Inaugural Address

by

Professor Ntebo Moroke
PhD in Statistics (NWU)
Professor of Statistics
Faculty of Economic and Management Sciences

Topic

"On Multivariate Analysis of High-dimensional data"

Wednesday, August 7, 2019

18:30 for 19:00

Programme

Academic procession
University Anthem

Scripture reading and prayer
Pastor ML Molefe

Word of welcome
by
Prof Sonia Swanepoel
Executive Dean: Faculty of Economic and Management Sciences

Introduction of Prof Ntebo Moroke
Prof Herman Van der Merwe: Deputy Dean Teaching and Learning

Inaugural lecture
"On Multivariate Data Analysis of High-dimensional data"

Presentation of certificate and congratulations
Prof M Setlalentoa
DVC: Community Engagement and Mafikeng Campus Operations

Closing and vote of thanks
Prof Babs Suruijlal: Deputy Dean Research and Innovation

Grace
Pastor ML Molefe

National Anthem
Academic procession

Dinner – Robinson Room

Prof ND Moroke - Bionote

Qualifications: Ntebogang Dinah Moroke was born at Zeerust, in Dinokana village of the North West Province. She attended her primary schooling at Keobusitse, and proceeded to Ramotshere and Ramatu High schools in Dinokana. Ntebo enrolled for her studies at the then University of North West in 1998. Her BCom degree in Economics and Statistics was awarded to her in 2001. She obtained her doctoral degree in Statistics in 2014 at the North West University, in the Faculty of Commerce and Administration. She took up a voluntary position as a mathematics and physical science tutor at Ramatu high school in 1997. In 2006 to 2011, she was appointed as a tutor by University of South Africa and taught some of the Statistics modules from first year to third year. In 2002 she was appointed as a part time junior lecturer at the University of North West in the Faculty of Commerce and Administration after completing her honours degree. She got a substantive position as a Lecturer in 2005 July, promoted to a Senior Lecturer in 2015, Associate Professor in 2016, and she was promoted to the ranks of a full Professor of Statistics in 2018.

Management: She was appointed as acting Head of the Department of Statistics from 2007 to 2011 and a Programme Leader from 2011 to 2015. Ntebo contributed immensely to curriculum development and review; and marketing of Statistics and Operations Research programs. Ntebo has contributed to the growth of the program both as an academic and a researcher. She played a key role in the external program evaluation in 2015, during which two commendations in relation to staff qualification improvement and research output were received. She was appointed as an acting Associate Research Professor in 2015 March to 2016 June, and was appointed substantively thereafter in 2016. Ntebo also acted as Deputy Dean responsible Community Engagement and Stakeholder Relations in 2017 in the Faculty of Economic and Management Sciences and got formal appointment in this position in 2018 to date.

Research supervision and publications: Since 2016, Ntebo has successfully promoted thesis of five doctoral students. Four of the candidates are members of staff in Statistics and Operations research program in the Faculty of Economic and Management Sciences. She has successfully delivered more than 10 Masters students (five are Statistics and Operations Research staff members). Ntebo is currently supervising 5 PhDs (3 staff) and 2 masters students. She has successfully mentored a number of staff members within the faculty, and from other faculties. Ntebo has published more than 30 articles in nationally and internationally accredited journals and has also delivered a number of papers at local and international conferences. Ntebo's research interest is drawn to analysing real life high-dimensional data applying Multivariate Data Analysis methods.

Committee membership: Ntebo is an active member of the South African Statistical Association (SASA). She served in several committees including Human Research Ethics committee, campus disciplinary committee, campus senate and institutional senate and campus resources audit committee. By virtue of her being the Deputy Dean, she is a member of Institutional senate, senate for research and innovation, senate for teaching and learning, a chairperson for community engagements committee, faculty board member, and other scientific committees in the faculty. Ntebo serves as a member of editorial board to review articles submitted to Journal of Statistics & Management Systems, Risk Governance and Regulations Journal, Scientific Research and Essays Journal, Journal of Modern Applied Statistical Methods, African Journal of Economic and Management Studies and African Journal of Information Systems. She is a scientific committee member of the International Business Information Management Association (IBIMA) conference since 2014 to date. She is a member of a panel that evaluate proposals submitted to National Research Fund (NRF) for rating and funding. She also serves as an external moderator and examiner for some of the local universities.

Awards and recognition: Two of the papers presented at international conferences received the most paper award in 2016 and best prize for journal award in 2017. Ntebo received knowledge share award in 2016 from NRF in collaboration with South African Statistical Association (SASA) for academic Statistics in crisis. She was a runner-up for the institutional most productive junior researcher in 2015, received a Rector's award for most productive junior researcher in 2015, received award for faculty's most productive junior researcher in 2014 and a recognition for emerging junior researcher in 2013. She also received an Institutional teaching excellence prestige award in 2012 and was nominated as the most inspiring lecturer sponsored by Rapport newspaper in 2012.

Table of Contents

Protocols

Vice Chancellor
Deputy Vice Chancellor,
Members of Council present,
The Executive Deans,
Pastor,
Deputy Deans,
School Directors and Deputy Directors,
Academic staff
Distinguished Guests,
Other functionaries present,
Students,
Ladies and Gentlemen,

Dumelang …Good afternoon…Goeie naand

I welcome you all to my inaugural address

1. Preamble

First, I give all the glory, honour and adoration to God for His manifold blessings upon my life and for granting me this opportunity to deliver this inaugural lecture. God has seen me through extremely trying times. I sincerely thank Him for providing me a safe passage through the vicissitudes of life and so many painful unwelcome episodes I have experienced to see this day. Most of all, I thank Him for reinventing me with excellent health. This occasion is a testimony to His goodness. I glorify Him for His Omnipresence from the cradle through the primary school and the university.

I welcome all of you to this inaugural lecture titled "*On Multivariate Analysis of High-dimensional data*". Before I proceed with my talk, I will like to acknowledge the presence of erudite and distinguished scholars who have presented their inaugural lectures from this podium. It is an honour and privilege that the NWU have bestowed this opportunity to stand before you to share my academic and research activities that landed me in this position. It is an honour for me to follow your paths and look forward to being part of more inaugural lectures in the near future.

I will in this lecture highlight a brief summary of my contributions to the field of statistics as an academician and a researcher. Madam Deputy Vice Chancellor, it is on this premise that I seek your permission to use this platform to share the activities of my career in statistics that led to my elevation and appointment as a Professor in Statistics.

It is a challenge for someone in my field to address a heterogeneous audience like this one without luring them to sleep with much of statistical expressions. I shall try to carry everybody along, specifically those whose background in statistics is on the periphery. Allow me to try as much as possible to make this talk less statistical, while at the same time urging you to up your residual appreciation of at least statistical symbols and the impressions they convey.

This lecture is organised into four sections. The second section looks into my journey into the Statistics field and I have titled this "the throw-back". The third section shares my academic, professional, research development and contribution in the field and I referred this to "throw in", and the last section titled "throw out" captures my contribution to staff development and marketing of the NWU. My future endeavours is outlined in Section 6 and Section 7 provides summary of the lecture.

2. Throw-back:

2.1 My encounter with and journey into Statistics profession

The reasons that led me to pursuing my career in statistics are very interesting. I came to the university with the intention to secure myself a spot in science. I was very much in love with the subject mathematics and wanted so badly to have a qualification in this field, but believe me, I had no further plans with this field for the future. Was told programs in Science faculty were full and had no other options. I was one of the students who did not submit application for admission on time, instead came to the university late in January during registration to submit my application. I became so devastated when we got turned away, fortunately on my way out of the university, I met someone I knew who suggested another option of a degree in Statistics.

She gave me an orientation around the calendar and told me that statistics is just like mathematics. Just the sound of the name "statistics" made me feel like I will never make it at the university. It was the first time I ever heard about such a name "statistics" and everyone advised me against registering for this qualification otherwise I will end up not graduating. Just because I did not want to go back home, I decided to apply and register the next day. I am a very good example of someone who followed a profession not prepared for. I realised with time that I have made the right choice of career because I was doing something very dear to my heart. I finished both my undergraduate and

honours within four years, and by then there was a serious staff turnover in the Department of Statistics, and I took an advantage of the situation.

In 2002, the Department of Statistics offered me a month's contract and three other junior lecturers were on three month's contract, reasons for this treatment still remain unknown to me. The then Dean kept telling me that I must try very hard to prove that I am worth it otherwise I will be out. Since that time to date, I have been working very hard to prove my worth and that is the most painful treatment I will never give anyone. I was on one month contract for six months, and in February 2003, I got an improved three months' contract which got renewed with another, and then two six months, obviously that came with a very heavy workload.

During that time, there was a staff complement of 5, all teaching full time and part time classes and supervising honours research projects. There were resignations and retirements and only three permanent staff were left. The department was still offering a taught masters course with six modules and mini dissertation. I decided to register masters in 2004 with four other students, taught by Prof Serumaga-Zake and part time Prof Arnab from University of Botswana. Everyone dropped out and I was the only one who persevered to the end. I finished and graduated in 2005 October and six months later was appointed on three year contract as a lecturer. I was appointed an acting Head of the Department in the interim in 2007, relieving a colleague who went on a study leave at the time. Upon his return, he decided not to continue with this position as he thought I was doing well. My appreciation to Dr Metsileng for his selflessness and for entrusting me with the responsibility of managing the department in his absence. I took up this responsibility, the name was changed to Program Leader after the merger. I requested to be relieved off my duties as Program Leader in 2015, as I wanted to focus on building research capacity in the program.

I enrolled for PhD in 2009, took a break for two years due to workload and supervisory issues. Thanks to Prof Sonia for joining the faculty and for giving me hope again. I managed to pick up the pieces of my studies and did PhD under very difficult conditions. I completed PhD in 2013 and graduated in 2014, and proud to say that I am the first person ever to graduate PhD in Statistics in this university on this campus. The grace of God located me again and found myself being the first and only female to be appointed as a lecturer and then a professor in the Department of Statistics. I am also the second longest serving staff in the department

3. Throw in:

3.1 Academic contribution

Madam Deputy Vice Chancellor, I feel a sense of duty to start within my constituency as my contribution to the growth of Statistics Profession; academically and professionally. The department used to enrol about 1200 students at first year including part time. We provided and still are providing service to other faculties across campus. Second and third year classes used to have very few students and that was worrisome for me, probably I did not understand the meaning of the description "scare skill" properly. We embarked on plans to attract more students into the program and that strategy saw us enrolling more than 60 and 30 students at third year and honours respectively, and the targets again declined due to issues of subsidy by the DHET. During those years, the university was not very strict on enrolment targets, hence departments could enrol as many students as they could. The workload was a bit overwhelming but that did not prevent us from obtaining good throughput rates. I taught modules across first year to honours, supervised honours projects, and was also responsible for the operations of the department.

