Development and Application of Multivariate Spatial Clustering Statistics

by

TIMOTHEUS BRIAN DARIKWA

THESIS

Submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

PROMOTER: Prof. Samuel Manda
(Biostatistics Research Unit, South Africa Medical Research Council)

CO-PROMOTER: Prof. ’Maseka Lesaoana
(School of Mathematical and Computer Sciences, University of Limpopo)

2021

Declaration

I, **Timotheus Brian Darikwa**, declare that the thesis hereby submitted to the University of Limpopo, for the degree of Doctor of Philosophy in Statistics has not been previously submitted by me for a degree at this or any university; that it is my own work in design and in execution, and that all material contained herein has been duly acknowledged.

Signature:.......................Date:.................................
Darikwa, T.B.

Abstract

In spatial statistics, several methods have been developed to measure the extent of local and global spatial dependence (clustering) in measured data across areas in a region of research interest. These methods are now routinely implemented in most Geographical Information Systems (GIS) and statistical computer packages. However, spatial statistics for measuring joint spatial dependence of multiple spatial measurement and outcome data have not been well developed. A naive analysis would simply apply univariate spatial dependence methods to each data separately. Though this is simple and straightforward, it ignores possible relationships between multiple spatial data because they may be measuring the same phenomena. Limited work has been done on extending the Moran's index, a commonly used and applied univariate measure of spatial clustering, to bivariate Moran's index in order to assess spatial dependence for two spatial data. The overall aim of this PhD was to develop multivariate spatial clustering methods for multiple spatial data, especially in the health sciences. Our proposed multivariate spatial clustering statistic is based on the fundamental theory regarding canonical correlations. We firstly reviewed and applied univariate and bivariate Moran's indexes to spatial analyses of multiple non-communicable diseases and related risk factors in South Africa. Then we derived our proposed multivariate spatial clustering method, which was evaluated by simulation studies and applied to a spatial analysis of multiple non-communicable diseases and related risk factors in South Africa. Simulation studies showed that our proposed multivariate spatial statistic was able to identify correctly clusters of areas with high risks as well as clusters with low risk.

Dedication To

To my wife Sheron, my children, my mom, my dad, my brothers, my sisters and my foremost mentors, with love and gratitude.

Acknowledgments

To begin with, I would like to express my gratitude to my academic promoter Professor Samuel Manda (South Africa Medical Research Council) and co-promoter Professor 'Maseka Lesaoana (University of Limpopo) for their patience, inspiration and guidance throughout my PhD studies and co-authoring papers based on this thesis. I would also like to thank Professor Yanguang Chen for his valuable feedback and critical comments on Chapter 4.

I am very grateful to Statistics South Africa for providing data used in the PhD thesis. The work reported herein was made possible through funding by the South African Medical Research Council through its Division of Research Capacity Development under the Biostatistics Capacity Development partnership with the Belgian Development Agency (Enabel) under its framework of Building Academic Partnerships for Economic Development (BAPED), grant no. 57042 (2019); Teaching development grant national collaborative project, no. APP-TDG-088 (2017). I am also greatly indebted to the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) of South Africa and VLIR who provided funding in the first year of my studies for travelling and tuition fees. I am also greatly indebted to the University of Limpopo for funding my attendance at South African Statistical Association (SASA) Conferences and their continuous support towards my PhD studies. Without the financial support I received from the above at various stages of my studies this, work would not have been possible.

My heartfelt thanks to my fellow University of Limpopo PhD students and colleagues, namely Reshoketswe Mokobane, Daniel Maposa, Alex Boateng, Philemon Nyamugure, Paul Moleko, Norman Maswanganyi and the Department of Statistics and Operations Research as a whole for their support and encouragement. Doctor Pholosho Kgopa I thank you for your words of encouragement and time to read some of my work. To the Almighty Lord, I am very grateful for the gift of life, good health and wisdom.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models. Most importantly, I wish to thank my loving and supportive wife, Sheron, and my four wonderful children, Panashe, Adelaide, Makanaka and Munashe BJ, who provide unending inspiration. I thank God, I have you!!

Contents

Declaration **i**

Abstract **ii**

Dedication **iii**

Acknowledgments **iv**

Table of Contents **vi**

List of Figures **x**

List of Tables **xiii**

List of Abbreviations and Acronyms **xvi**

Research Outputs **xviii**

1 Introduction **1**

1.1 Background 1

1.2 Overview of cardiovascular diseases 2

1.2.1 The problem of cardiovascular diseases 3

1.2.2 Risk and environmental factors for cardiovascular diseases 5

1.2.3 Spatial clustering of cardiovascular diseases 6

1.2.4 Spatial epidemiology of cardiovascular diseases in South Africa . . . 7
1.3 Research Problem . . . 9
1.4 Study aims and objectives . . . 9
1.5 Overview of thesis . . . 10

2 Review of spatial autocorrelation methods 9
2.1 Introduction . . . 9
2.2 Spatial autocorrelation methods . . . 10
2.2.1 Spatial weights . . . 10
2.2.2 Univariate spatial autocorrelation . . . 14
2.2.3 Global indexes of spatial autocorrelation . . . 14
2.2.4 Local indexes of spatial autocorrelation . . . 16
2.3 Linear regression-based Moran's index . . . 17
2.3.1 Framework for significance testing . . . 19
2.4 Summary of the chapter . . . 20

3 Application of univariate spatial autocorrelation methods to cardiovascular diseases in South Africa 21
3.1 Introduction . . . 21
3.2 Application to cardiovascular prevalence data . . . 22
3.2.1 Data . . . 22
3.2.2 Variable definitions . . . 23
3.2.3 Descriptive statistics of the variables . . . 24
3.2.4 Variables correlations . . . 25
3.2.5 Age-gender stratified clustering analysis . . . 26
3.2.6 Age-gender standardised spatial clustering analysis . . . 31
3.3 Application to cardiovascular mortality rates . . . 35
3.3.1 Data . . . 35
3.3.2 Estimation of mortality rates . . . 37

3.3.3 Spatial autocorrelation analysis 39
3.4 Application using a linear regression-based Moran's index 48
3.4.1 Analysis using distance based spatial weights 48
3.4.2 Sensitivity Analysis 58
3.4.3 Analysis using contiguity spatial weights 59
3.5 Summary of the chapter 62

4 Review and application of bivariate spatial clustering methods. 65
4.1 Introduction 65
4.1.1 Bivariate spatial autocorrelation measures 66
4.2 Application to cardiovascular prevalence data 70
4.2.1 Age-gender standardised joint spatial clustering analysis . 71
4.3 Application to cardiovascular mortality rates 74
4.3.1 Bivariate association of individual CVD maps over time . . 74
4.3.2 Bivariate spatial association between two CVDs at a point in time 76
4.4 Chapter summary 78

5 The proposed measure of statistical multivariate spatial autocorrelation 81
5.1 Introduction 81
5.2 Methods 84
5.2.1 Standardised coefficients for simple linear regression equations 84
5.2.2 The Moran's index as linear regression and the notion of a multivariate Moran's index 85
5.2.3 Canonical correlation analysis 89
5.2.4 Significance tests for canonical correlation 91
5.2.5 Multiple regression equivalence of canonical correlation . . 92
5.2.6 Using CCA to extend Moran's index to multivariate case . 95
5.2.7 Multivariate LISA clusters 97

5.3 Hypothetical example 97
5.3.1 Verification of univariate and bivariate spatial autocorrelation results 98
5.3.2 Multivariate spatial autocorrelation analysis 101
5.4 Chapter summary 104

6 Simulation Study and Application of the Proposed Multivariate Spatial Clustering Statistics 105
6.1 Introduction 105
6.2 Simulation study 106
6.3 Application to cardiovascular mortality 110
6.4 Chapter summary 115

7 Discussion and conclusions 117
7.1 Introduction 117
7.2 Summary of the findings 118
7.3 Limitations 125
7.4 Strength 127
7.5 Contribution to knowledge 128
7.6 Future direction 129
7.7 Conclusion 129

A Appendix A for Chapter 3 143
A.1 R CODE FOR CHEN'S REGRESSION APPROACH 149

B Appendix B for Chapter 5 198
B.1 R code for univariate and bivariate Moran's I using canonical approach 198
B.2 R code for multivariate Moran's I using canonical approach 200

List of Figures

3.1 The South African map showing the district names and sample sizes drawn from each district in the 2016 SADHS. 23

3.2 Maps of prevalence rates of the CVDs and their related risk factors across the district for the sample age group 15-64 years. 27

3.3 Univariate spatial clusters of CVDs and their risk factors with significant association for all participants, males and females using raw rates. 29

3.4 Maps of the district standardised incidence rates of the CVDs and their related risk factors. 32

3.5 Univariate LISA Maps of the district standardised incidence rates of the CVDs and their related risk factors. 34

3.6 Quantile maps showing the distribution of raw, smoothed and adjusted mortality rates for the year 2001. 41

3.7 Quantile maps showing the distribution of raw, smoothed and adjusted mortality rates for the year 2011. 42

3.8 Univariate Moran's I LISA maps showing the distribution of clusters of raw, smoothed and adjusted mortality rates for the year 2001. . 46

3.9 Univariate Moran's I LISA maps showing the distribution of clusters of raw, smoothed and adjusted mortality rates for the year 2011. . 47

3.10 Comparison of histogram and normal Q-Q plots of the residuals of the Moran's regression model using Poisson based ischaemic mortality adjusted rates in South Africa, 2011. 50

3.11 The scatter plots of the Moran's index using Poisson based ischaemic mortality adjusted rates in South Africa, 2011. 52
3.12 The univariate local Moran's spatial clusters for ishaemic mortality rates based on Chen's regression approach, for the inverse and exponential spatial weights. 54
3.13 The agreement charts for comparing ischaemic heart disease mortality LISA cluster map categories by three spatial weight matrices. . . 56
3.14 The univariate local Moran's spatial clusters for ishaemic mortality rates based on Chen's regression approach, for the Queen's spatial weights. 61
3.15 The univariate local Moran's spatial clusters for ishaemic mortality rates based on the original Moran's index, for the Queen's spatial weights. 62

4.1 Univariate and joint spatial clusters of CVDs and their risk factors with significant association for all participants using raw rates. . 71
4.2 Joint Spatial Clusters of CVDs and their risk factors with significant association for all participants, adjusted for the national age-sex distribution of the sample. 73
4.3 The raw, smoothed and adjusted mortality rates based Bivariate Moran's I LISA maps between same CVDs for the year 2001 and 2011. 76
4.4 The significant adjusted mortality rates based Bivariate Moran's I LISA map between two CVDs at a point in time, 2001 and 2011. 78

5.1 Hypothetical spatial data. 98
5.2 Univariate and bivariate LISA clusters for the hypothetical data. 100

6.1 Unique value based spatial distribution of the hypothetical data. 107
6.2 Univariate and bivariate LISA spatial patterns of the hypothetical data. 108

6.3 Multivariate LISA map for Y_1, X_1 and X_2 using the hypothetical data with 20 “hot-spots” clusters. 109

6.4 Univariate and bivariate LISA clusters for the hypothetical data. 113

6.5 Univariate and bivariate LISA clusters for the hypothetical data. 114

A.1 Quantile maps of prevalence rates of the CVDs and their related risk factors by gender. 144

A.2 Quantile maps of prevalence rates of the CVDs and their related risk factors by age. 145

A.3 Quantile maps of prevalence rates of the CVDs and their related risk factors by gender for ages 15-39 years. 146

A.4 Quantile maps of prevalence rates of the CVDs and their related risk factors by gender for ages 40-64 years. 147

A.5 Univariate spatial clusters of CVDs and their risk factors by age groups. 148

List of Tables

3.1 Summary statistics of the prevalence of CVDs and related risk factors across the districts. 25

3.2 Global univariate spatial autocorrelation Moran's I values for the CVDs and their risk factors for all participants split by age and gender. 29

3.3 Effects of age-gender stratification on the sample size used in the analyses. 31

3.4 Global univariate spatial autocorrelation association for the age-sex standardised incidence rates of CVDs and identified risk factors for all participants. 33

3.5 Distribution of number of deaths across age groups by year, South Africa. 36

3.6 Descriptive statistics of raw, EB smoothed and adjusted mortality rates across municipalities for CVA, IHD, HHD and DBT, South Africa. 40

3.7 Univariate global Moran's spatial autocorrelations for the model residuals, raw, smoothed and adjusted mortality rates due to CVA, IHD, DBT and HHD in 2001 and 2011. 44

3.8 The results of Chen's regression approach to Moran's index when applied to mortality rates due to ischaemic heart conditions in South Africa: 2011. 48

3.9 The Shapiro-Wilk's W test of normality on the residuals of ischaemic heart disease adjusted mortality rates for negative exponential and inverse power function weight matrices. 51
3.10 Classification Tables for cluster categories. 55
3.11 The concordance strength Bangdiwala test statistics. 57
3.12 Sensitivity Analysis of the effect of changing the inverse spatial weights matrix on the p-value in significance testing of Moran's index and s_f when applying Chen's regression approach. 58
3.13 Comparison of the global univariate Moran's index of spatial autocorrelation between Chen's approach and the original approach using a Queen weight matrix. 60

4.1 Global bivariate spatial autocorrelation association between the prevalence of CVDs and identified risk factors for all participants. 70
4.2 Global univariate and bivariate spatial autocorrelation association between the age-sex standardised incidence rates of CVDs and identified risk factors for all participants. 72
4.3 Bivariate global spatial autocorrelations measuring spatial dependence of individual CVD rates for raw, smoothed and adjusted mortality rates between the years 2001 and 2011. 75
4.4 Bivariate global spatial autocorrelations measuring spatial dependence of individual CVD rates between two time periods for raw, smoothed and adjusted mortality rates between the years 2001 and 2011. . 77

