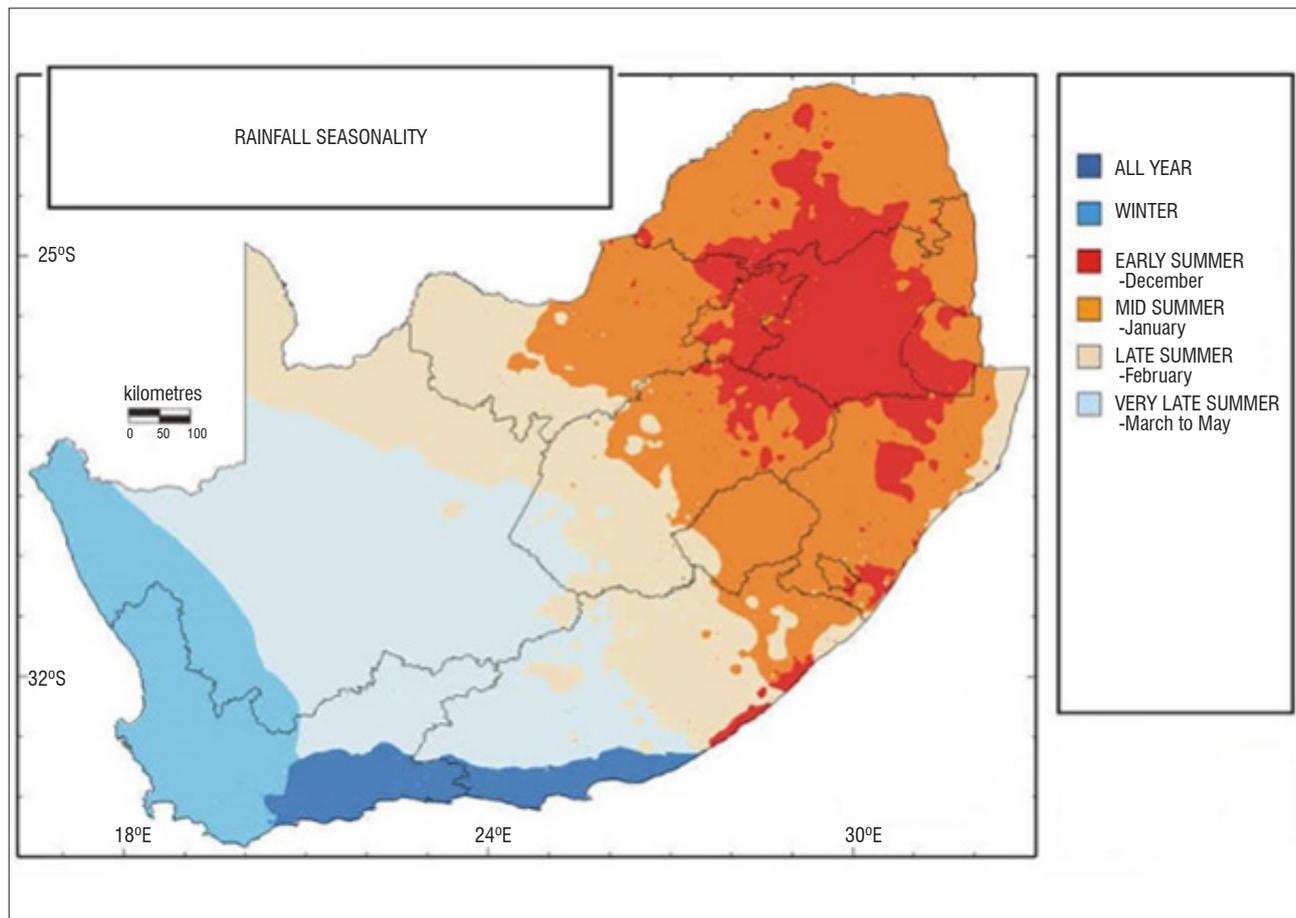
Source: EUMETSAT[35]

Figure 2:   Rainfall seasonality over South Africa.

In Equation 1, *Pi* corresponds to the amount of precipitation of the ith month. In this way, the PCI values were calculated for each year at each gridded point to detect the temporal variation of the index. The computed annual PCI can then be used to indicate the monthly variation in the distribution of precipitation on an annual basis. At seasonal timescale (i.e. DJF, MAM, JJA and SON), the PCI values were calculated based on Equation 2:

$$PCI_{seasonal} = \frac{\sum_{i=1}^{3} Pi^2}{(\sum_{i=1}^{3} Pi)^2} \text{ x } 0.25 \qquad \text{Equation 2}$$

Similarly, at supra-seasonal timescales (e.g. the wet season represented by October to March months and the dry season from April to September months), the PCI values were calculated according to Equation 3:

$$PCI_{supra\text{-}seasonal} = \frac{\sum_{i=1}^{6} Pi^2}{(\sum_{i=1}^{6} Pi)^2} \text{ x } 0.50 \qquad \text{Equation 3}$$

According to Oliver's classification, PCI values can be classified as ranging from uniform precipitation distribution to a strong irregular distribution; see a summary of this classification in Table 1.

This classification was also adopted in this study in order to assess changes of statistical parameters of PCI (including the median, mean, coefficient of variation (CV), standard deviation (std), trends and significant trends) with the aim of understanding the rainfall distribution and concentration in the period from 1998 to 2015. All the statistical moments generated in the analysis were rasterised and plotted using the level-plot graphics function of the Lattice package in R, the statistical computing and graphics software.

Table 1:   Classification of precipitation concentration based on precipitation concentration index (PCI) values

PCI value	Significance (temporal distribution)
PCI ≤ 10	Uniform precipitation distribution (low precipitation concentration)
10 < PCI ≤ 15	Moderate precipitation distribution
16 < PCI ≤ 20	Irregular precipitation distribution
PCI > 20	Strong irregular precipitation distribution

Table adapted from Oliver[13]

Results

Spatial characteristics of annual PCI

The spatial variability of the annual PCI statistics between 1998 and 2015 are presented in Figure 3. It is noted that the precipitation distribution is largely moderate (i.e. $11 \leq PCI_{\{mean/median\}} \leq 15$) along the east coastal regions of the country, covering parts of the Eastern Cape and KwaZulu-Natal (KZN) Provinces. The central and the eastern parts of the country depict PCI values ranging between 14 and 16 while the north, west and southern regions depict higher PCI values, with some areas reaching 25. These areas are therefore characterised by irregular to high precipitation distribution over a year. Overall, the pattern of the annual PCI values delineates South Africa into three major regions: moderate, irregular and high rainfall distribution regions. Based on the coefficient of variation and the standard deviation results, the PCI is highly variable towards the west and southwestern regions of the country, covering
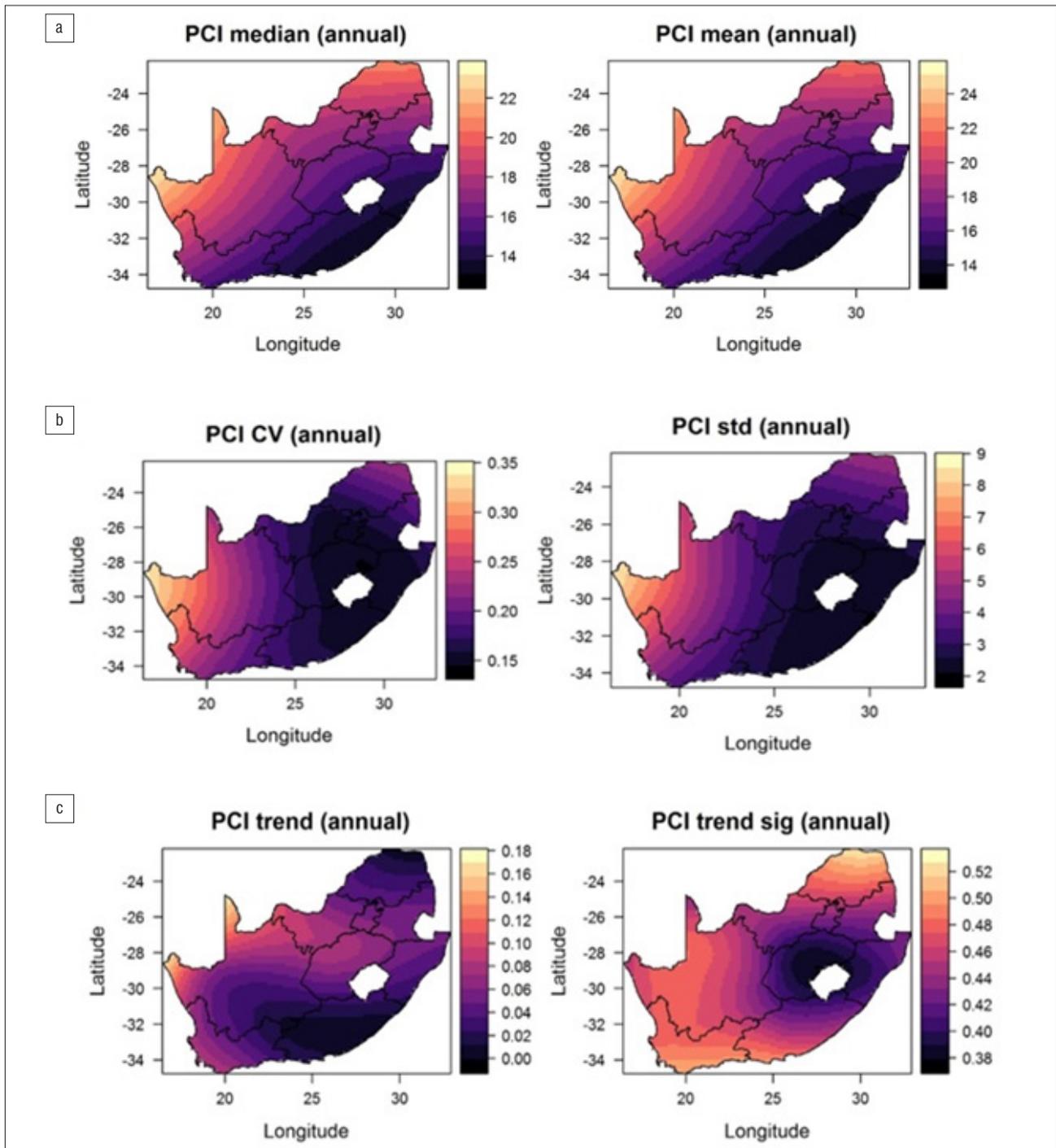
Figure 3: Spatial statistical characteristics of precipitation concentration index (PCI): (a) distribution of median and mean of PCI; (b) variability given by coefficient of variation (CV) and standard deviation (std) of PCI; and (c) distribution of non-linear pre-whitened PCI trends and its statistical significance.

the Northern Cape and Western Cape Provinces, and less variable over the northern regions of the country (e.g. Limpopo Province). The central interior region and parts of the eastern region of the country (i.e. Free State, KZN, Gauteng as well as some parts of the North West and Eastern Cape Provinces) depict lower PCI variability. During the period 1998 to 2015, South Africa exhibited positive, statistically insignificant PCI trends (less than 0.2 mm/year). According to Theil[33], PCI trends are considered significant if the estimated p-value falls below a critical value (e.g. $\alpha < 0.05$), indicating 95% confidence. Furthermore, negative trends are also observed in small regions of the Eastern Cape and Limpopo Provinces (see Figure 3c). These results corroborate the precipitation trends reported in Kruger and Nxumalo[34] for example.

Spatial characteristics of seasonal PCI

Results for seasonal PCI calculations are depicted in Figure 4. During the summer season, lower PCI values (≤ 10) are evident in the interior region (e.g. the Free State and parts of the North West) and towards the eastern part of the country, covering the Eastern Cape and KZN Provinces. Based on the PCI classification criteria, these regions are characterised by lower precipitation distribution at seasonal timescales. However, the PCI values increase towards the southwestern regions of South Africa, covering the Northern Cape and Western Cape Provinces. During the summer season, precipitation in these regions exhibits moderate to high seasonal precipitation distribution. The PCI values
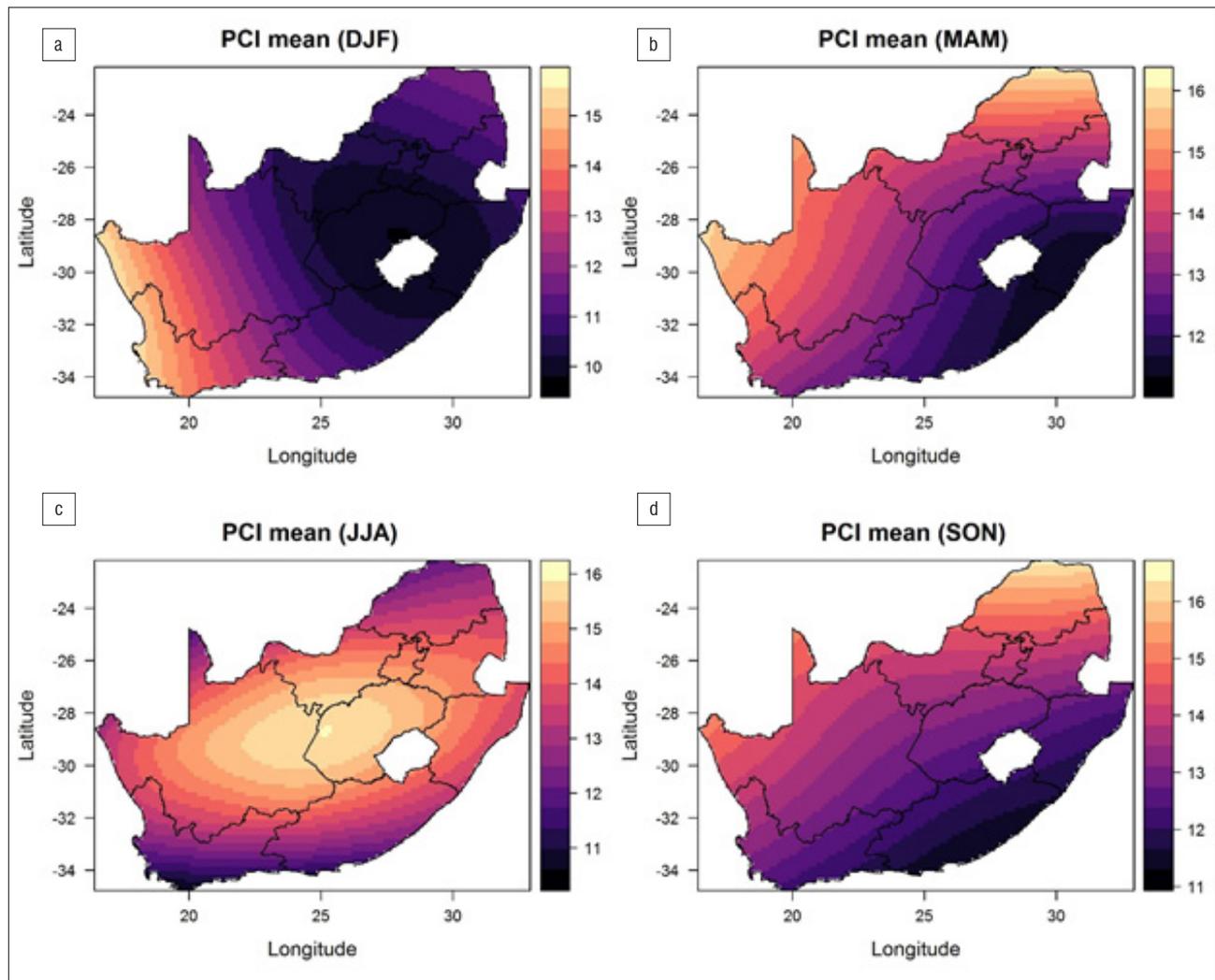
Figure 4: Spatial distribution of seasonal precipitation concentration index (PCI) averages for (a) December–January–February (DJF), (b) March–April–May (MAM), (c) June–July–August (JJA), (d) September–October–November (SON).

during autumn are low along the east coast while high in the north, west and southwestern regions, implying low precipitation and moderate to irregular precipitation distribution, respectively. Precipitation in winter is generally more spatially regularly distributed, with higher PCI values (up to 16) are observed in the central region of the country.

During the spring season, parts of the Western Cape, Eastern Cape and KZN Provinces exhibit largely low precipitation distribution while the central interior regions (including the Free State, Gauteng and Northern Cape Provinces) exhibit mostly moderate precipitation distribution.

Figure 5 depicts the spatial distribution of variability of PCI values. From Figure 5, the PCI values in summer are less variable in the central interior, while highly variable in the western areas of South Africa (e.g. large areas of the Northern Cape and Western Cape Provinces). During autumn, there is high variability in the precipitation distribution patterns in the Western Cape, Northern Cape, North West, parts of the Free State and Limpopo Provinces. On the other hand, low variability in precipitation distribution patterns is evident in large parts of KZN and the Eastern Cape. Similar spatial variability patterns were exhibited in winter and spring. From Figure 5, it can also be observed that the degree of JJA PCI variability is relatively twice that of SON PCI variability. During the winter season, the Northern Cape, North West and some parts of the Free State and Limpopo Provinces exhibit a higher degree of PCI spread from the mean. This variability is in contrast to the inherent spread in the PCI values in some southern parts of the Western Cape and the southwestern parts of the Eastern Cape Provinces. It can be concluded that, during the 1998–2015 period, the degree of PCI variability exhibited a noticeable seasonal dependence.

Figure 6 illustrates the seasonal PCI trends during the period of study. As shown in Figure 6, the central interior parts of South Africa, including some parts of the Northern Cape, Eastern Cape and Free State Provinces exhibit negative PCI trends while the rest of the country depicts positive trends, albeit statistically insignificant during DJF. During MAM and JJA, the vast majority of South Africa exhibits statistically insignificant positive trends while limited northeastern parts of Limpopo and Mpumalanga exhibit negative trends. In contrast, some parts of the Western Cape Province depict statistically insignificant negative trends while the rest of South Africa depicts positive (yet statistically insignificant) trends during SON.

Spatial characteristics of supra-seasonal PCI

Figure 7 depicts the PCI mean and standard deviation values during the generally wet and dry months across South Africa. During the wet months (see left upper panel in Figure 7), the PCI values exhibit an increasing southwest to northeast gradient. In addition, the precipitation distribution varies from irregular to highly distributed, from the south to the northern parts of the country. In particular, the coastal regions generally exhibit low PCI values. The standard deviation exhibits similar spatial pattern of precipitation distribution, as shown on the left lower panel of Figure 7. During the dry months (right upper panel in Figure 7), low to moderate precipitation distribution is evident mostly in KZN, the Free State and some parts of the North West, Gauteng and Mpumalanga Provinces. On the other hand, irregular precipitation distribution is observed in the Northern Cape and Western Cape Provinces. The spread of the PCI from the PCI mean exhibits similar spatial distribution patterns to those of the precipitation distribution during the wet and dry months.
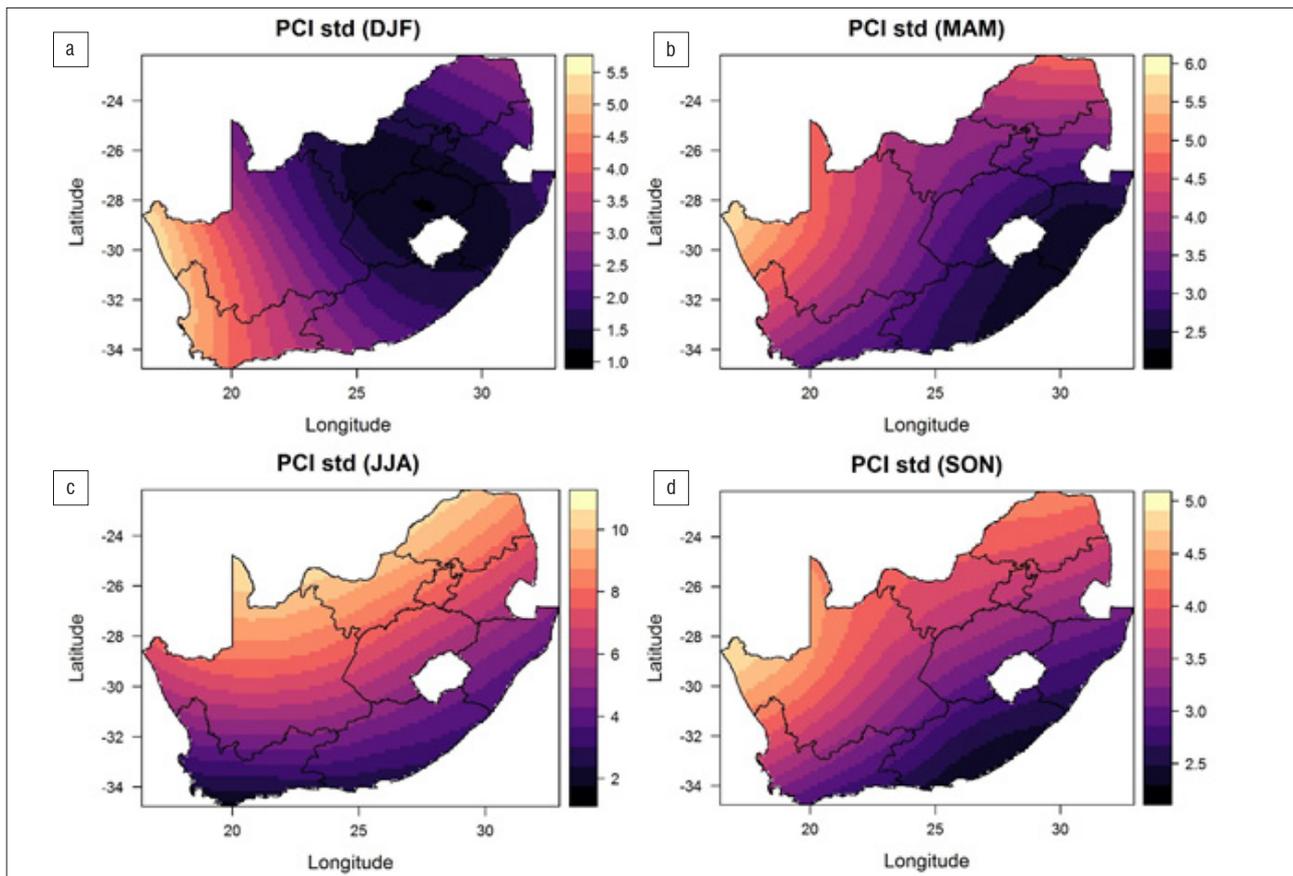
Figure 5: Spatial variability of seasonal precipitation concentration index (PCI) for (a) December–January–February (DJF), (b) March–April–May (MAM), (c) June–July–August (JJA), (d) September–October–November (SON).
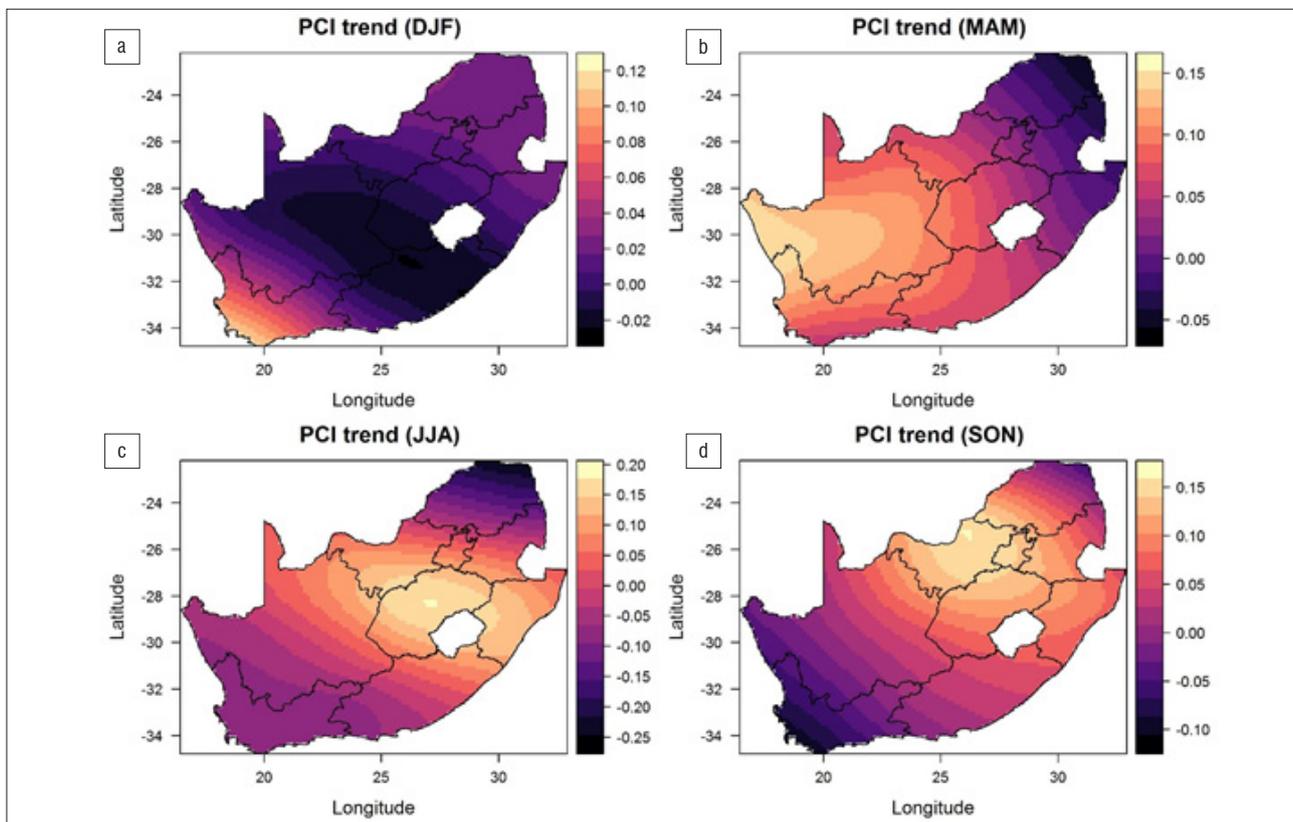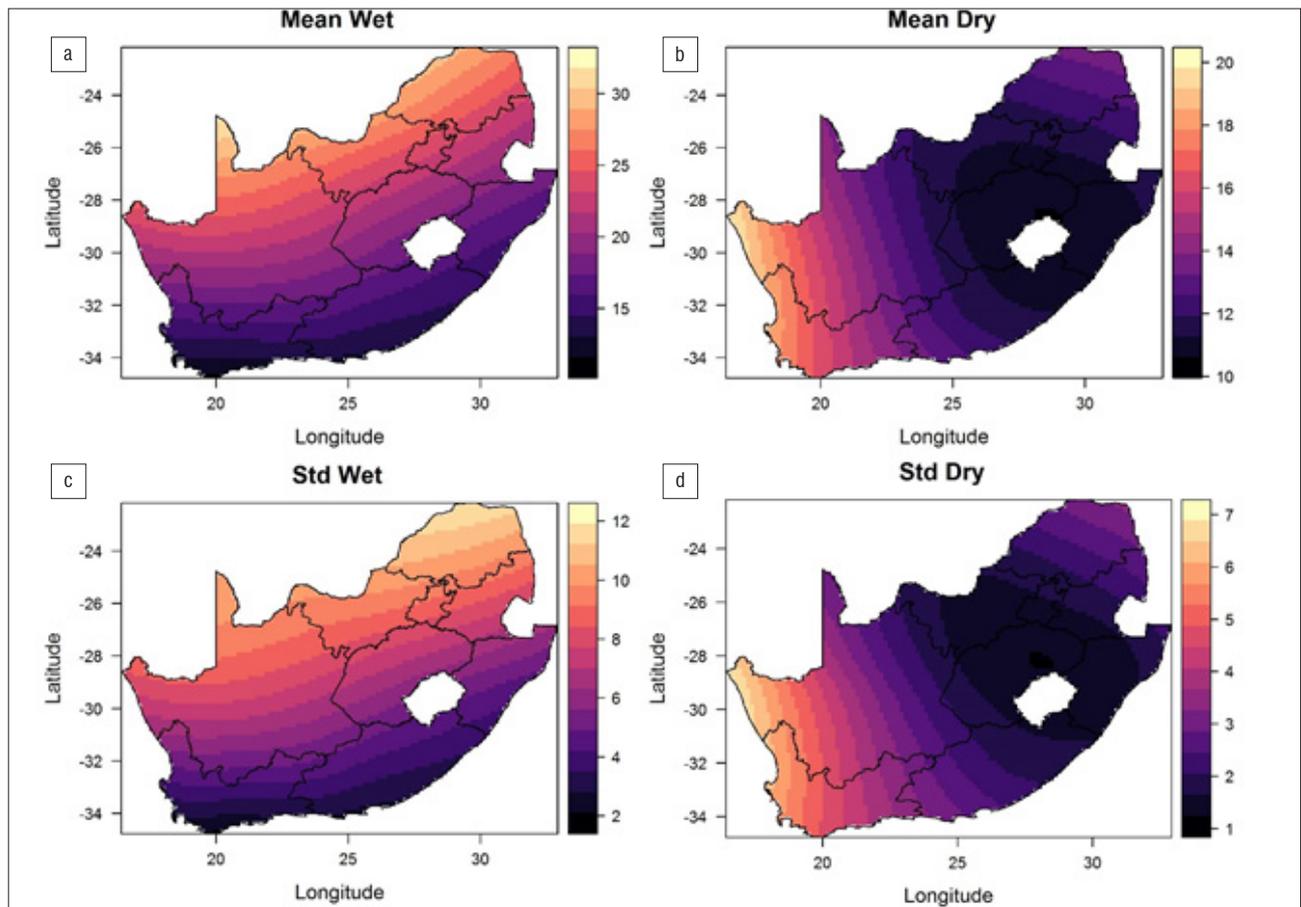