I have contributed to the lecturing, development and review of study guides and of modules such as Introduction to Descriptive Statistics, Inferential Statistics, Categorical Data Analysis, Multivariate Techniques, Multivariate Data Analysis, Applied Regression Analysis, Econometric Methods, Time Series Analysis and Financial Statistics. I developed and introduced Statistical Computing module in 2013. This module introduces students to the use of different commercial statistical software packages in data analysis such as SAS and SPSS. Without using a computer,

one cannot perform any realistic statistical analysis of big data set, and this is something that I still strongly feel that the university is owing me as an undergraduate student. I was taught only the theoretical part of the aforementioned modules and others, no practical at all, but I do not blame my poor lecturers as I now understand what they must have been going through. We used to rely only on on-line statistical calculators like Excel to perform statistical data analysis. This software have some challenges in that: they are slow and depend on the cyberspace connection, and the more serious problem is that they are very limited and are nowhere equal commercial to off-the-shelf statistical packages. Furthermore, the functions in Excel are poor, and they often return the answer "#NUM", which simply means that the algorithm Excel is using has crashed. I had to change this mentality and I remember having discussions with my colleagues during the departmental meeting that we need to prepare our students as data analysts and that can only happen if we incorporate practical into our curriculum.

I was also motivated and at the same time challenged by the comments from former students that their lack of application of SAS and SPSS is putting too much pressure on them as Statisticians. That alone made me feel like I am failing my own students. Statistical packages require a high speed functional computer and this is still a challenge we are facing as Statisticians on this campus. Available computers available do not have a capacity to accommodate the type of statistical packages we require as Statisticians. Even the available packages are incomplete with lots of advanced modules. Madam DVC, how do we prepare ourselves and students for the most topical "4th industrial revolution era" as Statisticians when we are still stuck in the second era?" Just to quote one other comment and recommendations from the EPE panel;

The panel had the following concerns*: "ICT in the school is lacking. The panel identified a lack of computer facilities, particularly in light of the fact that many of the students on the Mafikeng campus do not have their own computer. Students referred to problem with computer labs, including problems with maintenance and coordination, and lack of internet access. Students sometimes have to supply their own data bundles. Student referred to incidents of power going off and interrupting tests on eFundi".*

Recommendations *"A dedicated computer lab for statistics students is strongly recommended, with suitable advanced statistical software for Statistics students. Consideration could be given to moving to free software, as statistical software can be very expensive (e.g. SAS)."*

Some of the concerns raised have been sorted but the major one of dedicated computer lab has not been addressed to date. I am, very grateful to the university department of IT for their support with the Wi-Fi, SAS, Statistica, SPSS and some basic computers available. Also my appreciation to my former lecturer and colleague Dr Johnson Arkaah for his immense contribution in training students and staff to use SAS software.

As a former lecturer and program leader, I always tried to keep communication going with my former students to get feedback from them in terms of their work performance and to investigate if the skills we imparted on them are utilised efficiently. I also sought for advices from them about the relevance of our programs to the industry, and believe me, our programs grew from strength to strength due to students' comments and suggestions which we implemented during module and program reviews, hence the commendations received from the EPE panel. My former students are not just graduates from the Statistics and Operations Research programs, they are also ambassadors who help with marketing and selling these programs. The relationship we currently have with most of the partners in the industry is due to referrals and recommendations by our former students. They never forget where they came from. I do not have current statistics, but I can confirm that most of the students who graduated were absorbed by StatsSA, JSE, Eskom, SARB, SAS, some of the banks and some government departments as interns, and some on substantial higher positions and the feedback we received from these partners have always been very positive. I quote an observation and commendation from the EPE panel:

"**The panel observed**: *The panel was satisfied that the programme is appropriately balanced between theory and application and produces graduates with a rigorous background in statistical principles: probability, applied statistics methods, the development of data, and analytic and statistical skills.*
The department has been involved with organisations who have given presentations to the programme (staff and exit level students) for some time. The idea is to keep the programme relevant and expose the students to what industry wants. Site visits have been planned to some of these organisations: Eskom, Johannesburg Stock Exchange (JSE), ABSA, FNB, National Treasury and Statistics South Africa".

*"**Commendation**: The consultation with stakeholders to keep the course relevant to prospective employers is commendable"*
The panel was impressed by the quality of the students whom they interviewed. They are very committed and loyal, and consider themselves to be ambassadors of the programme. They are willing to be involved in recruitment initiatives".

My passion for teaching and sharing my time with students earned me some formal and informal recognition. I received an institutional teaching excellence prestige award in 2012 and was nominated as the most inspiring lecturer sponsored by Rapport newspaper in 2012.

3.2 Professional contribution

My professional contribution in here refers to formal and informal consultation services rendered with regards to my area of expertise. I started offering free statistical consultation services to students as soon as I graduated honours degree. I used this as an opportunity to further enhance my statistical knowledge as most of the services rendered required a vast application of different statistical methods to do data analysis. The kind of advices I gave to people contributed to building my confidence in conducting data analysis. Among other services provided, I used to advice and still am advising mostly postgraduate students and researchers on how to conduct their surveys; the issues related to the population and sample, the type of questions to ask, collection of responses, methodological issues up to data analysis and results. I also informally offered training to people on how to use SPSS since it was the only package available on campus; and that helped shape my analysis skills and self-reliance on the use of this package. Of late, time has been an impeding factor due to my growing responsibilities as a researcher and manager.

Myself and other postgraduate students were contracted by the North West Parks and Tourism board in 2002 to develop a survey instrument which was used to investigate the impact of tourism establishments in the growth of the North West Province. We were also involved in the survey as data collectors around the entire province, and rendered statistical services, compiled and presented the report. I was later approached to serve as an intern working with the statistician in this department on *ad hoc* basis and that was not a problem for me since I was still part time employed by the university. In 2006, I was contracted by Gaobotse consulting company as a Senior Statistician with other team members to conduct a survey in the villages around Mafikeng airport. This also served as a good training for me as questionnaire administrator and researcher.

I have always ensured that my services are available to anyone who need them. My interaction with aforementioned organisations and some colleagues at this level enhanced an appreciation for collaborative studies. The then School of Economic and Decision Sciences received award for most productive school for three consecutive years, 2014, 2015 and 2016. I am proud to say that the department of Statistics contributed more research units in the school during these years and these are the years in which I was acknowledged the faculty's emerging junior researcher, awarded most productive junior researcher, received nomination for institutional most productive junior researcher and also won the rector's award for most productive researcher. The interaction and collaboration with other people did not only help accelerate the progress and extend the breadth of my and their knowledge, it also helped

enhance the quality of my skills and extended the repertoire of my partners (students). Since my graduation in 2014, I have received and still am receiving invitations from different people to join their research groups and I am approached by many students and staff to act as their study leader and mentor.

Critically reading other researchers work has also sculpted me into becoming a better researcher. I have been serving as an editorial board member of several multidisciplinary and interdisciplinary journals and conferences since 2014 to date. I have been exposed to studies from diverse fields and that has abetted me in finding synergies in the application of different statistical methods to other fields. I am also kept abreast of latest trends in different methods researchers are using and in some instances I come across papers that have used statistics incorrectly. My interaction with different people has further exposed me to some myths and misconceptions, or shall I say poor perception of statistics and statisticians. I realised that there is little awareness on the need to involve the statistician at all stages of investigation; *i.e,* designing of survey tool, data collection, analysis, and presentation of research results. One of the challenges statisticians are faced with is when researchers come with finished experimental data, demanding analysis that should conform to their expected results, and could express disappointment if otherwise. In most cases, I found myself being confronted with researchers who bring a set of data and *demand* the use of a list of statistical tests and/or techniques for such data. Some simply want "certified P-value, reliability statistic, correlation coefficient from qualitative responses, etc.", while others even come for interpretation of already analysed data (at times wrongly analysed).

3.3 Research development

Due to my involvement in aforementioned survey studies, my attention was drawn to analysing complex data which obviously requires the application of a plethora of Multivariate Data Analysis methods. In most cases, basic descriptive and inferential statistics are used to summarise the responses and conduct basic statistical tests. I am very fascinated by the interplay of variables in multivariate data and by the challenge of unravelling the effect of each variable hence my focus has always been on multivariate data analysis. I am passionate about playing around with different multivariate techniques obviously taking their theories into account to analyse high-dimensional data. My continuing passion is to present the power and utility of Multivariate Analysis in the form of research. I also enticed this interest to my students who most of them if not all also applied these methods in their studies. This then became an area of research focus not only for me but also for the department. As such, my research evolves around exploring different multivariate data analysis techniques to high-dimensional or complex real life data from different fields of studies.

Other reasons that led to this focus area was and still is the few studies published in the area. In most studies, researchers treat variables one by one using conventional univariate designs. You can imagine the time one will take to perform such analyses on more than 10 variables individually instead of blending in all of them concurrently. In many instances, the variables are intertwined in such a way that when analysed individually, they yield little information about the system. Using multivariate analysis, the variables are inspected instantaneously in order to evaluate crucial features of the process that produced them. "Multivariate Analysis is conceptualised by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding" (Olkin, 2001).

As stated by Rencher (2003) and Hair *et al*. (2010), the multivariate approach enables the analyst to "explore the joint performance of the variables and to determine the effect of each variable in the presence of the others". Multivariate analysis has a provision both for the descriptive and inferential procedures. A search for patterns in the data or test of hypotheses about patterns of *a priori* interest can still be done with the application of multivariate analyses. One can, using multivariate descriptive techniques peer beneath the tangled web of variables on the surface and extract the essence of the system. Moreover, multivariate inferential procedures include hypothesis tests that process any number of variables without inflating the *Type I error rate* and further allow for whatever interconnections

present between the variables. Multivariate analysis also provide for methods useful when the researcher intends to do classification of objects or subjects, estimations of models and projections into the future.

All data collection processes yield high-dimensional multivariate data and much computational effort is required to do analysis. Pituch and Steven (2016) opined that the "use of multiple criterion measures can paint more complete and detailed description of the phenomenon under investigation". Since large data sets not only contain many observations, there are also multiple variables of different measurement scales which can easily and safely be handled with multivariate data analysis methods. Most multivariate methods afford an opportunity to examine the phenomenon under study by determining how multiple variables interface. Most researchers are comfortable to use modest, extensive pragmatic methods and circumvent from dabbling in complicated and sophisticated methods.

Few decades ago, understanding of multivariate analysis was obviously beyond the reach of all but the most highly mathematically trained educational researchers and this explains why there is dearth of literature around studies that applied multivariate analyses to date. This should no longer be the case since technological changes are advancing rapidly. There is an increased availability of sophisticated statistical software packages which my students and I are exploring to analyse high-dimensional data sets. Owing to the size and complexity of the underlying data sets, much computational effort is required. With the continued and intense growth of computational power, multivariate methodology plays an increasingly important role in data analysis, and multivariate techniques, once solely in the realm of theory, are now finding value in application. Even though technological advances are capacitating us to move beyond manual analysis of data, one should also take note of the following challenges which are associated with big data;

- they increase the time needed to capture all variables,
- they increase the cost of the investigation,
- they make the analysis of the data complex and at times impossible, and hence,
- the large number of variables adds another difficult conceptualization layer/level and interpretation level on the normally accepted and understood levels by a common human mind.
- the use of multidimensional data may render the whole investigation process difficult or worthless.