5.1 Univariate Moran's index using the canonical approach. 99
5.2 Bivariate spatial autocorrelation derived from the original Moran's index and the new multivariate approach. 101
5.3 Summary results for the canonical correlation analysis for the hypothetical spatial data. 102

5.4 Multivariate spatial autocorrelation analysis of the hypothetical spatial data, r = *0.6*. 103

6.1 Agreement table of the categories of multivariate joint clusters and simulated data when joint clusters equals to 20. 109
6.2 Simulation results 110
6.3 Bivariate spatial autocorrelation derived from the original Moran's index for Poisson regression adjusted cardiovascular mortality rates. 111
6.4 Summary results for the canonical correlation analysis. 112
6.5 Summary results for the multivariate Moran's index estimation procedure for cardiovascular mortality in South Africa. 112

A.1 Correlation analysis between CVDs and their risk factors for the whole sample, by age, gender and age-gender combinations. 143
A.2 Classification of spatial autocorrelation based on inverse power spatial weight, $\alpha = 1$. 156
A.3 Classification of spatial autocorrelation based on negative exponential spatial weight, $\alpha = 1$. 170
A.4 Classification of spatial autocorrelation based negative exponential spatial weight, $\alpha = 2$. 184

B.1 Multivariate spatial autocorrelation analysis of the hypothetical spatial data, r = *0.1*. 204
B.2 Multivariate spatial autocorrelation analysis of the hypothetical spatial data, r = *0.3*. 204
B.3 Multivariate spatial autocorrelation analysis of the hypothetical spatial data, r = *0.5*. 204
B.4 Multivariate spatial autocorrelation analysis of the hypothetical spatial data, r = *0.7*. 205

B.5 Multivariate spatial autocorrelation analysis of the hypothetical spatial data, r = *0.9*. . . . 205

List of Abbreviations and Acronyms

CCA	Canonical Correlation Analysis
CHD	Coronary Heart Disease
COD	Causes of Death
CVA	Cerebrovascular heart disease
CVD	Cardiovascular Disease
DBT	Diabetes
DNF	Death Notification Forms
EB	Empirical Bayes
GISA	Global Indicators of Spatial Autocorrelation
HA	Heart attack
HBC	High Blood Cholesterol
HHD	Hypertensive heart disease
ICD-10	International Classification of Diseases and Related Health Problems (10th revision)
IHD	Ischaemic Heart Disease
LISA	Local Indicators of Spatial Autocorrelation
MMI	Multivariate Moran's Index
NCD	Non Communicable Disease
NDoH	National Department of Health
RR	Raw rates

SD1	Standard deviation of spatial-lagged values of unstandardised criterion variable
SD2	Standard deviation of unstandardised criterion variable
SDRatio	Ratio between SD1 and SD2
SADHS	South Africa Demographic and Health Survey
SDG	Sustainable Development Goal
SIR	Standardised Incidence Ratio
SSA	Sub-Saharan Africa
Stats SA	Statistics South Africa
WHO	World Health Organization

Research Outputs

The following sections give a list of research outputs from this thesis.

Peer Reviewed Journal Publications

1. Darikwa, T. B., Manda, S. & Lesaoana, M., 2019 Assessing joint spatial autocorrelations between mortality rates due to cardiovascular conditions in South Africa. *Geospatial Health* **14**, 294–305.

2. Darikwa, T.B. & Manda, S.O., 2020 Spatial Co-Clustering of Cardiovascular Diseases and Select Risk Factors among Adults in South Africa. *Int. J. Environ. Res. Public Health* **17**, 3583.

Extended Abstracts in Conference Proceedings

1. Darikwa, T., Manda, S. O. & Lesaoana, M. 2015. Bivariate spatial autocorrelations (Theoretical Review). Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2015), 1-2 October 2015, Extended Abstracts, Bolivia Lodge, **Polokwane, South Africa**.

2. Darikwa, T., Manda, S. O. & Lesaoana, M. 2016. Small-Area Spatial Dependence Among Cardiovascular-Related Mortality rates in South Africa. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2016), 3-4 October 2013, Extended Abstracts, Bolivia Lodge, **Polokwane, South Africa**.

3. Darikwa, T., Manda, S. O. & Lesaoana, M. 2018. Assessing joint spatial autocorrelations between mortality rates due to cardiovascular conditions. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2018), 3-4 October 2013, Extended Abstracts, Bolivia Lodge, **Polokwane, South Africa**.

International Conferences

1. Darikwa, T., Manda, S. O. & Lesaoana, M. 2018. A New Approach to Assess Multivariate Spatial Autocorrelations- (Methodological Development). Southern Africa Mathematical Sciences Association Annual Conference (SAMSA 2018), 19-22 November 2018, **Palapye, Botswana**.

Other Conferences

1. Darikwa, T., Manda, S. O. & Lesaoana, M. 2019. Assessing Multivariate Joint Spatial Autocorrelations. Joint Conference of the Sub-Saharan Network (SUSAN) of the International Biometrics Society (IBS) and DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB)(SUSAN-SSACAB 2019), 8 - 11 September 2018, **Cape Town, South Africa**.

2. Darikwa, T., Manda, S. O. & Lesaoana, M. 2018. A New Method To Analyse Multivariate Spatial Autocorrelation - (Applications). South African Statistical Association (SASA 2018), 26 to 29 November 2018, **University of South Africa, South Africa**.

3. Darikwa, T., Manda, S. O. & Lesaoana, M. 2015. Investigating Bivariate Spatial Autocorrelations of Cardiovascular Mortality in South Africa: 2011. (SASA 2015), 29 November - 2 December 2015, **University of Pretoria, South Africa**.

4. Darikwa, T., Manda, S. O. & Lesaoana, M. 2014. Bivariate spatial autocorrelation measures. South African Statistical Association (SASA 2014), 27 - 30 October 2014, University of Rhodes, South Africa.

Chapter 1

Introduction

1.1 Background

In spatial statistics, spatial clustering statistical methods have long been used to group spatial objects into groups called clusters, so that objects in one cluster have similar characteristics compared to objects in other clusters. Most of the development in spatial clustering methods have focused on one areal health data (outcome), and the most widely used measure is the Moran's I index of spatial autocorrelation (Moran, 1950). Both local and global indexes are widely available and implemented in many geographic information system (GIS) software (Anselin, 1995; Anselin *et al.*, 2002; Waller & Gotway, 2004). The univariate Moran's I has recently been expanded to cases where there are two spatially measured health data (Lee, 2001; Anselin *et al.*, 2002).

Wartenberg (1985) was the first to derive a multivariate spatial measure using principal component and factor analysis. Anselin *et al.* (2002) extended the ideas of Wartenberg (1985) to develop a bivariate spatial association measure for both Moran's I local and global indexes. Alternative constructions of both the univariate and bivariate Moran's indexes have been developed (Lee, 2001, 2004; Chen, 2013, 2015). However, because of access to interrelated multiple geographic health data, there has been a need to develop spatial clustering statistical measures. For example, in studying the spatial epidemiology of cardiovascular diseases (CVDs) and risk factors, there has empirical evidence pointing to spatial co-recurrence in these (Ford & Highfield, 2016; Kandala *et al.*, 2013). In such situation, one can analyse each cardiovascular or risk factor separately using standard clustering statistics. However, such analyses have less power and could mean estimates not efficient. A multivariate measure of clustering for all the studied CVDs and risk factors are more appropriate as this would give a more powerful test of significance compared to individual analysis of the CVD conditions and their risk factors. In addition, estimating joint clustering of two or more CVDs may provide more evidence for an integrated intervention approach that targets all the modelled CVDs and risk factors instead of targeting only one CVD. Research work in this area of joint clustering is not well-developed, and to be best of our knowledge it is nonexistence for spatial epidemiology and public health where most of the health data measured are multivariate. This PhD is set in this context to develop multivariate spatial clustering methods and apply them to CVD-related conditions in South Africa.

1.2 Overview of cardiovascular diseases

Cardiovascular diseases data are going to be used in the applications of the autocorrelation methods that are going to be applied in this study. Thus, it is important to understand the epidemiology of the diseases so as to appreciate why we are using them as interrelated diseases. This section will first look at the increasing problem of cardiovascular diseases in the world and how it can negatively impact a country's development. This is followed by a subsection of risk factors of CVDs and then a subsection outlining why CVDs are spatial and has a tendency of clustering geographically.

1.2.1 The problem of cardiovascular diseases

The epidemic of non-communicable diseases (NCDs) that claim the majority of deaths in the world is led by CVDs. Cardiovascular diseases are comprised of a group of diseases relating to the heart or blood vessels. The biggest CVD killers in the world are cerebrovascular disease and ischaemic heart disease. Cerebrovascular disease (CVA) is a grouping of different conditions and diseases involving the blood vessels connecting to the brain. When these blood vessels are damaged or malformations accrue inside this may lead to the brain being damaged as it is deprived of oxygen and nutrients. The most common manifestation of CVA is a stroke. Ischaemic heart disease (IHD) or coronery heart disease occurs when disorders in the blood vessels of the heart result in the deprivation of oxygen and nutrients to the heart. The most common manifestation of IHD is a heart attack.

Global mortality attributed to NCDs has become so grave in recent times that NCDs have been included as one of the 17 sustainable development goal (SDG) targets where premature mortality (between 30-70 years of age) attributed to the diseases have to be reduced by one third by the year 2030 (WHO, 2015). It was estimated in 2013 that about a third of global deaths can be attributed to CVDs, while IHD, stroke and heart failure contribute about 80% of all CVD deaths (GBD Collaborators, 2017; Noubiap *et al.*, 2015). Recent studies have shown that the proportion of people dying from CVDs in the world is increasing, with WHO (2018) reporting that NCDs killed 41 million (71%) of the 57 million deaths in 2016, and most of these deaths (44% or 17.9 million) were attributed to CVDs. The 2016 NCD mortality of 41 million deaths represents a 16%

growth on 2006 deaths (Gaziano, 2007; WHO, 2018).

The increase in the NCD epidemic is being fuelled by marked rises in prevalence in the low and middle income countries (GBD Collaborators, 2017). Non-communicable diseases were generally regarded as a problem of European countries but of late low and middle income countries, particularly African countries, have recorded increases in prevalence, while Europe and and the Western countries have recorded decreases in NCD prevalence (Peer *et al.*, 2012; Wesonga *et al.*, 2016; Yaya *et al.*, 2018). Currently Africa is the region with the highest prevalence of hypertension (46%) in the world and, in 2016, IHD claimed the highest number of NCD deaths (511916, 5.8% of all deaths) on the continent, followed by stroke (373485, 4.2% of all deaths, with South Africa being the worst hit country in the region (GBD Collaborators, 2017).

In South Africa, about 69% of deaths due to NCDs occur before age 70 years in men compared to 54% in women (WHO, 2018). With more men economically active than women, the burden of NCDs, especially cardiovascular diseases (CVDs) and their related risk factors, has had a negative social and economic impact that affects the productive population. Absenteeism and deaths have cost companies lots of money in terms of productivity hours lost and replacement costs when recruiting new staff, while families lose out when a breadwinner is deceased. The country reportedly spends a quarter of its total public health service costs on trying to contain CVDs or 3% of its gross domestic product (Pestana *et al.*, 1996). It has also been reported that the cost of containing obesity and related CVDs in low- and middle-income countries is about 8% of their total public health service costs (Gaziano, 2007). Thus, a reduction in the prevalence of CVDs and related risk factors will not only reduce CVD mortality, but save companies and countries money, while at the same time boosting productivity and improvements in the quality of life.

1.2.2 Risk and environmental factors for cardiovascular diseases

Several studies in Africa have found high prevalence of risk factors of CVDs among the adult population (Alberts *et al.*, 2005; Matsha *et al.*, 2012; Neupane *et al.*, 2016; Njelekela *et al.*, 2009; Olawuyi & Adeoye, 2018; Peer *et al.*, 2012; Wesonga *et al.*, 2016; Yaya *et al.*, 2018).

The high prevalence of the risk factors of CVDs in Africa has been attributed to the nutritional transition taking place in Africa. In this transition, improved health systems means a reduction in communicable (infectious) diseases that are now giving way to increased NCDs as more people indulge in health-damaging lifestyles. Such lifestyles are exemplified by physical inactivity, smoking, heavy episodic alcohol (binge) drinking, low fruit intake, low vegetable intake, high salt intake and high intake of unhealthy meals. These are referred to as modifiable behavioural risk factors of CVDs and they usually give rise to modifiable biological or metabolic risk factors of CVDs. Modifiable metabolic risk factors include, among others, diabetes, raised blood pressure, raised plasma triglycerides or high cholesterol, and overweight or obesity (Alberti *et al.*, 2005; Pelzom *et al.*, 2017). A collection of these related modifiable biological risk factors of CVDs is referred to as metabolic syndrome (Alberti *et al.*, 2005). Metabolic syndrome (MetS) have the advantage that they are measurable and easily detectable in low-resourced settings compared to behavioural risk factors. As such, they tend to give a more accurate assessment of future cardiovascular related problems or mortality. Increased levels of MetS are indicative of future cardiovascular-related problems or mortality.