Figure 6: Distribution of seasonal precipitation concentration index (PCI) trends for (a) December–January–February (DJF), (b) March–April–May (MAM), (c) June–July–August (JJA), (d) September–October–November (SON).

Figure 7: Spatial variability of supra-seasonal precipitation concentration index for (a,c) wet and (b,d) dry supra-seasons.
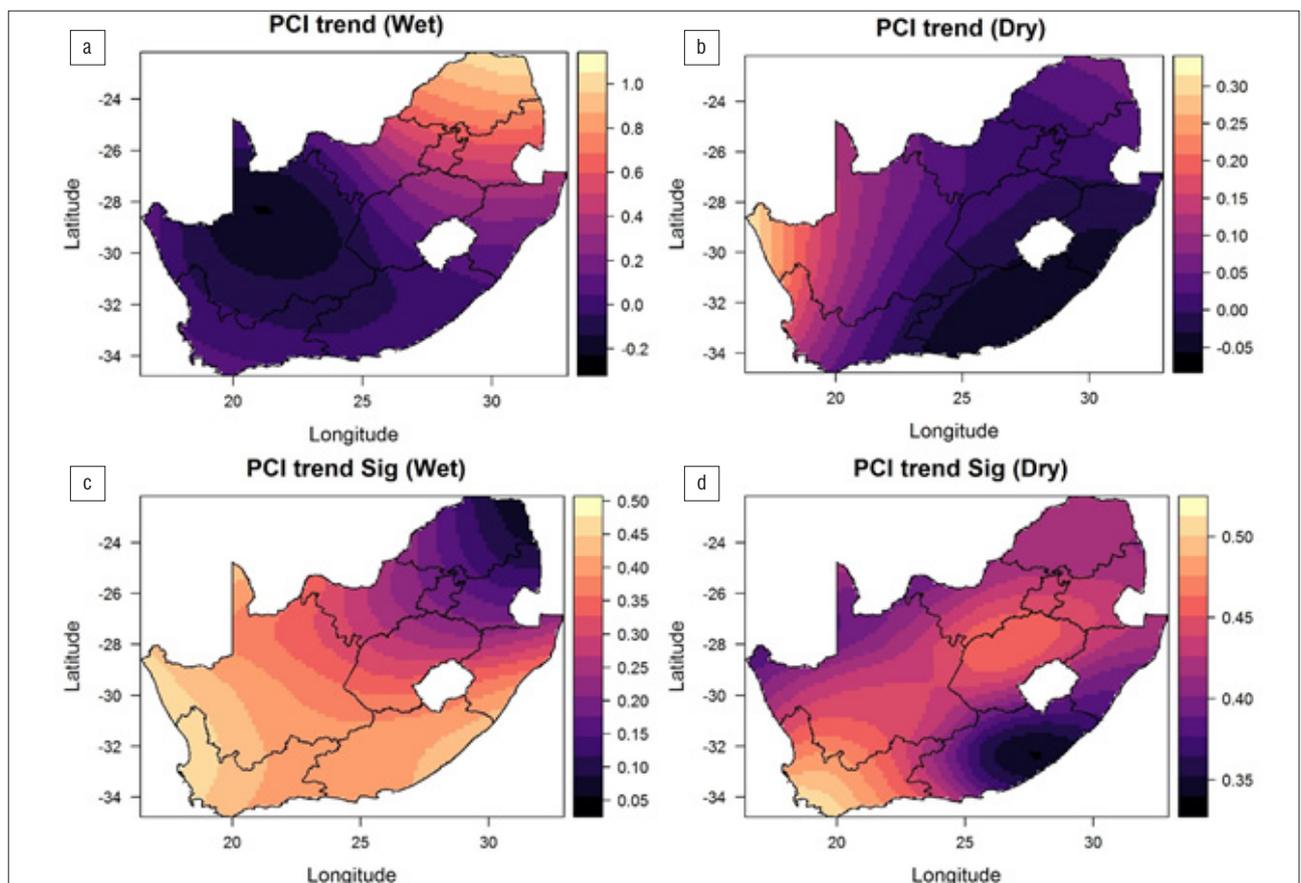
Figure 8: Supra-seasonal precipitation concentration index (PCI) trends for (a,c) wet and (b,d) dry supra-seasons.

Trends in PCI values during the wet and dry months are depicted in Figure 8. As depicted in Figure 8, statistically insignificant negative (see left lower panel) trends are observed during the generally wet months, mostly in the Northern Cape Province whereas noticeable statistically insignificant positive trends are observed in Limpopo, Mpumalanga, Gauteng, KZN, and some parts of the North West and Free State Provinces. Furthermore, the dry period exhibits statistically insignificant negative trends (see right lower panel) mainly in the Eastern Cape as well as some parts of KZN while the Northern Cape and Western Cape Provinces exhibit statistically insignificant positive trends.

Discussion

South Africa is one of the countries in Africa that receives less precipitation annually. Viewed as a water-stressed country, South Africa receives an average annual rainfall of less than 500 mm. Rainfall in South Africa is highly unpredictable and sporadic in most regions, with the intensity, frequency, duration and distribution always being a concern, particularly to farmers and water resources managers. Various regions in South Africa are currently experiencing prolonged drought duration events that are crippling water resources and also impacting on other sectors such as agriculture.[36-38] In particular, five of the nine South African provinces (i.e. the Western Cape, Free State, North West, KZN and Limpopo) are currently declared as drought disaster areas. The rainfall anomalies, particularly in these drought disaster provinces, have caused undesirable effects on crop production and have stressed water resources, which threaten the food security, health and economy of South Africa.[36]

The spatial distribution of seasonal and annual PCI values calculated in this study indicates that the vast majority of South Africa received contrasting precipitation (i.e. low to moderate distribution patterns) in the last two decades. Because poor-resourced farmers depend on rainfall for their crop and livestock production, the spatial-temporal distribution of PCI across South Africa suggests unfavourable conditions for their agriculture production. Given that rainfall is a vital climatic factor influencing crop growth, less concentrated rainfall has adverse effects on agricultural productivity.[39,40] In addition, rainfall is the main component that feeds water to most water reservoirs in South Africa. Therefore, inadequate rainfall will have a significant impact on water resources and water supply to the South African community. For example, drought attributed to low rainfall in the Western Cape Province has resulted in significant water level reduction in most dams in the province.[38] In addition, the Vaal Dam, which supplies water to approximately 25 million citizens, reached its lowest level during the 2015/2016 hydrological year, although summer rainfall has since improved the situation. Furthermore, changes in precipitation distribution, intensity and concentration may lead to soil erosion (gully erosion is dominant in the Northern Cape and Eastern Cape while sheet and rill erosion mostly occur in the eastern parts of South Africa), desertification and forest fires. As reported in Van Dijk et al.[41], for example, high concentration and intensity of rainfall attributed to heavy storms may cause greater soil loss through increased soil particle detachment. In addition, high-intensity rainfall often causes higher rates of infiltration, excess run-off and greater transport of eroded sediments.[42,43] These impacts are mostly unfavourable to agriculture and water sectors in South Africa.

Conclusions

We investigated the spatial-temporal distribution of precipitation concentration in South Africa for the period 1998–2015. The TRMM 3B42 version 7 satellite precipitation data sets were used to calculate PCI values at three timescales (i.e. annual, seasonal and supra-seasonal). Based on the results, precipitation concentration across South Africa exhibits noticeable spatial-temporal contrasts. The spatial variability of PCI values manifests as transition zones of the precipitation concentration that are in most cases aligned to the three climatological zones of the region. At annual timescales, the PCI values illustrate that the precipitation distribution ranges from relatively uniform (mainly in the central and southern interior) to highly irregular (especially to the northeastern and western parts of South Africa). At seasonal timescales, the

precipitation distribution is uniform during DJF, moderate during MAM and SON and highly irregular in JJA. Between 1998 and 2015, the annual PCIs across South Africa exhibit positive (but statistically insignificant) trends. Furthermore, during the SON season, large parts of South Africa exhibit generally negative PCI trends while during the MAM and JJA seasons, the PCI trends are generally positive. During the wet season, precipitation distribution is highly irregular across the Limpopo, Gauteng, North West and Northern Cape Provinces. On the other hand, the dry season exhibits largely moderate to low precipitation concentration over large parts of South Africa except for some parts of the western regions of the Northern Cape and northern regions of the Western Cape which depict irregular precipitation concentration. The PCI trends during the wet season are largely positive in the eastern and southwestern parts of South Africa while the interior parts exhibit negative PCI trends. The observed spatial pattern of PCI trends transition zones during the wet season mimic the climatological zones of South Africa, i.e. positive PCI trends correspond to the Mediterranean and sub-tropical climatic zones while the interior tropical climatological zone is characterised by negative PCI trends. Overall, given that precipitation concentration is an index of rainfall variability, water availability and rainfall erosivity, our results have practical applications in water resource planning and management as well as disaster preparedness. In addition, the spatial contrasts of PCI could provide information on water variability for the relevant government agencies. The implications of such observed change have strong influence on the natural processes of soil erosion, flooding, fluvial regimes and groundwater recharge and, therefore, serve as a warning tool for flooding and erosion within, for example, urban and peri-urban communities that are prone to these hydrological processes.

Author's contributions

C.M.B. conceptualised the research study and wrote the manuscript. J.O.B. was responsible for writing scripts to analyse the data and edited the manuscript. A.M.A. edited the manuscript.

References

1. Meehl GA, Washington WM, Santer BD, Collins WD, Arblaster JM, Hu A, et al. Climate change projections for the twenty-first century and climate change commitment in the CCSM3. J Clim. 2006;19:2597–2616. https://doi.org/10.1175/JCLI3746.1

2. IPCC, WGI. Climate change 2007: The physical science basis: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. In: IPCC fourth assessment report. Geneva: IPCC; 2007.

3. Mora C, Frazier AG, Longman RJ, Dacks RS, Walton MM, Tong EJ, et al. The projected timing of climate departure from recent variability. Nature. 2013;502:183–195. https://doi.org/10.1038/nature12540

4. Rouault M, Roy SS, Balling Jr RC. The diurnal cycle of rainfall in South Africa in the austral summer. Int J Climatol. 2013;33:770–777. https://doi.org/10.1002/joc.3451

5. Roy SS, Rouault M. Spatial patterns of seasonal scale trends in extreme hourly precipitation in South Africa. Appl Geogr. 2013;39:151–157. https://doi.org/10.1016/j.apgeog.2012.11.022

6. Pohl B, Fauchereau N, Reason CJC, Rouault M. Relationships between the Antarctic Oscillation, the Madden-Julian Oscillation, and ENSO, and consequences for rainfall analysis. J Clim. 2009;23:238–254. https://doi.org/10.1175/2009JCLI2443.1

7. Liebmann B, Blade I, Kiladis GN, Carvalho LMV, Senay GB, Allured D, et al. Seasonality of African precipitation from 1996 to 2009. J Clim. 2012;25:4304–4322. https://doi.org/10.1175/JCLI-D-11-00157.1

8. Pohl B, Rouault M, Roy SS. Simulation of the annual and diurnal cycles of rainfall over South Africa by a regional climate model. Clim Dyn. 2014;43:2207–2226. https://doi.org/10.1007/s00382-013-2046-8

9.   Hewitson BC, Crane RG. Consensus between GCM climate change projections with empirical downscaling. Int J Clim. 2006;26:1315–1337. https://doi.org/10.1002/joc.1314

10. Engelbrecht FA, McGregor JL, Engelbrecht CJ. Dynamics of the Conformal-Cubic Atmospheric Model projected climate-change signal over southern Africa. Int J Climatol. 2009;29:1013–1033. https://doi.org/10.1002/joc.1742

11. Shongwe ME, Van Oldenborgh GJ, Van Den Hurk BJJ. Projected changes in mean and extreme precipitation in Africa under global warming. Part I: Southern Africa. Amer Meteor Soc. 2009;22:3819–3837. https://doi.org/10.1175/2009JCLI2317.1

12. Pohl B, Macron C, Monerie P-A. Fewer rainy days and more extreme rainfall by the end of the century in southern Africa. Sci Rep. 2017;7(46466):1–7. https://doi.org/10.1038/srep46466

13. Oliver JE. Monthly precipitation distribution: A comparative index. Prof Geogr. 1980;32:300–309. https://doi.org/10.1111/j.0033-0124.1980.00300.x

14. Zhang Q, Xu CY, Marco G, Chen YP, Liu CL. Changing properties of precipitation concentration in the Pearl River Basin, China. Stoc Environ Res Risk Assess. 2009;23:377–385. https://doi.org/10.1007/s00477-008-0225-7

15. Michiels P, Gabriels D, Hartmann R. Using the seasonal and temporal precipitation concentration index for characterizing monthly rainfall distribution in Spain. Catena. 1992;19(1):43–58. https://doi.org/10.1016/0341-8162(92)90016-5

16. Apaydin H, Erpul G, Bayramin I, Gabriels D. Evaluation of indices for characterizing the distribution and concentration of precipitation: A case for the region of Southeastern Anatolia Project, Turkey. J Hydrol. 2006;328:726–732. https://doi.org/10.1016/j.jhydrol.2006.01.019

17. Martin-Vide J. Spatial distribution of a daily precipitation concentration index in Peninsular Spain. Int J Climatol. 2004;24:959–971. https://doi.org/10.1002/joc.1030

18. Li X, Jiang F, Li L, Wang G. Spatial and temporal variability of precipitation concentration index, concentration degree and concentration period in Xinjiang, China. Int J Climatol. 2011;31(11):1679–1693.

19. De Luis M, Gonzalez-Hidalgo JC, Brunetti M, Longares LA. Precipitation concentration changes in Spain 1946-2005. Nat Hazards Earth Syst Sci. 2011;11:1259–1265. https://doi.org/10.5194/nhess-11-1259-2011

20. Valli M, Sree KS, Krishna IVM. Analysis of precipitation concentration index and rainfall prediction in various agro-climatic zones of Andhra Pradesh, India. Int Res J Environ Sci. 2013;2(5):53–61.

21. Iskander SMd, Rajib MA, Rahman MdM. Trending regional precipitation distribution and intensity: Use of climate indices. Atmos Clim Sci. 2014;4:385–393.

22. Gocic M, Shamshirband S, Razak Z, Petkovic D, Ch S, Trajkovic S. Long-term precipitation analysis and estimation of precipitation concentration index using three support vector machine methods. Adv Meteorol. 2016;2016:1–11. https://doi.org/10.1155/2016/7912357

23. Kruger AC. Observed trends in daily precipitation indices in South Africa: 1910-2004. Int J Climatol. 2006;26:2275–2285. https://doi.org/10.1002/joc.1368

24. New M, Hewitson B, Stephenson DB, Tsiga A, Kruger A, et al. Evidence of trends in daily climate extremes over southern and west Africa. J Geophys Res. 2006;111,D14102, 11 pages. https://doi.org/10.1029/2005JD006289

25. Kruger AC, Sekele SS. Trends in extreme temperature indices in South Africa: 1962–2009. Int J Climatol. 2013;33:661–676. https://doi.org/10.1002/joc.3455

26. MacKellar N, New M, Jack C. Observed and modelled trends in rainfall and temperature for South Africa: 1960–2010. S Afr J Sci. 2014;110(7/8), Art. #2013-0353, 13 pages. https://doi.org/10.1590/sajs.2014/20130353

27. Botai MC, Botai JO, De Wit JP, Ncongwane KP. Trend characteristics of rainfall and streamflow in the Vaal River catchment. 33rd Annual Conference of the South African Society for Atmospheric Sciences. 2017:38.

28. Huffman GJ, Adler RF, Bolvin DT, Gu G, Nelkin EJ, Bowman KP, et al. The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. J Hydrometeorol. 2007;8:38–55. https://doi.org/10.1175/JHM560.1

29. Chiu LS, Liu Z, Vongsaard J, Morain S, Budge A, Neville P, et al. Comparison of TRMM and water district rain rates over New Mexico. Adv Atmos Sci. 2006;23:1–13. https://doi.org/10.1007/s00376-006-0001-x

30. Kummerow C, Barnes W, Kozu T, Shiue J, Simpson J. The tropical rainfall measuring mission (TRMM) sensor package. J Atmos Ocean Technol. 1998;15:809–817. https://doi.org/10.1175/1520-0426(1998)015<0809:TTRMMT>2.0.CO;2

31. Guo R, Liu Y. Evaluation of satellite precipitation products with rain gauge data at different scales: Implications for hydrological applications. Water. 2016;8(7), Art. #281, 20 pages. https://doi.org/10.3390/w8070281

32. Zhang L, Qian Y. Annual distribution features of the yearly precipitation in China and their interannual variations. Acta Meteorol Sin. 2003;17(02):146–163.

33. Theil H. A rank-invariant method of linear and polynomial regression analysis, I. Nederlands Akad Wetensch Proc. 1950;53:386–392.

34. Kruger AC, Nxumalo MP. Historical rainfall trends in South Africa: 1921 – 2015. Water SA. 2017;43(2):285–297.

35. EUMETSAT. Winter lightning associated with cut-off low over South Africa [webpage on the Internet]. c2011 [cited 2017 Oct 10]. Available from: http://www.eumetsat.int/website/home/Images/ImageLibrary/DAT_2351448.html

36. Araulo J, Abiodun B, Crespo O. Impacts of drought on grape yields in Western Cape, South Africa. Theor Appl Climatol. 2014;123(1):117–130.

37. Botai CM, Botai JO, Dlamini LC, Zwane NS, Phaduli E. Characteristics of droughts in South Africa: A case study of Free Sate and North West provinces. Water. 2016;8:439–462. https://doi.org/10.3390/w8100439

38. Botai CM, Botai JO, De Wit JP, Ncongwane KP, Adeola AM. Drought characteristics over the Western Cape Province, South Africa. Water. 2017;9:876–892. https://doi.org/10.3390/w9110876

39. Gbetibouo G, Hassan R. Economic impact of climate change on major South African field crops: A Ricardian approach. Glob Planet Change. 2005;47:143–152. https://doi.org/10.1016/j.gloplacha.2004.10.009

40. Kang Y, Khan S, Ma X. Climate change impacts on crop yield, crop water productivity and food security – A review. Progr Nat Sci. 2009;19:1665–1674. https://doi.org/10.1016/j.pnsc.2009.08.001

41. Van Dijk AIJM, Bruijnzeel LA, Rosewell CJ. Rainfall intensity – Kinetic energy relationships: A critical literature appraisal. J Hydrol. 2002;261:1–23. https://doi.org/10.1016/S0022-1694(02)00020-3

42. Rose CW. Erosion and sedimentation. In: Bonnell M, Hufschmidt MM, Gladwell JS, editors. Hydrology and water management in the humid tropics – Hydrological research issues and strategies for water management. Cambridge, UK: Cambridge University Press; 1993. p. 301–343. https://doi.org/10.1017/CBO9780511564468.018

43. Mohamadi MA, Kavian A. Effects of rainfall patterns on runoff and soil erosion in field plots. Int Soil Water Conserv Res. 2015;3:273–281. https://doi.org/10.1016/j.iswcr.2015.10.001

Cosmic ray neutrons provide an innovative technique for estimating intermediate scale soil moisture

AUTHORS:
Thigesh Vather[1,2] iD
Colin Everson[1,2] iD
Michael Mengistu[3] iD
Trenton Franz[4] iD

AFFILIATIONS:
[1]The Centre for Water Resources Research, University of KwaZulu-Natal, Pietermaritzburg, South Africa

[2]South African Earth Observation Network, Pietermaritzburg, South Africa

[3]South African Weather Service, Pretoria, South Africa

[4]School of Natural Resources, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

CORRESPONDENCE TO:
Thigesh Vather

EMAIL:
thigesh@saeon.ac.za

Soil moisture is an important hydrological parameter, which is essential for a variety of applications, thereby extending to numerous disciplines. Currently, there are three methods of estimating soil moisture: ground-based (in-situ) measurements; remote sensing based methods and land surface models. In recent years, the cosmic ray probe (CRP), which is an in-situ technique, has been implemented in several countries across the globe. The CRP provides area-averaged soil moisture at an intermediate scale and thus bridges the gap between in-situ point measurements and global satellite-based soil moisture estimates. The aim of this study was to test the suitability of the CRP to provide spatial estimates of soil moisture. The CRP was set up and calibrated in Cathedral Peak Catchment VI. An in-situ soil moisture network consisting of time-domain reflectometry and Echo probes was created in Catchment VI, and was used to validate the CRP soil moisture estimates. Once calibrated, the CRP was found to provide spatial estimates of soil moisture, which correlated well with the in-situ soil moisture network data set and yielded an R^2 value of 0.845. The use of the CRP for soil moisture monitoring provided reliable, accurate and continuous soil moisture estimates over the catchment area. The wealth of current and potential applications makes the CRP very appealing for scientists and engineers in various fields.

Significance:

- The cosmic ray probe provides spatial estimates of surface soil moisture at an intermediate scale of 18 hectares.

- A single cosmic ray probe can replace a network of conventional in-situ instruments to provide reliable soil moisture estimates.

- The cosmic ray probe is capable of estimating soil moisture in previously problematic areas (saline soil, wetlands, rocky soil).

- Cosmic ray probes can provide data for hydro-meteorologists interested in land–atmosphere interactions.

- The cosmic ray probe estimates can be promising for remote sensing scientists for product calibration and validation.

Introduction

There has been a continual need to monitor and measure the various parameters in land surface hydrology, in order to deepen the understanding of hydrological processes, their importance in the hydrological cycle and their interactions between each other.[1,2] Soil moisture is an important parameter in the hydrological cycle and impacts a variety of applications, including agricultural management, climate and weather applications, flood and drought forecasting and groundwater recharge.