Figure 1 below shows variety of Multivariate Analysis Techniques that are at the researcher's disposal;

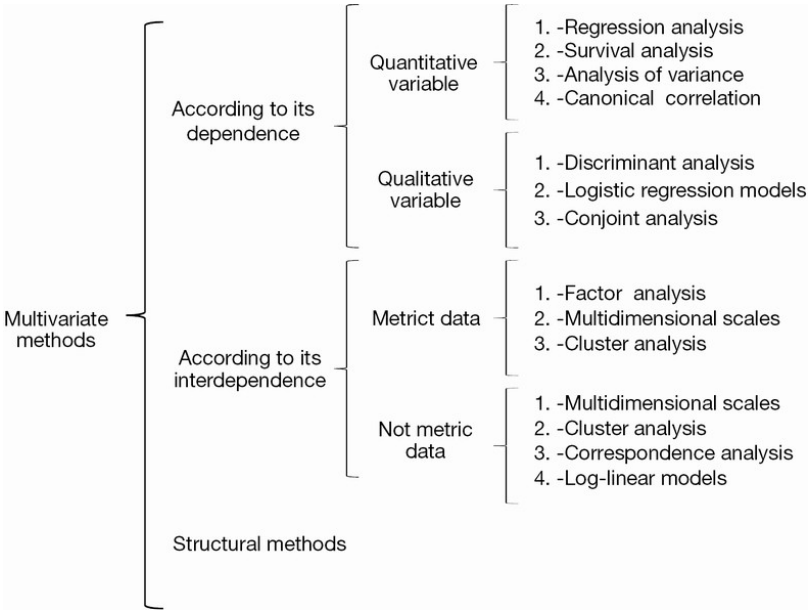

Figure 1 A Guide to using Multivariate Techniques

Each multivariate technique is chosen based on a specific research objective for which it is best suited. All statistical techniques have certain pros and cons that should be clearly understood by the analyst before attempting to interpret the results. Statistical packages such as SAS, SPSS, S-Plus, Python, Matlab, R, and others, make it progressively easy to run a procedure, but the results can be disastrously misinterpreted if the analysis is not attentive.

3.4 Contribution to the field of statistics

In this section, I give a glimpse of some of the papers published on my research focus area. I have explored the performance of numerous multivariate techniques with application of data from social, behavioural, economic and financial, marketing and education sciences. Summary of some of the papers published is given in subsequent sections. Prior to the fundamental analysis, it is important to consider the data, due to the effect the characteristics of the data may have upon the results. Data screening is imperative as decisions at the earlier steps influence decisions to be taken at later steps. Screening of the input data helps to assess the appropriateness of the use of a particular data set. This process aids in the isolation of data peculiarities and further allows the data to be adjusted in advance of further multivariate analysis. The checklist below isolates key decision points which need to be assessed to prevent poor data induced analysis problems. Consideration and resolution of problems encountered in the screening of a data set is necessary to ensure a robust statistical assessment.

Pertinent issues in the data:

- check effect size
- inspect the data for outliers and missing values, be aware of measurement scales
- identify and deal with non-normal variables
- evaluate homogeneity of variances and covariances within the groups
- check the data for non-linearity and heteroscedasticity
- evaluate variables for multicollinearity and singularity

The critical assumptions underlying multivariate analyses are more conceptual than statistical. From a statistical standpoint, the departures from *normality*, *homoscedasticity*, and *linearity* apply only to the extent that they diminish the observed correlations. Only normality is necessary if a statistical test is applied to the significance of the factors, but these tests are rarely used (Hair et al., 2005).

The purposes of the studies discussed in subsequent sections is to give an illustration of the more easily accessible multivariate analysis techniques in an effort to demonstrate the value of moving beyond the commonly used univariate techniques. The findings of the studies is manifold:

- to use the findings in providing recommendations for policy purposes and decision making,
- to set a baseline for scholars who are interested in analysing high-dimensional data,
- to provide readers with the supporting knowledge necessary for making proper interpretations
- to provide foundation to various researchers to select appropriate technique for their fields
- to create an understanding to scholars the strength and weaknesses of the techniques
- to contribute to the dearth of literature about the applicability of multivariate analysis methods

I was motivated by the comments of one of the external examiners in 2013 who along with students' reports, sent a copy of her paper that has just been published in the field of multivariate analyses. On her report, she recommended that I summarise the research reports into publishable research papers. Since I was still a novice when it comes to publishing, I wrote my first paper and requested a colleague of mine who was also my mentor to tear it apart and provide feedback. A big thank you to Prof Mavetera for his patience and constructive comments.

Overview of papers

The first paper under my name was a product of my collaboration (Moroke and Mavetera, 2013). We used Exploratory Factor Analysis (EFA) method to investigate inter-relationships amongst serious crimes in South Africa, with an ultimate aim of grouping similar crimes together. This study provided a modest way of employing the otherwise theoretical aspects of factor analyses (FA) in real life high-dimensional data. EFA was used to classify and group the observed crime variables into few unobservable proxies called factors. According to Manly (2001), in order to measure crime in a satisfactory manner, different categories of crime need to be classified and separated into groups of similar or comparable offences. Hair *et al*. (2010) also recommend the application of EFA as it considers all variables simultaneously "each relating to others but employing the concept of the variate".

Hair *et al.* define EFA as a multivariate technique for identifying whether the correlation between a set of observed variables stems from their relationship to one or more latent variables in the data, each of which takes the form of a linear model. Johnson and Wichern (2007) explain the purpose of EFA as being the description, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called factors. Rencher (2002) refer to this technique as one-sample procedure for applications to data with groups; that is, assuming a random sample $(y_1, y_2, \dots, y_n)$ from a homogeneous population with mean vector μ and covariance Σ matrix. Each variable$(y_1, y_2, \dots, y_p)$ in the random vector *y* is assumed to be a linear function of *m* factors $(f_1, f_2, \dots, f_m)$ with the accompanying error term to account for that part of the variable that is unique. The model showing linear combination of factors is written as:

$$y_1 = -\mu_1 + \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1m}f_m + \varepsilon_1 \quad y_2 = \mu_2 + \lambda_{21}f_1 + \lambda_{22}f_2 + \cdots + \lambda_{2m}f_m + \varepsilon_2$$
$$\vdots \qquad\qquad \vdots$$
$$y_p = \mu_p + \lambda_{p1}f_1 + \lambda_{p2}f_2 + \cdots + \lambda_{pm}f_m + \varepsilon_p. \qquad [1]$$

Exploratory Factor Analysis attempts to bring inter-correlated variables together under more general, underlying variables. From (1), the factors are unobservable and this distinguishes the factor model from the multivariate regression model. The error terms are independent of each other such that $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma_i^2$. The factors f_i are independent of one another and also do not depend on the error terms such that $E(f_i) = 0$ and $var(f_i) = 1$. The sample variance of a variable y_i is defined by:

$$\delta_{ii} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \Psi_1, \qquad [2]$$

where communality $h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \Psi_1$. Hatch (1994) defines communality as the variance in observed variables accounted for by common factors. The goal of EFA is to reduce "the dimensionality of the original space and to give an interpretation to the new space, spanned by a reduced number of new dimensions which are supposed to underlie the old ones" (Rietveld and Van Hout 1993), or to explain the variance in the observed variables in terms of underlying latent factors" (Habing 2003). Thus, EFA offers not only the possibility of gaining a clear view of the data, but also the possibility of using the output in subsequent analyses (Field 2000; Rietveld and Van Hout 1993).

Listed below are some of the benefits of applying dimensionality reduction methods such as EFA to a dataset:

- space required to store the data is reduced as the number of dimensions comes down
- less dimensions lead to less computation or training time
- some algorithms do not perform well when large dimensions are used, as such, reducing high dimensions necessitates the usefulness of the algorithm
- the problem of multicollinearity is taken care of by removing redundant features.
- it helps in visualizing data. It is very difficult to visualize data in higher dimensions so reducing the space to 2D or 3D may allow plotting and observing patterns more clearly

In another study published in 2015, I applied a *Two-Step Clustering Algorithm* to the same data to confirm the results obtained using EFA. Just like EFA, the use of this method allows an analyst to have a different perspective on the data with no preconceived ideas regarding profiles (unsupervised learning), similarities, or performance measures. Clustering methods can safely be used for pattern recognition and data segmentation. Consequently, this

algorithm reduces high dimensionality of the data by collecting the variables into fewer dissimilar clusters. Few studies such as those conducted by Thanassoulis (1996); Yin *et al*. (2007); Leonard & Droege (2008); Rege *et al*. (2008); Kim *et al*. (2009); and Thaler *et al*. (2010) Po *et al*. (2009) have applied different clustering methods in different fields of studies. Lattin *et al*. (2003) defines this technique as general element in stopping rules which measures the diversity of all the observations across all clusters.

When performing cluster analysis, the observations are grouped by taking distances and similarities into consideration (Rencher and Christensen, 2012). As variables are being clustered, the analysis becomes more descriptive than predictive as the main concern is relationships in the data set. Therefore, no condition for linearity of the relationships among variants is assumed (Atlas *et al*., 2013). Cluster analysis is not dynamic but rather a static method used for describing current situation. This method is therefore not convenient in estimation analysis.

The variables used in the two studies were the 29 serious crime ratios per 100 000 of the population across the 1119 police stations in all the 9 provinces in South Africa. The data was collected during the period of financial year 2009/2010. A total of 2 121 887 serious crime cases were registered in SA across all the nine provinces during the time. The source of this data is the national office of the South African Police Service website at www.saps.gov.za. The variables were measured in metric scale hence EFA and Two-step Cluster methods were used.

Preliminary analyses addressed pertinent issues, KMO_MSA (0.948), suggesting that the degree of common variance between the 29 variables is marvellous entailing that if FA is conducted, the factors extracted will account for a significant amount of variance. The test also confirmed that the sample used was adequate. Moreover, the overall Cronbach's alpha (0.912) was an indication of strong internal consistency among the crime variables. The corresponding determinant of the correlation matrix was greater than 0.000. A logical conclusion then was that the data was appropriate for EFA and that the degree of multicollinearity between these variables was not severe. The 29 observable variables were explained by four factors from the use of EFA and four clusters from a Two-step clustering method. These new composite variables can be used for further analysis with the application of available dependence Multivariate Analysis methods.

The SAPS may use the results of this study when reporting on national crime statistics as well as improving the forms used to report crime. The results can also help magistrates in determining appropriate sentences for crimes committed. Legal authorities may also refer to the findings when developing interventions tailored to meet the needs of individual cluster of crimes. More emphasise can be placed on crimes that pose a serious threat. It is noted that a lot of money, time, and resources may be saved if the results of this study are considered. The study also contributed to the existing literature in the field of Multivariate Data Analyses.