Factors associated with the presence of MetS and behavioural risk factors are known. Foremost among them are socio-economic and demographic factors such as gender, age, ethnicity, educational level, employment status, area or location of residence and wealthy or poverty status. Thus, primary prevention of CVDs that occurs at individual level mainly focuses on maintaining a healthy lifestyle. But an individual's healthy lifestyle, or lack of it, cannot be detached from the social, physical and geopolitical environment existing in their spheres of existence (van Rheenen, 2015). Hence a person's healthy lifestyle is a product of the community he or she lives in. Communities cannot be at the same level of nutritional or epidemiological transition as investments in healthy systems are bound to differ by region, so the intensity of the determinants of CVDs and their risk factors tends to differ between communities. The higher the intensity of the determinants, the higher the likelihood of the presence of the CVDs and their risk factors in the population of that community. Thus, a suitable prevention strategy must aim to reduce the average level or intensity of the determinants of CVDs in the population or in the community (van Rheenen, 2015).

1.2.3 Spatial clustering of cardiovascular diseases

A community-based approach to assumes that the level of CVDs is dependent on the distribution of the communities and the different intensities of the determinants within those communities (van Rheenen, 2015). Different communities are located in different regions, and evidence is such that CVD outcomes cluster geographically (Ford & Highfield, 2016). Additionally, CVDs share the same lifestyle risk factors, hence they tend to cluster together (Tsai *et al.*, 2009). Bradshaw *et al.* (1995) noted that the different lifestyles led by urban and rural persons, as well as contrasting cultures, genetic make-up and social classes of the population of South Africa, justifies doing analysis of the geographic distribution of CVDs.

1.2.4 Spatial epidemiology of cardiovascular diseases in South Africa

South Africa has one of the highest burdens of cardiovascular diseases in the region (Bradshaw *et al.*, 2006; Cappuccio & Miller, 2016). This has been fuelled by rapid urbanisation and changes in lifestyle (which are more sedentary), and high salt and fat and sugar diet dependency (Manning *et al.*, 1974; Steyn *et al.*, 2006). Prior to 1996, the country was reported to had spent a quarter of its budget (or 3% of its gross domestic product) in trying to contain CVDs (Pestana *et al.*, 1996). However, resources allocated to containing CVDs are constrained with a high emphasis placed on alleviating the problem of HIV/AIDs and other infectious diseases (Schutte, 2018). South Africa is a signatory to the World Health Organization's (WHO) Sustainable Development Goal (SDG) number 3 (WHO, 2013) that tasks Governments to proactively monitor, prevent and control NCDs. South African National Department of Health (NDoH) strategies aimed at reducing NCD morbidity, mortality and associated risk factors, have identified CVDs as a priority (NDoH, 2013).

Cardiovascular diseases and their risk factors are known to cluster geographically, depending on levels of deprivation (Ford & Highfield, 2016). In South Africa, the previously disadvantaged or deprived communities are exhibiting higher prevalence levels of metabolic syndrome, i.e., a collection of risk factors for CVDs and diabetes (Alberti *et al.*, 2005) than the advantaged communities, putting areas populated by the majority Blacks at high risk of mortality due to CVDs. This is attributed to, among other factors, a nutritional transition taking place in the country. It is fundamental to understand and monitor the changing spatial patterns of CVDs, and identify cluster areas of high mortality risk of NCDs if the SDG target is to be met. It is the objective of this study to assess joint clusters of CVDs in South Africa.

Investigation of the spatial variation and clustering of mortality in South Africa is not new. Descriptive risk maps have been used to describe the geographic variation of NCD mortality (Bradshaw *et al.*, 1995, 2006; Groenewald *et al.*, 2014), while univariate spatial autocorrelation measures have been used to determine the presence of spatial heterogeneity or variation in an area and to detect clusters of HIV/AIDS mortality (Tanser *et al.*, 2009), infant mortality (Sartorius *et al.*, 2011) and all-cause mortality (Sartorius *et al.*, 2010).

Univariate global indicators of spatial autocorrelation (GISA) are used to detect spatial heterogeneity or variation of cause-specific mortality for an area. They test the extent to which neighbours are similar or different in the region of study. One can use them to detect the presence of clusters, but they do not reveal the actual clusters (Waller & Gotway, 2004). Actual clusters are detected using local indices of spatial autocorrelation (LISA). Using univariate LISA to identify clusters entails investigating if high mortality risk of a single disease in a given area extends to neighbouring areas. Areas of high mortality risk that extend to nearby areas form a cluster known as “hot-spots ”, while areas of low mortality risk that extend to nearby areas form “cold-spots ” for the disease in question (Anselin *et al.*, 2002; Waller & Gotway, 2004).

Joint mapping of multiple disease outcomes has also been done in South Africa using the shared-spatial component method to establish ecological associations between HIV/AIDS and syphilis (Manda *et al.*, 2012), as well as multiple CVDs (Kandala *et al.*, 2014). The methods only permit, whether or not multiple diseases spatially co-exist, the measurement of one disease but do not measure the extent to which one disease in an area affect the burden of related diseases in adjacent areas. Sometimes, when dealing with interrelated diseases like CVDs, it is important to determine how they influence each other spatially. Univariate spatial autocorrelations only allow the measurement of spatial correlations of a disease in an area with itself in the adjacent areas. However, multivariate spatial autocorrelation techniques could be invaluable in providing useful insight into the spatial dependency of two or more interrelated disease outcomes.

1.3 Research Problem

Advancements in technology such as in Geographical Information Systems (GIS) and other computer software for data collection and storage has meant that geographically indexed data of interrelated health outcomes that were not readily available before are now ubiquitous. Nonetheless, methods of spatial autocorrelation that can handle multiple health outcomes are lacking. Bivariate measures have been developed for measuring the spatial association between two spatial data outcomes. However, spatial data with more than two outcomes are commonly collected in many different research studies. Thus, there may be a need to have a single measure that represents their joint spatial autocorrelation. We are not aware of such a spatial clustering statistics that has been developed for a general number of spatial data outcomes.

The development of multivariate spatial autocorrelation measures would permit applied researchers and users of spatial maps to show joint clusters for more than two interrelated spatial data outcomes. In public health, for example, such a display could support policies on integrated intervention approaches for two or more health outcomes.

1.4 Study aims and objectives

The primary aim is to develop multivariate spatial autocorrelation measures for interrelated spatial outcome data and then apply the derived measures to an analysis of multiple cardiovascular conditions, namely CVA, IHD,HHD and related risk factors in South Africa. These will be achieved through the following objectives:

Objective I: Review of bivariate spatial clustering methods and applying to cardiovascular diseases and associated risk factors in South Africa.

Objective II: Derivation and validation of multivariate spatial autocorrelation measure.

Objective III: A comparison of estimated clustering patterns in CVDs and risk factors in South Africa between univariate and multivariate spatial clustering methods.

1.5 Overview of thesis

This Chapter has given the background of the study, the burden of CVDs in South Africa, the research problem, the aim and objectives of the study. The next three chapters review and apply the currently available univariate and bivariate spatial autocorrelation measures. An illustration of the application of the spatial autocorrelation methods is done in Chapter 4, where applications are done to assess co-clustering of age-standardised incidence ratios and mortality rates of cardiovascular data and its risk factors. In the first illustration, particular attention is given to the following four cardiovascular diseases and their risk factors: stroke, heart attack, high blood cholesterol, hypertension and tobacco smoking. The second illustration involves application of Empirical Bayes approach in smoothing cardiovascular mortality rates and Poisson regression modelling in estimating mortality rates adjusted for age, race and poverty. Recently developed bivariate spatial autocorrelation by Lee (2001) and variants of the Moran's index were then used to identify pairwise co-clusters of Empirical Bayes smoothed rates and the Poisson regression adjusted mortality rates of cerebrovascular heart disease, ischaemic heart disease, hypertensive heart disease and diabetes. Chapters 5 and 6 provides a new multivariate spatial autocorrelation measure, based on canonical correlation analysis, that extends the analyses of the Moran's index of spatial autocorrelation to three or more health outcomes. A summary and conclusion are given in Chapter 7 with a discussion on the future direction of the study.

Chapter 2

Review of spatial autocorrelation methods

2.1 Introduction

This Chapter describes the foundation and theory of statistical spatial clustering measures. The univariate and bivariate Moran's indexes and a recently developed alternative construct of the Moran's index of spatial association are described.

Spatial autocorrelation is the correlation of a geo-referenced variable to itself geographically. If there is geographical interdependence between geo-referenced observed values then this data is said to exhibit spatial autocorrelation. When there are random spatial patterns then the data shows no spatial autocorrelation. Spatial autocorrelation measures the degree to which one area is similar or dissimilar to its geographically contiguous areas. Spatial autocorrelation, like Pearson's autocorrelation function, can be positive or negative. Positive spatial autocorrelation occurs when geographically contiguous areas are similar while negative spatial autocorrelation occurs when the geographically contiguous areas are not similar.

Moran (1950) was one of the early pioneers of "spatial correlation" which Cliff & Ord (1969) later referred to as "spatial autocorrelation." Cliff & Ord (1969) were the first to develop a framework for testing spatial autocorrelation in a given region under the null hypothesis of spatial randomness. The measures that are used to test the extent of spatial autocorrelation are divided into global and local measures of spatial autocorrelation Anselin (1995). Global measures look at the spatial autocorrelation for the whole area under study while the local measures look at spatial autocorrelations at local neighbourhoods. These measures will be looked at in the subsequent subsections.

2.2 Spatial autocorrelation methods

In this section we review spatial autocorrelation methods but before that we will define the notion of spatial weights which is important in the calculation of spatial autocorrelations.

2.2.1 Spatial weights

Analysing spatial autocorrelation requires one to quantify location. Knowledge of the neighbourhood structure of the regions under study is important for one to be able to quantify location in order to analyse spatial effects. Spatial effects here refer to geographical dependence and geographical heterogeneity. The neighbourhood structure is represented as a proximity matrix known as a spatial weight matrix, W. A spatial weight matrix, $W = \{w_{ij}\}_{i,j=1}^{n}$, is an $n \times n$ matrix that defines the closeness or connectedness of two areas A_i and A_j in space, where $\{w_{ij}$ is the ij^{th} element of the weight matrix. Spatial weight matrices can either be contiguity (neighbourhood) or distance based. A contiguity structure shows how one area is located in relation to others, whereas distance based structures show the relative spatial distance of one area from the others. In contiguity structures one would expect neighbours to have more spatial dependence than those that are far away. In distance based neighbourhood structures, spatial dependency is expected to decline as the distance between areas increases. Areas that are far from each other should exhibit spatial heterogeneity (dissimilar relationships), while those that are close should show similar relationships (Kosfeld, 2010).

The spatial contiguity matrices are the simplest there are in terms of neighbourhood structure definition and their contiguity based spatial weights are defined as follows:

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are close or connected or neighbours} \\ 0 & \text{otherwise}. \end{cases}$$

where areas A_i and A_j are said to be neighbours or connected if either (1) they share a border (rook contiguity/simple contiguity); or (2) they share a corner (bishop contiguity); or (3) they share either a border or a corner (queen contiguity) (Kosfeld, 2010). When neighbours are adjacent, as is the case when they share a boundary, then the weight matrix is referred to as first-order adjacency matrix. Elements in the diagonals of the weight matrix are zeros since an area is not considered to be its own neighbour.

The definition of neighbourhood can be expanded to higher-order adjacency matrices in which we incorporate the neighbours of a neighbour. This brings about the concept of second-order neighbours, then third-order neighbours, $etcetera$. In general a k^{th}-order adjacency weight matrix is defined as follows:

$$w_{ij} = \begin{cases} \omega_k & \text{if } i \text{ and } j \text{ are } k^{th}\text{-order neighbours and } k = 1, 2, ..., n \\ 0 & \text{otherwise.} \end{cases}$$

where the corresponding weight of each order can be written in vector form: $\omega = (\omega_{\mathbf{1}}, \omega_{\mathbf{2}}, ..., \omega_{\mathbf{n}})$. Weights are assigned to the neighbours such that the nearest neighbours have higher weights while the furthest neighbours have the lowest weights. The limitations of adjacency weight matrices is that they do not take into consideration differences in the sizes of the different areas. Alternatively, one can define the weight matrix based on distance.

The simplest of the distance-based spatial matrix, like the contiguity matrices, is also a binary connectivity matrix defined such that two areas A_i and A_j are neighbours if the distance between them, d_{ij} , is less than a specified distance, say δ , beyond which autocorrelation is not expected (Kosfeld, 2010). This structure is called the cross hatched or distance band contiguity. There are many ways of measuring distance but the most commonly used distance is the Euclidean distance between centroids of the areas (Waller & Gotway, 2004; Kosfeld, 2010). Similarly defined is the k-Nearest Neighbour contiguity, where area A_j is one of the k areas close to A_i (Waller & Gotway, 2004; Kosfeld, 2010).

Functional distance based spatial weight matrices have also been formulated. One such example is that based on the power function, $w_{ij} = d_{ij}^{\alpha}$, where α is the power parameter. When α is equal to negative 1 we have an inverse distance and when it is equal to 2 we have a quadratic inverse distance which is also known as the gravity model. The distance between two areas A_i and A_j , d_{ij}, can be measured from the centroid of the areas or from major cities or any points so chosen to be representative of the areas (Waller & Gotway, 2004; Kosfeld, 2010).

Functional distance based spatial weight matrices can be based on the inverse and exponential functions. The spatial weight matrix based on the inverse matrix is derived using Equation 2.1,

$$w_{ij} = \begin{cases} d_{ij}^{-\alpha} & \text{if } i \neq j \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

while that based on exponential function is given by Equation 2.2,

$$w_{ij} = \begin{cases} e^{-\alpha \frac{d_{ij}}{\bar{d}}} & \text{if } i \neq j \\ 0 & \text{otherwise,} \end{cases} \tag{2.2}$$

where d_{ij} is the distance between the centroid of area i and that of area j, while $\bar{d}$ is the average of all the distances between the areas under study. There are other forms of spatial contiguity not discussed here but one can see Waller & Gotway (2004) or Kosfeld (2010) for further reading.