Soil moisture is a difficult parameter to continuously monitor and measure at a catchment scale because of its heterogeneous characteristics. It varies both spatially and temporally and is thus a dynamic resource. Currently there are three methods of estimating soil moisture: (1) ground-based (in-situ) measurements, which are carried out using field instruments; (2) remote sensing based methods, which use specialised sensors on satellites and aircrafts and (3) land surface models, which use meteorological data as inputs, at a predefined spatial resolution.[3,4] Inherently, each of these methods possess their respective advantages and limitations, constraining their effectiveness for hydrological applications.[5]

In-situ measurements of soil moisture are the conventional methods used by several disciplines. The point measurements obtained cannot adequately represent the spatial characteristics of soil moisture and thus limit our understanding of feedbacks to the atmosphere. However, these point measurements play a key role in a variety of large-scale applications and are invaluable as both calibration and validation data.[6] Recently, the cosmic ray probe (CRP) method, which is an innovative in-situ instrument, has been developed and implemented in numerous countries across the globe.[7] Countries such as the USA, Australia, Germany and the United Kingdom have invested greatly in this technology and have created national soil moisture monitoring networks. The CRP technique obtains the area-averaged soil moisture, at an intermediate scale, by observing and measuring the cosmic ray neutrons above the soil surface.[8,9] The CRP has not been evaluated in South Africa, thus the aim of this study was to test and evaluate the suitability of the CRP to provide spatial estimates of soil moisture, using the Cathedral Peak Catchment VI as a study area. The Cathedral Peak area is important as it is the 'water towers' of the region, as much of our water is generated in this area. The CRP was set up in Catchment VI, as it is already heavily instrumented, and an added soil moisture instrument would add value to any potential research done in the area.

The potential applications of the CRP make it appealing to scientists in various fields, such as agricultural and ecological monitoring, climate science, drought and flood forecasting, as well as slope stability.[10] It should also be noted that the discipline of remote sensing can benefit greatly from this innovative technology by using CRP estimates for both the calibration and validation of sensors and data products, as it overcomes the spatial

limitations of conventional in-situ soil moisture estimates.[10] The use of the CRP across different continents results in measurement technique consistency, which reduces the uncertainties related to in-situ measurements that use a variety of probe types or methods.[11]

In-situ methods of soil moisture measurement

Conventional methods of soil moisture estimation

In-situ soil moisture measurements have played a key role for a variety of large-scale applications and have been invaluable as calibration and validation data for satellite-based products, sensors and models.[6] These include the gravimetric, neutron scattering and time-domain reflectance (TDR) methods.[12] The major limitation of all conventional in-situ methods is that they provide point measurements, which do not account for the spatial characteristics of soil moisture.[13] No single point measurement can be entirely representative of larger areas, because of the heterogeneity that exists in soil properties, topography, land cover and meteorological conditions.[6] To overcome this limitation, dense soil moisture networks can be set up. However, the high costs of operation and maintenance make the set-up of the network financially unfeasible,[6] particularly for the objective of establishing a national scale network.

Cosmic ray probe

The CRP is the only in-situ technique that can obtain the average soil moisture content over hundreds of square metres – something that would require a dense in-situ point measurement network.[9,14] The CRP estimates soil moisture hourly at a shallow measurement depth that ranges between 0.12 m and 0.72 m. The CRP system consists of neutron counters (moderated and bare tubes); a data logger which measures barometric pressure, humidity and temperature; a telemetry system with antenna to connect to an iridium satellite; and a battery and solar panel for powering the system.

Production of cosmic ray neutrons

Cosmic ray neutrons originate in space, where they are produced by the blast waves of exploding stars. Some of these high energy particle flows in space reach the earth's atmosphere, where they are affected by the earth's magnetic field.[15] The high energy particles are captured into the earth's atmosphere and collide with atmospheric nuclei to initiate a cascade of secondary cosmic rays.[14] Fast neutrons are created, as these secondary cosmic rays pass through the atmosphere and then through the top few metres of the biosphere, hydrosphere and lithosphere.[10] These fast neutrons undergo elastic collisions with nuclei present in the soil, thereby losing energy.[10] Some of the fast neutrons are adsorbed by the soil during the collision, whilst others will be scattered above the surface of the soil.[15] The cosmic ray neutrons lose energy with each collision, therefore high energy neutrons become fast neutrons (in the atmosphere), which further lose energy and become thermal neutrons (in the soil). As a result of fast neutrons being strongly moderated by hydrogen present in the environment, their measured intensities relate to changes in soil moisture, as well as other hydrogen sources at the earth's surface.[8,9,16]

Moderation of neutrons

According to Ochsner et al.[14] and Jiao et al.[15], the moderation process of cosmic ray neutrons depends on three factors:

1. The scattering probability or the elemental scattering cross-section.

2. The logarithmic decrement of energy per collision.

3. The number of atoms of an element per unit mass of material, which is proportional to the concentration of the element and to the inverse of its mass number.

The combination of the abovementioned factors defines the neutron stopping power of a material.[14] Hydrogen, which is found mainly as water in the soil, plays the most significant role in moderating cosmic ray neutrons in the soil. Hydrogen has by far the highest stopping power of any element ($\sim$25 times greater), as the hydrogen atom has a high probability of scattering a neutron as a result of its fairly large elastic

scattering cross-section.[15] Hydrogen is the most efficient element with regard to the decrement of energy per collision and has a low atomic mass and makes up a substantial portion of all the atoms in many soils, as a consequence of the presence of water in the soil.[15] The presence of water within the soil pores plays an important and central role in moderating the concentration of cosmic ray neutrons above the soil surface.[17]

Cosmic ray probe measurements

The fast neutrons that are produced in the air and soil travel in all directions between the air and soil, thus creating an equilibrium concentration of neutrons. This equilibrium concentration is shifted because of changes (addition or subtraction) in the hydrogen content of the media. The soil moisture content is estimated by the concentration of cosmic ray neutrons, which are generated within the soil and moderated predominantly by hydrogen, before being diffused back into the atmosphere.[16] The soil moisture content can therefore be inferred directly from these neutron fluxes.[10]

The CRP system consists of two sensors, which are the moderated and bare sensor. The moderated sensor is shielded by a 25.4-mm-thick polyethylene and measures the fast neutron intensity above the soil surface. The fast neutron intensity above the soil surface is inversely proportional to the soil moisture content.[7] The bare tube (unshielded) measures the thermal neutrons, which are attributed to the water above the soil surface (biomass and snow).

Measurement footprint and depth of the cosmic ray probe estimates

The CRP senses all hydrogen present within the distance that fast neutrons can travel in air, water, soil and other materials near the earth's surface. Thus, the measurement distance varies according to the density and chemical composition of the material.[14] The footprint (measurement area) of the CRP is defined as the area around the probe from which 86% of the counted neutrons arise, is primarily associated with the chemical and physical properties of the air and is inversely proportional to the air density.[15]

The radial footprint of the CRP is reliant on the neutrons' ability to travel hundreds of metres from their source, through the neutrons scattering in the air.[7] Hence the scattering properties of air significantly affect the diameter of the footprint.[15] When the CRP is placed in a static position a few metres above the ground, it has a radial footprint of $\sim$240 m ($\sim$480 m diameter) at sea level.[7]

The effective measurement depth is the thickness at which 86% of counted neutrons arise, which depends strongly on the soil moisture content.[14] The measurement depth ranges from 0.12 m to 0.7 m and is inversely proportional to the soil moisture content.[9] The decrease in the measurement depth as a result of an increase in soil moisture is nonlinear.

The cosmic ray technology consists of instruments that can be used either in a fixed position (CRP) or in a moving vehicle (cosmic ray rover). The CRP is used to obtain continuous measurements of an area which are normally summed to give hourly neutron counts, whilst the roving method can be used for mapping soil moisture over large areas.[18]

The CRP method has several advantages: the method is passive and non-contact (non-invasive); the system is easily automated and portable; and it has minimal power requirements. Applications are not limited to soil moisture, as the CRP can also be used to estimate above-ground biomass and snow depth. Most importantly, the measurement footprint is at an intermediate scale of around 18 ha. The method requires low data processing and is insensitive to soil texture, bulk density or surface roughness.[8-10,15-17,19,20]

The CRP does, however, require calibration, which is performed by obtaining corresponding measurements of area-averaged soil moisture and neutron intensity. The area-averaged soil moisture is obtained from ground-based point measurements, by collecting several soil samples within the CRP measurement area and determining the average soil moisture per calibration. The measured neutron intensities need to

be adjusted and corrected for variations in location, incoming high-energy particles, atmospheric pressure, absolute humidity and changes in biomass.[21]

It is recommended that a minimum of two field calibrations are conducted, to obtain reliable estimates of soil moisture. These two calibrations should be carried out in opposing seasons, to calibrate the instrument over contrasting soil moisture statuses (a dry calibration in winter and a wet calibration in summer).[18] Representative soil samples of the measurement area are required to be analysed, to correct the calibration function for lattice water (water of crystallisation not measured with conventional techniques) and water in organic matter.[14]

For absolute volumetric water content (VWC), studies have shown that a minimum of three calibrations should be conducted to obtain soil moisture estimates with an error less than 2–3%VWC. The CRP can provide useful data even if the calibration is not conducted. For use in flood forecasting or land surface model data assimilation, the relative changes (patterns) can be used from the corrected neutron counts directly.

Methodology

Study site

The Cathedral Peak Research Catchment VI lies within the KwaZulu-Natal Province of South Africa, in the Tugela Catchment, as shown in Figure 1. It comprises 15 gauged catchments that are situated on the little Berg, located below the Drakensberg escarpment, which creates a natural border between the northeastern side of Lesotho and South Africa.[22]

The Drakensberg mountain range is the highest mountain range in South Africa and gives rise to many of the rivers which are of great economic importance to the country.[23] Catchment VI has a catchment area of 0.68 km² and is located by latitude 28.99°S and longitude 29.25°E. It is moderately dissected by streams and has a stream density of 3.25 km/km².[23] The altitude ranges from 1860 m at the weir (northernmost point of the catchment) to 2070 m at the highest point of the catchment, with an average catchment slope of 19%.[24]

The land cover of Catchment VI is uKhahlamba basalt grassland.[25] The soils in the catchments are moderately weathered immature soils, which are primarily derived from basalt.[26] The soils in the catchment are classified as lateritic yellow and red earths, with heavy black soils occurring in saturated zones and along stream banks.[27] There is a contrast in soil properties among the soil layers. The topsoil has a friable consistency, which results in rapid infiltration, whilst the subsoil has a very high clay content, which results in poor infiltration. The topsoil has a high organic matter content (6–10%), which results in a high water-holding capacity.[24] The region is characterised climatically by its cold dry winters and hot wet summers. The mean annual precipitation is 1400 mm, with 85% falling between October and March.[22] Catchment VI has a mean annual precipitation of 1299 mm.[27]

Cosmic ray probe

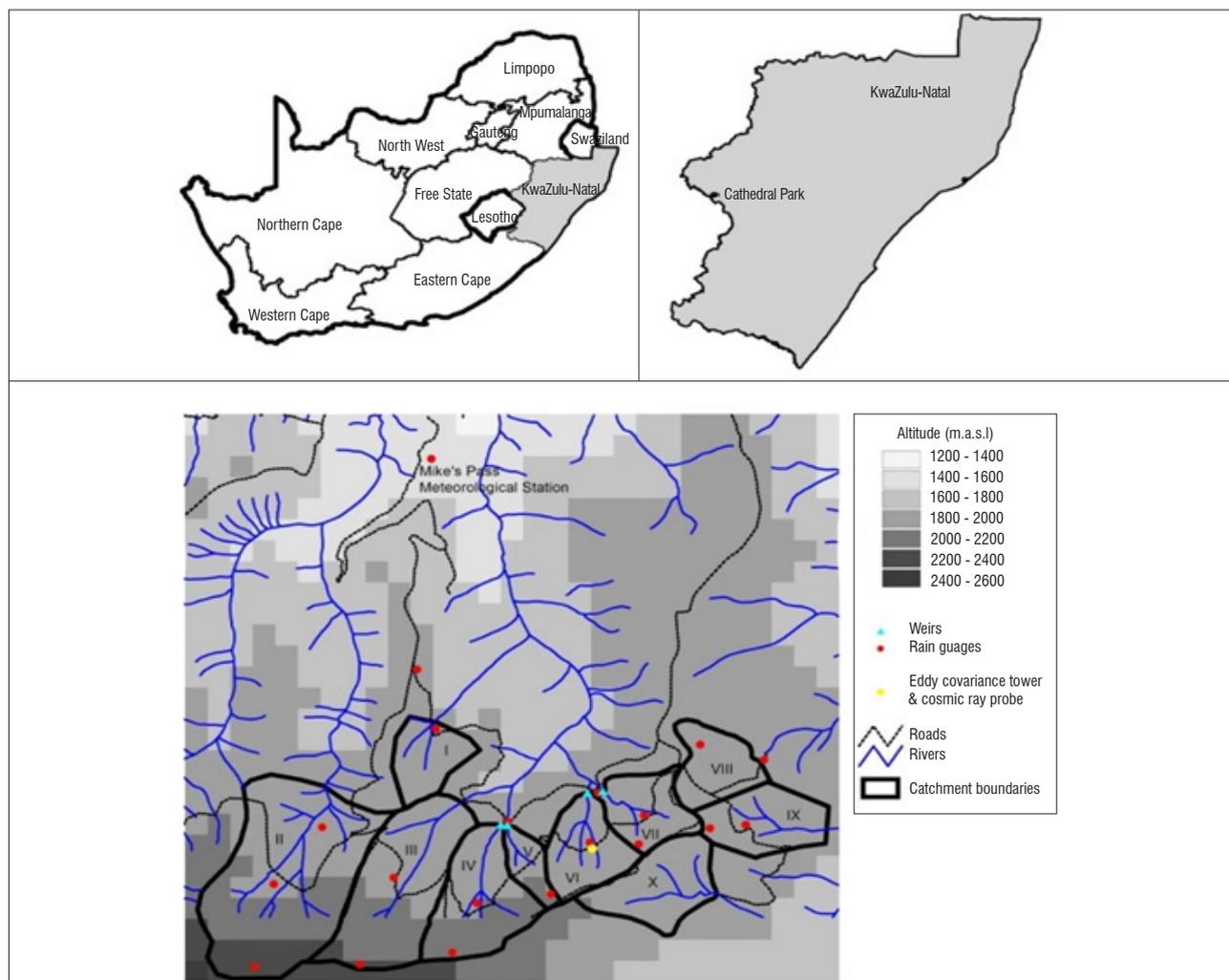The CRP was installed on 28 February 2014 in the Cathedral Peak Catchment VI (Figure 2).

Figure 1: Location of Cathedral Peak Catchment VI, within the Tugela Catchment, in the KwaZulu-Natal Province, South Africa.
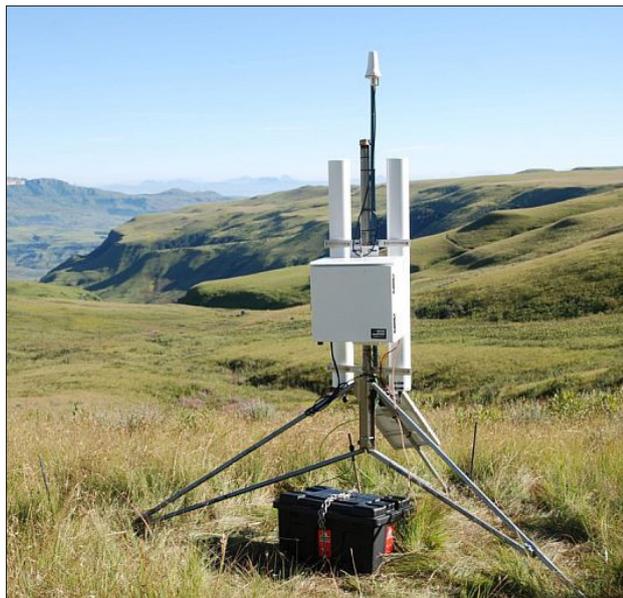
Photo: Thigesh Vather

Figure 2: The cosmic ray probe in Cathedral Peak Catchment VI.

The calibration procedure involves obtaining an estimate of the area-averaged soil moisture content over the CRP measurement footprint by gravimetric sampling and the subsequent neutron intensity. Soil samples were taken at three radial rings, extending outwards from the CRP. The radial rings were situated at distances of 25 m, 100 m and 200 m from the CRP (Figure 3). At each of the three rings, eight points were taken at an equal distance along the circumference of the ring. At each sample point, soil samples were taken at different depths.
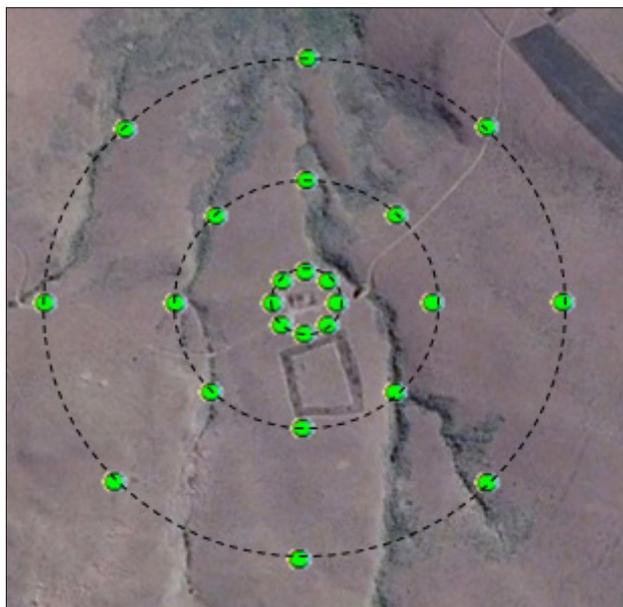
Figure 3: Calibration sampling points along three radial rings extending outwards from the cosmic ray probe (at 25 m, 100 m and 200 m). At each sample point, soil samples were taken at different depths.

Four calibrations were carried out over a period of 8 months. The calibration dates were 9 July 2014 (winter), 28 August 2014 (spring), 2 December 2014 (summer) and 22 January 2015 (late summer).

The gravimetric method was used to obtain soil samples within the CRP footprint, to determine the area-average soil moisture for the calibration of the CRP. The gravimetric soil moisture content is expressed, by

weight, as the ratio of the mass of water present to the dry weight of the soil sample (g/g). The CRP, however, measures the volumetric soil moisture content, which is expressed as the ratio of volume of water to the total volume of the soil sample (cm^3/cm^3). The calibration of the CRP requires a representative bulk density value to convert the gravimetric soil moisture content to volumetric water content. The bulk density determination was carried out by obtaining undisturbed soil cores at various locations and depths within the catchment. The average measured bulk density was 0.593 g/cm^3. This low bulk density value can be attributed to the soil cover, organic matter content, soil structure, porosity and the lack of compaction, as the catchment is situated in an undisturbed area.

Calibration

The final calibration was carried out once all four calibrations had been completed. The aim of the calibration was to determine the average N_o value (site-specific calibration parameter), which is the theoretical neutron intensity (counting rate) in air above dry soil (no moisture). Data from the CRP are sent hourly, via satellite link, to the COSMOS server: http://cosmos.hwr.arizona.edu/Probes/probemap.php. The calibration procedure followed that of Franz et al.[21,28]

The first step in the calibration procedure was to correct the neutron counts. This correction involved determining the neutron correction factors, using the following equations.[28]

$$N = \frac{N' * CP * CWV}{CI * CS},$$
Equation 1

where N is the corrected neutron counts per hour, N' is the raw moderated neutron counts, CP is the pressure correction factor, CWV is the water vapour correction factor, CI is the high-energy intensity correction factor, and CS is the scaling factor for geomagnetic latitude.[28]

$$CP = \exp\left(\frac{P_o - P}{L}\right),$$
Equation 2

where L is the mass attenuation length for high-energy neutrons (g/cm^2), P is the atmospheric pressure (mb) at a specific site and P_0 is the reference atmospheric pressure (mb).[9]

$$CWV = 1 + 0.0054 (P_{vo} - P_{vo}^{ref}),$$
Equation 3

where P_{vo} is the absolute humidity of the air (g/m^3) and P_{vo}^{ref} is the reference absolute humidity of the air (g/m^3).

$$CI = \frac{I_m}{I_{ref}},$$
Equation 4

where I_m is the selected neutron monitoring count rate at any particular point in time and I_{ref} is the reference count rate for the same neutron monitor from an arbitrary fixed point in time. The neutron flux data was obtained through the neutron monitoring database (www.nmdb.eu), which provides real-time data from a global network of monitoring stations.

$$CS = f(x,y,z,t),$$
Equation 5

where x, y, z is location and elevation, and t is time.

The following calibration function was then used to determine the N_o value for each calibration[21]:

$$(\theta_p + \theta_{LW} + \theta_{SOC}) = \frac{0.0808}{\frac{N}{N_o} - 0.372} - 0.115$$
Equation 6

Rearranging the calibration function to determine N_0:

$$N_o = \left(\frac{N}{\frac{0.0808}{(\theta_p + \theta_{LW} + \theta_{SOC}) + 0.115} + 0.372}\right)$$
Equation 7

Rearranging the calibration function to determine VWC:

$$VWC = \left(\frac{0.0808}{\left(\frac{N}{N_0}\right) - 0.372} - 0.115 - (\theta_{LW} + \theta_{SOC}) \right) P_{bd}, \qquad \text{Equation 8}$$

where θ_p is the gravimetric water content (g/g), θ_{LW} is lattice water content (g/g), θ_{soc} is soil organic carbon water content (g/g), P_{bd} is dry soil bulk density (g/cm³), N is the corrected neutron counts per hour, and N_0 is an instrument-specific calibrated parameter. We note that soil moisture is often expressed in units of volume per cent ($VWC(\%) = \theta p \times P_{bd} \times 100$).

θ_{soc} was not determined, but was given a value of 0.01 g/g based on published values.[8] θ_{lw} was determined to be 0.154 g/g. A 50-g representative soil sample was sent to Activation Laboratories in Canada for θ_w determination by combustion at 1000 °C.

There is a need to correct for biomass.[21]

$$(\theta_p + \theta_{LW} + \theta_{SOC}) = \frac{0.0808}{\frac{N}{N_0(BWE)} - 0.372} - 0.115, \qquad \text{Equation 9}$$

where BWE is the biomass water equivalent (mm). The biomass calculation is done for vegetation types whose biomass changes with their growing stage. Because the vegetation of Catchment VI is short grassland, the biomass was small[29] and the change in biomass in this context was therefore insignificant and a biomass correction was not required.

The neutron count (N) for each calibration was determined as the average neutron count during which the soil samples for that calibration were obtained. These counts were used to determine the N_0 value for each calibration, using the rearranged calibration Equation 7 (Table 1).

The average N_0 value for the calibrations was calculated to be 3249.324. This calculated N_0 value was used in the rearranged calibration function equation (Equation 8) to determine the hourly VWC (Figure 4). The hourly CRP data (grey line) were converted into daily (black line) average to smooth the data and for comparison with other daily-derived soil moisture products. The results show that soil moisture was higher in the summer and lower in the winter (Figure 4), as expected in the summer rainfall region of South Africa.

The effective measurement depth of the CRP over the 1-year period between March 2014 and March 2015 (Figure 5), showed that the effective measurement depth ranged from 0.084 m to 0.148 m, with an average effective measurement depth of 0.117 m.

Creating a soil moisture network

It was essential to create an in-situ soil moisture network within the CRP measurement area, to obtain data that could be used to validate the calibrated CRP estimates. The in-situ soil moisture network consisted of three types of measurement instruments, which were installed on 9 July 2014: one soil pit with TDR probes inserted horizontally at depths of 0.05, 0.10, 0.15, 0.20 and 0.30 m; five wireless TDR sensors (0.12 m) inserted vertically into the soil surface; and eight Echo probes inserted

horizontally at a depth of 0.10 m. The five wireless TDR probes and eight Echo probes were placed within the CRP's footprint area. All three sets of data were merged by weighting each point within the catchment equally, to create a representative in-situ soil moisture estimate. For consistency, only the 0.10 m and 0.15 m TDR pit data were used, which coincide with the two other in-situ measurement method depths and the effective measurement depth of the CRP. The CRP has a varying effective depth, which could be validated against the TDR pit, which has sensors at different depths; however, the aim was to test the suitability of the CRP to provide spatial estimates of soil moisture, which required a spatial validation data set and not a single point measurement.

Results and discussion

Soil moisture varies both spatially and temporally and fluxes in soil moisture content occur over short time periods and distances. The key input to the soil moisture content in Cathedral Peak Catchment VI is rainfall. Therefore, rainfall data are necessary to support and explain changes in soil moisture. Rainfall data from a rain gauge situated within Catchment VI were obtained for the 1-year period from March 2014 to March 2015 (Figure 5). The rainfall distribution showed that most of the rainfall (73%) occurred in the summer months.

Validating the CRP

A time series of the CRP and in-situ soil moisture data (Figure 6) showed the dependency of the soil moisture fluctuations on rainfall. There were smaller fluctuations in winter as a result of lower rainfall than in summer, during which the fluctuations in soil moisture were greater, because of more rainfall events. The CRP estimates correlated better in the wetter periods (slope=0.976) when the soil moisture values were above 30%, compared to the drier periods (slope=0.691). The CRP estimated soil moisture on average 3.45 VWC (%) higher throughout the validation period, when compared to the in-situ data set. This result is ascribed to differences in the measurement depths between the CRP and the in-situ instruments. Overall, the CRP-derived soil moisture followed the general trend of the in-situ soil moisture variations.

A graph of in-situ (x-axis) versus the CRP (y-axis) soil moisture estimates revealed that the majority of the points were above the 1:1 line (dotted) (Figure 7). There were no extreme outliers. The positive y-intercept of 12.024 indicated an over-estimation by the CRP of the lower values (when the soil was drier). There was a good agreement between both data sets, shown by the high R^2 (0.845).

A graph of the residuals (differences between the in-situ and CRP data sets) against time was plotted (Figure 8). The ΔVWC (%) was the difference between the in-situ (independent) and CRP (dependent) data sets. This graph is plotted to illustrate how the difference in variables change over time. Most of the residuals are negative, which indicates that the CRP over-estimates soil moisture when compared to the in-situ data set. In the drier periods, the residuals range from -0.52 to -11.54 VWC (%) (with an average of -6.25 VWC (%)), which indicates that the CRP over-estimates soil moisture throughout these periods. In the wet periods, the residuals vary from 6.16 to -7.08 VWC (%) (with an average of -0.90 VWC (%)). The CRP performed better in the wetter periods, when compared to the in-situ measurements, as the average absolute residual value is less than that of the corresponding drier periods.