Montshiwa and Moroke (2014) used factor analysis to assess the reliability and validity of student-lecturer evaluation questionnaire used at the North West University. It was impressed on us that no statistical testing was done on this Optical Character Recognition (OCR) based questionnaire before implementation and we took advantage of this study. The questionnaire was only piloted during the second semester of the year 2011 after being round robin to academics across the University for their inputs before it could be finalised. A 26 item OCR based questionnaire was distributed to 442 registered and available statistics undergraduate students towards the end of the academic year of 2013. The collection of data was reinforced by a designated member of the then Academic Development Centre, now called the Centre for Teaching and Learning, who explained to students the purpose of the questionnaire and also helped in administering it. In order to encourage honest responses to a somewhat sensitive subject, students were assured that their anonymity would be observed and that the results of the study would be used for research purposes only.

A return rate of about 68 % was achieved. Preliminary data analysis results provided enough evidence to conclude that the selected sample was adequate, the constructs conformed to reliability issues, and no issues of multicollinearity were noted. Exploratory factor analysis re-arranged the student-lecturer evaluation questionnaire collecting the 26 statements into four factors instead of the original five (preparation, presentation, relationship with

students, assessment and subject content). The exploration of the scale properties provided important an information on the NWU OCR instrument used for lecturer evaluation. The findings revealed that the new factors as they are designed have an acceptable level of internal consistency. Even though the survey was multidimensional, each variable appeared to focus on a particular latent factor. The results obtained in this study were presented to ADC and it was upon the office to take further steps. The application of EFA on the NWU OCR instrument provided ADC division more insight into the instrument used for evaluating lecturers. This method was recommended for use by different organisations or researchers who intend evaluating their survey instruments.

One other study by Moroke and Pulenyane (2014) used the Hierarchical Agglomerative Clustering (HAC) technique to determine clusters of leading death causes in South Africa. The paper presented an exploratory method for investigating the structure underlying the data. The AHC algorithm is effective when applied to metric or a metric or mixture of the two variables. Apart from identifying similarities and developing subgroupings of homogenous entities, HAC method may also be worthwhile in finding the true groups that are assumed to truly exist and may also be useful for data. This involves determining differences and similarities among multidimensional space. Inconsistencies and complexities in the data may be addressed when this tool is used as a precursor in data analysis. Owing to the inherent complexity in multivariate data, it is often desirable to find relationships among a suite of variables from which patterns or structures can be determined. This may be done either to gain a thorough understanding of outcome variables or to develop groups that can be subjected to further analyses (Cross, 2013).

Although SA has a functioning death registration system, the quality of cause of death data has been questioned. This is due to data from demographic surveillance studies using verbal autopsies to determine cause-specific mortality according to Adjuik *et al*. (2006). The available system does not indicate or identify the more deadly death causes or does not show groupings of these causes accordingly. World Health Organization (WHO)[1] is constantly monitoring improvements of data on causes of death. Some of the causes may be similar but the ICD-10 coding system is unable to identify them. As a result, this study used HAC algorithm to investigate these causes and attempted to identify the similarities and differences between the diseases. The HAC help in eliminating the duplication of the recordings and the new clusters may also help doctors and other responsible authorities when finding cure for certain group of deadly diseases.

The data used in this study is records on mortality and causes of death in SA collected by the Department of Home affairs. After verifying and validating the data using the framework proposed by Mahapatra *et al*. (2007), Statistics South Africa (Stats SA) head office published the data with compact disc and has it available to users on request. The original list consisted of 1079 mortality and death causes which took about 572 673 lives in the country. The data was recorded from January to December of the year 2009 by age, sex, population group, marital status, place or institution of death occurrence, province of death occurrence and province of usual residence of the deceased (Stats SA release, 2010). Prior to data analyses, the data was standardised using z-scores. This is a relevant transformation when the data is measured on continuous space. There were about 50% cases with blank spaces implying that no death was recorded against that disease in that particular year. After filtering the data, only 527 variables were left and used in the analysis.

A dendogram of a single linkage method from the HAC revealed the five clusters of diseases formed from the 537 leading death causes that claimed lives of 572 673 people in South Africa during January to December 2009. These death causes were collected in clusters according to their significant impact. Respiratory tuberculosis and pneumonia appeared to be the main leading causes of death followed by diarrhoea, stroke and heart failure.

[1] *The United Nations agency coordinates international health activities and helps governments improve health services. This agency checks the quality of the identification of causes of death and the coding system used in a particular country.*

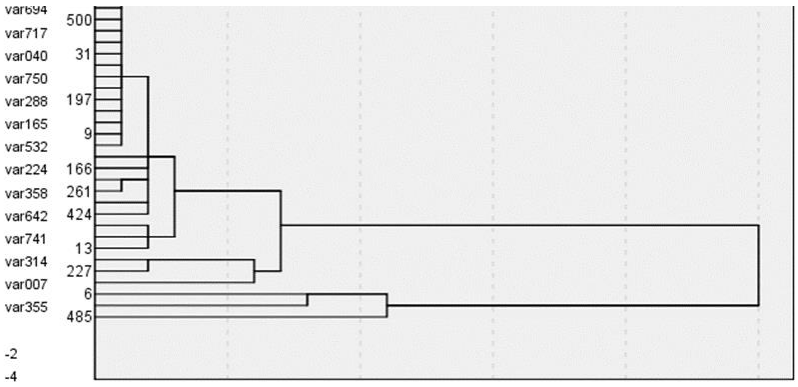

Figure 2 A Dendogram of leading death causes

Discriminant analysis was used in this study to confirm the convergence and classification validity of the 537 dearth causes into identified clusters. This Multivariate Analysis method helps in confirming if variables converge together as expected (convergent validity) and as reported by the analysis. Furthermore, a convergent validity statistic provides a basis for verifying the statistical significance of each cluster. The convergent validity statistic ranges between zero and one, with a value closer to zero denoting a high level of significance of that cluster.

Table 1 Wilk's Lambda from leading death causes

Test of clusters (s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	.001	3448.420	48	.000
2 through 4	.026	1916.343	33	.000
3 through 4	.152	994.488	20	.000
4	.588	280.044	9	.000

Only the first four clusters were reported to be highly significant according to the observed probabilities associated with Wilk's Lambda in Table 1.

Table 2 Eigenvalues from leading death causes

Cluster	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	17.254[a]	67.5	67.5	.972
2	4.741[a]	18.5	86.0	.909
3	2.874[a]	11.2	97.3	.861
4	.700[a]	2.7	100.0	.642

a. First 4 canonical discriminant functions were used in the analysis.

From Table 2, the canonical correlation coefficient of the first cluster implies that there is about 97% of convergence between the variables in that cluster also suggesting the closeness of these variables to one another. Moreover, the variables in the first cluster explains about 68% of total variation (contribution of the variables in this cluster). Conversely, variables in cluster four contribute about 2% in that cluster. Upon observing the classification status of the diseases to clusters, an apparent error rate (AER) of 0.04%, implies that most of these diseases are correctly classified (correct classification rate of 99.6%) as members to the suggested respective clusters. One of the diseases was incorrectly classified in the second cluster.

While long-term plans can be secured for death causes in the fifth cluster, it is important to pay special attention to diseases in the first four clusters urgently, more specifically those in the first cluster. The Department of Health in South Africa can channel more resources towards the alleviation of these deadly diseases. This may reduce death rates in the country. Other Multivariate Analysis methods may be used to further the analyses.

Another study by Mahole, Moroke and Mavetera (2014) used Multidimensional Scaling approach to study "poverty levels among Local Municipalities in the Ngaka Modiri Molema District Municipality" This district municipality is one of the four district municipalities of the North West Province of RSA comprising of five local municipalities, namely: Mafikeng, Ratlou, Ramotshere Moiloa, Ditsobotla and Tswaing. The Ngaka Modiri Molema District Municipality (NMMDM) is a predominantly rural region, where the majority of its population live. The communities residing in these

rural areas are more severely affected by aspects such as poverty and unemployment. These communities suffer from low levels of income and spending power, which results in very low standards of living. The identification and formulation of appropriate strategic plans and policies of poverty eradication by the NMMDM is determined by the appropriateness of the available information. However, the NMMDM does not have at hand data and information regarding the poverty levels in these municipalities. In addition, they do not know, if any, the similarities amongst these municipalities with regard to poverty levels. It is proposed that Multidimensional Scaling (MDS) approach be used to classify five local municipalities of the NMMDM into groups according to their similarities in poverty levels.

Multidimensional scaling, just like factor and cluster analyses, is an exploratory data analysis tool used to condense a large amount of data and presenting it in a simple spatial map. This map communicates important relationships in the most economical manner (Mugavin, 2008). The author further emphasized MDS as having several advantages such as modelling relationships among variables. Svetlana *et al*. (2013) defines MDS as a set of techniques for analysis of proximities (similarities or dissimilarities) that reveal the structure and facilitates visualization of high-dimensional data. The technique is descriptive and the idea of statistical inference is almost completely absent.

Multidimensional scaling technique does not require adherence to multivariate normality and have been found effective in extracting typical information in data exploration according to Johnston (1995) and Steyvers (2002). Giguère (2006) and Tsogo *et al*. (2000) suggest MDS when the study sought to find structure in the data. MDS is useful when the researcher wishes to find a spatial configuration of objects. The underlying dimensions extracted from the spatial structure of the data are thought by Ding (2006) to reflect hidden structures, or important relationships within it. This procedure is achieved by rescaling a set of dissimilarities measurements into distances assigned to specific locations in a spatial configuration. The more the points are closer together on the spatial map, the more similar are the objects. As a result, a visual representation of dissimilarities (or similarities) among objects, cases, or more broadly observations, will be provided (Jaworska and Chupetlovska-Anastasova, 2009).

The first step when doing MDS analysis is to produce matrix of distances between n objects. The number of dimensions for the mapping of objects is fixed for a particular solution. The following are general steps for carrying out a MDS analyses:

Step 1: Set up the n objects in p dimensions, where coordinates $x_1, x_2, \dots, x_p$ are assumed for each object in p-dimensional space. In this study, n = 10 variables measuring poverty: personal income, per capita income, per household income, disposable income, unemployment, formal dwelling backlog, sanitation backlog, water backlog, no electricity and no formal refuse removal, and p is determined from the analysis of the results. Poverty levels of the five local municipalities in the NMMDM were studied as a reflection of poverty levels in the whole population of the North West Province. Stratified random sampling method was used, where the population was partitioned into non-overlapping strata, which were independently sampled. The value of p is unknown at this stage.

Step 2: Euclidian distances between the objects are calculated for the assumed configuration using the formula:

$$d_{ij}^2 = \sum_{i=1}^{p} (x_{ij} - y_{ij})^2, \qquad [3]$$

where d_{ij}^2 is a square of Euclidian distance between points i and j, x_{ij} and y_{ij} are the coordinates on an axis. These objects are grouped according to the similarities between them.