There are times when some areas have or are suspected to have more neighbours than others. This can occur with irregular polygons where certain areas may be smaller or bigger in size than others and thus have more neighbours than others. One may want to adjust for this fact by creating proportional weights for the number of neighbours for an area (Waller & Gotway, 2004; Kosfeld, 2010). This is achieved through the creation of a row-standardised weight matrix (Kosfeld, 2010) whose entries will be given by:

$$w_{ij}^{std} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

This standardisation is appropriate for this study in which irregularly shaped municipalities of South Africa are considered as the unit of analysis. In most instances the spatial weight matrix, $\mathrm{W} = [w_{ij}]_{n \times n}$, is standardised to make it a unitary matrix such that $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$. Standardisation also leads to easier interpretation of results.

The spatial weight matrix, $\mathbf{W} = [w_{ij}]$, is assumed to be unitary with the following three properties:

- It is symmetric. i.e., $w_{ij} = w_{ji}$ or $\mathbf{W} = \mathbf{W}^T$;
- Its diagonal elements are all zeros, ie., $w_{ij} = 0$ for all i; and
- It must satisfy the normalisation condition, i.e., $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$.

Chen (2013) provided two ways in which a contiguity weights matrix with zero diagonal elements, $\mathbf{V} = [v_{ij}]_{n \times n}$, can be made a unitary spatial weights matrix:

1. $w_{ij} = \frac{v_{ij}}{\sum_i^n \sum_j^n v_{ij}}$, or
2. $w_{ij}^{*} = \frac{n \cdot v_{ij}}{\sum_i^n \sum_j^n v_{ij}}$.

2.2.2 Univariate spatial autocorrelation

2.2.3 Global indexes of spatial autocorrelation

Global indexes of spatial autocorrelation (GISA) are used to determine the extent to which neighbours are similar in the study region. GISA can only detect the presence of clusters, but do not identify where the clusters are located (Waller & Gotway, 2004). There are a number of Global tests such as the quadrat method, the nearest neighbour method, Geary's C and the global Moran I test (Waller & Gotway, 2004). The most popular of the GISA is the global Moran I statistic Griffith (1987). It is this method which is at the centre of this PhD study and is discussed next.

Global Moran index

The global univariate Moran I statistic measures the extent of the linear relation between the observed geo-referenced data $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ and their corresponding spatial lags (a weighted average of neighbouring values). The global Moran's I using the standardised spatial weights is given by:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \cdot (x_i - \bar{x}) \cdot (x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\sum_{j=1}^{n} w_{ij} \cdot (x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.3}$$

where w_{ij} is the ij^{th} element of the spatial weight matrix, which is a measure of the spatial proximity between municipalities i and j. This may also be written as

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{2.4}$$

which can be expressed in matrix or quadratic form as follows:

$$I = \mathbf{z}^T \mathbf{W} \mathbf{z}, \tag{2.5}$$

where $\mathbf{z}^T = [z_i] = [(x_i - \bar{x})/\sigma_x]$ is a vector of the standardised z-score values of the x_i's, $\mathbf{W} = [w_{ij}]$ is the spatial weight matrix and σ_x is the standard deviation of the x_i's.

In order to test if the null hypothesis of spatial randomness or no spatial autocorrelation is significant, one can assume that sampling is from areas whose spatial process realisations are normally distributed with constant mean and constant variance for each area. Otherwise, a randomisation approach is implemented.

2.2.4 Local indexes of spatial autocorrelation

Having established the presence of an underlying pattern or spatial clustering in the data using GISA such as the Moran's I discussed in the previous section, one may be interested in detecting clusters that gave rise to a significant GISA. This is done using local indicators of spatial autocorrelation (LISA). Furthermore, one can also identify outliers using LISA. "Hot-spots" and "cold-spots" are associated with positive spatial autocorrelation. When the sign of local spatial autocorrelations negates that of global spatial autocorrelation then that area is an outlier. For instance, when the global statistics are saying that there is positive spatial autocorrelation, then local areas with negative spatial autocorrelations will be spatial outliers. Although there are various LISA such as the Getis G statistic (Ord & Getis, 1995), this PhD study only consider the most widely used local Moran's I. The local Moran's I statistic is given by:

$$I_i = \frac{n(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij} \cdot (x_j - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2}. \tag{2.6}$$

Equation 2.6, just as with the global Moran's I, can be rewritten as a function of the standardised z-scores as:

$$I_i = z_i \cdot \sum_{j=1}^{n} w_{ij} z_j, \ i \neq j. \tag{2.7}$$

The local Moran's I can be written in matrix notation as:

$$I_i = \mathbf{z}^T \mathbf{W}_i \mathbf{z}, \tag{2.8}$$

where $\sum_{i=1}^{n} I_i = I$ and $\mathbf{W}_i$ is a global spatial weight matrix whose entries are zero with the exception of the entries in the i^{th} row.

2.3 Linear regression-based Moran's index

Chen (2013) developed a regression approach to univariate Moran's index with the objective of making it simpler to implement and interpret. This was achieved by taking into consideration certain relationships and assumptions that will be outlined in the subsequent paragraphs.

Firstly, it was shown that

$$\mathbf{z}^T\mathbf{z} = \sum_{i=1}^{n} z_i^2 = n, \tag{2.9}$$

since the norm or length of z is given by:

$$\|\mathbf{z}\| = \sqrt{\sum_{i=1}^{n} z_i^2} = \sqrt{\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{\sigma_x})^2} = \sqrt{\frac{n}{\sigma_x^2} \cdot \frac{1}{n} \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{n}{\sigma_x^2} \cdot \frac{1}{n} \cdot \sigma_x^2} = \sqrt{n}. \tag{2.10}$$

Chen (2013) adopted the unitary weight matrix approach which will also be applied in this study.

An ideal spatial weight matrix (ISWM), M*, was derived by pre-multiplying both sides of Equation 2.5 by z to obtain

$$\mathbf{z}I = \mathbf{z}\mathbf{z}^T\mathbf{W}\mathbf{z}. \tag{2.11}$$

Since I is a scalar Equation 2.11 can be rewritten as

$$\mathrm{M}^*\mathbf{z} = \mathbf{z}\mathbf{z}^T\mathbf{W}\mathbf{z} = I\mathbf{z}, \tag{2.12}$$

where

$$\mathbf{M}^* = \mathbf{z}\mathbf{z}^T\mathbf{W} = \begin{pmatrix} z_1 \sum_{j=1}^n w_{1j}z_j & z_1 \sum_{j=1}^n w_{2j}z_j & \cdots & z_1 \sum_{j=1}^n w_{nj}z_j \\ z_2 \sum_{j=1}^n w_{1j}z_j & z_2 \sum_{j=1}^n w_{2j}z_j & \cdots & z_2 \sum_{j=1}^n w_{nj}z_j \\ \vdots & \vdots & \ddots & \vdots \\ z_n \sum_{j=1}^n w_{1j}z_j & z_n \sum_{j=1}^n w_{2j}z_j & \cdots & z_n \sum_{j=1}^n w_{nj}z_j, \end{pmatrix} \tag{2.13}$$

is the ISWM. The diagonal of M* consists of the Moran's I based local indicators of spatial autocorrelation, I_i. Additionally, the trace of M* gives the global Moran's I, i.e., $Tr(\mathrm{M}^*) = \sum_{i=1}^n I_i = I$. Chen (2013) defined $f^* = \mathrm{M}^*\mathbf{z}$ so that

$$f^* = \mathrm{M}^*\mathbf{z} = \mathbf{z}\mathbf{z}^T\mathbf{W}\mathbf{z} = I\mathbf{z}. \tag{2.14}$$

It follows that the Moran's index is the gradient obtained when f^* is regressed on z. This way of calculating the Moran's index is simple to apply and easy to discern.

Using the fact that the maximum eigenvalue of the matrix zz^T was its dimension, n, Chen (2013) showed that

$$f = \mathbf{M}\mathbf{z} = \mathbf{z}\mathbf{z}^T\mathbf{W}\mathbf{z} = n\mathbf{W}\mathbf{z} \tag{2.15}$$

where M, the so called Real Spatial Weight Matrix (RSWM).

Note that Mz is approximately close to M*z when the spatial autocorrelation is high but significantly different when otherwise (Chen, 2013). On the one hand, the relationship between f^* and z is a regression line fit whose gradient is an estimate of the Moran's index, as alluded earlier. On the other hand, the relationship between f and z is the actual observed spatial pattern. Thus, a graph of f^* on z with coordinates (z_i, f_i^*) will give points following a straight line while a graph of f on z with coordinates (z_i, f_i) will have points showing an irregular pattern. The gradient of a fitted line to either of these two plots will give an estimate of the univariate Moran's index. The plots described in this paragraph represent the revised Moran's I scatter-plot by Chen (2013).

In order to assess the adequacy of the Moran's index derived using Chen (2013), a diagnostic check of the residuals needs to be performed. The residuals of the spatial autocorrelations,e_f, in the formulation by Chen (2013) is given by

$$\mathbf{e}_f = f - f^* = \mathbf{Mz} - \mathbf{M}^*\mathbf{z}, \tag{2.16}$$

and the standard error, s_f, of the residuals is defined by Chen (2013) as

$$\mathrm{s}_f = \sqrt{\frac{1}{n} e_f^T e_f}. \tag{2.17}$$

The diagnostic check of the residuals is two-fold. First, we check if the residuals follow a normal distribution, failing of which an adjustment must be done to the weights matrix, otherwise a new weights function has to be chosen. Second, the standard error of the residuals could have an upper-bound of 0.15, i.e., $s_f < 0.15$ (Chen, 2013).

2.3.1 Framework for significance testing

Lee (2004) provided a framework for significance testing of indicators of spatial association measures which is based on the permutation tests of Mantel (1967). The number of possible permutations as in the Mantel (1967) proposal can be prohibitively large even in this age of high speed computers leading researchers consider sampling from the permutations (Kosfeld, 2010). In Monte Carlo simulation method locations are randomly reordered for all the cases. A permutation of the randomly reordered locations is randomly selected and the locations are assigned to the cases. Assuming that the locations are not related to the cases then all the possible permutations have an equal chance of being selected.

In this framework an observed indicator of spatial association measure, $\Gamma_{(obs)}$, say, is determined for the health outcome of interest. Sampling of the permutations is repeated many times, say for $l = 1, 2, .., L$ permutations, and each time a corresponding test statistic $\{\Gamma_{(l)} : l = 1, 2, ...L\}$ is calculated using the spatial autocorrelation measure. The number of permutations will depend on the number of cases and the significance level of the test. While a large value of l is required for the empirical distribution to give a good approximation of the null distribution it is recommended that a minimum sample of 99 permutations for a 5% significance test level and a minimum of 999 samples for a 1% significance test level (Waller & Gotway, 2004). A proportion of $\Gamma_{(l)}$'s that are greater than the observed $\Gamma_{(obs)}$ statistic is determined and a $p - value$ is calculated as $p - value = P(\Gamma_{(l)} > \Gamma_{(obs)})$. High $p - value$s suggest that there is no evidence of spatial clustering in the data. This is the same framework that has been used in the significance testing of the original Moran's index.

2.4 Summary of the chapter

In this chapter the relevant literature pertaining to univariate spatial autocorrelation measures has been presented. First, the widely used Moran's index was presented. Then the an alternative construct of the Moran's index of spatial autocorrelation measure based on a regression approach was reviewed. In this linear regression based approach, Chen (2013) intention is to present the Moran's index in way that is easy to appreciate and interpret while simultaneously seeking to give a basis for the choice of the spatial weight matrix for use in the analysis. The approach suggested using the regression standard errors as a basis for determining the spatial weight to use. A suitable spatial weight matrix would give a regression standard error of less than 0.15 (Chen, 2013). The approach by Chen (2013) provides an innovative way of displaying the scatter plots and deriving the Moran's index.

The Moran's index has been widely applied in spatial epidemiology, but linear regression based approach by Chen (2013) is yet to be extensively applied to real life problems. These two methods, i.e. the traditional way and the Chen way, will be applied to two sets of CVD data in South Africa in the next chapter.

Chapter 3

Application of univariate spatial autocorrelation methods to cardiovascular diseases in South Africa

3.1 Introduction

This Chapter presents applications of the univariate Moran's index and the linear regression approach to the Moran's described in Chapter 2 to real life data. An application of the univariate spatial autocorrelation measures is done to two different South African health outcomes datasets recorded at municipality level. The first illustration is done on self-reported prevalence of two CVDs stroke and heart attack, and three risk factors of interest namely raised blood pressure, raised cholesterol and smoking. A second illustration was in detecting individual clustering of mortality attributed to cerebrovascular, ischaemic hypertensive heart diseases and diabetes in South Africa.

The data used are recorded at local municipal level. There are 52 district municipalities and 234 local municipalities in South Africa (2011 boundaries). These municipalities are in the form of irregular polygons. Municipalities were then considered to be fixed and countable areal units. Thus, areal data spatial autocorrelation measures were considered for analysis.

3.2 Application to cardiovascular prevalence data

3.2.1 Data

Secondary data collected as part of the South African Demographic and Health Survey in 2016 (SADHS 2016) were used. The SADHS 2016's adult health module recorded information that included, among others, self-reported prevalence of two CVDs stroke and heart attack, the three risk factors of interest raised blood pressure, raised cholesterol and smoking for both male and female adults aged 15 years and older. A total of 12 717 adults were targeted for this adult health module, but only 10 336 responded. In this application, we used district for the spatial analysis.