Table 1: Calibrations, dates, gravimetric soil moisture, bulk density, neutron counts and calculated N_0 values

Calibration (date)	Moisture status	Gravimetric water content (g/g)	Bulk density (g/cm³)	Volumetric water content (cm³/cm³)	Neutron count (count/h)	N_0
1 (09/07/2014)	Dry	0.490	0.593	0.291	1731.684	3250.573
2 (28/08/2014)	Dry	0.438	0.593	0.259	1761.408	3242.507
3 (02/12/2014)	Wet	0.647	0.593	0.384	1652.600	3255.973
4 (22/01/2015)	Wet	0.741	0.593	0.439	1611.059	3248.243
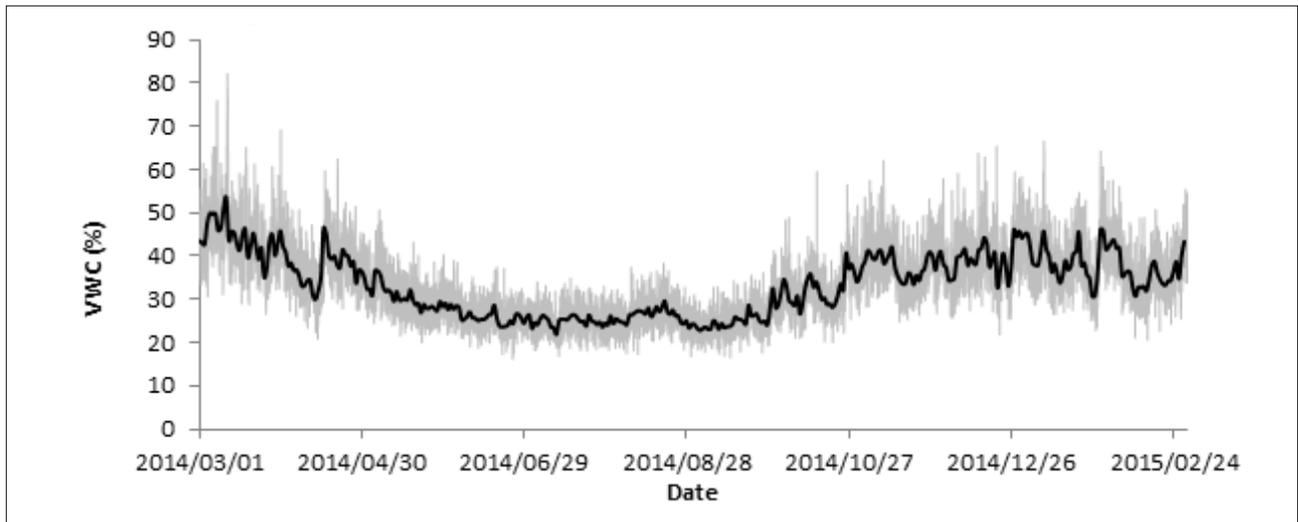
VWC, volumetric water content

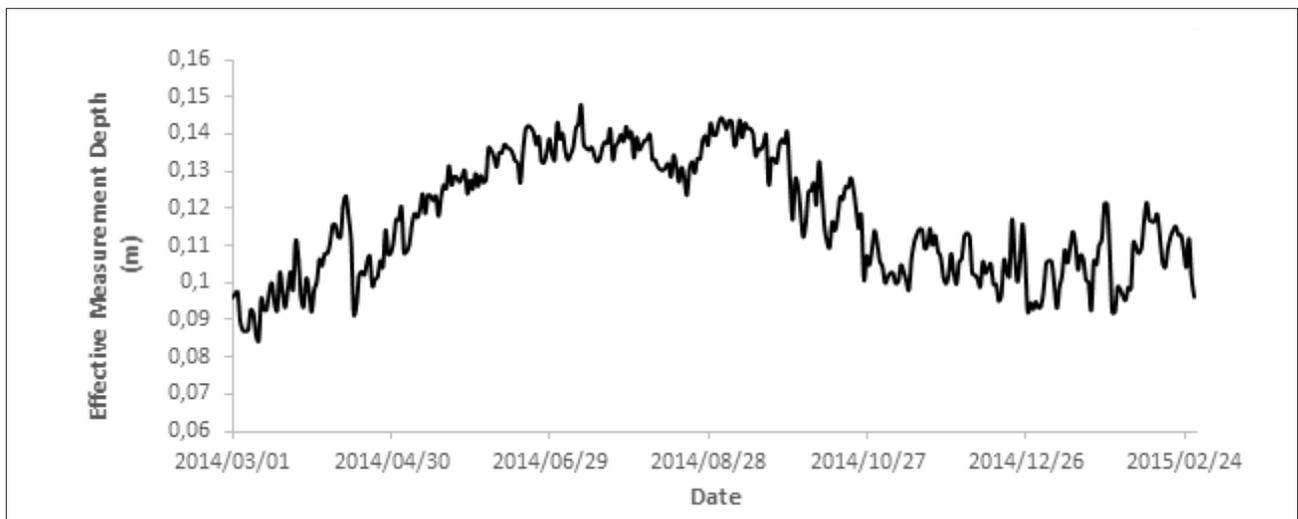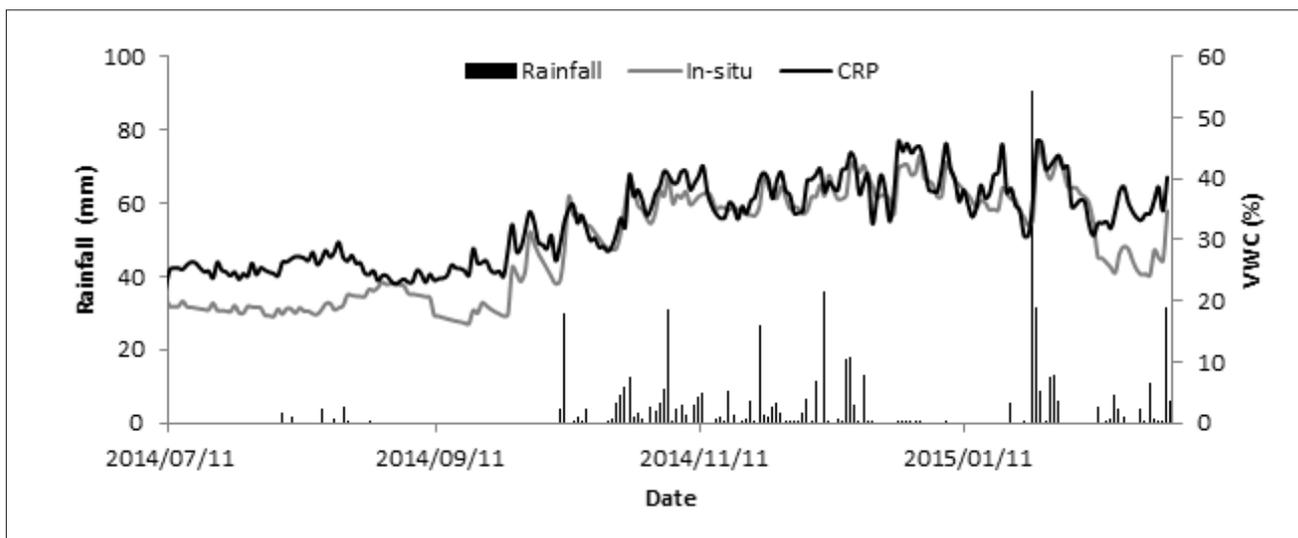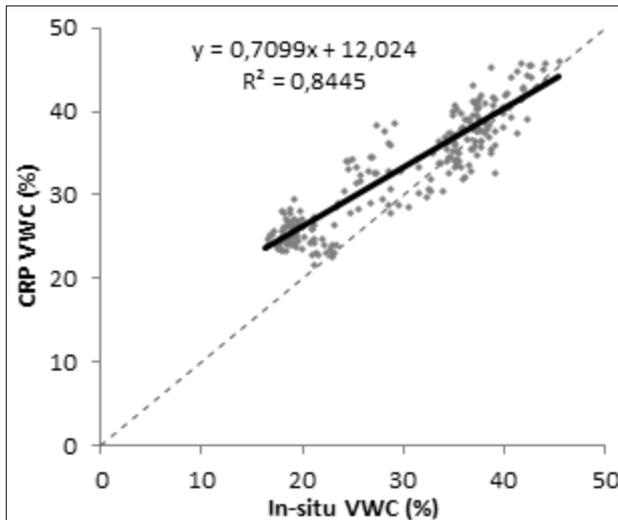**Figure 4:** Hourly (grey line) and daily (black line) soil moisture estimates using the calibrated cosmic ray probe.

Figure 5: The cosmic ray probe's effective measurement depth (m).

VWC, volumetric water content

Figure 6: Daily soil moisture estimates for Catchment VI using in-situ and cosmic ray probe (CRP) methods plotted against rainfall.

VWC, volumetric water content

Figure 7:   Scatterplot of in-situ versus cosmic ray probe (CRP) soil moisture estimates for the period 11 July 2014 to 1 March 2015. The dotted line indicates the 1:1 line.

A paired *t*-test (Table 2) revealed that the mean of the CRP was 3.5% higher than the in-situ data (negative *t*-statistic). The variance for the in-situ soil moisture was higher than the CRP by 30 and suggests the higher variation of point measurements when compared to the large area average of the CRP. The absolute value of the *t*-statistic (4.827) was greater than the critical two-tail value (1.965) and the *p*-value (1.919x10^{-6}) was less than the alpha value (0.05) showing that there was a significant difference between the methods.

The differences between the in-situ and the CRP soil moisture estimates can be explained by the following. The in-situ data set was not the actual average of the catchment, but the average of the sensors in the catchment. If the sensors were placed in different locations in the catchment, the in-situ data set could be different.

The area covered by the in-situ sensors was smaller than that covered by the CRP. The in-situ soil moisture sensors were a maximum of 200 m away from the CRP; however, the CRP has a measurement footprint that exceeds 200 m in radius.

Table 2:   Results of *t*-test comparing in-situ and cosmic ray probe (CRP) soil moisture estimates

	In-situ	CRP
Mean	29.557	33.005
Variance	74.464	44.433
Number of observations	233	233
Hypothesised mean difference	0	
d.f.	436	
t-statistic	-4.827	
p(T<=t) two-tail	1.919x10^{-6}	
t:Critical two-tail	1.965	

The CRP was not measuring soil moisture at a constant depth throughout the validation period, as its measurement depth depends on the soil moisture content. When the soil was dry, the CRP probe was measuring at a deeper depth (0.14 m) than when the soil was wet 0.10 m). The representative in-situ data set was measuring constantly at an average depth of 0.12 m. Therefore, the dry periods did not correlate as well as the wet periods, as the measurement depths were different.

Conclusions

Understanding the spatial and temporal variability of soil moisture at different scales is of great importance in many land surface disciplines, such as hydrology. Soil moisture is a key hydrological variable, as it impacts the water and energy balance at the land surface–atmosphere interface and is the main water source for natural vegetation and agriculture. It is difficult to quantify and assess the soil moisture content at an intermediate scale, because of the heterogeneity in soil and land-cover properties, climate drivers and topography.

The CRP, once properly calibrated, is suitable for providing spatial estimates of soil moisture, as the measurements correlated well with the representative in-situ soil moisture data set. The CRP calibration procedure described here is adequate for predicting soil moisture at a scale of around 18 ha; however, potential errors can be introduced throughout the procedure, which range from selecting the sample points
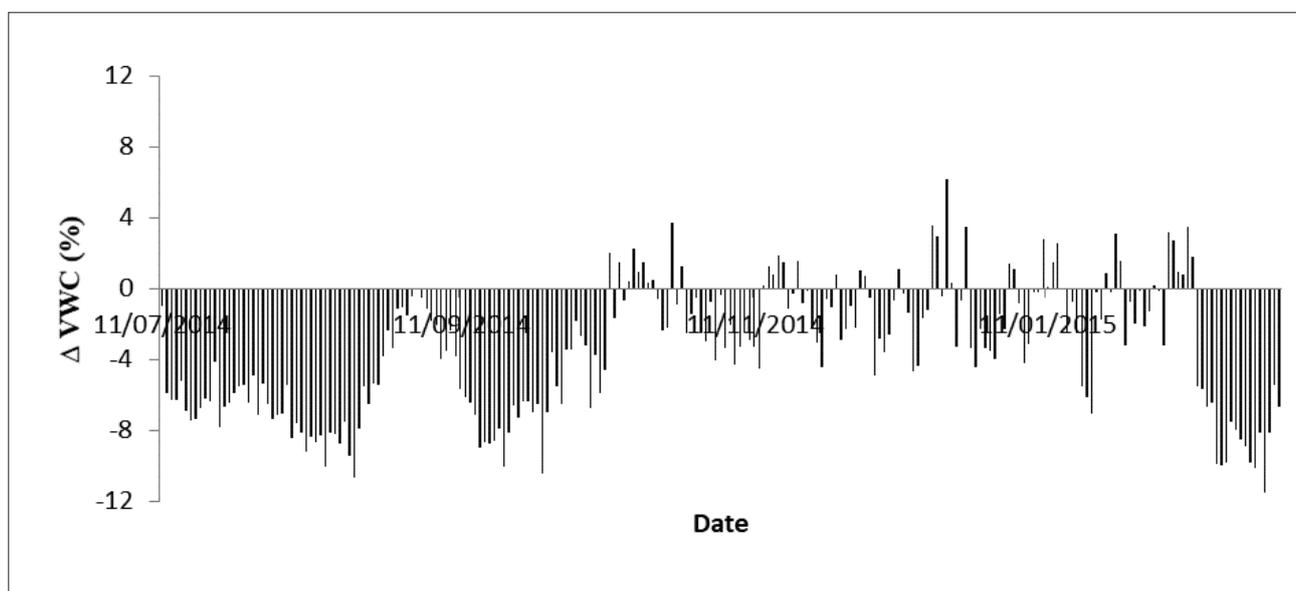
VWC, volumetric water content

Figure 8:   Residual graph of in-situ soil moisture estimates against those made using the cosmic ray probe.

and determining a representative bulk density, to determining the average neutron count (N_0) value. Therefore, proper procedures must be adhered to in order to minimise potential errors. One calibration is sufficient for soils with sparse vegetation, but multiple calibrations obtained at different soil moisture contents provide a more accurate calibration.[30]

The validation of the CRP with a representative catchment soil moisture data set, from the in-situ soil moisture network, showed that the CRP is suitable for providing continuous spatial soil moisture estimates.

Recommendations

The following recommendations can be used to address the main limitations that were experienced in this research study. These recommendations will provide assistance for future research studies.

The calibration of the CRP is time and labour intensive, as numerous gravimetric soil samples within the measurement footprint at several depths are required. The use of a TDR Hydro-Sense probe, which obtains instantaneous measurements of volumetric soil moisture when inserted into the soil, can potentially be used to obtain the necessary soil moisture measurements for the calibration. Thus, it would greatly reduce the time and labour required, as well as eliminate the need to determine the bulk density which is needed to convert gravimetric soil moisture to volumetric soil moisture.

The validation of the CRP should ideally be done with in-situ soil moisture sensors at a variety of depths. Ideally, several TDR pits in the CRP measurement volume could be used. This approach would improve the validation, as the CRP does not measure at a constant depth. However, this approach would be very capital, time and labour intensive, as a dense network of TDR pits would be required.

The number of calibrations carried out against the accuracy of the calibrated CRP estimates needs to be considered. In theory, more calibrations carried out would result in more reliable estimates of soil moisture. In practice, each calibration requires time, labour and capital to conduct. Therefore, although we conducted four calibrations, reliable estimates could have been obtained with two calibrations, one each in opposing seasons.

Authors' contributions

T.V.: Setting up and calibrating the cosmic ray probe, writing the manuscript, editing the manuscript. C.E.: Supervision, setting up the cosmic ray probe, planning the logistics of the fieldwork, editing the manuscript. M.M.: Supervision, setting up the cosmic ray probe, calibrating procedures, editing the manuscript. T.F.: Cosmic ray probe teaching, calibration training, editing the manuscript.

References

1. Fang B, Lakshmi V. Soil moisture at watershed scale: Remote sensing techniques J Hydrol. 2014;516:258–272. http://dx.doi.org/10.1016/j.jhydrol.2013.12.008

2. Jackson TJ, Cosh MH, Bindlish R, Starks PJ, Bosch DD, Seyfried M, et al. Validation of advanced microwave scanning radiometer soil moisture products. IEEE Trans Geosci Remote Sens. 2010;48(12):4256–4272. http://dx.doi.org/10.1109/TGRS.2010.2051035

3. Albergel C, De Rosnay P, Gruhier C, Mu˜noz-Sabater J, Hasenauer S, Isaksen L, et al. Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. Remote Sens Environ. 2012;118:215–226. http://dx.doi.org/10.1175/JHM-D-11-0107.1

4. Mecklenburg S, Drusch M, Kerr Y, Font J, Roger O. ESA's soil moisture and ocean salinity mission: Mission performance and operations. IEEE Trans Geosci Remote Sens. 2012;50(5):1354–1366. http://dx.doi.org/10.1109/TGRS.2012.2187666

5. Ni-Meister W. Soil moisture initialization for climate prediction: Characterization of model and observation errors. J Geophys Res. 2005;110:11–29. http://dx.doi.org/10.1029/2004JD005745

6. Gruber A, Dorigo WA, Zwieback S, Xaver A, Wagner W. Characterizing coarse-scale representativeness of in situ soil moisture measurements from the International Soil Moisture Network. Vadose Zone J. 2013;12(2), Art. #170, 16 pages. http://dx.doi.org/10.2136/vzj2012.0170

7. Kohli M, Shron M, Zreda M, Schmidt U, Dietrich P, Zacharias S. Footprint characteristics revised for field-scale soil moisture monitoring with cosmic-ray neutrons. Water Resour Res. 2015;10:5772–5790. http://dx.doi.org/10.1002/2015WR017169

8. Franz T, Zreda M, Ferre T, Rosolem R. An assessment of the effect of horizontal soil moisture heterogeneity on the area-average measurement of cosmic-ray neutrons. Water Resour Res. 2013;49:6450–6458. http://dx.doi.org/10.1002/wrcr.20530

9. Zreda M, Shuttleworth WJ, Zeng X, Zweck C, Desilets D, Franz T, et al. COSMOS: The COsmic-ray Soil Moisture Observing System. Hydrol Earth Syst Sci. 2012;16:4079–4099. http://dx.doi.org/10.5194/hess-16-4079-2012

10. Desilets D, Zreda M, Ferre T. Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. Water Resour Res. 2010;46, W11505, 7 pages. http://dx.doi.org/10.1029/2009WR008726

11. Kim S, Liu YY, Johnson FM, Parinussa RM, Sharma A. A global comparison of alternate AMSR2 soil moisture products: Why do they differ? Remote Sens Environ. 2015;161:43–62. http://dx.doi.org/10.1016/j.rse.2015.02.002

12. Walker JP, Willgoose GR, Kalma JD. In situ measurement of soil moisture: A comparison of techniques. J Hydrol. 2004;2963:85–99. http://dx.doi.org/10.1016/j.jhydrol.2004.01.008

13. Qin J, Yang K, Chen Y, Zhao L, Han M. Spatial up-scaling of in-situ soil moisture measurements based on MODIS-derived apparent thermal inertia. Remote Sens Environ. 2013;138:1–9. http://dx.doi.org/10.1016/j.rse

14. Ochsner TE, Cosh MH, Cuenca RH, Dorigo WA, Draper CS, Hagimoto Y, et al. State of the art in large-scale soil moisture monitoring. Soil Sci Soc Am J. 2013;77(6):1888–1919. http://dx.doi.org/10.2136/sssaj2013.03.0093

15. Jiao Q, Zhu Z, Du F. Theory and application of measuring mesoscale soil moisture by cosmic-ray fast neutron probe. 2014 IOP Conf Ser Earth Environ Sci. 2014;17, Art. #012147, 8 pages. http://dx.doi.org/10.1088/1755-1315/17/1/012147

16. Zreda M, Desilets D, Ferre TPA, Scott RL. Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. Geophys Res Lett. 2008;35(21), L21402, 5 pages. http://dx.doi.org/10.1029/2008GL035655

17. Desilets D, Zreda M. Footprint diameter for a cosmic-ray soil moisture probe: Theory and Monte Carlo simulations. Water Resour Res. 2013;49:3566–3575. http://dx.doi.org/10.1002/wrcr.20187

18. Dutta R, D'este C. Virtual calibration of cosmic ray sensor: Using supervised ensemble machine learning. Int J Adv Comput Sci Applic. 2013;4(8):104–110. http://dx.doi.org/10.14569/IJACSA.2013.040815

19. Franz T, Zreda M, Rosolem R, Ferre T. A universal calibration function for determination of soil moisture with cosmic-ray neutrons. Hydrol Earth Syst Sci. 2013;17:453–460. http://dx.doi.org/10.5194/hess-17-453-2013

20. Franz T, Zreda M, Rosolem R, Ferre T. Field validation of a cosmic-ray neutron sensor using a distributed sensor network. Vadose Zone J. 2012;11(4), Art. #0046, 10 pages. http://dx.doi.org/10.2136/vzj2012.0046

21. Franz T, Wang T, Avery W, Finkenbiner C, Brocca L. Spatiotemporal characterization of soil moisture fields in agricultural areas using cosmic-ray neutron probes and data fusion. EGU General Assembly 2015; 2015 April 12–17; Vienna, Austria.

22. Gush M, Scott DF, Jewitt GPW, Schulze RE, Lumsden T, Hallowes L, et al. Estimation of streamflow reduction resulting from commercial afforestation in South Africa. WRC report TT 173/02. Pretoria: Water Research Commission; 2002.

23. Everson CS, Molefe GL, Everson TM. Monitoring and modelling components of the water balance in a grassland catchment in the summer rainfall area of South Africa. WRC report no: 493/1/98. Pretoria: Water Research Commission; 1998.

24. Kuenene BT, Van Huyssteen CW, Hensley M. Soil water saturation in the Cathedral Peak VI catchment, KwaZulu-Natal. Paper presented at: SAGA Biennial Technical Meeting and Exhibition; 2009 September 16–18; Swaziland. p. 565–570. Available from: http://www.earthdoc.org/publication/publicationdetails/?publication=51554

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

26. Scott DF, Prinsloo FW, Moses G, Mehlomakulu M, Simmers ADA. A re-analysis of the South African catchment afforestation experimental data. WRC report no. 810/1/00.Pretoria: Water Research Commission; 2000.

27. Jarmain C, Govender M, Everson CS. Improving the basis for predicting total evaporation from natural veld types in South Africa. WRC report no. 1219/1/04. Pretoria: Water Research Commission; 2004.

28. Franz T. Stationary probe data calculations. Lincoln, NE: Franz Hydrogeophysics Lab Group, University of Lincoln-Nebraska; 2014. p. 1–4.

29. Everson CS, Everson TM. The long-term effects of fire regime on primary production of montane grasslands in South Africa. Afr J Range For Sci. 2016;33(1):33–41. http://dx.doi.org/10.2989/10220119.2015.1124922

30. Zreda M. Land-surface hydrology with cosmic-ray neutrons: Principles and applications. J Jpn Soc Soil Phys. 2016;132:25–30.

Community composition and functions of endophytic bacteria of Bt maize

AUTHORS:
Asnath R. Mashiane[1,2]
Rasheed A. Adeleke[1,2] (iD)
Cornelius C. Bezuidenhout[2] (iD)
George J. Chirima[3]

AFFILIATIONS:
[1]Microbiology and Environmental Biotechnology Research Group, Agricultural Research Council, Pretoria, South Africa

[2]Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa

[3]Center for African Ecology, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Rasheed A. Adeleke

EMAIL:
adeleker@arc.agric.za

We investigated the potential effects of genetic modification of Bt maize on the community composition and functions of bacterial endophytes associated with transgenic maize (Bt MON 810) in comparison with its isogenic parental line at two developmental stages. Bacterial isolates were obtained from transgenic (Bt) and non-transgenic (non-Bt) maize at 50- and 90-day-old developmental stages. Isolated bacterial endophytes were screened for their capabilities in phosphate solubilisation, nitrogen fixation, production of antifungal metabolites and production of indole acetic acid. After molecular identification, 60 isolates were obtained and clustered into 19 and 18 operational taxonomic units from 50- and 90-day-old maize, respectively. The isolates belonged to the genera *Bacillus, Pantoea, Serratia, Yersinia, Enterobacter, Pseudomonas, Acinetobacter* and *Stenotrophomonas.* Functional attributes and diversity of the isolated endophytes at both developmental stages were not significantly different for both maize varieties. However, functional attributes were significantly affected by plant growth stage. Isolates from younger plants were more efficient producers of indole acetic acid, but exhibited little or no capabilities for nitrogen fixation, phosphate solubilisation and antifungal activity in both maize genotypes. Based on these outcomes, Bt modification in maize does not seem to affect the community composition or functional attributes of bacterial endophytes.

Significance:
- Bt modification in maize does not affect the ecological guild or functional attributes of cultivable bacterial endophytes.