We were generally interested in uncovering any structure or pattern that may be present in the levels of poverty among the five local municipalities of the NMMDM, in particular to identify the dimensions on which MDS makes its similarity in judgements. A number of studies applied MDS in different fields such as marketing (Steffire, 1969; Neidell, 1969), computer studies (Green and Carmone, 1969; Venna and Kaski, 2006), data mining and exploration (Silva and Tenenbaum, 2004; Groenen and Velden; 2004, Zhang, 2010) among others.

The preliminary results and assumptions of Multivariate Analysis were observed. The MDS results according to Table 3 revealed that the five municipalities can be grouped into two dimensions according to their poverty levels.

Table 3 Dimensionality of poverty variables

Dimension	Eigen Values	Cumulative %
1	7.104	71.039
2	2.376	94.800
3	0.464	

The two poverty dimensions accounted for 94% of variance according to the eigenvalues. To judge the goodness-of-fit of MDS, the study used an observed Standardized Residuals Sum of Squares (STRESS 1) (0.00027). This value in Table 4 was found to be appealing since it was less than conventional levels of significance. Tucker's Coefficient of Congruence (0.99986) and the dispersion accounted for (0.99973) are close to 1, implying that the two dimensions are sufficient to map the 5 municipalities according to their poverty levels.

Table 4 STRESS and Fit measures for poverty variables

Normalized Raw Stress	**0.00027**
Stress-1	**0.01650**[a]
Stress-II	0.03977[a]
S-Stress	0.00076[b]
Dispersion Accounted For (D.A.F.)	**0.99973**
Tucker's Coefficient of Congruence	**0.99986**

The proximities matrix was generated for the five local municipalities using a Euclidian distance measure. The coefficients play a very important task in deciding about the contribution/location of the municipality. A perpetual map was used to determine the positioning of the five municipalities according to their poverty levels. These maps can be used if the researcher intends to avoid statistical concepts such as the p-values, confidence intervals, hypothesis testing, etc. Perpetual maps can be used as a primary means by a reader to assess the situation or arrive at a conclusion. Three helpful things to look at when reading a perpetual map are directions, regions and clustering points.

Table 5 Standardised optimal coordinates for poverty variables

Local Municipality	Dimension	
	1	2
Mahikeng_L_M	**.999**	-.220
Ratlou_L_M	-.454	-.331
Ramotshere_Moiloa_L_M	-.018	.099
Ditsobotla_L_M	.021	**.560**
Tswaing_L_M	-.549	-.108

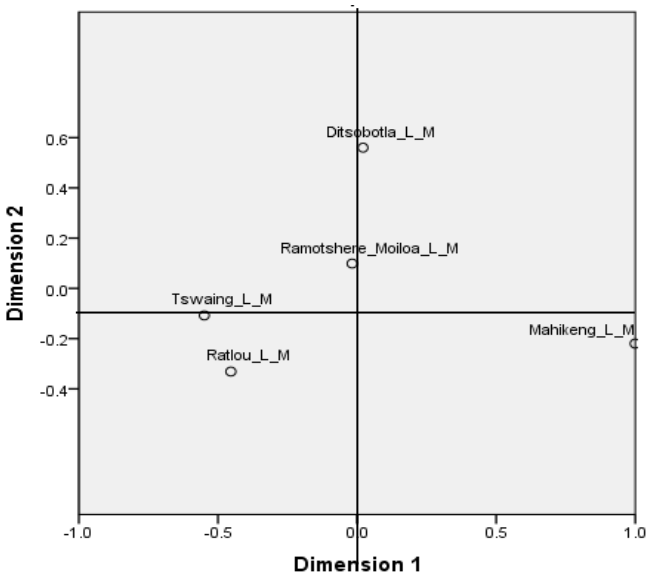

Figure 3 Perpetual map of NMMDM

Both the optimal matrix and the optimal perpetual map showed the Mafikeng_LM is associated with extremely high poverty levels followed by Ditsobotla_LM. These two local municipalities are positioned on the same spatial space of high poverty levels. Ratlou and Tswaing local municipalities are associated with extremely low poverty levels as opposed to Ramotshere-Moiloa local municipality with moderately low levels of poverty. It is evident that MDS has successfully mapped the five municipalities into two-dimensional space. The straight line (Euclidean) distances between the points match the observed distances.

A recommendation from the study was a concerted increase in distribution of resources according to the clusters. NMMDM with high poverty levels should be closely monitored, more specifically the Mafikeng local municipality. The government may embark on implementing strategies that target poverty reduction on this municipality and later eradication. Relevant municipal managers from these poverty stricken municipalities may also investigate measures the municipalities who are not highly poverty stricken are using to implement them in theirs. This could be a solution to the entire district municipality.

Since the main aim of the district municipality is to improve the quality of its community's standard of living through the implementation of formulated policies, the study on this note therefore make another recommendation to the district municipality to encourage the local municipalities with similarities to work together in the development of strategic plans and policies that will assist them in eradicating extreme poverty and hunger effectively and efficiently. This may be achieved by developing realistic projects that will benefit communities in need based on the resources allocated. It was further recommended that the resources be distributed according to the levels of poverty and community needs in the five local municipalities. With Ditsobotla local municipality identified as the least serviced and most impoverished local municipality, the NMMDM should identify it as the second local municipality where most of the municipal funds should be spent. The above recommendations may also assist the district municipality to efficiently account for the municipal funds.

Moreover, Moroke (2014) applied a Metric Multidimensional Scaling approach on time series data. The study sought to profile some the dire determinants of household debts in South Africa. MDS was used firstly as a preliminary method and also to produce a model of household debts. I later on did a confirmatory analysis using multivariate econometric methods. Owing to high dimensionality of data, MDS approach does perform a statistical significance of individual determinant. MDS tend to give cumbersome and complicated results. Despite the suitability of Metric MDS to analysis of time series data, this method has not been applied in this area. There is therefore a need for a study that explores effective frameworks that may provide the results which non statisticians may also find easy and interesting to read and understand. A Metric MDS provides a guiding map that helps in reducing the complexity inherent to the proximities by combining the determinants according to the type and the extent of the effect in household debts. Furthermore, the study fills an undermined gap in the literature by using this novel mapping technique. The application of Metric MDS method by this study is to further lure scholars who are analysing high-dimensional time series data.

This study used data collected from the South African Reserve Bank and Statistics South Africa for the period 1990 Q1 to 2013 Q1 consisting of ten macroeconomic and financial determinants of household debts each with 93 observations. Literature suggested numerous theories which explain household indebtedness. This study consulted two of these theories and related literature to help in identifying the determinants of household debt. It is reported in the literature that the level of household indebtedness is determined by supply and demand. Meng *et al*. (2011) highlighted that most of the households *enter into debt due to the availability of funding*, by, for instance, credit providers. As a result, this study analysed the factors affecting borrowing and/or lending and adopted the approach used by Meng *et al*. (2011). The following were identified as potential household debt determinants; house prices (HP), consumer prices (CP), household income (INC), interest rates (IR), gross domestic product (GDP), household consumption (HC), household savings (HS), unemployment rates (UR) and exchange rates (ER) and tax rates (TAX).

Prior to the application of metric multidimensional scale method, I addressed issues of stationarity since time series data was used. This was also to ensure the robustness of the results and to avoid spurious regression results. Spurious regression arises when time series variables are non-stationary and independent. In the presence of spurious regression, it has been proven that, *inter alia*, the "Ordinary Least Squares parameter estimates and the R^2 converge to the functional of Brownian motions such as the "t-ratios diverge in distribution, and the Durbin-Watson statistic converges in probability to zero". Moreover, the corresponding results for some common tests for normality and homoskedasticity of the errors are derived in a spurious regression. For the sake of this study, pertinent model assumptions about the error term were accounted for before analysing data. For ease of understanding, consider the simple univariate regression model, estimated by OLS:

$$y_t = \alpha + \beta x_t + \mu_t. \qquad [4]$$

The regression becomes "spurious" because both the criterion variable and the regressor follow independent $I(1)$ processes:

$$y_t = y_{t-1} + v_t \;\; ; \;\; v_t \sim iid(0, \sigma_v^2), \qquad [5]$$
$$x_t = x_{t-1} + w_t \;\; ; \;\; w_t \sim iid(0, \sigma_w^2), \qquad [6]$$

with v_t and w_t independent for all t, and (without loss of generality) $v_0 = w_0 = 0$. In fact [(Phillips (1986, p.313)] v_t and w_t may have heteroskedastic variances, so, the true parameter values are $\alpha = \beta = 0$.

"Many of the basic pitfalls associated with the use of non-stationary data in regression analysis have been well documented. In particular, Phillips (1986) exposed the underlying reasons for several observed empirical features of "spurious regressions". Among other things, the author showed that the standard t-test and F-test statistics diverge as $T \uparrow \infty$, and the serial correlation statistic converges to zero in probability. Thus, each of the associated null hypotheses will be rejected with increasing probability as the sample size grows, even though in fact they are actually true". It is therefore advisable to test the data for stationarity (and possible cointegration) prior to embarking on the estimation of a time series regression. It is however important to note that stationarity (unit root) tests have low power and are sensitive to structural breaks in the data.

Upon addressing the important assumptions, the analysis was carried out and a scree plot of normalised STRESS produced converged at two suggesting that the ten household debts determinants can be profiled according to two dimensions.

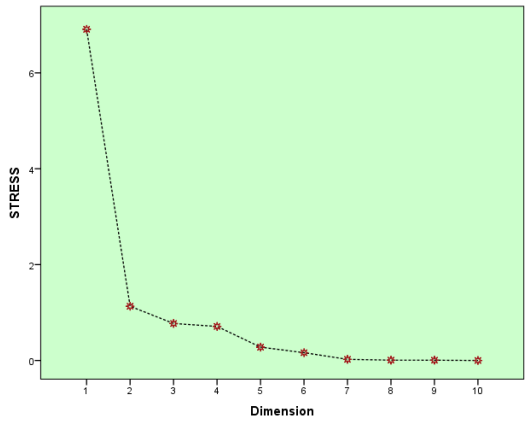

Figure 4 Scree plot of household debts determinants

A STRESS 1 measure was calculated as 0.00077, confirming the best fit of Metric MDS model and the Tucker's Coefficient of Congruence implied that 99.9% of variance in the model is accounted for by the two dimensions. This was also a confirmation that the ten selected determinants can better be represented in a two dimensional perpetual map.

As stated earlier, MDS method can also be used for model fitting purposes. Table 3 gives a summary results for the profiles and coefficients coupled with the direction of association. One drawback about interdependence techniques is that they do not provide enough space to do model diagnostic tests, hence their limited use to dimension reduction and data summarisation. MDS depends more on the graphs than tests.