Figure 3.1 shows the map of the 52 districts of South Africa and the number of the sampled adults, which ranged from 23 to 544 per district, with an average of 203 subjects. Due to the sample design, Central Karoo, which has a very sparse population was not included in the sample, and hence excluded from the analyses. For the purposes of our study, the data were stratified by gender (male and female) and age (15-39 years (young adults) and 40-64 years (adults). A cut off point of 40 years was used as it has been observed that the burden of CVDs increased significantly after the age of 40 years (Petoumenos *et al.*, 2014).

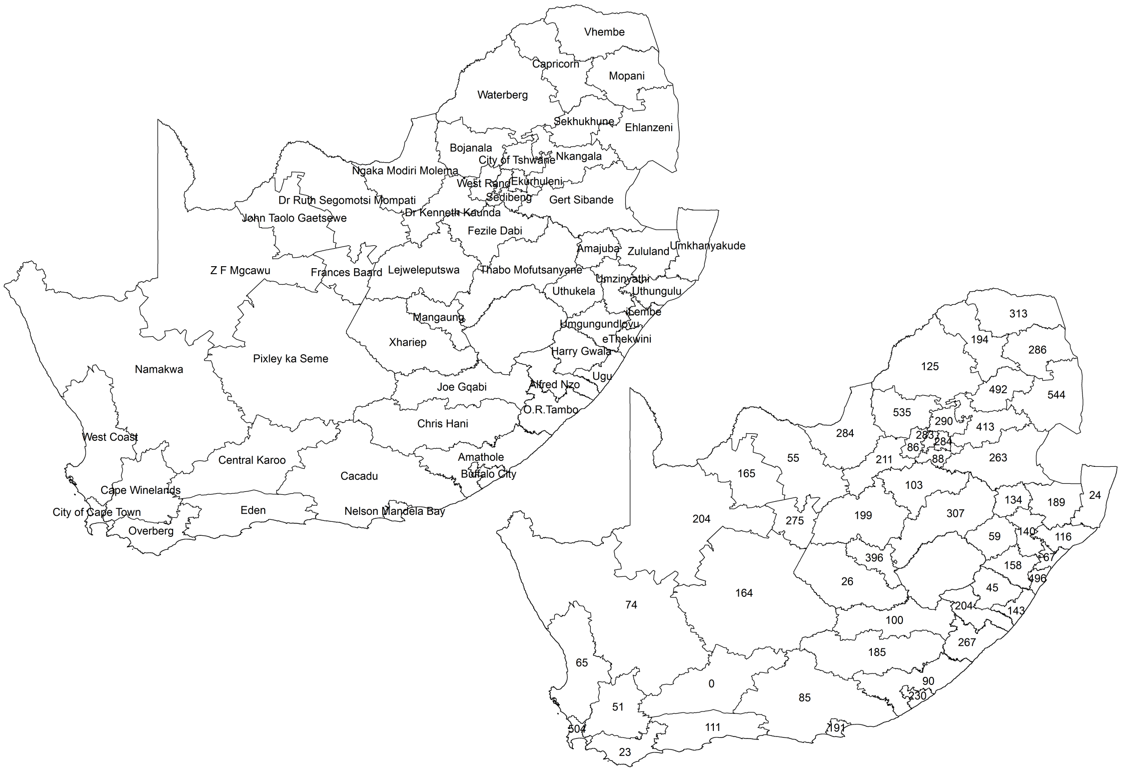


Figure 3.1: The South African map showing the district names and sample sizes drawn from each district in the 2016 SADHS.

3.2.2 Variable definitions

The CVD variables considered in this study were stroke and heart attack and are defined below.

- Stroke: A dichotomous variable in which a person who self-reported to have been diagnosed with stroke is assigned a value 1 and zero otherwise.
- Heart attack : A dichotomous variable in which a person who self-reported to have been diagnosed with heart attack is assigned a value 1 and zero otherwise.

Three risk factors of CVDs considered in this study are smoking, hypertension and high blood cholesterol. These are defined below.

- Smoking: A dichotomous variable in which a respondent who stated that he/she smokes daily or occasionally is assigned a value 1 and zero otherwise.
- High blood cholesterol (HBC): A dichotomous variable in which a person who self-reported to have been diagnosed with high cholesterol is a success and is assigned a value 1 and zero otherwise.
- Hypertension : This was defined as a systolic BP measurement of at least 140 mmHg or diastolic BP measurement of at least 90 mmHg or self-report of hypertension diagnosis as hypertensive or on hypertension medication.

3.2.3 Descriptive statistics of the variables

A total of 9 154 participants aged between 15 and 64 years were sampled, of which about 5 337 (58%) were females, and 5 848 (64%) were aged between 15 and 39 years. Table 3.1 shows summary statistics of the prevalence of variables used in the analysis by districts. The summary statistics were derived for all the data combined, gender, age groups and by both age-gender. District level prevalence rates range from 0% to 100% across the CVDs and related risk factors. On average, the prevalence of reported heart attack (2.4%) is twice the prevalence of strokes (1.2%). Heart attack prevalence rates are higher for females (2.8%) than among males (1.8%).

An average of 22% of those sampled described themselves as daily or regular smokers. The higher proportion of the smokers are among males (39%) than females (10%). Hypertension has a very high national prevalence of 34%, which is higher among women (36%) than men (31%). The prevalence of the

CVDs and their risk factors are increasing with an increase in age, as would be expected.

Table 3.1: Summary statistics of the prevalence of CVDs and related risk factors across the districts.

Sub-Group	CVD or Risk Factor	Minimum	First Quartile	Median	Mean	Third Quartile	Max	Sample size
All	Stroke	0.0%	0.0%	1.1%	1.2%	1.6%	9.1%	9,154
	Heart attack	0.0%	1.1%	2.2%	2.4%	3.5%	9.1%	
	Smoking	0.0%	15.8%	20.8%	21.9%	25.0%	53.2%	
	HBC	0.0%	0.7%	1.5%	2.2%	3.0%	9.1%	
	Hypertension	9.5%	27.6%	32.9%	34.3%	41.2%	65.4%	
Male	Stroke	0.0%	0.0%	0.0%	1.1%	1.2%	20.0%	3,817
	Heart attack	0.0%	0.0%	1.1%	1.8%	2.2%	20.0%	
	Smoking	0.0%	31.8%	37.8%	39.0%	46.1%	81.8%	
	HBC	0.0%	0.0%	0.7%	1.8%	2.9%	20.0%	
	Hypertension	0.0%	22.4%	29.6%	31.2%	40.8%	80.0%	
Female	Stroke	0.0%	0.0%	1.3%	1.3%	2.2%	5.4%	5,337
	Heart attack	0.0%	1.1%	2.6%	2.8%	4.1%	10.7%	
	Smoking	0.0%	1.7%	4.8%	9.9%	12.3%	41.7%	
	HBC	0.0%	0.2%	1.6%	2.4%	2.9%	12.5%	
	Hypertension	15.4%	29.9%	35.7%	36.3%	41.6%	66.7%	
15-39	Stroke	0.0%	0.0%	0.0%	0.5%	0.8%	2.9%	5,848
	Heart attack	0.0%	0.0%	0.3%	0.9%	1.4%	4.6%	
	Smoke	0.0%	14.4%	19.6%	20.1%	23.4%	46.6%	
	HBC	0.0%	0.0%	0.0%	0.6%	0.8%	5.7%	
	Hypertension	0.0%	15.3%	20.8%	21.0%	27.6%	40.9%	
40-64	Stroke	0.0%	0.0%	1.5%	2.2%	3.5%	10.0%	3,306
	Heart attack	0.0%	0.6%	4.5%	4.5%	6.5%	14.1%	
	Smoking	0.0%	15.7%	21.8%	23.4%	29.6%	66.7%	
	HBC	0.0%	1.3%	3.3%	4.6%	5.7%	19.1%	
	Hypertension	25.0%	47.2%	56.8%	55.5%	63.9%	92.3%	
Male 15-39	Stroke	0.0%	0.0%	0.0%	0.1%	0.0%	2.0%	2,574
	Heart attack	0.0%	0.0%	0.0%	0.7%	1.3%	5.6%	
	Smoking	0.0%	27.4%	35.7%	35.4%	44.5%	66.7%	
	HBC	0.0%	0.0%	0.0%	0.4%	0.0%	5.9%	
	Hypertension	0.0%	15.2%	21.2%	22.2%	29.7%	50.0%	
Female 15-39	Stroke	0.0%	0.0%	0.0%	0.7%	1.3%	4.4%	3,274
	Heart attack	0.0%	0.0%	0.0%	1.1%	1.9%	7.3%	
	Smoking	0.0%	1.2%	3.9%	8.0%	7.9%	40.5%	
	HBC	0.0%	0.0%	0.0%	0.8%	1.1%	10.3%	
	Hypertension	0.0%	15.2%	20.9%	19.8%	25.4%	37.7%	
Male 40-64	Stroke	0.0%	0.0%	1.2%	2.1%	2.5%	25.0%	1,243
	Heart attack	0.0%	0.0%	0.0%	3.3%	5.9%	25.0%	
	Smoking	0.0%	35.0%	42.9%	45.2%	52.1%	100.0%	
	HBC	0.0%	0.0%	0.0%	3.7%	5.7%	25.0%	
	Hypertension	0.0%	36.5%	50.0%	46.9%	61.3%	100.0%	
Female 40-64	Stroke	0.0%	0.0%	1.2%	2.1%	4.7%	8.3%	2,063
	Heart attack	0.0%	0.0%	4.9%	5.1%	7.6%	18.8%	
	Smoking	0.0%	0.8%	6.9%	11.5%	17.1%	46.2%	
	HBC	0.0%	0.0%	3.1%	4.9%	5.3%	27.8%	
	Hypertension	33.3%	52.1%	61.4%	60.3%	67.6%	100.0%	

Key: CVD, cardiovascular disease; HBC, high blood cholesterol.

3.2.4 Variables correlations

The correlations between district level prevalence data were assessed using Pearson correlation coefficient and the results are shown in Table A.1 (Appendix

A). Correlation is important in measuring point-to-point association between datasets and captures the relationship of the health outcomes within a district. The correlations were assessed for overall sample, by age, gender and age-gender. When all the data is used, it was observed (results shown in Table A1) that there is a strong association (±0.5 to ±1) between stroke and heart attack (0.85); stroke and HBC (0.82); heart attack and HBC (0.71); smoking and HBC (0.55); smoking and hypertension (0.73); and HBC and hypertension (0.53). Removing the effects of gender, there is a weak association between the CVDs and their risk factors for females, but strong positive association for males for stroke and heart attack (0.81); stroke and HBC (0.66); heart attack and HBC (0.64).

Generally, the correlations, with a few exceptions, reduce with each further division: gender or age and then age-gender. It is further observed that the direct association between HBC and the two CVD outcomes of stroke and heart attack are more pronounced in the ages 40-60 years than in the 15-40 year age group. Similarly, in a district with higher proportion of individuals smoking, there are also a relatively higher proportion of stroke (especially in ages 40-64 years). However, the direct association between hypertension and the CVD outcomes did not show similar trends, with relatively small correlations between them.

3.2.5 Age-gender stratified clustering analysis

Maps of raw prevalence rates

The distribution of the prevalence of the CVDs and their risk factors over different districts of South Africa shed more light on the similarities or dissimilarities in their spatial variation across the country. Figure 3.2 shows the quintile maps of the raw prevalence rates for the CVDs and their risk factors over different districts of South Africa. Central Karoo has undefined values and is

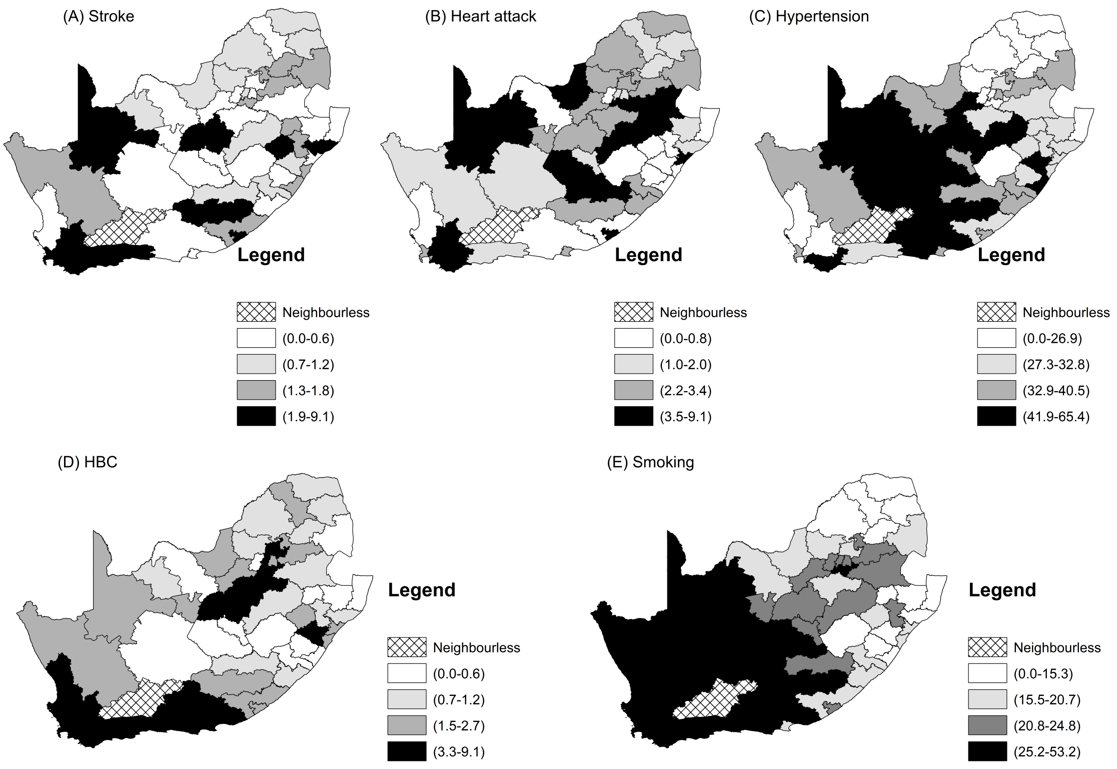


Figure 3.2: Maps of prevalence rates of the CVDs and their related risk factors across the district for the sample age group 15-64 years.

shown as neighbourless. The darker the colour, the higher the raw prevalence in the district. Stroke (see Figure 3.2 A) and smoking have black colours dotted across the country. High values of heart attack in Figure 3.2 B are concentrated in the centre of the country, stretching from the north to the south. Hypertension (Figure 3.2 C) has a belt of high prevalence values stretching from the west coast to the south coast of the country. The high rates for high blood cholesterol stretch from the central districts of the country all the way to the west. It is in these high value areas that high risk clusters for each given CVD or risk factor are expected to form.