Introduction

Maize is one of the most important crops in the world, and is consumed as a staple food as well as animal feed in both developed and developing countries.[1] Globally, maize production is threatened by a number of factors, especially the outbreak of pests such as stalk borer insects.[1] The tissue damage caused by the stem borers allows pathogenic microbes to colonise open areas leading to leaf and cob rots as well as mycotoxin accumulation.[1] The introduction of genetically modified Bt maize, which contains a gene from the soil bacterium *Bacillus thuringiensis* (Bt), has given rise to the production of highly resistant varieties that are toxic to major insect pests of the orders Lepidoptera and Coleoptera.[2,3] On the other hand, the overall impact of such modification in maize cannot be predicted.[4] For example, such genetic modification could adversely affect non-target organisms, especially the composition and functions of the natural microbial community of maize plants, which includes endophytic, ectophytic and rhizospheric microbes.[5,6]

Endophytes are bacteria, archaea or fungi that live inter- or extracellularly without causing any symptoms of disease to the host plant.[7] They are present in virtually all plants studied to date.[8-10] Their diversity varies from plant to plant as a result of many factors that may include, among others, plant species, genotype, tissue, growth stage and differences in colonisation pathway.[6,11] Bacterial endophytes have the capacity to promote plant growth because of their participation in nutrient cycling – phosphorus solubilisation and nitrogen fixation, hormone production (indole acetic acid) as well as suppression of pathogens (biocontrol agents).[12-16] Gaining a more detailed understanding of the microbial community in genetically modified maize is imperative for evaluating the resilience of varieties as well as the potential implication of the modification on ecosystem functioning. Few studies have delved into this important aspect of endophyte biology.[3,17-19] In this study, we hypothesised that genetic modification of maize plants (such as Bt maize) might influence the diversity and functional attributes of endophytic bacteria associated with the plant. This led us to investigate the potential impacts of genetic modification of maize on the community composition and functions of cultivable bacterial endophytes from different plant parts that included leaves, stems, cobs and tassels/husks at two different plant growth stages. Such an approach will improve our knowledge in this field, particularly that relating to: (1) the relationship between maize phyllosphere and community composition of bacterial endophytes and (2) potential impacts of genetic modification of Bt maize on the community composition and functional roles of the bacteria endophytes at different plant growth stages.

Methods and materials

Sample collection

Samples were collected at the Agricultural Research Council – Grain Crop Institute in Potchefstroom in the North West Province of South Africa (26°43'39.2'S, 27°04'48.8'E). The study was conducted using a single variety of transgenic Bt maize (MON 810) and its isogenic parental line (non-Bt), which served as a reference or control. The Bt and non-Bt maize fields were cultivated using disc ploughing. The maize was hand planted in 1.2-m rows. The soil type for both plant genotypes was the same as the plots were next to each other. Plants were irrigated weekly as needed. Field temperatures varied between 24 °C and 32 °C. Plants were sampled 'destructively' at two developmental stages: pre-flowering (50 days after emergence) and post-flowering (90 days after emergence). A total of 20 maize plants (10 Bt and 10 non-Bt) were sampled and analysed at each developmental stage. The

plant parts were severed with sterile scissors, placed in a plastic bag and transported to the laboratory. Collected explants of healthy leaves, stem, tassels and seeds were stored separately for analysis.

Isolation of endophytes

Stored samples of healthy leaves, stem, tassels and seeds were cut into pieces of approximately 20 mm^2 segments (explants). The explants were surface sterilised using a three-step approach that involved immersion in 70% (v/v) ethanol for 60 s, followed by rinsing with distilled water and subsequent sterilisation in 3% (v/v) sodium hypochlorite for 60 s and finally in 70% (v/v) ethanol for 30 s. Samples were further washed in sterile distilled water three times, for 60 s each.[20] Nutrient agar (Merck (Pty) Ltd, Johannesburg, South Africa), tryptone soy agar (Merck), and nutrient broth media were used for the isolation of bacteria from the explants. The process involved the inoculation of explants obtained from different parts of the maize plants on the three different media. The explants were inoculated at the centre of the plates containing these three media. All plates were incubated at 27 °C in duplicate for 24 h. Sub-culturing was done until pure isolates were obtained. Pure isolates from the same plant parts but different replicates were grouped together in subsequent analyses.

Colony PCR and sequencing

An aliquot of pure single colony culture was transferred to a 1.5-mL microcentrifuge tube containing 200 μL sterile milliQ water and homogenised using a vortex (Labnet International, Edison, NJ, USA). DNA amplification of the isolates was done directly using the suspended cells in a colony polymerase chain reaction (PCR).[21,22] The PCR targeted the partial 16S rDNA gene of the bacterial isolates. The amplification was done using primer sets 341F (5' CCTACGGGAGGCACCAG3') and 907R (5' CCGTCAATTCCTTTGATTT3').[23] Primers were synthesised by Inqaba Biotech (South Africa). The 20 μL reaction mixture included 2X PCR master mix (0.05 U/μL Phusion Flash II DNA polymerase, 4 mM MgCl$_2$ and 0.4 mM dNTPs (Thermo scientific, USA), specific primers (100 pmole/μL) and distilled water. Reagents were mixed by brief centrifugation at 6500 rcf (Cencom I, Barcelona, Spain). The PCR programme involved an initial denaturation step of 98 °C for 60 s, 35 cycles of 98 °C for 30 s, 50 °C for 30 s and 72 °C for 60 s, with a final extension step of 72 °C for 5 min. The integrity of PCR amplicons was verified with the aid of gel electrophoresis on a 1% agarose gel using a 100-bp molecular marker (Biolabs). Amplicons ($\approx$600 bp) were further purified and sequenced bi-directionally with the same set of primers. Forward and reverse sequences were inspected, edited and assembled using Bio-Edit.[24] Sequence data obtained were submitted to Genbank and given accession numbers KT120070–KT120099 and KT459755–KT459782.

Taxonomic assignment and phylogenetic reconstruction

All the 16S rDNA gene sequences were assigned to bacterial taxa using SeqMatch on the Ribosomal Database Project website (http://rdp.cme.msu.edu/index.jsp).[25] This assignment was followed by multiple sequence alignments and clustering into operational taxonomic units (OTUs) using the mothur software.[26] The assignment was based on a 99% similarity level between OTUs. Matched sequences, one for each OTU, were later obtained from the National Center for Biotechnology Information (NCBI)'s Genbank using the accession numbers. These sequences alongside the OTU representatives were used to construct a library. All sequences were aligned using the multiple sequence alignment software MAFFT version 7.[27] Mega6 software was used to generate a neighbour-joining phylogenetic tree consisting of representative OTUs and their close relatives (matched sequences).[28]

Nitrogen fixation and indole acetic acid assay

Isolates were streaked on a Burk's nitrogen-free culture medium. The medium comprised 10 g glucose, 0.52 g K$_2$HPO$_4$, 0.41 g KH$_2$PO$_4$, 0.05 g Na$_2$SO$_4$, 0.2 g CaCl$_2$, 0.1 g MgSO$_4$·7H$_2$O, 0.005 g FeSO$_4$·7H$_2$O, 0.0025 g Na$_2$MoO$_4$·2H$_2$O, 15 g agar per litre.[29] Inoculated plates were incubated at 28 °C for 7 days to allow growth. Bacterial isolates showing visible growth on the plates were considered to be positive nitrogen fixers.[30]

For indole acetic acid (IAA) production, the cultures (in triplicate) were grown on tryptophan broth and incubated at 30 °C for 48 h on a shaker incubator at 180 rcf.[31] Bacterial cells were separated from the supernatant by centrifugation at 10 000 rcf for 10 min in a centrifuge (Biocen 22 R, Orto Alresa, Madrid, Spain). IAA was measured by mixing 1 mL broth with 2 mL Salkowsky's reagents (2% of 0.5 M FeCl$_3$ + 35% HClO$_4$), and the resulting suspension was incubated for 20 min at room temperature. Absorbance was measured at 530 nm using a spectrophotometer (V-1100D, Xinke Instruments Co., Ltd, Sichuan, China). The final amount of IAA was calculated using the standard of pure IAA (Sigma-Aldrich Ltd, Johannesburg, South Africa) prepared separately. IAA production was indicated by development of a pink colouration.

Phosphate solubilisation

The phosphate solubilisation ability of the bacterial isolates was assessed by plate assay using the Indian National Botanical Research Institute's phosphate growth medium.[32,33] The medium contained (g/L): MgSO$_4$.7H$_2$O–0.25; (NH$_4$)$_2$SO$_4$–0.10; MgCl$_2$.6H$_2$O–5.00; KCl–0.20; Ca$_3$(PO$_4$)$_2$–2.5; glucose-10 and agar-20. A clear halo around the bacterial colony was considered positive for phosphate solubilisation. Each of the isolates was inoculated in triplicate and grown at 30 °C for 8 days.

The negative control was without the inoculum.

Solubilisation efficiency = (diameter of halo)/ (diameter of colony) X 100

Antifungal activity assay

The maize pathogen used for the antifungal activity assay in this study was *Fusarium verticillioides* 10025, obtained from the Agricultural Research Council – Plant Protection Research Institute. The test was performed using potato dextrose agar (Merck (Pty) Ltd, South Africa) medium that had a 5-mm disc of fungi mycelia (*F. verticillioides* 10025) placed at the centre of the plate, with a single streak of bacterial culture 3 mm away from the fungal disc. The plates were grown for 5 days at 25 °C and the inhibition activity was evaluated by comparing the radius of the fungal growth treated with bacteria against the control. The inhibition estimation was calculated by:

$$\% \text{ Inhibition in radial growth} = \frac{r1 - r2}{r1} \text{ X } 100,$$

where r1 is the radial mycelia growth in the control and r2 is the radial mycelia growth in the treatment. The antifungal activity of each isolate was tested in triplicate.

Statistical analyses

Assessments of potential impacts of genetic modification of the BT maize on the type, composition and functions of endophytes prevalent at each growth stage were determined by multivariate analysis.[34-36] For this analysis, data for all plant parts were pooled per growth stage.

Multivariate analysis combines different measurements from the same sample and can recognise correlations and interactions between factors; it is therefore a good tool to understand external influences on species composition as well influences on functions simultaneously.[34] For this reason, many studies aiming to quantify the effects of transgenic plants on microbial-associated communities have used this instrument.[37]

Community composition as well as capabilities of the endophytes for solubilisation of phosphate, fixation of atmospheric nitrogen, production of antifungal metabolites, and production of IAA were compared between Bt and non-Bt maize genotypes for the pre- and post-flowering stages. Principal component analysis (PCA) was used to determine how the isolated endophytes could be related directly to each of the identified functional attributes. This approach not only allowed the detection of the variation in community composition but also showed their functional capabilities at both the pre-flowering stage (50 days) and post-flowering stage (90 days). The PCA was applied in MS Excel version 2013. Data were arranged such that the four variables – nitrogen fixation, % inhibition, % phosphate efficiency and IAA production – were mapped as arrows. In the resulting bi-plot, endophyte types were represented by points, circles, triangles etc., while function variables were represented

by arrows. Bi-plots represented an endophyte's position along an arrow representing a capability to perform that particular function, because arrows representing each function were plotted in the direction of their maximum change. Thus, long arrows indicated high capability for that function. Numbers (diversity) of isolated endophytes from Bt versus non-Bt maize did not differ if points occurred close together or in the same positions along the arrows. Endophyte species occurring in positions close to or beyond the tip of the specific arrow were strongly and positively correlated with that functional capability. A perpendicular from the arrow to an endophyte point indicated the position of that species in relation to values of that function. Those species at the opposite end of the arrow were less strongly affected.[38] Thus, in addition to presenting the variation in community composition, the bi-plot also accounted for variation in ability to solubilise phosphate, fix atmospheric nitrogen, and produce antifungal metabolites.[36,39] In this way, the PCA approach allowed a quick appraisal of how community composition and functions vary between the pre- and post-flowering stages and between the Bt and non-Bt maize varieties.[39]

Statistical tests were run on SPSS version 19.0 comparing these capabilities during the two periods – pre-flowering and post-flowering. For the capacity to fix nitrogen during pre-flowering versus post-flowering, a score of 1 was applied if nitrogen fixation was present and 0 if absent, and the analysis was conducted using chi-square tests. For solubilisation of phosphate, production of IAA, and antifungal activity at 50 days, independent *t*-tests were applied (Supplementary table 1). Because endophyte activity at 90 days was not independent of endophyte activity at 50 days, paired *t*-tests were used to compare these activities between the two treatments at 90 days (Supplementary table 2). To broadly assess functional efficiencies of the species between plant growth stages, a one-way repeated measures analysis of variance comparing the capabilities for specific function at each growth stage was applied.

Results

Bacterial isolates

The amplified DNA products were 600 bp in length. The sequences were clustered into OTUs at 99% similarity. The homology sequence and phylogenetic analyses of the 16S rDNA of the bacterial endophytes associated with stems, leaves, tassels and seeds of Bt and non-Bt maize indicated that they belong to eight genera: *Bacillus, Pantoea, Stenotrophomonas, Yersinia, Serratia, Pseudomonas, Enterobacter* and *Acinetobacter* (Figures 1 to 4; Table 1).

Table 1 indicates the number of isolates obtained from both Bt (transgenic) and non-Bt maize plants with their accession numbers and OTU representatives. The diversity of cultivable bacterial endophytes (Figures 3 and 4) in the Bt variety was not significantly different from that of the non-Bt maize at 50 days. The prevalence of species at 50 days in Bt maize was 48.3%, whilst in non-Bt maize was 51.7%. The relative prevalence of species in Bt maize increased as a function of growth of the plants from 48.3% to 66% at 90 days (Figures 1 and 2). The diversity of endophytes obtained in the present study were of the phyla Firmicutes (28%) and Proteobacteria (72%). From the endophytic bacterial species obtained from both maize genotypes, *Bacillus* was the most recurrent from the leaves (17%), stem (12%) and seeds (8%) while *Pantoea* was isolated more from the leaves (12%). The genus *Bacillus* was found in all the plant parts investigated while *Pantoea* and *Enterobacter* were found in the stems and leaves. *Stenotrophomonas* and *Serratia* were identified from the seeds and stems. Lastly, *Yersinia* was identified in the leaves and tassels while *Acinetobacter* was only obtained from the stems (Table 1).

Bacterial isolates and functions

Bacterial community composition between Bt and non-Bt maize was not different, suggesting that Bt modification may not have a negative impact on the bacterial endophyte populations. Specifically, there was no significant difference in terms of endophyte community composition between Bt and non-Bt for both plant growth stages, pre-flowering and

post-flowering (Figure 5). Lower capability for nitrogen fixation was strongly associated with the pre-flowering stage (i.e. young plants) for both Bt and non-Bt maize (Figure 6). However, all isolates from the post-flowering stage were able to fix nitrogen in both Bt and non-Bt maize (Figure 6). Antifungal activity and phosphate solubilisation capabilities of the isolates were positively associated with the post-flowering stage and weakly associated with the pre-flowering stage.

For the pre-flowering plants, on both non-Bt and Bt maize, the highest IAA production was particularly by BT4_50S (*Pantoea* species) and NBT5_50S (*Acinetobacter* species). However, there were three species with exceptions to the growth stage trend as indicated in Figure 5a. These endophytes exhibited high antifungal activity and phosphate efficiency.

Screening for potential atmospheric nitrogen fixers

There were no significant differences in nitrogen fixation between Bt and non-Bt maize at 50 days ($p=0.779$). The number of nitrogen fixers increased in both Bt and non-Bt varieties from 50 days to 90 days; the increase was significant in the Bt variety ($p=0.003$).

Indole acetic acid assay

Figure 7 shows the activity of IAA production at both 50 days and 90 days for Bt and non-Bt maize plants. There were no significant differences in IAA production between Bt and non-Bt maize at 50 days ($p=0.555$). However, there was a significant difference in IAA production between 50 and 90 days' developmental growth stages for both Bt and non-Bt maize varieties, with isolates obtained from younger plants exhibiting an elevated capacity for production of IAA ($p=0.008$).

Phosphate solubilisation

The distribution of phosphate solubilisation is presented in Figure 8. Phosphate solubilisation was not significantly different for Bt and non-Bt maize varieties compared laterally at both 50- and 90-day periods *($p=$ 0.582)* (compared independently)*.* However, when compared across growth stages, isolates obtained from older plants (post-flowering) were significantly more efficient in phosphate solubilisation ($p=0.0027$).

Antifungal activity

Figure 9 illustrates the bacterial endophytes with potential inhibition against the test fungal phytopathogen (*F. verticillioides*). There was no significant difference between Bt and non-Bt maize in terms of antifungal activity at 50 ($p=0.627$) and 90 days ($p=0.652$) when compared separately. However, when plant age was considered for both categories of Bt and non-Bt maize, antifungal activity was highest in older plants (i.e. post-flowering) ($p=0.001$).

Discussion

Transgenic *Bacillus thuringiensis* (Bt) maize is the most widely grown crop in the world.[40] The transgenic plants have functional genes inserted into their genome which are expressed in all the tissues and stages during plant growth. When such genetic modification exists in the plant tissues, non-target organisms such as endophytes, epiphytes and rhizospheric microbes could be at risk. Unfortunately, plants greatly depend on these microbes for health and growth as they play a crucial role in nutrient mineralisation, biological control, hormone production and resistance to stress.[5] Possible effects of the genetic modification (Bt maize) on the community composition and functional attributes of endophytes during two developmental plant growth stages in comparison to the non-Bt maize isogenic parental line were evaluated. Remarkably, there was no significant difference in the composition of the culturable bacterial community of the two maize genotypes. This result could be related to maize plants having the same physiological characteristics, thus hosting the same endophytic group. Similarly, another study done by Saxena and Stotzky[5] did not find any difference in bacterial communities between transgenic and non-transgenic maize. However, shifts in microbial density of Bt maize were observed during the reproductive stage (90 days old), when compared to non-Bt maize, which could be related to plant age and soil type.[11,41]
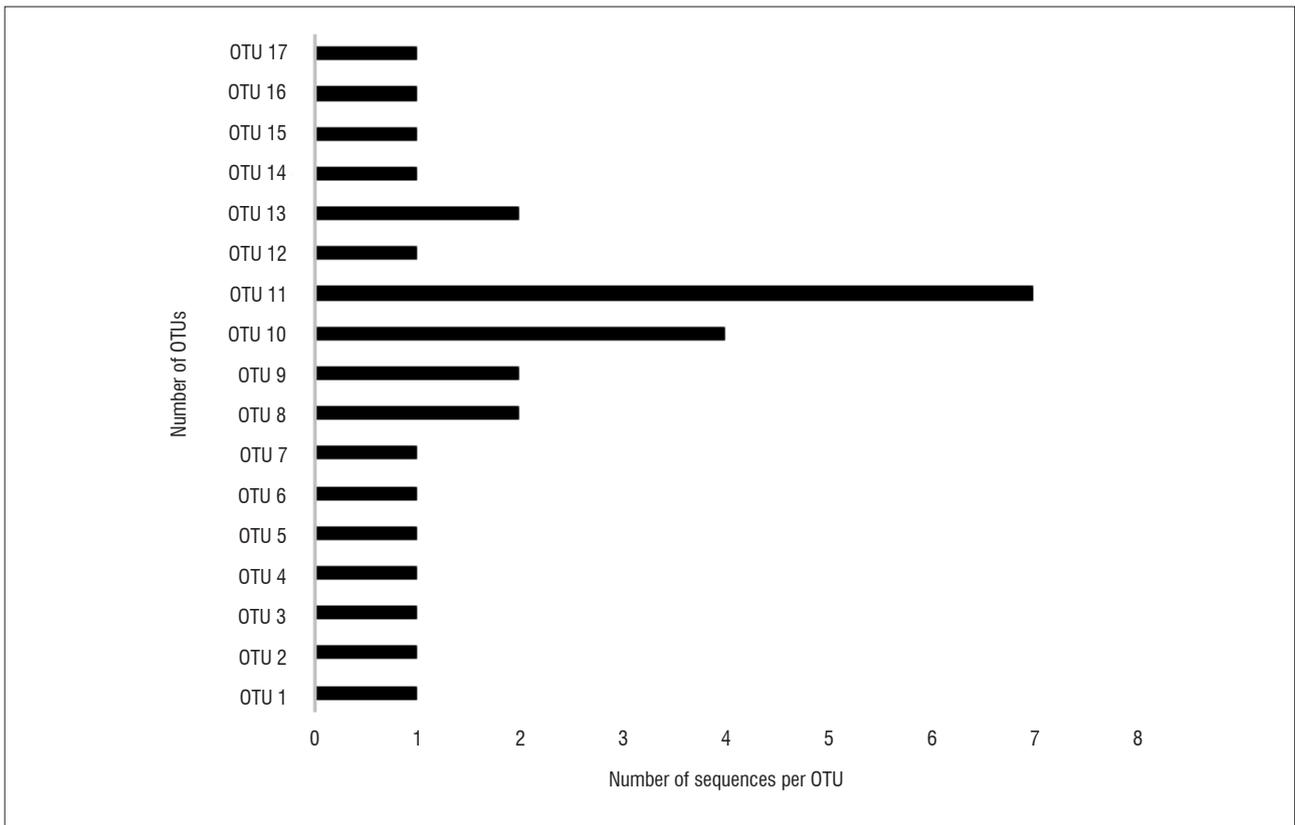
Figure 1: Operational taxonomic units (OTUs) of 16S bacterial sequences obtained from 50-day-old Bt and non-Bt maize plants.
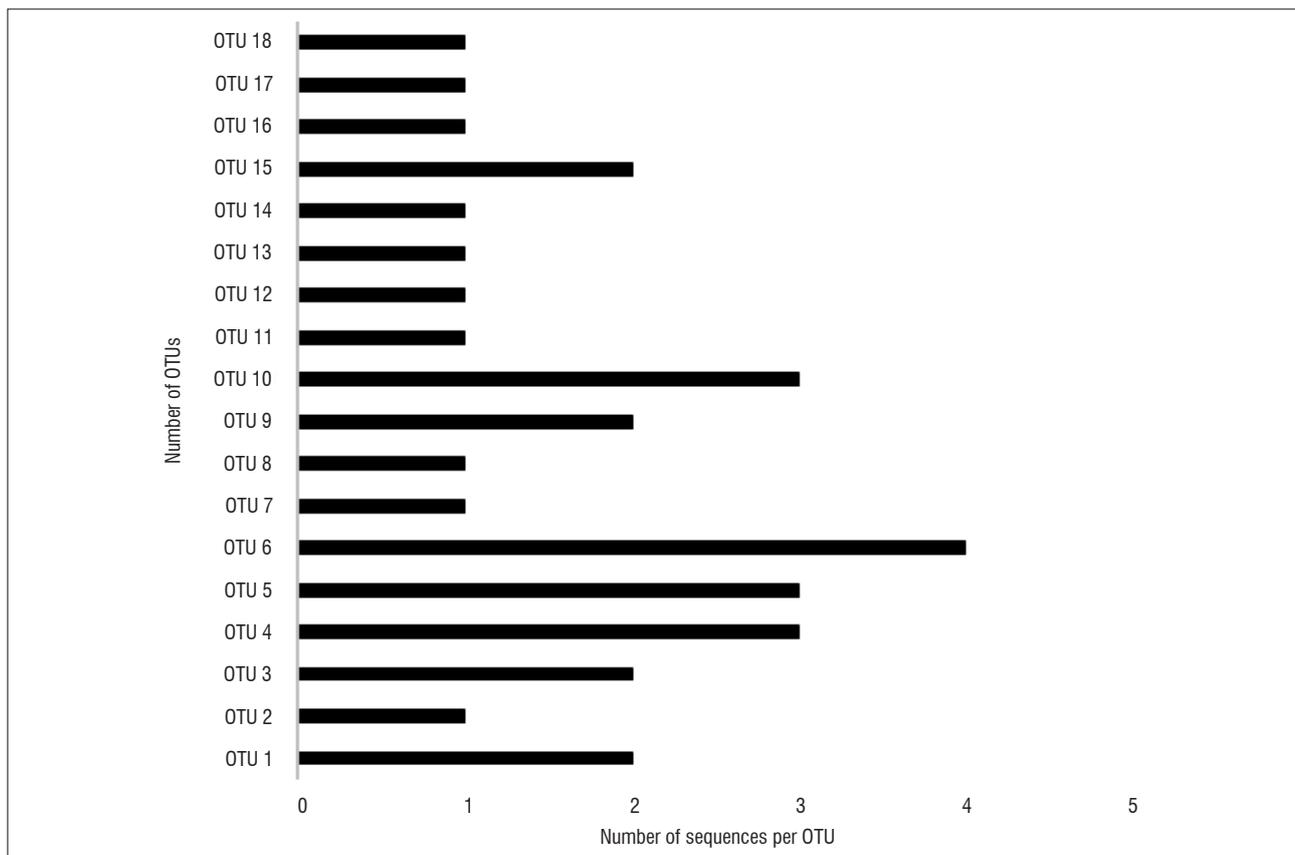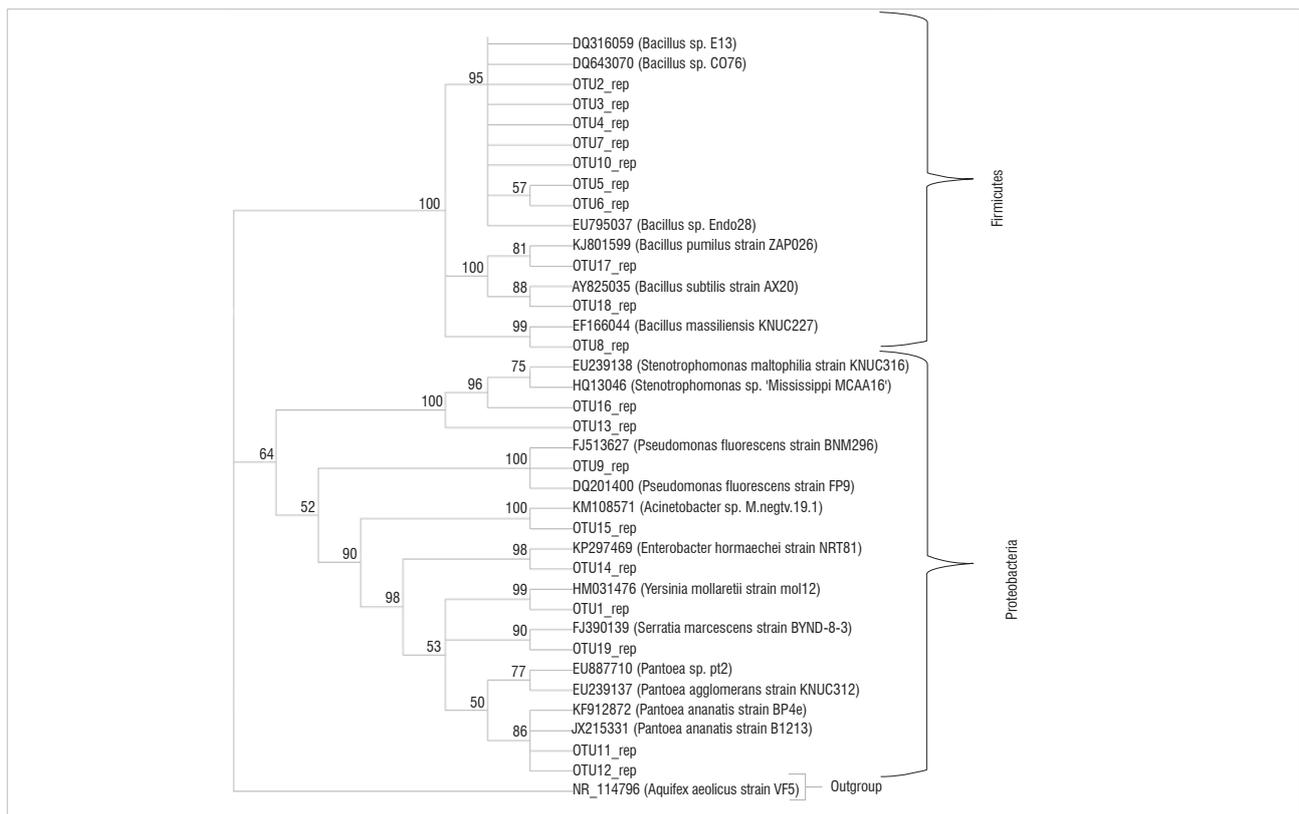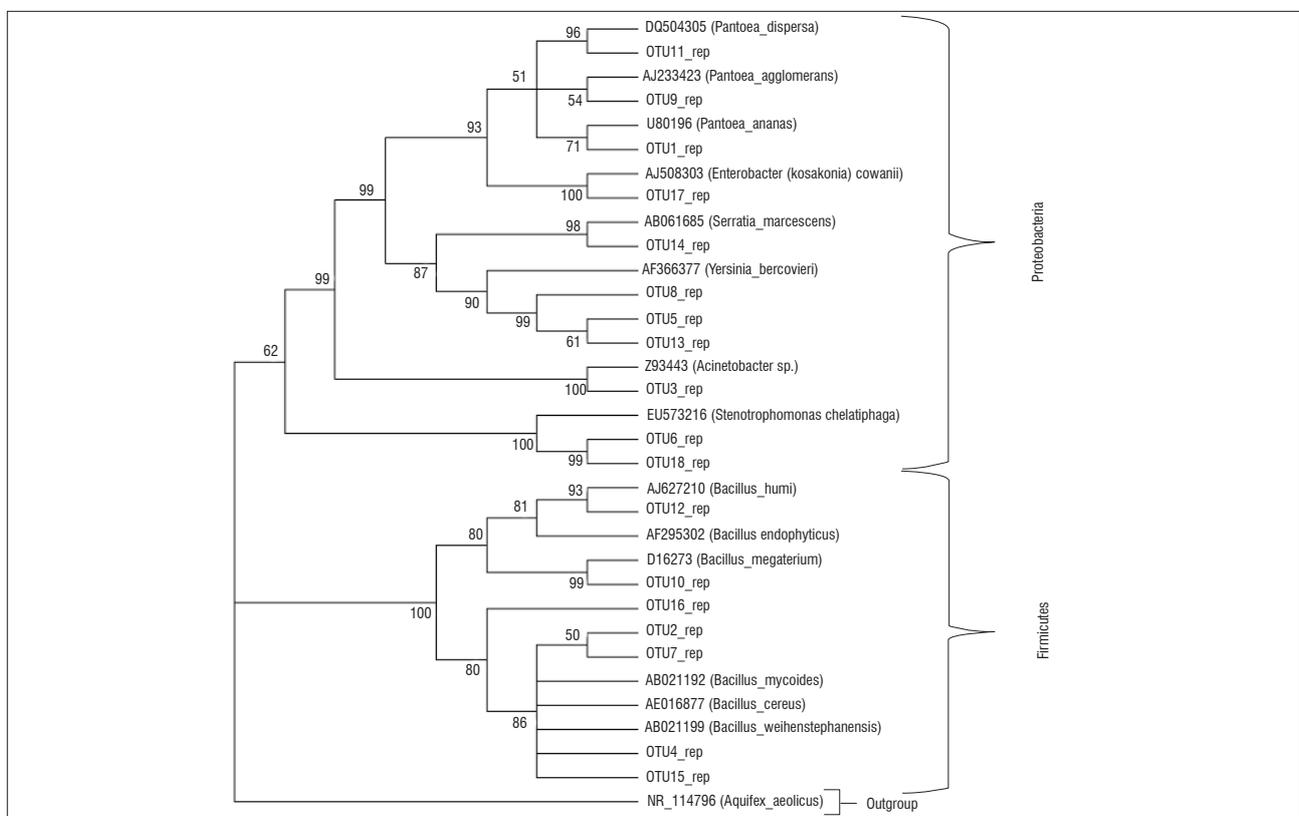