A model with theoretical signs for the ten determinants of household debt expressed in regression form is:

$$y_t = \alpha + \underbrace{\beta_1 HP_t}_{+} + \underbrace{\beta_2 CPI_t}_{-} + \underbrace{\beta_3 INC_t}_{+} + \underbrace{\beta_4 IR_t}_{-} + \underbrace{\beta_5 GDP_t}_{+} + \underbrace{\beta_6 HC_t}_{+} + \underbrace{\beta_7 HS_t}_{-} + \underbrace{\beta_8 ER_t}_{-} + \underbrace{\beta_9 UR_t}_{-} + \underbrace{\beta_{10} TAX_t}_{-} + \varepsilon_t \quad [7]$$

Table 6 Standardised aggregate coordinates for household debts

	Dimension	
	1	2
HP	.209	.085
CPI	-.481	-.044
INC	**.705**	.089
IR	-.481	-.044
GDP	**1.405**	-.305
HC	**.566**	.394
HS	-.481	-.044
ER	-.481	-.044
UR	-.481	-.044
TAX	-.481	-.044

The findings revealed two profiles of household debts. Gross domestic product had extremely high positive contribution to household debts during the study period. Household income (INC) and household consumption (HC) also affected household debts positively and significantly. House Prices had low positive effect. The figure reveals less congruence with the clustering in other dimensions which clumped in dimension one with estimated coefficients between 0 and -0.5. These variables were associated with extremely low levels of household debts. A follow-up study by Moroke *et al*. (2014) used a Multivariate cointegration by Johansen and the Toda-Yamamoto causality testing approaches. The findings of the two studies were generally in agreement. The findings were also in agreement with those by reported by Debelle (2004), Subhanij (2007), Prinsloo (2002) and Kotzé and Smit (2008). On the contrary, household income (INC) had a negative contribution to household debts according to Vector Error Correction Model disqualifying Meng et al. (2011) contention.

A visual mapping of the standardised aggregate coordinates gives a clear picture about the variables.

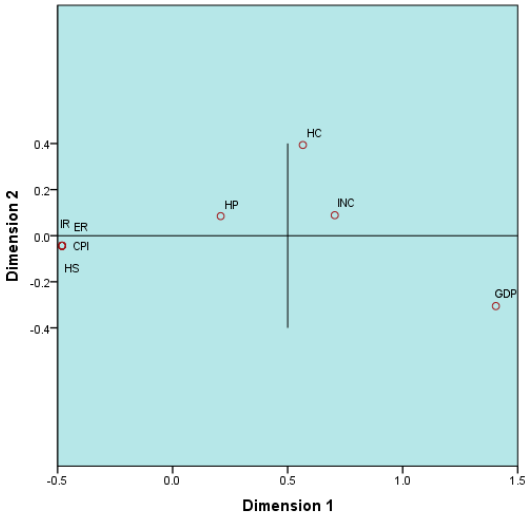

Figure 5 Perpetual map of household debts determinants

Just to satisfy our curiosity, the coefficients are plotted on a scatter plot to further assess the goodness of fit of the model. Illustrated in this plot is a departure from linearity measured by the STRESS and Tucker's Coefficient of Congruence.

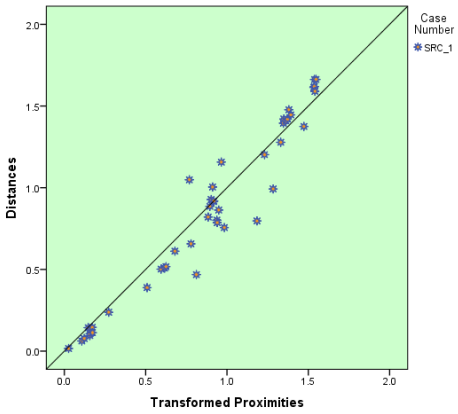

Figure 6 Transformed proximities residual plot of household debts

All the points lie on a diagonal line on Figure 4, implying that a model exhibit a perfect fit. Few of the observations reveal some points vertical departing from diagonal line. These departures represent the residual of the corresponding observation.

MDS demonstrated its effectiveness in classifying household debt determinants according to theory. Also revealed is that an MDS is a useful tool to use in quantifying the ubiquitous, but slimy, notion of similarity. MDS helped in segregating those determinants associated with high levels from those associated with low levels of household debts. The findings have policy implications.

Nwanamidwa and Moroke (2016) used Multiple Discriminant Analysis and Multinomial Logistic Regression models to evaluate the convergence of these models when applied to a field of education. We modelled the association of the performance of the four final year modules offered to students who enrolled for Statistics program and used the models to assess if students registered for the said program are fit to progress with their studies the next academic year. The application of these techniques in education sciences is not evident. The findings of this study could be useful by faculty or university administration when planning ahead for enrolment, for student recruitment and retention. At-risk students are also likely identified and monitored through the use of these techniques. A permission to use the students' performance and to conduct the study was sought from the university.

A discriminant analysis (DA) and logistic regression (LR) are multivariate statistical techniques used for evaluating the association between various covariates and categorical outcome variable. These methods can also be used for predicting and categorizing a set of observations (students in our case) into predefined groups or categories (Abledu *et al*., 2016; Sen, 2010 & Green *et al*., 2000). The groups for both techniques are assumed to be mutually exclusive and collectively exhaustive. While a DA is used to determine which continuous variables best discriminate between two or more naturally occurring groups, a LR performs similar task and can also take into account non-metric independent variables. The DA and LR focus on the association between multiple independent variables and a categorical dependent variable by forming a composite of the independent variables. One can derive a classification model for predicting the group membership of new observations (Antonogeorgos *et al*. 2009).

Third year statistics students' performance for 2013 to 2015 was used. Enrolments for these years were: 2013 (50), 2014(44), 2015(44) and 2016(49). These four groups were regarded as polychotomous dependent variable and student's marks on the four modules were used as independent variables. Multivariate normality was not observed since students' performance is unlikely to follow a normal distribution. This assumption was also overruled according to central limit theorem since the sample used (187) was in excess of 30. Logistic regression is not prone to non-normality and heteroscedasticity since non-metric variables may also be factored in the model. Since enrolments numbers for the four years was not the same, it was unlikely to obtain a homogeneous variance-covariance matrices. Take note that lack of homogeneity of variance may yield unreliable results, especially when small sample size is used. No multicollinearity was present between the variables according to the Tolerance and VIF statistics.

Table 7 Collinearity statistics for MDA and MLR

	Tolerance	*VIF*
Module1	.706	1.417
Module2	.820	1.219
Module3	.976	1.025
Module4	.685	1.460

The two methods have similar functional form but they differ in methods of coefficient estimation. In most cases a discriminant model is less robust as opposed to the logistic regression model (non-linearity between variables, no issues of multicollinearity and heteroscedasticity, continuous and non-metric predictive variables).

Both the MDA and MLR models were fitted and their parameter estimates passed the goodness-of-fit tests.

Multiple Discriminant Analysis model:

$$D_{2013} = -3.019 - 0.005module_1 + 0.086module_2 - 0.007module_3 - 0.022module_4 \quad [8]$$

$$D_{2014} = -4.139 - 0.005module_1 - 0.016module_2 + 0.083module_3 - 0.001module_4 \quad [9]$$

$$D_{2014} = -6.192 + 0.020module_1 + 0.009module_2 - 0.017module_3 + 0.083module_4 \quad [10]$$

Multinomial Logistic Regression model:

$$logit(P_{2013}) = -5.708 - 0.023module_1 + 0.059module_2 + 0.064module_3 - 0.012module_4 \quad [11]$$

$$logit(P_{2014}) = -0.0275 + 0.002module_1 - 0.113module_2 + 0.002module_3 + 0.035module_4 \quad [12]$$

$$logit(P_{2015}) = -3.506 + 0.000module_1 - 0.011module_2 + 0.024module_3 + 0.040module_4 \quad [13]$$

The sensitivity, specificity and accuracy of the models also assessed on the same set of data. The predictive power of the models were later checked to produce the results in Table 8.

Table 8 Classification results of MDA and MLR

	MDA	MLR
2013	68%	68%
2014	75.0%	75.0%
2015	31.8%	22.7%
2016	57.1%	59.2%
Overall classification	58.3%	56.7%

The overall correct classification rate was about 58% and 57% for MDA and MLR respectively, implying that in principle both models have similar predictive powers. In most cases, the two methods converged in results. Better classification rates could have been recorded if students' performance in the four modules was monitored earlier and relevant strategies could have been devised to obtain improved performance rates. Recommendation for future studies is factoring of other variables that affect students' performance to obtain improved results. This study is significant as it highlights differences and similarities in the efficiency of MDA and MLR models. It further provides an insight to scholars on the reliability of these models.

Montshiwa and Moroke (2017) explored the efficiency of six count data models to analysing marriage data. Sample size is a key to most multivariate data analyses techniques and as such we were interested in investigation what effect it has on Categorical Multivariate Analysis methods of count data such as the Poisson regression model (PRM), Negative binomial regression model (NBRM), Zero-inflated Poisson (ZIP), Zero-inflated negative binomial (ZINB), Poisson Hurdle model (PHM) and Negative binomial hurdle model (NBHM). Count data models belong to the Generalised Linear Models (GLM) family.

The most important part of Multivariate Analysis is selecting the right technique for a particular type of data set in order to obtain more realistic, valid and reliable results. Recently, the fundamental differentiator of methods used in modelling count data is the distributional assumptions of the mean and variance, the presence of excess zeros and the theoretical knowledge of the data set. There are fragmented conclusions in literature in terms of which count data model is the best. The performance of the extensions of PRM which are meant to address its limitations such as its non-applicability to under-/ over-dispersed data and excess zeros is continuously questioned and new models are

iteratively established in order to improve such extensions of PRM. This study sought to understand whether or not the sample size variations can improve the efficiency of count data models relative to under- or over- dispersion and excess zeros without further iterative re-parameterisation of the known models.

The data used in this study were sourced from Data First and were collected by Statistics South Africa through the Marriage and Divorce database. The ten randomly selected samples ranging from 4392 to 43916 and differing by 10% in size were used. Since the interest of the current study was to explore the performance of count data models under different sample sizes. The merged dataset for 2010 and 2011 (N=43916) were further divided into ten random samples in multiples of 10% until 100% (N). The categorical variables used in this study were: Male Race, Female Race, Male Occupation, Female Occupation, Male Status (Marital status of husband), Female Status (Marital status of wife), Male No Times Married, Female No Times Married, Solemnisation, and Marriage Type. The continuous variables were Male Age, Female Age, No of Children and Duration of Marriage (dependent variable).

Important assumptions of count data models equi-dispersion and no serious multicollinearity were addressed at preliminary data analysis stage. Empirical findings relative to the within-sample comparison revealed that for sample sizes of 10% (4392) to 50% (21958), NBHM outperformed all models whereas NBRM was favoured by most comparison criteria for sample sizes of at least 60% (25107). The between-sample comparison revealed that generally, the preferred models from the within-sample comparison (NBHM for at most 50% sample size and NBRM for at least sample size) become less effective as the sample size increases. ZINB did not converge when the sample size is at least 50%. The problem of the non-convergence of ZINB was also noted in the study by Famoye and Singh (2006). As such, one may remark that ZINB has a disadvantage of not converging especially as the sample size becomes large. NBHM for the smallest sample size under study was selected as the most effective model for fitting the Duration of Marriage and was found to be significant in overall.