Figure A.1 in Appendix A shows the maps of prevalence rates by gender, although there are some overlap in terms of common districts of high values.

It is observed that the geographic distributions differ by gender and also differ from the distribution when all the data are used. For example, the high rates of HBC in Figure A.1 C cover some districts in the middle of the country and some in the south-western part of the country, but the high rates for males in Figure A.1 G cover a few districts in the south western part of the country. Similarly, females have a different spatial distribution for HBC as shown in Figure A.1 H, but almost similar when all genders are combined. The distribution of the CVDs and their risk factors also differs by age as shown in Figure A.2 (Appendix A). It was further revealed that the spatial distribution of the CVDs changes for different age-gender combinations as shown in Figure A.3 and Figure A.4.

Clustering using data for different age-gender combinations

The global univariate spatial autocorrelation indexes for the prevalence of CVDs and identified risk factors for all participants, and also split by age or gender were calculated and are shown in Table 3.2. In the case of all participants, there is evidence of spatial clustering at 5% significance level with the exception of heart attack (Moran's I=-0.013). Stroke spatial clustering is significant for the male (0.136) and 40-60 years (0.150) categories. There is no evidence of spatial clustering in the data of smoking for males and HBC for age group 15-39 years olds.

Figure 3.3 shows the clusters for CVDs and their risk factors that exhibit significant spatial dependents at district level in South Africa, for all participants and for the different genders. The key shows "hot-spots " (High-High) in the black colour and "cold-spots " (Low-Low) in the light grey colour. Smoking spatial clustering is well pronounced in the with "hot-spots " clusters in western part of the country (Figures 3.3 B, F and J). In Figures 3.3 D, H and L, it can be observed that the "hot-spot " cluster of hypertension are in the central part of the country.

Table 3.2: Global univariate spatial autocorrelation Moran's *I* values for the CVDs and their risk factors for all participants split by age and gender.

	Stroke	Heart attack	Smoking	HBC	Hypertension
Male	0.136**	-0.044†	0.142†	0,239**	0.150**
Female	-0.171†	0.018†	0.787**	0.246**	0.351**
15-39 years	-0.033†	0.074†	0.294**	0.012†	0.282**
40-64 years	0.150**	-0.057†	0.600**	0.299**	0.317**
All	0.203**	-0.013†	0.662**	0.503**	0.329**

Key: HBC, high blood cholesterol; †, insignificant at 5% level; **, significant at 5% level.

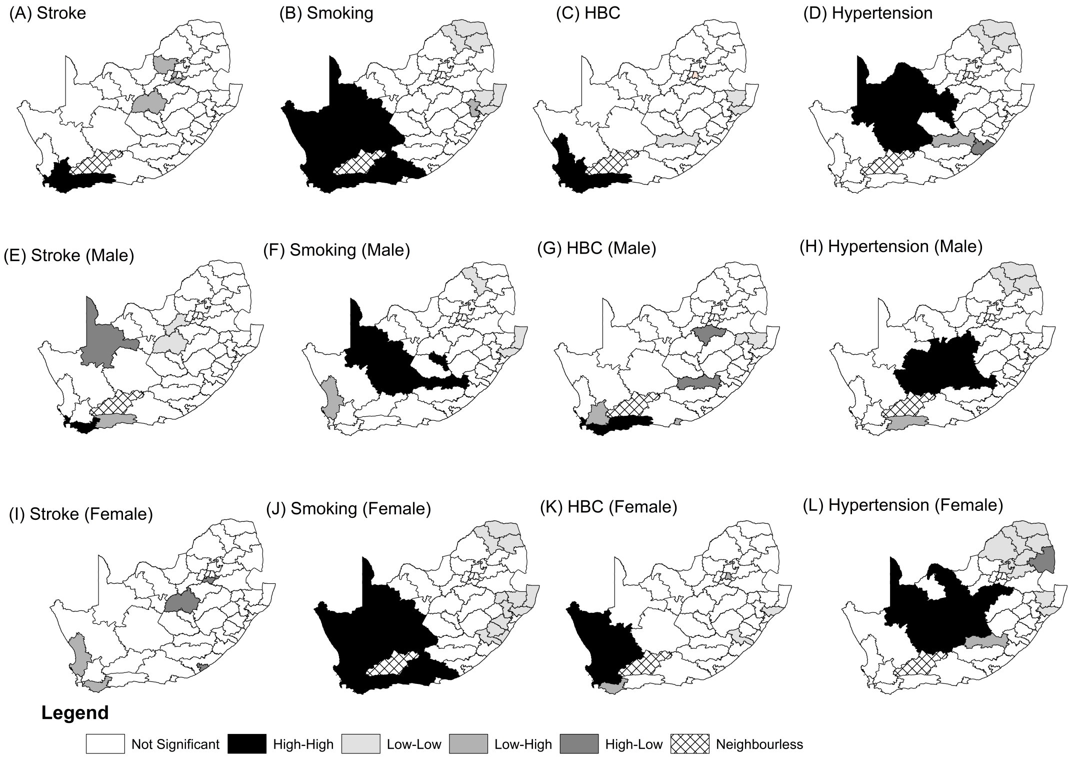


Figure 3.3: Univariate spatial clusters of CVDs and their risk factors with significant association for all participants, males and females using raw rates.

Generally stroke and HBC have "hot-spots" clusters found in the south western part of the country and comprises of City of Cape Town, Cape Winelands, Overberg and Eden Districts. Clustering of stroke for males and females were found not significant and it can be seen in Figures 3.3 E and I that the data is generally randomly distributed with no clusters present. There are clear differences in spatial clustering by gender for example for HBC “hot-spots” clusters for males (Figure 3.3 G) are in the southern most part of the country but those for females (Figure 3.3 K) are in the western part of the country.

Similarly, it was observed in Figure A5 (Appendix A) that spatial clustering is also dependent on age groups. We also approached the analyses using observed prevalence data within each age-gender groupings, namely males aged 15-39 years; females aged 15-39 years; males aged 40-64 years; and females aged 40-64 years (the results are not presented here).

It has been shown in this section that the spatial clusters vary when different age-gender combinations are analysed using raw prevalence of the two cardiovascular diseases and the three associated risk factors. While spatial analysis employing age-gender combinations is a novel way of analysing age-gender effects on spatial co-clustering it is also fraught with some inherent problems. Key among them is the fact that the sample sizes reduced when stratified by age and gender. This makes the conclusions possibly less reliable, especially when we simultaneously stratify by both age and gender, and we get average sample sizes of order 20 participants, with some districts having one observation. This can be seen in Table 3.3. The use of standardised incidence ratios, in this case, is preferable. The approach has an advantage that it does not reduce the sample sizes, and consequently, does not reduce power of testing, while simultaneously accounting for the effects of age and gender.

Table 3.3: Effects of age-gender stratification on the sample size used in the analyses.

Sub-Group	Mean	Std Error	Median	Minimum	Maximum	Range	Sample size
All	203	20	185	23	544	544	10,336
All (15-64 years)	179	18	157	11	503	492	9,154
Male (15-64 years)	75	8	58	5	252	247	3,817
Female (15-64 years)	105	10	89	6	288	282	5,337
Male 15-39	50	5	42	1	159	158	2,574
Female 15-39	64	6	55	1	193	192	3,275
Male 40-64	24	3	20	1	93	92	1,243
Female 40-64	40	4	36	3	125	122	2,063

3.2.6 Age-gender standardised spatial clustering analysis

In the previous section, our analyses used the raw prevalence of the two cardiovascular diseases and the three associated risk factors for the whole sample. However, the estimated level of spatial clustering may be misleading because of confounders such as age and gender that have an important effect on CVDs and their risk factors. We even calculated the age-gender adjusted prevalence; however, the district age-gender specific prevalence would be less reliable and unstable because of smaller districts samples and observed cases, which resulted in huge amount of random error (see Table 3.3). On the other hand, age-gender specific prevalence calculated from the overall adult sample should be much more stable because of the larger sample size. In this section, we used the age-gender specific prevalence obtained from whole SADHS adults (15-64 years) to estimate the expected number of CVD and risk factor cases based on the age-gender distribution of each district to obtain standardised incidence ratios (SIRs).

The SIR is simply a ratio of observed number of cases of a condition divided by expected number of cases. We use SIRs here for the main spatial autocorrelation analyses. Figure 3.4 shows the standardised incidence ratios by district. The quantile map has four categories: lower quartile is hollow; second quartile is light grey; third quartile is dark grey; and upper quartile is black. One district was not sampled in the South African Demographic and Health Survey of 2016, and it is indicated by “neighbourless” in the legend. The darker the area, the higher the SIRs.

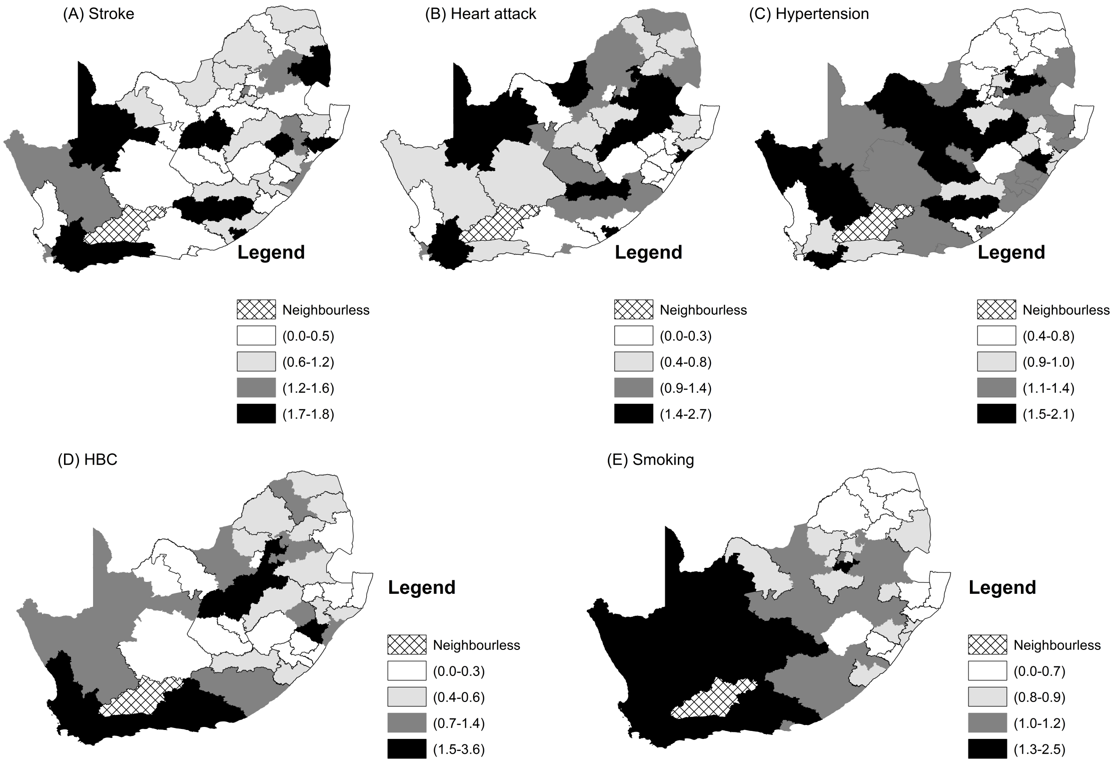


Figure 3.4: Maps of the district standardised incidence rates of the CVDs and their related risk factors.

Lower rates of all the two CVDs and three risk factors were seen in the more rural upper north-eastern part of the country, while higher rates of smoking and high blood cholesterol were observed in the more south-western parts. All of the five CVD measures were relatively high in the urban areas of the western part of the country, even though stroke and heart attack showed an even fluctuation. Higher rates of hypertension were more concentrated in middle part of the country along the south-north belt.

The calculated values of univariate global Moran's I values for the SIRs of CVDs and identified risk factors are presented in Table 3.4. The SIRs for heart attack do not show any spatial patterns with non-significant univariate Moran's index. However, stroke (at 10%) and the three risk factors of smoking, HBC and hypertension, are exhibiting spatial significance at 5% significance level.