Figure 2: Operational taxonomic units (OTUs) of 16S bacterial sequences obtained from 90-day-old Bt and non-Bt maize plants.

Figure 3: Phylogenetic relationship of representative operational taxonomic units (OTUs) of 50-day-old Bt and non-Bt bacterial isolates based on 16S rDNA gene and closely related sequences (Jukes–Cantor algorithm and neighbour-joining tree). Bootstrap values are based on 1000 replicates and are indicated in branches; *Aquifex aeolicus* was used as an outgroup.


Figure 4: Phylogenetic relationship of representative operational taxonomic units (OTUs) of 90-day-old Bt and non-Bt bacterial isolates based on 16S rDNA gene and closely related sequences (Jukes–Cantor algorithm and neighbour-joining tree). Bootstrap values are based on 1000 replicates and are indicated in branches; *Aquifex aeolicus* was used as an outgroup.

Table 1: Isolates obtained from explant tissues of 50- and 90-day-old Bt and non-Bt maize

| Developmental stage	Source	OTU representative	Most significant alignment	% Similarity
50-day isolates				
NBt8_50L(KT459767)	Leaf	OTU 1 (KT459767)	*Yersinia mollaretii*	99%
NBt5_50L(KT459762)	Leaf	OTU 2 KT459762)	*Bacillus* sp.	99%
NBt4_50L(KT459764)	Leaf	OTU 3 (KT459764)	*Bacillus* sp.	99%
Bt7_50L(KT459760)	Leaf	OTU 4 (KT459760)	*Bacillus* sp.	99%
Bt5_50L(KT459758)	Leaf	OTU 5 (KT459758)	*Bacillus* sp.	99%
NBt9_50S(KT459779)	Stem	OTU 6 (KT459779)	*Bacillus* sp.	99%
NBt3_50L(KT459763)	Leaf	OTU 7 (KT459763)	*Bacillus* sp.	99%
NBt10_50L(KT459769)	Leaf	OTU 8 (KT459769)	*Bacillus massiliensis*	99%
NBt10_50L1(KT459764)	Leaf			
NBt2_50L(KT459762)	Leaf	OTU 9 (KT459762)	*Pseudomonas fluorescens*	99%
NBt7_50L(KT459766)	Leaf			
Bt9_50S(KT459776)	Stem	OTU 10 (KT459777)	*Bacillus* sp.	99%
NBt4_50S(KT459777)	Stem			
NBt10_50S(KT459783	Stem			
Bt9_50L(KT459761	Leaf			
NBt7_50S(KT459780)	Stem	OTU 11 (KT459756)	*Enterobacter hormaechei*	99%
Bt2_50S(KT459772)	Stem			
Bt4_50S(KT459774)	Stem			
Bt2_50L(KT459755)	Leaf			
Bt3_50L(KT459756)	Leaf			
NBt6_50S(KT459779)	Stem			
Bt6_50L(KT459759)	Leaf			
Bt4_50L(KT459757)	Leaf	OTU 12 (KT459757)	*Pantoea ananatis*	99%
Bt5_50S(KT459775)	Stem	OTU 13 (KT459775)	*Stenotrophomonas maltophilia*	99%
NBt8_50S(KT459781)	Stem			
NBt5_50S(KT459778)	Stem	OTU 14 (KT459778)	*Acinetobacter* sp.	99%
Bt3_50S(KT459773)	Stem	OTU 15 (KT459773)	*Bacillus pumilus*	99%
Bt1_50S(KT459771)	Stem	OTU 16 (KT459771)	*Bacillus subtilis*	99%
Bt8_50S(KT459770)	Stem	OTU 17 (KT459770)	*Serratia marcescens*	99%
90-day isolates				
Bt6L1(KT120083)	Leaf	OTU 1 (KT120089)	*Pantoea ananatis*	99%
Bt2L2(KT120089)	Leaf			99%
Bt10H2(KT120093)	Tassel	OTU 2 (KT120093)	*Bacillus mycoides*	99%
NBt2S(KT120080)	Stem	OTU 3 (KT120080)	*Acinetobacter* sp.	99%
Bt9C**(KT120071)	Seed			
Bt4C*(KT120075)	Seed	OTU 4 (KT120075)	*Bacillus weihenstephanensis*	99%
Bt4L2*(KT120090)	Leaf			
Bt4C(KT120072)	Seed			
NBt3L*(KT120084)	Leaf	OTU 5 (KT120084)	*Yersinia bercovieri*	99%
NBt10H*(KT120094)	Tassel			
Bt8L2(KT120086)	Leaf			
Bt10S(KT120081)	Stem	OTU 6 (KT120081)	*Stenotrophomonas chelatiphaga*	99%
Bt5H2(KT120096)	Tassel			
NBt2C(KT120073)	Seed			
Bt6L(KT120087)	Leaf			
Bt1H(KT120095)	Tassel	OTU 7 (KT120095)	*Bacillus mycoides*	99%
Bt8L(KT153621)	Leaf	OTU 8 (KT153621)	*Yersinia* sp.	99%
Bt8S(KT120082)	Stem	OTU 9 (KT120082)	*Pantoea agglomerans*	99%
Bt2L(KT120091)	Leaf			
NBt10H(KT120098)	Tassel	OTU 10 (KT120098)	*Bacillus megaterium*	99%
NBt10C2(KT120078)	Seed			
Bt3H(KT120097)	Tassel			
Bt4L2(KT120092)	Leaf	OTU 11 (KT120092)	*Pantoea dispersa*	99%
Bt4C**(KT120077)	Seed	OTU 12 (KT120077)	*Bacillus humi*	99%
Bt9H(KT120099)	Tassel	OTU 13 (KT120099)	*Yersinia* sp.	99%
Bt1C(KT120076)	Seed	OTU 14 (KT120076)	*Serratia marcescens*	99%
NBt3L(KT120088)	Leaf	OTU 15 (KT120088)	*Bacillus cereus*	99%
NBt2S1*(KT120079)	Stem			
NBt10C1(KT120070)	Seed	OTU 16 (KT120070)	*Bacillus cereus*	99%
NBt9L(KT120085)	Leaf	OTU 17 (KT120085)	*Enterobacter cowanii*	99%
NBt6C1(KT120074)	Seed	OTU 18 (KT120074)	*Stenotrophomonas* sp.	99%

Keywords: NBt ↔ non-Bt
Bt ↔ Bt (transgenic) maize
OTU, operational taxonomic unit
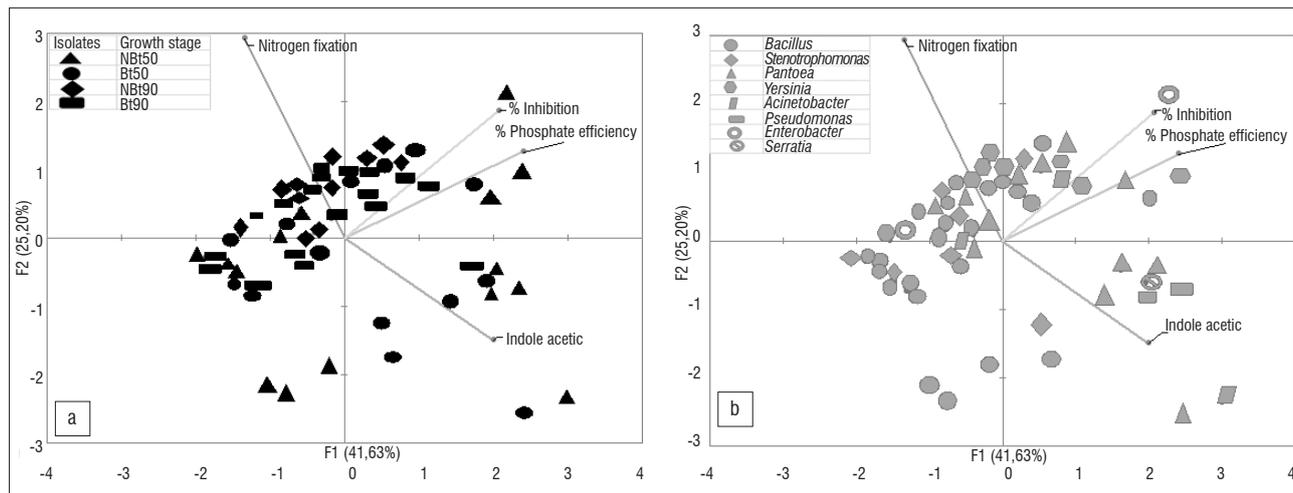
Figure 5: Clustering relationship between (a) endophytic isolates and (b) microbial genera and functional traits based on principal component analysis. (a) The distribution of endophytes isolated from Bt and non-Bt maize varieties at 50 and 90 days. Canonical correspondence analysis ordination diagram with endophytes isolated (circles = Bt 50 days, triangles = non-Bt 50 days, diamond = non-Bt 90 days, rectangles = Bt 90 days) and endophyte function variables (arrows). (b) The distribution of endophytes isolated in Bt and non-Bt maize varieties at 50 and 90 days. Canonical correspondence analysis ordination diagram with endophytes represented by various shapes (circles, triangles, squares, rectangles and diamonds) and endophyte function variables (arrows).
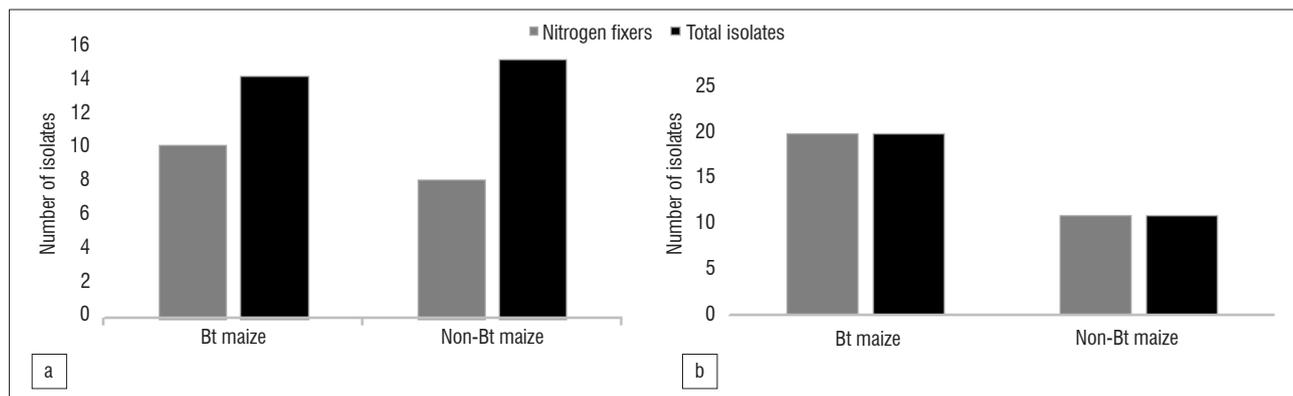

Figure 6: Total number of isolates which were shown to be potential nitrogen fixers in (a) 50-day-old maize isolates and (b) 90-day-old maize isolates.
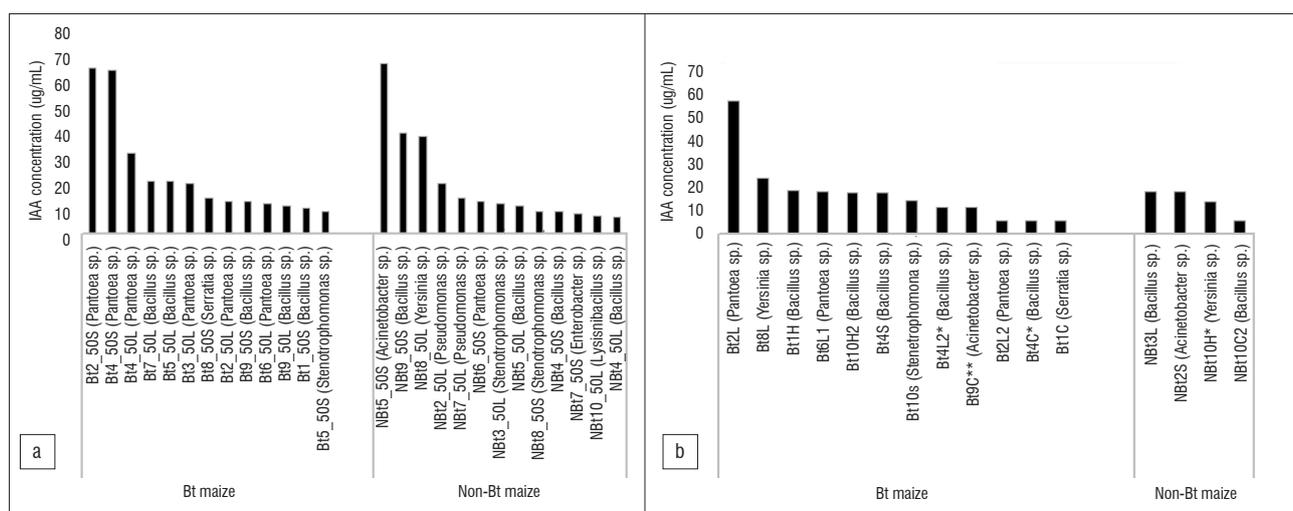
Figure 7: Estimated indole acetic acid (IAA) production of Bt and non-Bt maize endophytes in (a) 50-day-old maize isolates and (b) 90-day-old maize isolates.
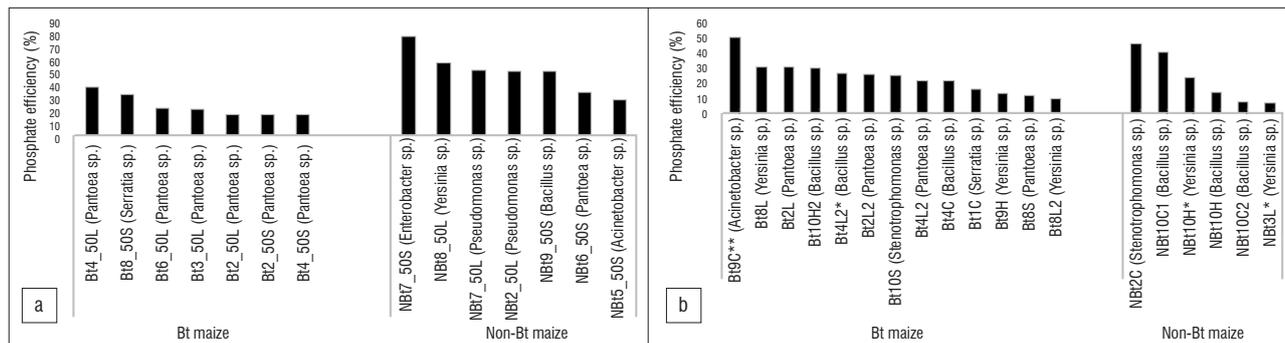
Figure 8: Phosphate efficiency of the bacterial isolates obtained from Bt and non-Bt maize plant shoots in (a) 50-day-old maize isolates and (b) 90-day-old maize isolates.
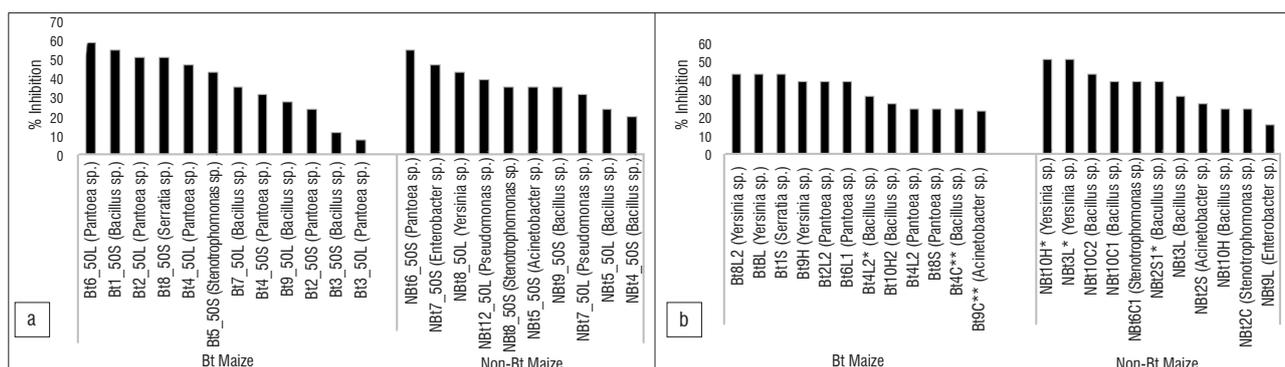
Figure 9: The effect of bacterial isolates on the growth of *Fusarium verticillioides* 10025 as measured by percentage inhibition of the radial growth of the colony: (a) Isolates from 50-day-old maize and (b) isolates from 90-day-old maize.

The phylogenetic pattern of culturable endophytes obtained in the present study revealed that they belong to the phyla Firmicutes and Proteobacteria. This finding is in agreement with previous studies linking these bacterial phyla to maize.[18,42] It has also been reported that, generally, Gamma- and Alphaproteobacteria are the dominant bacterial community inhabiting the phyllosphere, although the Firmicutes and Betaproteobacteria can also be present in large numbers. In this study, the dominant bacterial inhabitants were the Gammaproteobacteria and Firmicutes (Figures 3 and 4). The high density of endophytes on leaves compared to other plant parts might be because leaves have a large surface area and natural openings (stomata), allowing them to be the preferred point of tissue entry, as observed by Kumar and Hyde[43]. Furthermore, it has been shown that different species are commonly found on multiple plant tissues and there are some with a preference for the leaves.[15,19,44] This shows that there is a correlation between endophytes and plant parts harbouring them. In the present study, Bt and non-Bt maize cultivars harboured unrelated genera that are commonly found as maize endophytes such as *Pantoea, Bacillus, Enterobacter, Serratia, Yersinia, Stenotrophomonas, Pseudomonas* and *Acinetobacte*r (Table 1).[18,42] The results revealed the predominance of *Bacillus* and *Pantoea*, as well as the overall importance of all the isolated endophytes in plant growth processes. PCA clearly confirms that there is no differentiation between Bt and non-Bt maize. Endophyte functional diversity, which represents the capacity of microorganisms to perform different biological and ecological processes, is an important indicator of system disturbance and development. The PCA showed that the microbial community in the Bt and non-Bt maize at 50 days and at 90 days was not distinct.

Plant growth is sustained by the soil fertility state and a fertile soil is defined by the presence of important nutrients such as nitrogen, phosphorus and potassium.[45] Although nitrogen fixation has always been associated with legumes, it has been shown that non-leguminous plants such as maize also benefit from nitrogen fixers.[46] The most important constraint in maize production is low soil nitrogen, which contributes to a loss in production of about 30%.[47] Hence the presence of nitrogen fixers in the soil serves as a potential source of replacement for such loss. Furthermore, most of the potential nitrogen fixers in the rhizosphere have additional beneficial roles. For instance, in addition to their nitrogen-fixing capabilities, nitrogen fixers can also participate in root expansion, bioremediation and nutrient cycling.[46,48,49] In the present study, the number of isolates with nitrogen-fixing capabilities for both maize varieties (Bt and non-Bt) increased significantly between the two developmental stages (Figure 5a). The PCA showed that isolates with nitrogen-fixing capabilities for both maize varieties (Bt and non-Bt) clustered together (Figure 5b), with more isolates from the 90-day-old developmental stage. This probably occurs because more nitrogen is needed during maximum plant growth for reproduction purposes and as the plants continue to grow, available nitrogen becomes depleted.[47] The genera which had significant influence in fixing atmospheric nitrogen were *Bacillus* and *Pantoea*.

IAA production by isolates was significantly higher at 50 days than at 90 days in both maize varieties. As shown in Figure 5a, only 50-day isolates of both maize varieties grouped together. The major reason for clustering or high IAA at 50 days appears to be associated with plant age because it is a growth hormone needed by the plants at an early stage for root and stem growth regulation.[48] Similarly, the same pattern was observed with regard to phosphate solubilisation and antifungal activity, with an increase in both functional traits associated with isolates at the 90-day developmental stage.

The results obtained in this study are important for both scientists and farmers, especially as they relate to different biotechnological applications. Globally, the ultimate goal of farmers is to cultivate maize of desirable traits. Important scientific information that indicates that Bt maize does not have negative impacts on the ecosystem will be welcomed. This information may in turn translate to improvements in acceptability and marketability of the maize being produced by the farmers, as well as the ability to invest in environmentally friendly fertilisation approaches (e.g. biofertiliser) which rely on microbial activities.

A limitation of the study is the number of genotypes used as well as the sole use of a culture-based method. However, the culture-based approach was important in this study because it allowed us to assess the functional traits of the isolated bacterial endophytes (Figure 5b). In comparison, Mashiane et al.'s[49] metagenomics analytical approach provided more information about the abundance and diversity of the bacterial endophytes.