Negative binomial regression model was used for further analysis to predict the variable of interest and the results are summarised on Table 9.

Table 9 NBHM parameter estimates of Predictors of DurationOfMarriage

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
b0(Intercept)	0.06068	0.09542	2110	0.64	0.5249
b1(MaleOccupation)	0.002605	0.003476	2110	0.75	0.4537
b2(FemaleOccupation	0.001169	0.003252	2110	0.36	0.7192
b3(MaleStatus	-0.08319	0.03879	2110	-2.14	0.0321*[2]
b4(FemaleStatus	-0.1727	0.035	2110	-4.93	<.0001*
b5(MaleNoTimesMarried	-0.2775	0.05708	2110	-4.86	<.0001*
b6(FemaleNoTimesMarried)	-0.1484	0.04941	2110	-3	0.0027*
b7(MaleAge2)	0.2258	0.02086	2110	10.82	<.0001*
b8(FemaleAge2)	0.4526	0.02131	2110	21.24	<.0001*
b9(Solemnisation)	0.0271	0.02409	2110	1.12	0.2608
b10(MarriageType)	-0.03291	0.01159	2110	-2.84	0.0046*
b11(NoOfChildren)	0.1655	0.01126	2110	14.7	<.0001*
b12(CoupleRace)	0.08745	0.008914	2110	9.81	<.0001*

Table 9 presents parameter estimates of the predictors of DurationOfMarriage using NBHM. The results revealed that all the 12 predictors have a role to play as far as DurationOfMarriage is concerned. The key predictors of DurationOfMarriage such as (FemaleStatus), (MaleNoTimesMarried) (FemaleNoTimesMarried), (MaleAge2), (FemaleAge2), (MarriageType), (NoOfChildren) and (CoupleRace) were all significant at 5% level of significance. It is

[2] *significant at 5%

worth noting that female and male occupations did not play a significant role in DurationOfMarriage. Solemnisation is another predictor of less concern in marriage life span.

The findings of this study contribute to literature and have policy implications. Recommendations were made to marriage counsellors to consider variables such as the age of the couple (both male and female) and the number of times married (both male and female) when advising the couples about reducing the risk of divorce.

4. Throw out

4.1 Staff development

I formally started supervising masters and doctoral students from 2014 after graduating PhD. My first three masters students graduated in 2015 and first two PhD candidates in 2017. I have successfully supervised more than 12 masters students (5 staff members) and 5 doctoral candidates (4 staff members) since 2015 to date. I have five PhDs in progress (four staff members). All my students' work is along the focus area "Multivariate analysis of high-dimensional data".

Photo Gallery of PhD graduates from 2017 to 2019

Dr Montshiwa, T.V. 2017 - An Experimental Analysis of the Effect of Sample Size on the Efficiency of Count Data Models

Dr Mupundu, T. 2017-Spatial Distribution of HIV/AIDS in Botswana

Dr Tsoku, J.T. 2018- The implementation of statistical arbitrage strategy on the JSE top 40 index option

Dr Metsileng, L.D. 2018 -The performance of conditional heteroskedastic VAR enhanced Multivariate GARCH models on time-varying integrated data

Dr Mpeta, K. 2019- Efficiency of the Structural Equation Model and related models in validating the Theory of Planned Behaviour

Madam Deputy Vice Chancellor, I appreciate the efforts by the university to develop staff in areas of teaching and learning, and research. My biggest challenge is retaining these staff members especially young staff members from scarce skill fields of study. Most of the Statisticians are doing very well (in terms of salaries) in the industry, recruiting them into academia comes with a very high price, hence I have taken it upon myself to mainly focus on developing and strengthening staff capacity in the program. I sincerely request your office to help the faculty keep these staff members. We have for many years struggled to fill vacancies due to death of statisticians in the country and the issue of location and salary is affects us negatively and significantly. I quote another concern and recommendation from EPE panel;

Comments about staffing: *"The panel had concerns about high workload, and acknowledges the impact of the three vacant positions in this regard. In particular,* ***the panel was concerned about the impact of the single staff in the programme having to supervise all 17 masters and doctoral students****.*
The efforts to recruit at least two senior staff members with a doctorate are affirmed, and the panel appreciates the need to suspend any further masters and doctoral enrolments until the appointments are made. In recruiting for these positions, ***statistics could be treated as a scare skill with associated scarce skill allowance"****.*

Commendation*: "The staff are dedicated, hardworking and committed to improving their qualifications. Despite their high workload, they are managing to produce research output.*

That being said Madam Deputy Vice Chancellor, I can confirm that despite the heavy supervision and lecturing workload, I have single-handedly managed to successfully supervise all the 17 students between 2015 and now.

4.2 Marketing the NWU

Publications:

- My other accomplishments besides successful supervision of postgraduate students, is the research publications that resulted in some recognition.

- My application for NRF funding in 2016 was evaluated by the South African Statistical Association based on my study leading, article publication and conference proceedings records.
- I have published more than ±30 peer reviewed articles between 2013 and 2019
- More than 12 papers published on peer reviewed conference proceedings

Conferences proceedings:

- RAIS conference (2018),
- Annual Conference of South African Statistical Association (2018),
- World Academy of Science, Engineering and Technology (2017),
- International Conference on Business, Economics, and Financial Sciences, Management (2017),
- International conference on Sciences, Technology and Social Sciences (2017),
- IBIMA (2017, 2016, 2015, 2014),
- International Conference on Economics and Social Sciences (2017),
- Annual Eurasian Business Research Conference (2017),
- International Institute of Social and Economic Sciences (2016),

5. Way forward

- Broadening the focus area by international collaborations and through research chairmanship
- Continued mentorship and support to students
- NRF rating for the group
- Preparations for 4^{Tth} industrial revolution era
- Application of more recent Multivariate Analysis methods
- Exploration of more statistical packages to high-dimensional data

6. Conclusion

This lecture provided an intuitive grasp of multivariate analyses by application to different fields of studies. Multivariate data analysis methods have a fascinating way of unravelling the interplay of multiple variables. These methods can be applied by researchers and practitioners from all fields of study that measure several variables on experimental unit. Multivariate analyses have proven to be applicable methods in exploring the joint performance of the variables in a system and further to determine how each variable affects a criterion variable in the presence of the others. A researcher can select one or more techniques appropriate to the high-dimensional data at hand and the study objectives. Multivariate Data Analysis and High-dimensional data cannot be separated.

To conclude, I want to give you a take home message by quoting:

> *"If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical method, and many ways in which they can break down. They are a valuable aid to the interpretation of data, not sausage machines automatically transforming bodies of numbers into packets of scientific fact" ~ F.H.C. Mariot*

That being said, please use statistics with care. Take note of the assumptions and stop abusing and misusing statistics. Let us refrain from using statistics to bolster weak arguments. Statistics is definitely not one of the lies.

7. Appreciation

Madam Deputy Vice-Chancellor, it is not possible for me to conclude this lecture without giving my profound gratitude to my one and only concoction. I have to thank God for being patient with me when I knew I wasn't committing to him

the way I was supposed to. I thank Him for the trials and tribulations I faced throughout my studies, because they helped me to put him first. Without God I know I wouldn't have made it through the university. I am what I am today because of God's infinite mercies.

To the Vice-Chancellor the prophet of this time and Deputy Vice Chancellors of the NWU, I thank you all for giving me the platform upon which this lecture is delivered barely two years of attaining the rank of a professor. May God of mercies bless the fruits of your hands and give you the strength to nurture this institution and realise Vision 2025, the prophetic verdict that we are all running with.

The then Dean of the Faculty of Commerce and Administration, and this lady still remains the Super Executive Dean of the faculty of Economic and Management Sciences, Prof Sonia Swanepoel. When you joined the faculty, there were very few staff with PhDs and the research output was extremely low. Your super powers transformed all of us and started seeing research in a different perspective. Your perseverance, integrity and people-loving nature are just a few of your qualities that continue to inspire me. I salute you for your tenacity of purpose and outstanding leadership qualities. Thank you for your words of encouragement and support. I will remain forever grateful.

To my study leaders, the late Prof Petersen in collaboration with his wife, thank you for your guidance and support throughout my studies. Professor Phillip Serumaga-Zake, my former lecturer, study leader of my masters dissertation; you also played a very significant role on my academic development. Through you I met Prof Maseka Lesaoana who gave me confirmation about my students' work and motivated me to consider publishing it. Indeed, my exposure and interaction with colleagues beyond the country is due to all of you guys.

My late granny, my parents Mr. & Mrs. Moroke, who strived to see me through my academic pursuits particularly my primary, high school and undergraduate studies, thank you for investing in my education, for the love you gave me and the sacrifice you made in ensuring you both give me the right foundation for the future that I now enjoy.

Dr Dan Metsileng, you have no idea how much you inspired us. The tutorial lessons you gave us and the passion you had about statistics when delivering those lessons drew me nigh and lured me to grow fonder of statistics career. Your words specifically, I quote, *"take the road less travelled".* These words always kept me and still are keeping me thinking. I channelled my "*attitude*" towards my work as you always advised us to. Thank you very much for your uplifting words.

My former lecturers, specifically Mr Sedupane and Maruma who introduced me to statistics at undergraduate, I cannot thank you enough for sharing your knowledge with us. The foundation I received paved a way for me in an unimaginable way. You are nothing less than a blessing from God! Please accept my vehement protestations of gratitude. I'd love to express my gratitude for your generous gift.

To my mentors, Proffs Burger Van Lill and Nehemiah Mavetera, you are wonderful teachers, bosses, leaders, and friends. You are everything one could look for in a good mentor. You groomed me to be a sound professional and made working with you an interesting and memorable experience. I will always be grateful to you for your support and kindness. Not only have you been a fantastic mentors to me, but you have taught me how to mentor other people. Thanks for opening my eyes to new stages of opportunity and strength. I will forever be grateful for your guidance and kindness.

I extend my sincere gratitude to my colleagues in Statistics for agreeing to share workload and to my friends and everyone that I have come across in every aspect of my career. Each one of you have in one way or the other, influenced my total being, thus making this day possible. Thank you very much for affecting my life.

Mrs Disekwane Rebecca Mheta and her daughter Bonolo, I appreciate your love, kindness, support, and generosity you have shown to me and the kids. I am dearly thankful for the support you gave me and my kids when I was studying. You took care of Koki and helped her do her homework.

I acknowledge the team that put their efforts and time into planning this occasion. Estie, Esme, Nombulelo, Connie and members in your unit, when it comes to you guys, the word "TEAM" is actually the acronym for "Terrific Enthusiastic Ambitious and Motivating". Thanks for being the best TEAM ever and for your creativity in ensuring that everything is in place. Estie, thank you for your continued support and for helping me prepare emotionally for this day.