Table 3.4: Global univariate spatial autocorrelation association for the age-sex standardised incidence rates of CVDs and identified risk factors for all participants.

	Stroke	Heart Attack	Smoking	HBC	Hypertension
Moran's I	0.128*	-0.015†	0.606**	0.355**	0.236†

Key: HBC, high blood cholesterol; †, insignificant at 5% level; **, significant at 5% level ; *, significant at 10% level..

We also estimated univariate local indicators of spatial autocorrelations (LISA) for the five CVDs and risk factors. These are shown in Figure 3.5. Clusters of a high prevalence of smoking in districts that are surrounded by districts with high prevalence of smoking are in Figure 3.5 E. They form the largest "hot-spots " cluster stretching from the north through the central districts up to the south-western districts of the country. Ten districts constitute this "hot-spots " cluster. These are Cacadu (Eastern Cape Province), Namakwa (Northern Cape), Pixley ka Same (Northern Cape), ZF Mgcawu (Northern Cape), Frances Baard (Northern Cape), City of Cape Town (Western Cape), West Coast (Western Cape), Overberg (Western Cape), Cape Winelands (Western Cape) and Eden District (Western Cape).

There are some "cold-spots " clusters of smoking that are comprised of Zululand, Uthungulu, Umkhanyakhude (all in KwaZulu-Natal Province), Capricorn and Mopani District (in Limpopo Province). These "cold-spots " are generally clustered around rural districts. Hypertension has "hot-spots " cluster that is made up of 7 districts, namely Xhariep, Lejweleputswa, Mangaung (all in Free State Province), Pixley ka Seme, ZF Mgcawu, Frances Baard, and Dr. Ruth Segomotsi District (North West). The "cold-spots " are comprised of Capricorn, Vhembe, Mopani (Limpopo) and Johannesburg District in Gauteng Province.

The “hot-spots ” clusters of stroke and HBC in Figure 3.5 A and D, respectively, are concentrated in the Western Cape Province. They both share the “hot-spots ” districts of City of Cape Town, Eden, Overberg and Cape Winelands. In addition, the “hot-spots ” cluster of HBC includes West Coast District. The global univariate Moran’s index for heart attack was not significant but we included the LISA map shown in Figure 3.5 B. It shows a significant “hot-spot ” of one district called Gert Sibande in Mpumalanga and a “cold-spot ” in Umgugundlovu in KwaZulu Natal.

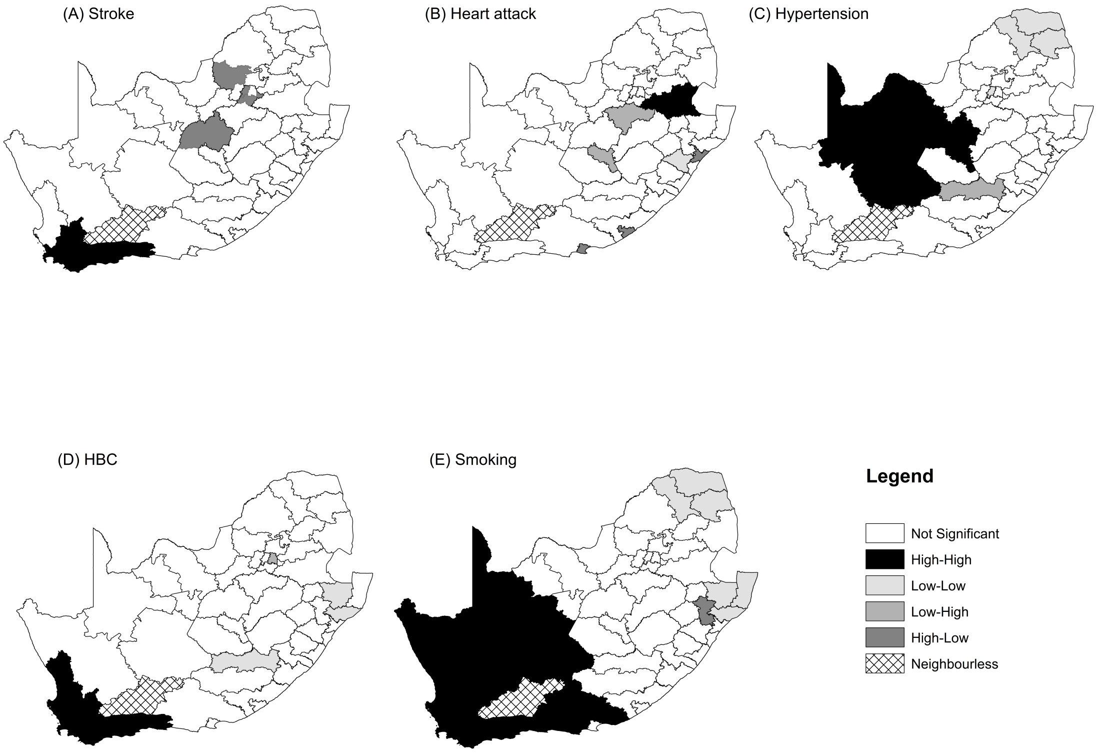


Figure 3.5: Univariate LISA Maps of the district standardised incidence rates of the CVDs and their related risk factors.

The second largest “hot-spots ” cluster is that of hypertension (Figure 3.5

C), which stretches from the central districts of the country up to the north western districts of the country and is made up of seven districts. These districts are: Xhariep, Lejweleputswa, Mangaung (all in Free State Province), Pixley ka Seme, ZF Mgcawu, Frances Baard, and Dr. Ruth Segomotsi District (in North West). The "cold-spots " are comprised of Capricorn, Vhembe, Mopani (in Limpopo) and Johannesburg District in Gauteng Province. The "hot-spots " clusters of stroke and HBC in Figures 3.5 A and D, respectively, are concentrated in the Western Cape Province. They both share the "hot-spots " districts of City of Cape Town, Eden, Overberg and Cape Winelands. In addition, the "hot-spots " cluster of HBC includes West Coast District.

3.3 Application to cardiovascular mortality rates

3.3.1 Data

Causes of death (COD) data from South Africa's vital registration system were used in this section. The data are collected using the death notification forms (DNFs). Medical personnel and other approved certifying authorities are allowed to complete the DNFs. Information collected is kept by the South African Department of Home Affairs, who in turn allow Stats SA to collate the COD data for onward distribution to the public. Stats SA uses revision number ten of the International Statistical Classification of Diseases and Related Health Problems [ICD-10] to code and classify the COD data as stipulated by the World Health Organization (2004). The quality of the data used is discussed by Joubert *et al.* (2013), but there have been vast improvements over the years as regional analysis has been made possible with plausible results (Day *et al.*, 2014; Groenewald *et al.*, 2014).

This section only considers ICD-10 defined broad groups of COD data due to three leading CVDs causing mortality in South Africa, and DBT which is a well-known biomarker for CVDs. The data are for the years 2001 and 2011. In terms of nomenclature, CVAzy, IHDzy and HHDzy will represent mortality due to cerebrovascular, ischaemic and hypertensive heart conditions in the year zy, respectively. Here zy takes values 01 and 11, representing the years 2001 and 2011, respectively.

Table 3.5 shows the distribution of deaths in South Africa for the years 2001 and 2011 by the age groups 0-29, 30-70 and 71 years and over. Overall, the total deaths due to HHD increased the most from 10769 in 2001 to 15609 deaths in 2011, an increase of 44.9%. It can also be seen that DBT increased by almost the same percentage (44.2%) from 14568 deaths in 2001 to 21056 deaths. CVA deaths increased by 14.6% (from 22590 to 25983), while IHD increased by only 2.1% (from 11779 to 12023) over the same period. In the age group 30-70 years, Table 3.5 shows that there has been a slight decrease in the number of deaths for CVA (-0.4%) and IHD (-5.8%) between 2001 and 2011, while HHD increased by about 22.3%. It is in this 30-70-year age group that premature mortality needs to be reduced and analysis will be done for this age group.

Table 3.5: Distribution of number of deaths across age groups by year, South Africa.

Year	Age Group	CVA		HHD		IHD		DBT	
		Number	**Percentage**	**Number**	**Percentage**	**Number**	**Percentage**	**Number**	**Percentage**
2011	0-29	485	1,90%	148	0,90%	165	1,40%	271	1,29%
	30-70	12196	47,10%	7180	46,00%	6183	51,40%	12063	57,29%
	71+	11946	46,10%	7561	48,40%	5248	43,60%	7736	36,74%
	Missing	1266	4,90%	720	4,60%	427	3,60%	986	4,68%
	Total	25893	100,00%	15609	100,00%	12023	100,00%	21056	100,00%
2001	0-29	593	2,60%	159	1,50%	129	1,10%	250	1,71%
	30-70	12241	54,20%	5873	54,50%	6564	55,70%	9185	62,92%
	71+	9756	43,20%	4735	44,00%	5074	43,10%	5145	35,25%
	Missing	0	0,00%	2	0,00%	12	0,10%	17	0,12%
	Total	22590	100,00%	10769	100,00%	11779	100,00%	14597	100,00%

Key: DBT, Diabetes; CVA,Cerebrovascular heart disease; HHD, Hypertensive heart disease; IHD, Ischaemic heart disease.

The data quality issues associated with DNF data are well known. Problems associated with these data include, among others, garbage codes, misclassification and incompleteness of death registration (Joubert *et al.*, 2013; Pillay-van Wyk *et al.*, 2011). Adjustments have to be made to these data to minimise bias that may be attributed to these quality issues.

Correcting the rate of mortality usually involves using the age or sex-specific death rates of standard population to which the mortality rates of interest are adjusted (Birnbaum *et al.*, 2011). There are two problems with this approach. Firstly, the choice of standard to use is usually arbitrary and subjective (Birnbaum *et al.*, 2011). Secondly, the standardised mortality rates assume that the characteristics of small and large areas are the same and the resulting estimates have been criticised for not being representative enough of the geographic distribution of rates (Clayton & Kaldor, 1987; Sarndal, 1984). Thus, alternative techniques have been sought to estimate rates at a local level for compromised data. These techniques are briefly described in the next sub-section.

3.3.2 Estimation of mortality rates

The EB approach and the Poisson regression model were considered for estimating the mortality rates at municipal level. In the EB approach, the number of observed deaths in municipality i and due to disease j, O_{ij} is allowed to follow a Poisson distribution with both the mean and variance equal to the product of P_i, the population at risk in municipality i and π_{ij} the unknown underlying risk of mortality due to disease j in municipality i. It follows that the observed deaths are conditioned on the varying underlying risk of mortality, and we write:

$$O_{ij}|\pi_{ij} \sim Poisson(\pi_{ij}P_i) \tag{3.1}$$

Additionally, the mortality risk, π_{ij}, is allowed to follow a Gamma distribution with shape parameter α and scale parameter ϕ. That is

$$\pi_{ij} \sim Gamma(\alpha, \phi) \tag{3.2}$$

where $E(\pi_{ij}) = \frac{\alpha}{\phi}$ and $Var(\pi_{ij}) = \frac{\alpha}{\phi^2}$. According to Bayes theorem, the following proportionality holds:

$$Pr(\pi_{ij}|O_{ij}) \propto Pr(O_{ij}|\pi_{ij}) \times Pr(\pi_{ij}) \tag{3.3}$$

and, importantly, the conditional posterior also follows a Gamma distribution with shape parameter $\alpha + O_{ij}$ and scale parameter $P_i + \phi$. It follows that

$$\pi_{ij}|O_{ij} \sim Gamma(\alpha + O_{ij}, P_i + \phi) \tag{3.4}$$

Since $E(\pi_{ij}|O_{ij}) = \frac{O_{ij}+\alpha}{P_i+\phi}$, it can be deduced that the raw rates, $\widehat{\pi_{ij}} = \frac{O_{ij}}{P_i}$ can be adjusted using posterior distribution, $Pr(\pi_{ij}|O_{ij})$ if α and ϕ can be derived from the prior distribution, $Pr(\pi_{ij})$. In-fact, it can be shown that the EB estimate of the underlying mortality is the expected value of the distribution of the conditional posterior:

$$\widehat{\pi_{ij}}^{EB} = E(\pi_{ij}|O_{ij}) = \frac{O_{ij} + \alpha}{P_i + \phi} \tag{3.5}$$

where the parameters α and ϕ are determined from the observed data.

In the Poisson regression approach, the expected mean of $O_{ij} (= \pi_{ij} P_i)$, denoted by μ_{ij} (the expected number of deaths in municipality i dying a premature death (between 30-70 year) for a given disease, j), is modelled as

$$\mu ij = \log(\pi_{ij} P_i) = log(P_i) + \alpha + \beta_1(p_{age}) + \beta_2(p_{race}) + \beta_3(poverty)(\pi_{ij}|O_{ij}) = \frac{O_{ij} + \alpha}{P_i + \phi} + \varepsilon_{ij} \tag{3.6}$$

where p_{age} is the proportion of the age group 30-70 that are aged 50 to 70 in the population of municipality i, p_{race}, is the proportion of a given race in municipality i for the given age group, and $poverty$, is the level of poverty in municipality i measured by the official South African multidimensional poverty index obtained from the 2001 and 2011 census data (Statistics South Africa, 2014).

A descriptive summary of the raw, smoothed and adjusted rates for the age group 30-70 years of interest to this study, is given in Table 3.6. Generally, we have mean rates of the same order for all the three rates. The major difference in the rates, however, is found in the ranges, where observed raw rates have the highest range in all instances owing to very high maximum values. Further investigations revealed that the municipalities with the smallest populations are also the ones with the highest (as well as smallest) mortality rates. The raw rates are sensitive to small population counts, resulting in instability. This is a well-documented problem when using raw mortality rates. Empirical Bayes rates are known to alleviate this problem (Leyland & Davies, 2005; Marshall, 1991). Adjusting for covariates also managed to alleviate the problem by reducing the maximum values and increasing the minimum values of the observed rates.