In conclusion, we have demonstrated that maize phyllosphere harbours different types of bacterial endophytes but their composition is not affected by the Bt genetic modification of the maize plant. Similarly, functional roles of the bacterial endophytes are not affected by the genetic modification. However, there was a significant increase in endophyte density from the 50-day to the 90-day developmental stage, suggesting that developmental stages of both Bt and non-Bt maize could drive the composition of the endophytic bacterial community. The beneficial characteristics of the endophytic bacteria in this study are important in agriculture. Thus, further biotechnological investigation needs to be conducted under field conditions to confirm the efficiency of these bacterial isolates in nutrient cycling and plant protection.

Acknowledgements

Authors' contributions

A.R.M.: Methodology, data collection, data analysis, sample analysis, validation, data curation, writing the initial draft. R.A.A.: Conceptualisation, methodology, data collection, data analysis, sample analysis, validation, data curation, writing the initial draft and revisions, student supervision, project leadership, project management, funding acquisition. C.C.B.: Methodology, sample analysis, validation, writing the initial draft and revisions. G.J.C.: Data analysis, validation.

References

1. Keetch DP, Webster JW, Ngqaka A, Akanbi R, Mahlangu P. Bt maize for small scale farmers: A case study. Afr J Biotechnol. 2005;4(3):1505–1509. http://doi:10.4314/ajfand.v4i13.71805

2. Ostry V, Ovesna J, Skarkova J, Pouchova V, Ruprich J. A review on comparative data concerning *Fusarium* mycotoxins in Bt maize and non-Bt isogenic maize. Mycotoxin Res. 2010;26(3):141–145. http://dx.doi.org/10.1007/s12550-010-0056-5

3. Prischl M, Hackl E, Pastar M, Pfeiffer S, Sessitsch A. Genetically modified Bt maize lines containing cry3Bb1, cry1A105 or cry1Ab2 do not affect the structure and functioning of root-associated endophyte communities. Appl Soil Ecol. 2012;54:39–48. http://doi.org/10.1016/j.apsoil.2011.12.005

4. Sticher L, Mauch-Mani B, Métraux JP. Systemic acquired resistance. Annu Rev Phytopathol. 1997;35(1):235–270. http://doi.org/10.1146/annurev.phyto.35.1.235

5. Saxena D, Stotzky G. Insecticidal toxin from *Bacillus thuringiensis* is released from roots of transgenic Bt corn in vitro and in situ. FEMS Microbiol Ecol. 2000;33(1):35–39. https://doi.org/10.1111/j.1574-6941.2000.tb00724.x

6. Strobel G, Daisy B, Castillo U, Harper J. Natural products from endophytic microorganisms. J Nat Prod. 2004;67(2):257–268. https://doi.org/10.1021/np030397v

7. Das A, Varma A. Symbiosis: The art of living. Symbiotic fungi. In: Varma A, Kharkwal AC, editors. Symbiotic fungi principles and practice. Berlin: Springer; 2009. p. 1–28. https://doi.org/10.1007/978-3-540-95894-9_1

8. Nair DN, Padmavathy S. Impact of endophytic microorganisms on plants, environment and humans. Sci World J. 2014;2014:1–11. http://dx.doi.org/10.1155/2014/250693

9. Ryan PR, Germaine K, Franks A, Ryan DJ, Dowling DN. Bacterial endophytes: Recent developments and applications. FEMS Microbiol Lett. 2008;278(1):1–9. https://doi.org/10.1111/j.1574-6968.2007.00918.x

10. Lee S, Flores-Encarnación M, Contreras-Zentella M, Garcia-Flores L, Escamilla JE, Kennedy C. Indole-3-acetic acid biosynthesis is deficient in *Gluconacetobacter diazotrophicus* strains with mutations in cytochrome *c* biogenesis genes. J Bacteriol. 2004;186(16):5384–91. https://doi.org/10.1128/JB.186.16.5384-5391.2004

11. Vieira PD, Motta CM, Lima D, Torres JB, Quecine MC, Azevedo JL, et al. Endophytic fungi associated with transgenic and non-transgenic cotton. Mycology. 2011;2(2):91–97.

12. Berg G, Krechel A, Ditz M, Sikora RA, Ulrich A, Hallmann J. Endophytic and ectophytic potato-associated bacterial communities differ in structure and antagonistic function against plant pathogenic fungi. FEMS Microbiol Ecol. 2005;51(2):215–229. https://doi.org/10.1016/j.femsec.2004.08.006

13. Lowman S, Kim-Dura S, Mei C, Nowak J. Strategies for enhancement of switchgrass (*Panicum virgatum* L.) performance under limited nitrogen supply based on utilization of N-fixing bacterial endophytes. Plant Soil. 2016;405(1–2):47–63. https://doi.org/10.1007/s11104-015-2640-0

14. Maida I, Chiellini C, Mengoni A, Bosi E, Firenzuoli F, Fondi M, et al. Antagonistic interactions between endophytic cultivable bacterial communities isolated from the medicinal plant *Echinacea purpurea*. Environ Microbiol. 2015;18:2357–2365. http://dx.doi.org/10.1111/1462-2920.12911

15. Rijavec T, Lapanje A, Dermastia M, Rupnik M. Isolation of bacterial endophytes from germinated maize kernels. Can J Microbiol. 2007;53:802–808. https://doi.org/10.1139/W07-048

16. Rodríguez-Blanco A, Sicardi M, Frioni L. Plant genotype and nitrogen fertilization effects on abundance and diversity of diazotrophic bacteria associated with maize (*Zea mays* L.). Biol Fert Soils. 2015;51(3):391–402. https://doi.org/10.1007/s00374-014-0986-8

17. Araújo JM, Silva AC, Azevedo JL. Isolation of endophytic actinomycetes from roots and leaves of maize (*Zea mays* L.). Braz Arch Biol Technol. 2000;43(4):1–2. http://dx.doi.org/10.1590/S1516-89132000000400016

18. Liu Y, Zuo S, Xu L, Zou Y, Song W. Study on diversity of endophytic bacterial communities in seeds of hybrid maize and their parental lines. Arch Microbiol. 2012;194(12):1001–1012. http://dx.doi.org/10.1007/s00203-012-0836-8

19. Roesch LFW, Camargo FAO, Bento FM, Triplett EW. Biodiversity of diazotrophic bacteria within the soil, root and stem of field-grown maize. Plant Soil. 2008;302(1–2):91–104. http://dx.doi.org/10.1007/s11104-007-9458-3

20. Arnold AE, Henk DA, Eells RL, Lutzoni F, Vilgalys R. Diversity and phylogenetic affinities of foliar fungal endophytes in loblolly pine inferred by culturing and environmental PCR. Mycol Res. 2007;99(2):185–206. http://dx.doi.org/10.1080/15572536.2007.11832578

21. Güssow D, Clackson T. Direct clone characterization from plaques and colonies by the polymerase chain reaction. Nucleic Acids Res. 1989;17:4000

22. Obi LU, Atagana HI, Adeleke RA. Isolation and characterisation of crude oil sludge degrading bacteria. SpringerPlus. 2016;5(1), Art. #1946, 13 pages. https://doi.org/10.1186/s40064-016-3617-z

23. Muyzer G, De Waal EC, Uitterlinden AG. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Appl Environ Microbiol. 1993;59(3):695–700.

24. Hall TA. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser. 1999;41:95–98.

25. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 2009;37(suppl 1):141–145. https://doi.org/10.1093/nar/gkn879

26. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–7541. https://doi.org/10.1128/AEM.01541-09

27. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780. https://doi.org/10.1093/molbev/mst010

28. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–2729. https://doi.org/10.1093/molbev/mst197

29. Wilson M, Knight D. Methods of plant pathology. Tuite J, editor. London: Academic; 1952.

30. Rao NSS. Soil microorganisms and plant growth. New Delhi: Oxford and IBH Publishing Co; 1977.

31. Mwashasha RM, Hunja M, Tani A. Screening of bacterial and fungal isolates for their plant growth promoting activities. In: Proceedings of the 2012 JKUAT Scientific, Technological and Industrialization Conference; 2012 November 15–16; Nairobi, Kenya. Nairobi: Jomo Kenyatta University of Agriculture and Technology, Research Production and Extension Division; 2012. http://journals.jkuat.ac.ke/index.php/jscp/article/view/868

32. Nautiyal CS. An efficient microbiological growth medium for screening phosphate solubilizing microorganisms. FEMS Microbiol Lett. 1999;170(1):265–270. https://doi.org/10.1111/j.1574-6968.1999.tb13383.x

33. Adeleke R, Cloete TE, Khasa DP. Culturable microorganisms associated with Sishen iron ore and their potential roles in biobeneficiation. World J Microbiol Biotechnol. 2012;28(3):1057–1070. https://doi.org/10.1007/s11274-011-0904-2

34. Andreote FD, Azevedo JL, Araújo WL. Assessing the diversity of bacterial communities associated with plants. Braz J Microbiol. 2009;40:417–432. https://doi.org/10.1590/S1517-83822009000300001

35. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007;62:142–160. https://doi.org/10.1111/j.1574-6941.2007.00375.x

36. Ter Braak CJF, Smilauer P. CANOCO reference manual and CanoDraw for Windows user's guide: Software for canonical community ordination (Version 4.5). Ithaca, NY: Microcomputer Power; 2002.

37. Andreote FD, Rossetto PB, Mendes R, Avila LA, Labate CA, Pizzirani-Kleiner AA, et al. Bacterial community in the rhizosphere and rhizoplane of wild type and transgenic eucalyptus. World J Microbiol Biotechnol. 2009;25:1065–1073. https://doi.org/10.1007/s11274-009-9990-9

38. Whittaker RH, Levin SA, Root RB. Niche, habitat, and ecotope. Amer Nat. 1973;107(955):321–338. http://doi.org/10.1086/282837

39. Ter Braak CJ. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. FEMS Microbiol Ecol. 1986;67(5):1167–1179. http://doi.org/10.2307/1938672

40. James C. Global status of commercialized biotech/GM crops. ISAAA brief no.39. Ithaca, NY: ISAAA; 2010.

41. Saxena D, Flores S, Stotzky G. Bt toxin is released in root exudates from 12 transgenic corn hybrids rep resenting three transformation events. Soil Biol Biochem. 2002;34(1):133–137. http://doi.org/10.1016/S0038-0717(01)00161-4

42. Pereira P, Ibáñez F, Rosenblueth M, Etcheverry M, Martínez-Romero E. Analysis of the bacterial diversity associated with the roots of maize (*Zea mays* L.) through culture-dependent and culture-independent methods. FEMS Microbiol Ecol. 2011;2011, Art. #938546, 10 pages. http://doi.org/10.5402/2011/938546

43. Kumar DS, Hyde KD. Biodiversity and tissue-recurrence of endophytic fungi in *Tripterygium wilfordii*. Fungal Divers. 2004;17:69–90. http://hdl.handle.net/10722/223074

44. Montañez A, Blanco AR, Barlocco C, Beracochea M, Sicardi M. Characterization of cultivable putative endophytic plant growth promoting bacteria associated with maize cultivars (*Zea mays* L.) and their inoculation effects in vitro. Appl Soil Ecol. 2012;58:21–28. http://doi.org/10.1016/j.apsoil.2012.02.009

45. Doebereiner J, Pedrosa FO. Nitrogen-fixing bacteria in nonleguminous crop plants. New York: Science Tech; 1987.

46. Lynch JM. Microbial metabolites. In: Lynch JM, editor. The rhizosphere. Chichester: Wiley; 1990. p. 177–206.

47. Pandey RK, Maranville JW, Admou A. Deficit irrigation and nitrogen effects on maize in a Sahelian environment: I. Grain yield and yield components. Agric Water Manage. 2000;46(1):1–3. http://doi.org/10.1016/S0378-3774(00)00073-1.

48. Teale WD, Paponov IA, Palme K. Auxin in action: Signalling, transport and the control of plant growth and development. Nat Rev Mol Cell Biol. 2006;7(11):847–859. http://dx.doi.org/10.1038/nrm2020

49. Mashiane RA, Ezeokoli OT, Adeleke RA, Bezuidenhout CC. Metagenomic analyses of bacterial endophytes associated with the phyllosphere of a Bt maize cultivar and its isogenic parental line from South Africa. World J Microbiol Biotechnol. 2017;33(4):80. https://doi.org/10.1007/s11274-017-2249-y

AUTHORS:
Taryn de Beer[1]
Rachel Wynberg[1]   iD

AFFILIATION:
[1]Department of Environmental
and Geographical Science,
University of Cape Town,
Cape Town, South Africa

CORRESPONDENCE TO:
Taryn de Beer

EMAIL:
taryn.debeer@gmail.com

Developing and implementing policy for the mandatory labelling of genetically modified food in South Africa

Like many other countries, South Africa has come under public pressure to introduce mandatory labelling for genetically modified (GM) foods. Although there is increased understanding of the social and political implications of GM labelling in developed countries, implications for the Global South are still poorly understood. South Africa, as a country that consumes, produces and trades GM food, represents a fitting case study of these dimensions in the context of a developing economy. Via policy analysis, stakeholder interviews and document inquiry we offer an overview of the evolution of GM food labelling, investigate the central influences on its development and implementation, determine the critical issues and identify the factors impeding or facilitating implementation. Our findings reveal that many significant events and decisions influenced the policy on mandatory GM food labelling in South Africa. They also suggest that several pertinent and problematic issues arose during its development as a result of (1) the contentious nature of GM food labelling; (2) stakeholder opinions, influences, and conflicted positions; and (3) its practical complexity. Key implementation issues included divergent interpretations, and thus high levels of ambiguity; an inefficient National Consumer Commission; a lack of recourse for non-compliance; and the absence of a government-enforcement agency. Lower capacity in developing countries underscores the importance of a participation process that is believable by and inclusive of all actors. Stakeholders' opinions about the policy development process were affected by their predetermined viewpoints about GM organisms (GMOs). Findings emphasise the significance on participatory processes of larger policy debates about the acceptability of GMOs, and the importance of contextualising GM food labelling policies within such debates.

Significance:

- The first review of the evolution of mandatory GM food labelling policy in South Africa is provided.

- A knowledge gap with regard to GM food labelling in developing countries is filled.

- The importance of procedural fairness in determining the degree of stakeholder satisfaction with policy decisions is revealed.

Introduction

The labelling of genetically modified (GM) food is regarded as one of the most disputed food issues of the 21st century.[1,2] As the land under GM crops has grown – reaching over 185 million hectares in 2016 from its inception in 1996 – GM food has increasingly entered our food chains.[3] Worldwide, there is ongoing controversy over genetic modification, alongside consumer concerns about the safety and risks associated with GM food, both to human and animal health, and to the environment.[4] This controversy placed public pressure on the European Union to introduce the first labelling policies for GM food in 2003. Since then, numerous other countries have introduced and implemented labelling legislation, policies, regulations or requirements for GM food.[5] The characteristics of these approaches and the degree to which they are implemented differ significantly.[2,6] Some countries have opted for a voluntary labelling approach, with guidelines, while almost 40 other countries, South Africa included, have approved mandatory labelling requirements.[4,6]

In a highly contested process, mandatory GM labelling was introduced in South Africa in 2009 through the *Consumer Protection Act (No. 68 of 2008)* (CPA), with associated regulations, which came into effect in 2011. Despite media attention[7-13] there has been little research to explore the evolution of the policy process and its implementation. Moreover, although there is increased understanding of the social and political implications of GM food labelling in industrialised countries, those for countries of the Global South remain poorly comprehended.[13] As a developing country that consumes, produces and trades GM food, South Africa is a fitting case study to fill this lacuna. We aim in this paper to provide an in-depth account of the development and implementation of GM food labelling policy in South Africa. As the policy's progression was examined and assessed, contentious and problematic issues were identified. Our investigation of these pertinent issues of GM food labelling in a developing country context can be used to inform and equip other developing countries that are considering mandatory labelling of GM foods. Through the analysis, we set out to suggest various possible improvements to South Africa's process of developing legislation by recommending that the government engage in best-practice participation for policy development.

Methodology

Policy development and implementation processes are complex and multifaceted and involve the participation of an array of stakeholders. To analyse these processes we used qualitative and quantitative research methods, framed around stakeholder perceptions. We collected, reviewed and analysed a range of relevant documents including emails, policy and legal documents, official government records, reports, press releases, newsletters,

website materials and organisational publications. Most documents were obtained from the Department of Trade and Industry (DTI) through a request in terms of the *Promotion of Access to Information Act, No. 2 of 2000* (PAIA) for access to the records of public bodies. In total, 429 documents were received.

Document analysis was complemented by semi-structured interviews, specifically tailored for different stakeholder groups, to draw out more considered, in-depth and complex responses. More than 100 stakeholders were contacted, identified by their engagement in the policy and through snowball sampling. Of these, 27 in-depth interviews were conducted with representatives from different sectors of society. These included the biotechnology and food industry (15), government (3), non-governmental organisations (NGOs) (8), and the academic and scientific community (1). Interviews with stakeholders from consumer groups and trade unions were not conducted, as representatives from these two groups were not available. The stakeholders who were interviewed and are referenced throughout the paper (by superscript code) are listed in Table 1.

Table 1: Interviewed stakeholders

Respondent code	Respondent organisation	Interview date
ACB1	African Centre for Biosafety; now the African Centre for Biodiversity	31 January 2013
ACB2	African Centre for Biosafety; now the African Centre for Biodiversity	25 March 2013
AFB	AfricaBio	20 March 2013
AGB	Agricultural Business Chamber	1 February 2013
BIOW	Biowatch SA	11 December 2012
CGC	Consumer Goods Council of South Africa	30 January 2013
DOH	National Department of Health	21 January 2013
DST	Department of Science and Technology	23 January 2013
H&H	Hahn & Hahn Inc. and Hahn & Hahn Attorneys	26 March 2013
MON	Monsanto South Africa	25 January 2013
SAF1	South African Freeze Alliance on Genetic Engineering	8 November 2012
SAF2	South African Freeze Alliance on Genetic Engineering	8 November 2012

Questions were prepared and piloted prior to the interviews. These consisted of yes/no questions, choosing from lists, and ranking and rating. Topics ranged from knowledge of GM organisms (GMOs) in agriculture and GM food labelling to positions on GMOs in agriculture and on GM food labelling. Open-ended questions focused on policy development, implementation and compliance. Interview responses and document excerpts were analysed using coding and thematic analysis. Data were triangulated so as to verify sources and provide a comprehensive narrative of stakeholder participation. Participation was confidential and anonymous. Ethical clearance was obtained through the Faculty of Science Research Ethics Committee at the University of Cape Town (SFREC 37_2012).

Development of a regulatory framework for GM food labelling

South Africa has a history of strongly promoting modern biotechnology.[14] Field trials for Bt cotton (referring to genes from the soil bacterium *Bacillus thuringiensis* which encode crystalline protein endotoxins that kill certain insect species)[15,16] were first undertaken in 1992, followed by commercial plantings of GM cotton and maize.[17,18] Eight years later, Bt white maize for human consumption (the staple food eaten by most South Africans) was commercialised.[17] In 2016, South Africa planted 2.66 million ha of commercial GM crops.[3] Three of these GM crop types were: maize (2.16 million ha, or 90% of the country's production), soybean (494 000 ha, 95% of production) and cotton (9000 ha, 100% of production).[3,19] GM canola may be imported as a commodity under a permit but may not be produced in the country.[5] Commercially grown GM crops in South Africa, as in the rest of the world, comprise only two main traits: tolerance to the herbicide glyphosate, and pest resistance through incorporation of genes from *Bacillus thuringiensis* (Bt). Stacked traits, which combine two or more genes of interest into a single plant, are increasingly incorporated into GM seed, and maize in particular.[19]

Many South African consumers are unaware of the extent of GM crops in the country and of their own consumption of GM foods.[5,20] Although a small group of NGOs and consumer groups began lobbying for mandatory labelling when commercial production of GM crops commenced in 1997, South Africa only started to make provision for GM food labelling in 2004 – seven years later.[7,21;SAF1]

Table 2 presents a chronology of the development of a regulatory framework for GM food labelling, setting out key milestones in the controversial history of this policy process. These milestones centred on (1) the adoption of voluntary labelling regulations in 2004 through the *Foodstuffs, Cosmetics and Disinfectants Act No. 54 of 1972*, administered by the Department of Health (DOH); (2) the promulgation of the CPA by the DTI in 2008, requiring mandatory labelling for GM goods, ingredients and components in food; and (3) regulations to give effect to the CPA.

Introduction of proposals for GM food labelling

In 2004, the DTI published a Green Paper on consumer policy, in which they proposed a new law to protect consumers and provided a first glimpse of proposals for GM food labelling.[22] Voluntary labelling was subsequently introduced under the *Foodstuffs, Cosmetics and Disinfectants Act*, which required labelling if food differed significantly from its conventional (non-GM) counterpart in terms of nutritional composition, storage requirements, preparation or cooking, or if it contained an allergen or a human or animal gene.

Two years later, the draft Consumer Protection Bill was released for comment, and included a clause requiring the mandatory labelling of GM ingredients. The draft Bill also made provision for product liability, in terms of which any producer, distributor or supplier would be liable for any damage caused by their product, including that which might arise from consuming GM food.

The Bill was opposed by the Department of Agriculture, Forestry and Fisheries (DAFF) and the DOH, on the grounds that it would send out a confusing message to stakeholders.[7,8,23,24] Concerns were also expressed about the cost of labelling and its impact on food prices, and the technical expertise required to ensure compliance.[24] In response, the GM labelling clause was removed from the draft Bill in 2006, with a proposal to address policy issues under the *Genetically Modified Organisms Act, No. 15 of 1997*, and elsewhere.[7,24] For the next two years, the South African Freeze Alliance on Genetic Engineering, an NGO network, campaigned to increase public awareness around GMO labelling[25,26,SAF1], mobilising a wide range of organisations to lobby the DTI to reinstate the labelling clause[7,25,26,SAF1,SAF2].

The first version of the Consumer Protection Bill [B 19—2008] that was introduced in Parliament in May 2008 confirmed the exclusion of GM labelling and also rendered those producing GMOs exempt from any liability for potential damage.[27(s61),28] There were divergent stakeholder opinions as to why the labelling clause had been retracted. NGOs argued that a sudden turn of events had occurred, in which the industry had lobbied the Acting President to rescind the clause.[SAF1,AGB] The industry's perspective was that GM labelling had never been envisaged for the CPA, and that the DTI, after consultation with industry members, believed it would be impractical to implement GM labelling and therefore withdrew it.[AGB,MON]

Table 2: Chronology of significant events and decisions during the policy development and implementation of mandatory GM food labelling in South Africa

Date	Policy event
1992	South Africa introduces the first field trials of GM crops.
1997	The first commercial GM crops are grown in South Africa. The *Genetically Modified Organisms Act 15 of 1997* (*GMO Act*) is promulgated.
Early 2000s	The South African Freeze Alliance on Genetic Engineering begins lobbying for mandatory labelling of GM foods.
2004	Voluntary GM food labelling is introduced as Regulation 25 of 2004 under the *Foodstuffs, Cosmetics and Disinfectants Act 54 of 1972*, administered by the Department of Health.[32] The Department of Trade and Industry publishes a Green Paper on consumer policy, proposing a new law to protect consumers.[22]
March 2006	A labelling clause is introduced in the draft Consumer Protection Bill (CPB).
September 2006	The labelling clause is removed from the CPB, with a proposal to address policy issues under the *GMO Act* and elsewhere.[12,58]
2008	A clause is introduced in the CPB that exempts producers or suppliers of GMOs from liability.[37,38] A series of public hearings and provincial briefings follows.
April 2009	In response to concerns, the GM labelling clause is reinstated[13,15,38] and the *Consumer Protection Act 68 of 2008* (CPA) is promulgated requiring mandatory labelling for GM goods, ingredients or components in food.[44] Liability is also addressed.
April 2011	The *CPA* comes into effect. Regulations are promulgated.
October 2011	Regulations governing the mandatory labelling of imported and locally produced GM foods come into effect, whereby goods, among other things, are defined as 'anything marketed for human consumption'.
2011	An industry-aligned working group is established to help coordinate efforts and to interpret labelling regulations and compliance measures.[50,MON,AGB]
2012	The National Consumer Commission develops a task team to draft GMO labelling guidelines.
2011–2012	Non-governmental organisations (NGOs) lay complaints for non-compliance with regulations.
2012	The industry threatens litigation against NGOs.
October 2012	Draft amendments to the regulations are published for comment.
July 2014	The Department of Trade and Industry hosts a consultative conference.
2014 – present	The public comment period is extended, but amendments to the regulations have yet to be finalised.

A series of public hearings and provincial briefings followed, with several provinces, particularly the Western Cape, arguing for the reinstatement of the clause.[28,ACB2] This engagement was largely thanks to the efforts of the South African Freeze Alliance on Genetic Engineering. Their widespread consumer awareness outreach programme had made an impact,[SAF1,AGB,SAF2,ACB1] alongside their testing of several random food products, which showed relatively high percentages of GM content.[8] The Congress of South African Trade Unions (COSATU), the country's main trade union federation which forms part of the tripartite alliance alongside the governing African National Congress and the South African Communist Party[17,29], also presented its case in support of mandatory labelling[30]. Without the support and influence from COSATU, GM labelling might well not have been included in the CPA.[SAF1,ACB1]

In an about-turn, the promulgation of the CPA saw the reinstatement of the GM labelling clause[8,9,28], including mandatory labelling for GM goods, ingredients, or components in food[31]. The controversial liability clause was also removed, with liability for any harm caused by a defective product, irrespective of any negligence, now placed at the door of producer, importer, distributor or retailer of goods.[28,32]

Justifying this change of direction, the DTI noted that there were 'no substantial cost implications anticipated by introducing GMO labelling, as the Bill [did] not prescribe how labelling should be done'[28]. Opposition from the DOH, DAFF, the Department of Science and Technology (DST) and the food industry, remained, based on the cost of labelling and its impact on food prices, and the technical expertise required to ensure compliance.[24,DST] The reinstatement of the clause was considered a great achievement by advocacy organisations and consumers, and 'a victory for consumer rights'[5(p.387)].