Pastor Molefe, I think in this room you know my journey more than anyone. You've been a wonderful blessing to me and my kids. I appreciate the time you have dedicated to giving me counselling. I wasn't sure I could trust anyone with my story. Being able to share with you has helped me tremendously. Thank you for the prayers and spirit filled words from the bible.

My lovely kids, for appreciating the dictum that "it is not enough for mother to be all and for you to be nothing". Thank you for your patience and support.

8. References

ABLEDU, G. K., BUCKMAN, A., ADADE, T., & KWOFIE, S. 2016. Comparison of logistic regression and linear discriminant analyses of the determinants of financial sustainability of rural banks in Ghana. American Journal of Theoretical and Applied Statistics, Vol. 5 No. 2, pp. 49-57.

ADJUIK, M., SMITH, T., CLARK, S., TODD, J., GARRIB, A., KINFU, Y., KAHN, K., MOLA, M., ASHRAF, A., MASANJA, H., ADAZU, U., SACARLAL, J., ALAM, N., MARRA, A., GBANGOU, A., MWAGENI, E. & BINKA, F. (2006). Cause-specific mortality rates in sub-Saharan Africa and Bangladesh. Bulletin of the World Health Organization, Vol. 84, pp, 181-188.

ALTAS, D., KUBAS, A. AND SEZEN, J. (2013). Analysis of environmental sensitivity in Thrace region through Two-Step cluster, Trakia Journal of Science, Vol. 11 No. 3, pp. 318-329.

ANTONOGEORGOS, G., PANAGIOTAKOS, D.B., PRIFTIS, K. N., & TZONOU, A. (2009). Logistic regression and linear discriminant analyses in evaluating factors associated with asthma prevalence among 10 to 12 years old children, divergence and similarity of the two statistical methods. International Journal of Pediatrics, pp. 1-6.

CROSS, L.C. (2013). Statistical and Methodological Considerations When Using Cluster Analysis in Neuropsychological Research. Springer Science and Business Media. New York. 978-1-4614-6744-1.

DEBELLE, G. (2004). Household debt and macroeconomy. BIS Quarterly Review, March, pp. 51-64, http://www.bis.org/publ/qtrpdf /r_qt0403e.pdf.

DING, C. S. (2006). Multidimensional scaling modelling approach to latent profile analysis in psychological research. International Journal of Psychology, Vol. 41 No. 3, pp. 226-238.

FAMOYE, F. & SINGH, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. Journal of Data Science, Vol. 4, pp. 117-130.

FIELD, A. (2000). Discovering Statistics Using SPSS for Windows. Sage publications London Thousand Oaks, New Delhi.

GIGUÈRE, G. (2006). Collecting and analysing data in multidimensional scaling experiments: A guide for psychologists using SPSS. Tutorial in Quantitative Methods for Psychology, Vol. 2 No. 1, pp. 27-38.

GREEN, P. E. & CARMONE, F. J. (1969). Multidimensional scaling: An introduction and comparison of nonmetric unfolding techniques. Journal of Marketing Research, Vol. 6 No. 3, pp. 330–341.

GROENEN, P. J. F. & VELDEN, M. (2004). Multidimensional scaling, Erasmus School of Economics (ESE).

HABING, B. (2003). Exploratory Factor Analysis, retrieved from http://www.stat.sc.edu/habing/courses/530EFA.pdf

HAIR, J, JR., BLACK, W., BABIN, B., ANDERSON, R., & TATHAM, R. (2010). *Multivariate Data Analysis*. Pearson Education, Inc., Upper Saddle River, New Jersey.

JAWORSKA, N. & CHUPETLOVSKA-ANASTASOVA, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. Tutorials in quantitative methods for psychology, Vol. 5, pp. 1-10.

JOHNSON, D.E. (1998). Applied multivariate methods for data analysis. Brooks/Cole publishing: United States of America.

KIM, J., YANG, J. & O'LAFSSON, S. (2009). An optimization approach to partitional data clustering, Journal of the Operations Research Society, Vol. 60 No. 8, pp. 1069-1084.

LEONARD, S.T. & DROEGE, M. (2008). The uses and benefits of cluster analysis in pharmacy research. Research in Social and Administrative Pharmacy, Vol 4, pp. 1-11.

MAHAPATRA, P., KENJI, S., ALAN, D.L., FRANCESCA, C., FRANCIS, C.N. SIMON, S. on behalf of the Monitoring Vital Events (MoVE) writing group (2007). Civil registration systems and vital statistics: successes and missed opportunities. The Lancet, Vol. 370 No. 10, pp. 16531663.

MAHOLE, K.C., MOROKE, N.D. & MAVETERA, N. (2014). Poverty levels among Local Municipalities in the NMMD of South Africa: A Multidimensional Scaling Approach. Mediterranean Journal of Social Sciences, Vol. 5 No. 2, pp. 549.

MANLY. B.F.J. 2001. Multivariate Statistical Methods, a Primer. Chapman and Hall/CRC.

MENG, S., HOANG, T. N. & SIRIWARDANA, M. (2011). The determinants of Australia household debt: a macro Level study. Business, economics and public policy working papers, 2011-2014.

MONTSHIWA, V.T. & MOROKE, N.D. (2014). Assessment of the reliability and validity of student-lecturer evaluation questionnaire: A Case of North West University, Mediterranean Journal of Social Sciences, Vol.5, No. 14, 352-364.

MONTSHIWA, T.V. & MOROKE, N.D. (2017). The Effect of Sample Size on the Efficiency of Count Data Models: Application to Marriage Data, Journal of Economics and Behavioural Studies, Vol. 9, No. 3, pp. 6-18.

MOROKE, N.D. & MAVETERA, N. (2013). European Journal of Social Sciences, Vol. 41 No. 3, pp. 386 - 401.

MOROKE, N.D. & PULENYANE M. (2014). Clusters of leading death causes in South Africa: Application of hierarchical agglomerative clustering technique. Mediterranean Journal of Social Sciences. Vol. 5 No. 20, pp. 220-226.

MOROKE, N.D. (2014). Profiling some of the dire household debt determinants: A metric multidimensional scaling approach. Journal of Economic and Behavioural Studies, Vol. 6 No. 11, pp. 858-867.

MOROKE, N.D. (2015). A Two-step clustering algorithm as applied to crime data of South Africa. Corporate Ownership and Control, Vol. 12 No. 2, 488-496.

MOROKE, N.D., MUKKUDEM-PETERSEN, J. & PETERSEN, M. (2014). Exploring Dynamic Relations between Macroeconomic Variables and Household Debts in South Africa: An Application of Toda-Yamamoto causality. International Research Journal of Economics and Finance, Vol 120, pp. 47-62.

MUGAVIN, M. E. (2008). Multidimensional scaling: A brief overview. Nursing Research, Vol. 57 No. 1, pp. 64-68.

NWANAMIDWA, S.T & MOROKE, N.D. (2016). Comparative study of Multiple Discriminant and Multinomial Logistic Regression analyses on Statistics students' performance. A dissertation submitted to NWU.

OLKIN, A.R.S. (2001). Multivariate analysis: an overview. International Encyclopedia of the Social & Behavioral Sciences, pp. 10240-10247.

PHILLIPS, P. C. B. (1986). Understanding spurious regressions in econometrics, Journal of Econometrics, Vol. 33, pp. 311-340.

PITUCH, K.A & STEVENS, J.P. (2016). Applied multivariate analysis for social sciences: Analyses with SAS and IBM's SPSS. Routledge, Taylor & Francis Group, New York and London.

PO, R.W., GUH, Y.Y. AND YANG, M.S. (2009). A new clustering approach using data envelopment analysis, European Journal of Operations Research, Vol. 199, pp. 276-284.

PRINSLOO, J. W. (2002). Household debt, wealth and saving. South African Reserve Bank Quarterly Bulletin.

Kotzé, L. and Smit, A. V. A. (2008). Personal finances: What is the possible impact on entrepreneurial activity in South Africa? Department of Business Management, Faculty of Economic and Management Sciences, University of the Free State. Available at Southern African Business Review, www.myunisa.ac.za, No. 226, pp. 63-78.

REGE, M., DONG, M. & FOTOUHI, F. (2008). Bipartite isoperimetric graph partitioning for data co-clustering, Data Mining and Knowledge Discovery, Vol. 16 No. 3, pp. 276-312.

RENCHER, A. C. (2003). Methods of multivariate analysis. John Wiley and Sons, New York.

RENCHER, A.C. & CHRISTENSEN, W.F. (2012). Methods of Multivariate Analysis, (3rded) Wiley-Interscience, ISBN 9780470178696.

RIETVELD, T. & R. VAN HOUT (1993). Statistical Techniques for the Study of Language and Language Behaviour". Mouton de Gruyter Berlin, New York.

ŞEN, A.B. (2010). Factors that Discriminate Between Domestic and Foreign Banks Operating in Turkey. İktisadi ve İdari Bilimler Dergisi, Vol. 28 No. 1, pp. 445-462.

STATISTICS SOUTH AFRICA (2010/11). Mortality and causes of death in South Africa, 2009: Findings from death notification. Statistical release P0309.3. Pretoria: Statistics South Africa.

STEFFIRE, V. J. (1969). Market Structure Studies: New Products for Old Markets and New Markets (Foreign) for Old Products, in Application of the Sciences in Marketing, F. M. Bass, CW. King, and E. A. Pessimer, eds. New York: John Wiley & Sons, 251-68.

STEYVERS, M. (2002). Multidimensional scaling. In: Encyclopaedia of cognitive science. Nature Publishing Group, London, UK.

SUBHANIJ, T. (2007). Some lessons from Securitisation Crisis. Bangkok Post, Business New, (13 July, 2007).

SVETLANA V.S., JING, W. & DOUGLAS. W. H. (2013). Examining Similarity structure: Multidimensional Scaling and Related Approaches in Neuroimaging. Computational and Mathematical Methods in Medicine, pp. 1-9.

THALER, N. S., BELLOW, D. T., RANDALL, C., GOLDSTEIN, G., MAYFIELD, J., & ALLEN, D. N. (2010). IQ profiles are associated with differences in behavioral functioning following pediatric traumatic brain injury. Archives of Clinical Neuropsychology, Vol. 25 pp. 781–790.

THANASSOULIS, E. (1996). A data envelopment analysis approach to clustering operating units for resource allocation purposes, Omega, Vol. 24, pp. 463-476.

TSOGO, L., MASSON, M. H. & BARDOT, A. (2000). Multidimensional scaling methods for many object sets: A review. Multivariate Behavioural Research, 35 No. 3, pp. 307-319.

VENNA, J. & KASKI, S. (2006). Local multidimensional scaling. Neural Networks, 19(6-7), 889–899.

YIN, X., HAN, J. & YU, P.S. (2007). Crossclus: Userguided multi-relational clustering, Data Mining and Knowledge Discovery, Vol. 15 No. 3, pp. 321-348.

ZHANG, Z. (2010). Customer education introduction to research methods. New York, NY: Pearson Learning Solutions.