Following a series of stakeholder workshops and an array of submissions[21,32,AGB], regulations were promulgated in 2011 to give effect to the law. These regulations provided for the mandatory labelling of imported and locally produced GM foods, whereby goods, among other things, were defined as 'anything marketed for human consumption'[31].

Implementation of the Consumer Protection Act: 2011−2017

A third phase of stakeholder involvement began after promulgation of the regulations. This phase included establishment of an industry-aligned working group to help coordinate efforts and to interpret labelling regulations and compliance measures.[33,MON] Central concerns were to establish how to quantify the stipulated 5% GM content and to clarify and define what was meant by 'contains at least 5% of GMOs'[34,35]. The industry appeared unanimous on the matter, and lodged complaints with the National Consumer Commission (NCC), seeking clarity on implementation.[MON] Inundated with requests, the NCC referred the matter to the Minister of Trade and Industry[36] and, the NCC, together with an interdepartmental government task team (DOH, DAFF, DST and DTI), planned to help the industry by publishing guidelines.

Implementation challenges led to very low compliance initially by industry, especially regarding maize.[36,MON,DST,DOH] Two of South Africa's largest retail chains – Pick 'n Pay and Woolworths – were among those advised by the commissioner to place implementation on hold until full consideration had been given to the issue.[37] In turn, consumers were urged by NGOs such as the African Centre for Biosafety to lay complaints with the NCC for the labelling of GM food.[38,SAF2] In 2012, the African Centre for Biosafety carried out tests on several food products, with

many testing positive for GMOs and with high levels of GM content – all were unlabelled.[39] Biowatch South Africa achieved similar results that year, finding only one product with a compliant label.[BIOW] The four non-compliant companies were Nestlé, Pioneer Foods (Bokomo), Futurelife and Premier Foods.[39]

Considerable confusion arose about the scope and reach of the CPA. For example, the Consumer Goods Council of South Africa, on behalf of industry, sought clarity from the NCC on whether 'the members must label only the four varietals of maize, cotton, soybean and rape seed (canola), according to the provisions, or is it the intention that they label even products of which these varietals are ingredients or components?'[DST,DOH,AFB,H&H] Receiving no response, the Consumer Goods Council of South Africa assumed[40,CGC] the NCC's position to be that it would not enforce the Act as 'clarity [had] not been reached and all loopholes [had] not been closed'.[40] Furthermore, a task team appointed by the commission to 'clarify all legal uncertainties and ambiguities, which may rise to interpretation problems' remained uncertain.[CGC] As such the Consumer Goods Council of South Africa advised its members not to label until they were given clarity.[CGC]

Simultaneously, the African Centre for Biosafety and consumers lodged formal complaints with the NCC, after an undertaking to start labelling in the allotted time was breached by non-compliant companies.[10,41] Several large food producers reacted by threatening legal action against the African Centre for Biosafety if claims continued that these companies were in contravention of the labelling law.[11] Nestle and Pioneer Foods stated that they were 'fully compliant' and were 'waiting the outcome of a meeting between the NCC and industry representatives for a ruling on the GM labelling legislation'.[11] Premier Foods noted it would be 'phasing in new packaging for its mealie-meal to include GM labels'.[11]

Various stakeholders encountered obstacles with the NCC and blamed its failure on capacity, resource and capability constraints.[33;SAF2,DOH,AFB] A number of technical issues also thwarted effective implementation. For example, a discrepancy was noted between the terminology 'genetically modified organism' used in the regulations and 'genetically modified ingredients or components of those goods' in Section 24(6) of the Act.[41] To standardise terms, draft amendments to the regulations were published for public comment.[42] These proposed that all local and imported food products, including processed products, containing 5% or more GM ingredients or components must be labelled as 'contains genetically modified ingredients or components' to enable consumers to make informed choices.[43] Numerous comments were received on the amendments, but to date the amended GM labelling regulations have not been finalised.[44]

At present, the monitoring of the Act continues to rest largely on the shoulders of civil society organisations.[SAF1,AGB,BIOW] While there is an increase in the number of food companies labelling products as GM, tests conducted on maize- and soya-based food products, including common maize meal and bread brands, reveal many labelling claims regarding percentage of GM content to be incorrect.[12] Many of the tested products also remain unlabelled, with some containing high GM content.[12]

Discussion

The chronology described above reveals the critical role played by different stakeholders in the policy process. Prominent stakeholders included the biotechnology and food industries, government, NGOs, academia, trade unions and consumer organisations. Each played a role in agro-food politics and policymaking, forming part of networks or alliances in which there were mutually understood interests and values. For the most part they reflected the conflicting positions of pro-GMO lobbyists, the anti-GMO network and other actors, such as those in government, who formed their own sets of alliances with these networks.[45]

However, the polarisation of views and positions occurred not only between industry, researcher and NGO interest groups, but also between certain government departments, which in turn created difficulties in reaching a consensus.[29,46;DOH] A central tension arose between a policy position supporting voluntary labelling and limited liability for those

manufacturing GM foods, versus one advocating mandatory labelling. Although the liability clause was later amended, the challenges of implementing mandatory labelling and a lack of clarity about the scope and reach of the CPA presented a further set of issues, accompanied by low levels of compliance by industry.[36]

Consultation or active participation?

The question remains as to whether the process was inclusive, fair and consultative, or exclusive, unfair and conflictual. The complex, dynamic, uncertain and multiscalar nature of today's environmental problems, as well as the various stakeholders and organisations they affect, calls for 'flexible and transparent decision-making that embraces a diversity of knowledge and values'.[47(p.2417)] Consequently, stakeholder participation has gradually become more entrenched in environmental decision-making, across national and international policies.[47(p.2417)] Policies which have successfully incorporated a stakeholder guided approach to their development, especially those intended to protect the environment and public interests, hold important lessons for GM labelling in South Africa.

Rowe and Frewer[48] describe a typology of stakeholder engagement focused on the 'nature', instead of the 'degree', of participation.[48] This defines different forms of public engagement by the way in which 'communication migrates'.[48(p.2419)] Collecting or soliciting input from participants is deemed 'consultation', while 'active participation' is seen as a two-way communication, with information exchanged through dialogue or bargaining.[48(p.2419),51]

In the GM labelling process, document analysis reveals that the term 'consultation' was used repeatedly by stakeholders, and refers to the gathering of information during the public hearings and providing comments on published drafts. Other more participatory mechanisms used included workshops, meetings and conferences, which opened up spaces for dialogue and more meaningful interaction. Innes and Booher[49(p.426)] assert that involving stakeholders to 'jointly recommend regulations', as in this case, is part of a collaborative participation process. Rowe and Frewer[48] call it 'active participation'. Active participation implies a process that is inclusive of stakeholders and places dialogue at the centre, rather than one that merely produces comments on already set-down proposed regulations. We can conclude that both consultation and active participation mechanisms – with open dialogue and two-way information exchange – were used by the government, through the DTI, during the participation process in developing policy governing the labelling of GM foods.

Inadequate representation

It is well recognised that the full range of stakeholders should be represented in policy development to lessen the probability that those on the 'periphery of the decision-making context or society are marginalised'.[47(p.2420)] This full range was not realised during the development of the mandatory labelling policy, with inadequate representation of farmer organisations and consumer groups. As noted by Parkins and Mitchell[50], such exclusions can well weaken such a process. In addition, representation in the process was restricted to a reduced set of interests comprising industry, government and NGOs that held a stake in and/or had knowledge of GM food labelling and dominated the policy space.[50,51,]

An interesting exception was the mobilisation of public interest NGOs to represent the diffuse and under-represented interests of consumers, and to ensure that there was a balance of lay versus expert participation.[51] As noted by Aerni[52(p.465)], the involvement of public interest groups in 'protest events…attracts the attention of the mass media and build[s] up public pressure on politicians to respond to [their] concerns'. Clearly, public opinion can influence and shape politics in developing countries such as South Africa, although, as Aerni[52(p.465)] argues, it may well be 'the opinion of academic, political, economic and traditional elites rather than the public at large that matters in such elite democracies'. In Kenya, for example, the food industry and politicians dominated the policy space by opposing new mandatory labelling regulations, cautioning that such labelling could raise food prices and reduce food security in the country.[43]

Perceived fairness polarised

Herian et al.[53] confirm that citizens who are given information about public participation procedures during policy development perceive the process to be fairer. If the participatory mechanisms used are believed to be 'transparent and consider [all the] conflicting claims and views', this could then boost 'public trust in decisions and civil society'[47(p.2420)].

In the case of GM labelling in South Africa, interesting schisms developed between NGOs and the industry, and, significantly and unusually, between government departments themselves. For example, the DOH and the DST, as well as the industry, distrusted the DTI, believing the process had not always been transparent, and that their views and input had not been considered. The weak and even dissonant relationships between the industry and the DTI, as well as between the DTI and other government departments, added to perceptions that the policy process had been unfair. NGOs, in contrast, perceived the process to have been transparent, for the most part, as they believed that they had been 'listened' to, and that their views and inputs had been considered. This belief enhanced trust in decisions and policymakers.[47]

While it is clearly important for public participation processes to be fair, transparent and inclusive, it is telling that perceptions of procedural fairness were limited to policy outcomes, rather than the policy process itself. This suggests that if schisms on controversial issues such as GMOs are not resolved prior to policymaking, then participatory measures in developing policies are doomed to fail from the start.

Towards improved policy development

Given the contested nature of GM food labelling, it is appropriate to ask what approaches might have worked better. A more deliberative democratic participatory policy process might have transformed the adversarial relationships between stakeholders. Many stakeholders, for example, reported experiencing consultation fatigue, because of perceptions that their involvement was ineffectual and had limited capacity to 'influence decisions'[47(p.2420)]. The consultations were believed to be 'talk shops', which produced uncertainty and hindered definitive activity.[47(p.2420)] The conflict between certain stakeholders obstructed authentic participation and discussion and resulted in ambiguity and delayed decisive collective action or consensus.

Democratic participation, as observed by Parkins and Mitchell[50], is where participation is restricted to 'voting and where public deliberation is severely limited to issue "sound bites" and popularity contests'[50(p.530)], which occur, for example, during public hearings. Under such circumstances, greater attention could have been given to creating deliberative spaces that deepened the discourse, improved the quality of decisions and produced concepts and knowledge that enhanced both expertise and understanding. As noted by Chambers (2003, cited in Parkins and Mitchell[50]), deliberative democratic participation offers and solicits debate and dialogue that can bring about rational, knowledgeable views, in which the participants are prepared to amend preferences in view of 'discussion, new information, and claims made by fellow participants'[50(p.530)]. By offering a collaborative and deliberative democratic participation process, the DTI could have provided an opportunity for meaningful stakeholder debate, personal reflections and an exchange of informed stakeholder opinion – an approach that could have transformed adversarial relationships between the stakeholders. An alternative and more cynical view is that any process would have met with resistance, because stakeholder positions and interests were founded on principle more than rationality. If stakeholders are not willing to negotiate their positions and interests and identify common objectives, then conflictual relationships will not be transformed.

There were certainly elements of the process that could have been improved. For example, the DTI could have replaced their 'tool-kit approach', which stresses picking the applicable tools for the task, with a method that looks at 'participation as a process'. This approach might have taken into account how to involve relevant stakeholders, the most suitable time to do so, and approaches for stakeholders to 'fairly and effectively shape environmental decisions'.[47(p.2422)] Best-practice participation can ensure that stakeholders have the power to influence policy decisions and have the technical competence to engage effectively.[47(p.2422),53-55] This issue is relevant to both developed and developing economies, which share poor knowledge about GMOs, but is likely exacerbated in a developing country context with low levels of literacy.[20] In this regard, efforts could be made to educate stakeholders, and develop their understanding and confidence.

Early interventions prior to the setting of policy, which could include one-on-one interactions, and active in-depth engagement between stakeholders and policymakers, would certainly help to strengthen the process, both through its design and stakeholder involvement. As noted by Reed[47], if participatory processes are to bring about 'high quality and durable decisions', then stakeholder engagement from the very start of the process is critical.[56,57] A stakeholder analysis is an important first step to methodically identify and represent those actors relevant to the decision-making processes.[47,57] The level of participation in the analysis should extend to active engagement in which there is a two-way transfer of information between stakeholders and policymakers.[47]

Conclusions

Stakeholders have a democratic right to participate in environmental decision-making, such as the policymaking process of mandatory GM food labelling in South Africa, as this policy could affect their everyday lives. At present, the CPA mandates the labelling of all GM foods in South Africa, because of a belief that a policy is needed to protect the rights of consumers. The findings of this paper, which are in line with Herian et al.[53(p.1)], show that perceptions of procedural fairness among stakeholders are a strong predictor of satisfaction with policy decisions. Similarly, the extent to which positions are accommodated influences perceptions on procedural fairness. Those that lobbied for labelling – NGOs and consumers – largely believed the process to be fair and inclusive, as well as consultative, although this view changed when their interests were not addressed. In a similar vein, those stakeholders against labelling – the industry, DAFF, DOH and DST – perceived the process to be fair when their point of view was accommodated, but unfair and exclusive, and the DTI's activities to be unacceptable, when their positions were not incorporated. Lower levels of capacity in developing countries, both with regard to the technical considerations of GM crops and the implementation of labelling legislation, underscore the importance of a public participation process that is believable by and inclusive of all actors. Our findings also emphasise the need to resolve larger policy questions regarding the adoption or rejection of GMOs, before the start of the participatory process.

Acknowledgements

Authors' contributions

T.d.B. conducted the research under the supervision of R.W. The research was co-conceptualised and co-written by both authors. The grant that enabled the research was held by R.W.

References

1. Klintman M. The genetically modified (GM) food labelling controversy: Ideological and epistemic crossovers. Soc Stud Sci. 2002;32(1):71–91. https://doi.org/10.1177/0306312702032001004

2. Bansal S, Ramaswami B. Labels for GM foods: What can they do? Econ Polit Wkly. 2010;45:26–27.

3. James C. Global status of commercialized biotech/GM crops. ISAAA brief no. 52. Ithaca, NY: International Service for the Acquisition of Agri-biotech Applications (ISAAA); 2016.

4. Zainol ZA, Nordin R, Akpoviri FI. Mandatory labelling of genetically modified (GM) foods. Int Environ Agreem-P. 2015;15(2):199–216. https://doi.org/10.1007/s10784-013-9230-5

5. Viljoen CD, Marx GM. The implications for mandatory GM labelling under the Consumer Protection Act in South Africa. Food Control. 2013;31(2):387–391. https://doi.org/10.1016/j.foodcont.2012.10.026

6. Gruère GP, Rao SR. A review of international labelling policies of genetically modified food to evaluate India's proposed rule. AgBioForum. 2007;10(1):51–64.

7. South African Freeze Alliance on Genetic Engineering (SAFeAGE). Input for Consumer Protection Bill (CPB). Letter to the portfolio committee of the DTI. 2008 August 08.

8. South African Freeze Alliance on Genetic Engineering (SAFeAge). Consumer rights recognised: GM foods will be labelled in South Africa [press release]. 2008 September 17 [cited 2017 Dec 08]. Available from: http://db.zs-intern.de/uploads/1221732671-Safeage-GELabelling.pdf

9. Gosling M. GM labelling clause reinstated in Bill. Cape Times. 2008 September 19 [cited 2013 Feb 20]. Available from: http://www.iol.co.za/index.php?set_id=1&click_id=125&art_id=vn20080919053532

10. African Centre for Biosafety (ACB). GMO watchdog to lodge complaint [press release]. 2012 March 20 [cited 2014 Jun 15]. Available from: http://www.acbio.org.za/index.php/media/64-media-releases/374-gmo-watchdog-to-lodge-complaint

11. Gosling M. Food fight over GMO labelling. Cape Times. 2012 April 10.

12. African Centre for Biosafety (ACB). Industry employing bullying tactics to scupper GM food labelling in South Africa [press release]. 2014 August 01 [cited 2014 Aug 10]. Available from: http://acbio.org.za/wp-content/uploads/2015/02/GM-Labelling_july_20141.pdf

13. De Beer TL. Stakeholder involvement in the development of genetically modified (GM) food labelling policy in South Africa [dissertation]. Cape Town: University of Cape Town; 2015.

14. Paarlberg RL. Governing the GM crop revolution: Policy choices for developing countries. (Food, agriculture, and the environment discussion paper 33). Washington DC: International Food Policy Research Institute; 2000. Available from: http://cdm15738.contentdm.oclc.org/utils/getfile/collection/p15738coll2/id/125379/filename/125380.pdf

15. Barton JE, Dracup M. Genetically modified crops and the environment. Agron J. 2000;92(4):797–803. https://doi.org/10.2134/agronj2000.924797x

16. Ahmad P, Ashraf M, Younis M, Hu X, Kumar A, Akram NA, et al. Role of transgenic plants in agriculture and biopharming. Biotechnol Adv. 2012;30(3):524–540. https://doi.org/10.1016/j.biotechadv.2011.09.006

17. Scoones I. Mobilizing against GM crops in India, South Africa and Brazil. J Agrar Change. 2008;8(2–3):315–344. https://doi.org/10.1111/j.1471-0366.2008.00172.x

18. Wynberg R, Fig D. A landmark victory for justice: Biowatch's battle with the South African state and Monsanto. Durban: Biowatch South Africa; 2013.

19. International Service for the Acquisition of Agri-biotech Applications (ISAAA). Biotech country, facts and trends: South Africa [infosheet on the Internet]. Ithaca, NY: ISAA; 2016. Available from: http://www.isaaa.org/resources/publications/biotech_country_facts_and_trends/

20. Botha GM, Viljoen CD. South Africa: A case study for voluntary GM labelling. Food Chem. 2009;112:1060–1064. https://doi.org/10.1016/j.foodchem.2008.06.050

21. African Centre for Biosafety (ACB). Understanding the impact of genetically modified crops in Africa: An activist's handbook. Johannesburg: ACB; 2012. Available from: http://acbio.org.za/wp-content/uploads/2015/02/ACB-Activist-Handbook.pdf

22. Department of Trade and Industry (DTI). Draft Green Paper on the consumer policy framework. Pretoria: DTI; 2004. Available from: https://www.gov.za/sites/default/files/26774_1.pdf

23. Department of Health (DOH). Viewpoint of the Department of Health on mandatory labelling of foodstuffs containing GMOs [press release]. 2008 September 18.

24. Department of Trade and Industry (DTI). Consumer Protection Bill: Response to public submissions [minutes]. Parliamentary Monitoring Group, South Africa; 2008 September 10. Available from: https://pmg.org.za/committee-meeting/9608/

25. South African Freeze Alliance on Genetic Engineering (SAFeAGE). Labelling of products and foods containing GMO's in South Africa [webpage on the Internet]. No date [cited 2012 Mar 17]. Unavailable.

26. Viljoen CD, Dajee BK, Botha GM. Detection of GMO in food products in South Africa: Implications of GMO labelling. Afr J Biotechnol. 2006;5:73–82.

27. Consumer Protection Bill B19–2008. Pretoria: Department of Trade and Industry; 2008. Available from: https://www.thedti.gov.za/parliament/bills/ConsumerProtectionBill.pdf

28. Department of Trade and Industry (DTI). Consumer Protection Bill: Presentation to the Select Committee on Economic and Foreign Affairs [presentation]. 2008 October 16. Available from: http://slideplayer.com/slide/9866348/

29. Wolson RA. Assessing the prospects for the adoption of biofortified crops in South Africa. AgBioForum. 2007;10(3):184–191.

30. Congress of South African Trade Unions (COSATU). COSATU submission on the Consumer Protection Bill [B19B–2008]. Letter to the Portfolio Committee on Trade and Industry; 2008 September 03.

31. Consumer Protection Act No. 68 of 2008. Government Gazette. 2008;526(32186). Available from: http://www.thedti.gov.za/business_regulation/acts/consumer_protection.PDF

32. Woker T. The Consumer Protection Act and genetically modified foods [presentation]. University of KwaZulu-Natal; 2009 Jun 1.

33. Industry Working Group. Industry comments on GM-labelling clause (Clause 24(6)), as included in proposed regulations for Consumer Protection Act (Act Nr. 68 of 2008), as published in Government Gazette, Vol. 545 of 29 Nov 2010, Nr. 33818 [private correspondence with DTI – PAIA request form document]. 2011 January 20.

34. Consumer Goods Council of South Africa (CGCSA). Comments and proposals on regulations required in terms of the Consumer Protection Act [private correspondence with DTI – PAIA request form document]. 2010 September 08.

35. AfricaBio. Annexure 1: Consumer Protection Act (Act Nr. 68 of 2008), proposed regulations pertaining to Clause 24(6) of the Act (Government Gazette, Vol. 545 of 29 November 2010, Nr. 33818). [Private correspondence with DTI – PAIA request form document]. 2011 January 28.

36. Mohlala M. Request for the Department of Trade and Industry's view on implementation of the genetically modified organisms provisions of the Consumer Protection Act and its regulations. Letter to Rob Davies, Minister of Trade and Industry; 2012 February 10.

37. Bega S. Foot-drag on GM labelling regulations is slammed. Saturday Star. 2012 February 18 [cited 2013 Nov 14]. Available from: https://www.pressreader.com/south-africa/saturday-star/20120218/281612417315437

38. Label GM foods. GM labelling saga update [Facebook post]. 2012 March 26 [cited 2013 Jun 24]. Available from: https://www.facebook.com/LabelGMFoods?fref=ts

39. African Centre for Biosafety (ACB). Unlabelled food tests positive for GM [press release]. 2012 February 15 [cited 2014 Jun 15]. Available: http://www.acbio.org.za/index.php/media/64-media-releases/373-unlabelled-food-tests-positive-for-gm

40. Consumer Goods Council of South Africa (CGCSA). Consumer Protection Act (CPA) and its reference to genetic modification (GM) labelling [press release]. 2012 March 16 [cited 2014 Mar 20]. Available: http://www.nestle.co.za/common/nestleimages/publishingimages/media/press_releases/20120316_consumer_protection_act.pdf

41. Vananda N. Request for corrections to be effected on regulation 7: Product labelling and trade description: Genetically modified organisms, of the Consumer Protection Act 68 of 2008. Letter to Rob Davies, the Minister of Trade and Industry; 2012 June 21.

42. Consumer Protection Act (CPA), No. 68 of 2008. Draft amendment regulations: Section 7. Government Gazette 568(35776): Government notice no. R.824. Pretoria: Government Printer; 2012.

43. Oh J, Ezezika OC. To label or not to label: Balancing the risks, benefits and costs of mandatory labelling of GM food in Africa. Agric Food Sec. 2014;3(8):1−8. https://doi.org/10.1186/2048-7010-3-8

44. Sikuka W. Republic of South Africa: Food and agricultural import regulations and standards – narrative. FAIRS Country Report. Global Agricultural Information Network (GAIN) report. Pretoria: USDA Foreign Agricultural Service; 2015.

45. Freidberg SE, Horowitz L. Converging networks and clashing stories: South Africa's agricultural biotechnology debate. Afr Today. 2004;51(1):3−25. https://doi.org/10.1353/at.2004.0063

46. Morris EJ, Thomson JA. Genetically modified crops commercialized in South Africa. In: Wambuga F, Kamanga D, editors. Biotechnology in Africa: Emergence, initiatives and future. Cham, Switzerland: Springer; 2014. p. 53−65. https://doi.org/10.1007/978-3-319-04001-1_4

47. Reed MS. Stakeholder participation for environmental management: A literature review. Biol Conserv. 2008;141(10):2417–2431. https://doi.org/10.1016/j.biocon.2008.07.014

48. Rowe G, Frewer LJ. Public participation methods: A framework for evaluation. Sci Technol Hum Val. 2000;25(3):3−29. https://doi.org/10.1177/016224390002500101

49. Innes JE, Booher DE. Reframing public participation: Strategies for the 21st century. Plan Theory Pract. 2004;5(4):419−436. https://doi.org/10.1080/1464935042000293170

50. Parkins JR, Mitchell RE. Public participation as public debate: A deliberative turn in natural resource management. Soc Nat Resour. 2005;18(6):529−540. https://doi.org/10.1080/08941920590947977

51. Dür A, De Bièvre D. The question of interest group influence. J Public Policy. 2007;27(1):1–12. https://doi.org/10.1017/S0143814X07000591

52. Aerni P. Stakeholder attitudes towards the risks and benefits of genetically modified crops in South Africa. Environ Sci Policy. 2005;8(5):464−476. https://doi.org/10.1016/j.envsci.2005.07.001

53. Herian MN, Hamm JA, Tomkins AJ, Zillig LMP. Public participation, procedural fairness, and evaluations of local governance: The moderating role of uncertainty. J Public Adm Res Theory. 2012;22(4):815−840. https://doi.org/10.1093/jopart/mur064

54. Richards C, Blackstock KL, Carter CE. Practical approaches to participation. SERG policy brief no. 1. Aberdeen: Macauley Land Use Research Institute; 2004.

55. Weber N, Christophersen T. The influence of nongovernmental organisations on the creation of Natura 2000 during the European policy process. Forest Policy Econ. 2002;4(1):1−12. https://doi.org/10.1016/S1389-9341(01)00070-3

56. Reed MS, Graves A, Dandy N, Posthumus H, Hubacek K, Morris J, et al. Who's in and why? A typology of stakeholder analysis methods for natural resource management. J Environ Manage. 2009;90(5):1933−1949. https://doi.org/10.1016/j.jenvman.2009.01.001

57. Lynam T, De Jong W, Sheil D, Kusumanto T, Evans K. A review of tools for incorporating community knowledge, preferences, and values into decision-making in natural resources management. Ecol Soc. 2007;12(1), Art. #5, 15 pages. https://doi.org/10.5751/ES-01987-120105

58. African Centre for Biosafety (ACB). Comments on the draft government notice on the consumer protection general regulations [private correspondence with DTI – PAIA request form document]. 2010 August 30.