3.3.3 Spatial autocorrelation analysis

The quantile maps of raw, smoothed and adjusted mortality rates at municipal level for each of the four disease conditions studied in South Africa for the years 2001 and 2011, are shown in Figures 3.6 and 3.7, respectively. Municipalities in the upper quantile indicate areas that experienced high rates of mortality, while those in the lower quantile indicate areas with low rates of mortality. The higher the quantile, the darker the colour (ranging from quantile 1 with white colour to quantile 4 with black colour). It follows that areas with the darker shade indicate areas of higher mortality than those of a relatively lighter shade.

Generally, the quantile maps reveal some form of clustering. Consider

Table 3.6: Descriptive statistics of raw, EB smoothed and adjusted mortality rates across municipalities for CVA, IHD, HHD and DBT, South Africa.

Model	Mean	SD	Minimum	Maximum
		CVA 2001		
Adj	86,37	18,96	41,26	157,10
EB	85,96	47,25	3,02	296,56
RR	88,75	59,63	0,00	389,89
		CVA 2011		
Adj	72,90	17,33	37,08	151,76
EB	76,24	48,18	7,20	377,59
RR	79,66	60,55	0,00	477,90
		IHD 2001		
Adj	48,81	33,25	6,63	147,98
EB	43,11	37,79	1,49	362,80
RR	44,65	45,24	0,00	407,68
		IHD 2011		
Adj	37,48	20,76	10,71	116,24
EB	34,31	28,91	2,09	259,34
RR	36,32	38,41	0,00	309,69
		HHD 2001		
Adj	43,56	183,69	12,96	149,08
EB	38,68	27,52	1,77	226,84
RR	38,69	33,72	0,00	244,37
		HHD 2011		
Adj	44,36	17,34	12,32	129,73
EB	44,01	29,33	5,37	164,24
RR	45,89	36,51	0,00	180,63
		DBT 2001		
Adj	64,72	18,04	21,70	122,78
EB	53,06	39,04	2,32	403,50
RR	51,66	46,52	0,00	457,39
		DBT 2011		
Adj	71,78	17,76	29,68	132,86
EB	66,49	43,99	7,73	353,56
RR	67,06	53,37	0,00	405,13

Key: SD = Standard deviation.

Figures 3.6 A-C to see the effects of smoothing and adjustment for covariates. The quantile map of the observed raw rate (CVA01-RR) in Figure 3.6 A and the smoothed rate (CVA01-EB) in Figure 3.6 B are almost similar in terms of their

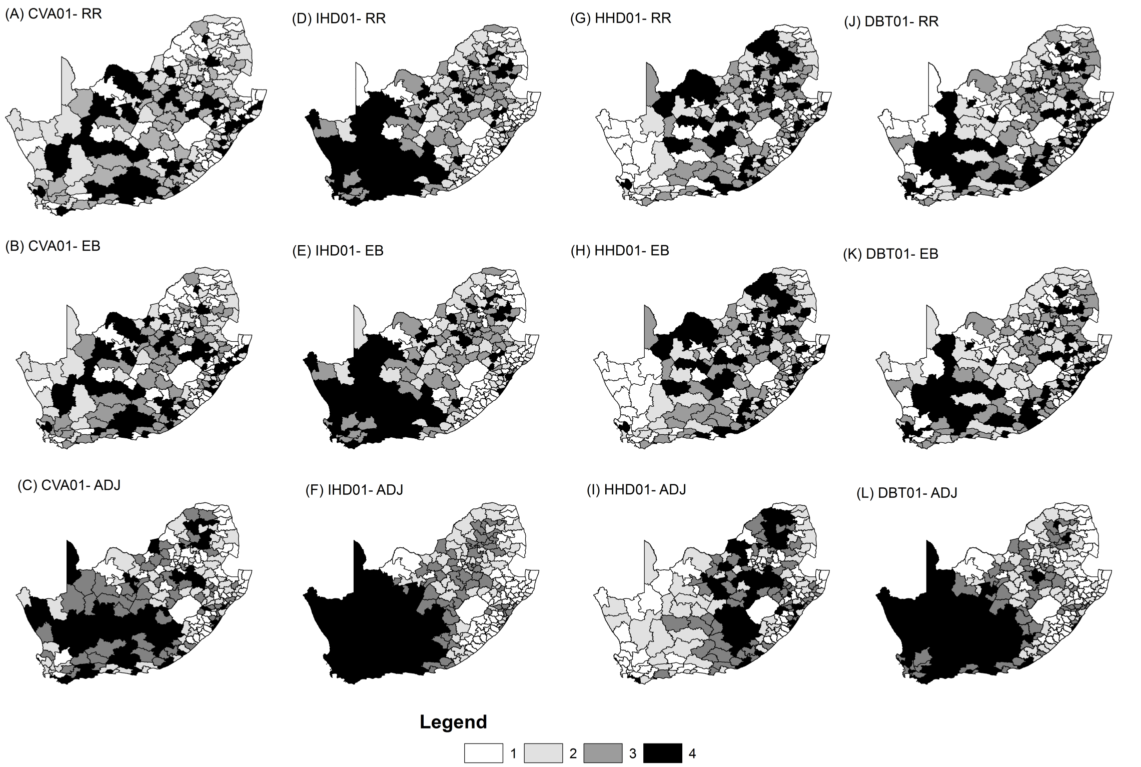


Figure 3.6: Quantile maps showing the distribution of raw, smoothed and adjusted mortality rates for the year 2001.

spatial distributions. There is not much difference between the distribution of mortality rates before and after smoothing. It seems that the effects of stabilising the crude rates with the EB approach has not, based on the evidence of the quantile maps, improved the ability to discern areas of higher mortality risk.

Adjusting for covariates, as the case of CVA01-Adj in Figure 3.6 C, results in a more defined cluster in the south-west part of the country when compared with raw and smoothed rates in Figures 3.6 A-B. This is the general pattern with all the other disease conditions, with dark colours more noticeable for adjusted rates than for raw and smoothed rates, and are mostly concentrated in the western part of the country. Only HHD clustering seems to stretch

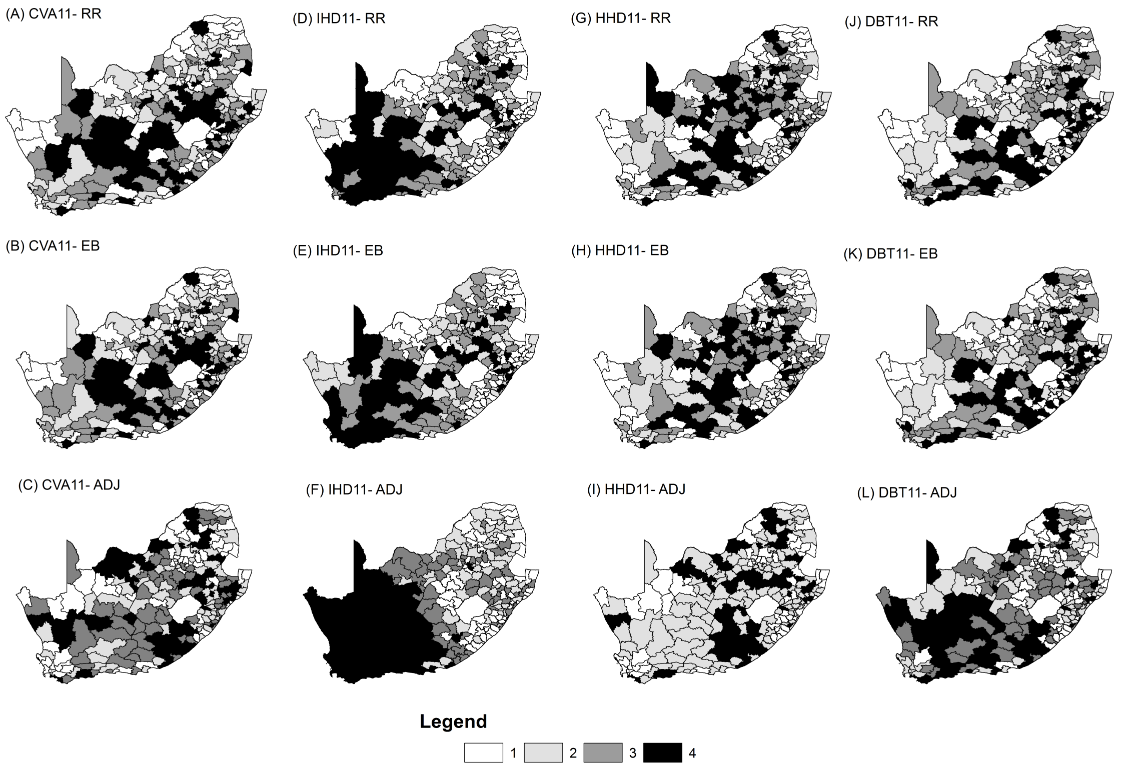


Figure 3.7: Quantile maps showing the distribution of raw, smoothed and adjusted mortality rates for the year 2011.

from the middle of the country towards the eastern part of the country. The spatial patterns exhibited in Figure 3.6 for the year 2001 are similar to the spatial patterns exhibited by the corresponding mortality rates in Figure 3.7 for the year 2011. In the next section the statistical significance tests of spatial autocorrelations were done and discerned clusters mapped.

Univariate global spatial autocorrelation

The choropleth maps in Figures 3.6 and 3.7 of the geographical variations for both crude and smoothed rates has shown evidence of clustering in CVD outcomes. In order to formally investigate spatial association, we measured the association in a formal way by using univariate clustering statistic. Table

3.7 presents the derived values for each CVD for the whole of South Africa for the years 2001 and 2011. For comparison purposes, the derivations were done using raw, smoothed and adjusted rates.

The univariate Moran's *I* test in Table 3.7 confirms that the distribution of the four conditions IHD, CVA, DBT and IHD varies geographically, when adjusted rates are used (p-value < 0.05). Both raw and smoothed rates failed to detect clusters of DBT, while the raw rate further failed to detect any significant clustering for CVA01 (p-value >0.05). The geographic variation based on adjusted rate, is significant for both the years 2001 and 2011. In all the cases, the calculated statistics for Moran's I are all positive and significant across the years. This means that the likelihood of the spatial patterns generated by mortality due to each of the three CVDs being due to random chance is negligibly small (less than 5%). Thus, one can conclude that the probability is high that municipalities that are nearer to each other tend to have comparable baseline mortality rates than the distant municipalities. In other words, there is some form of clustering exhibited by all three CVDs at the 5% significance level. This is a reflection of what is seen in the quantile maps in Figures 3.7 and 3.8.

Table 3.7 also shows the spatial autocorrelation statistic calculated for the residuals of the smoothed and adjusted rates for each of the CVDs for the years 2001 and 2011. The statistical autocorrelation of the residuals, based on the Moran's index, were found to be insignificant for some of the fitted models for adjusted rates (CVA11, IHD11 and HHD11) and the smoothed rates (CVA01, HHD01, CVA11 and IHD11). This statistical autocorretion analysis of the residuals is not a criterion for diagnostic checks for generalised linear models or EB approach but it would be preferable if residual spatial autocorrelations were not significant. This is because the presence of spatial autocorrelations in the residuals suggests that the model is not adequately specified. That

Table 3.7: Univariate global Moran's spatial autocorrelations for the model residuals, raw, smoothed and adjusted mortality rates due to CVA, IHD, DBT and HHD in 2001 and 2011.

Model	Moran's I (Estimates)	p-value	Moran's I (Residuals)	p-value
CVA 2001				
Adj	0,422	<0,001	0,038	†
EB	0,021	<0,001	0,068	<0,05
RR	0,029	†		
CVA 2011				
Adj	0,297	<0,001	0,109	<0,05
EB	0,078	<0,05	0,121	<0,05
RR	0,088	<0,05		
IHD 2001				
Adj	0,849	<0,001	-0,003	†
EB	0,108	<0,001	0,318	<0,05
RR	0,251	<0,001		
IHD 2011				
Adj	0,821	<0,001	0,149	<0,05
EB	0,093	<0,001	0,150	<0,05
RR	0,176	<0,05		
HHD 2001				
Adj	0,445	<0,001	0,066	†
EB	0,144	<0,001	0,045	†
RR	0,218	<0,001		
HHD 2011				
Adj	0,329	<0,001	0,112	<0,05
EB	0,135	<0,001	0,054	†
RR	0,101	<0,05		
DBT 2001				
Adj	0,684	<0,001	0,063	†
EB	0,005	†	0,003	†
RR	0,006	†		
DBT 2011				
Adj	0,316	<0,001	0,064	†
EB	0,038	†	0,003	†
RR	0,030	†		

Key: † = Insignificant p-values.

is to say there may exist some unmeasured covariates not specified in the model that may help in explaining the variation of mortality rates across the municipalities. Introducing spatial random effects or an eigenvector spatial filter (Griffith & Chun, 2014) did not remove the residual spatial autocorrelations, so the original specified Poisson regression model with covariates only was returned using the rule of parsimony. In the next section we looked at the LISA maps for all rates for visual comparison purposes only, irrespective of whether they are significant or not.

Univariate "hot-spot" analysis

The local indicators of autocorrelation based on Moran's I were used to determine the actual clusters at municipal level. The resulting univariate LISA maps for raw, smoothed and adjusted rates for CVA, IHD, DBT and HHD are shown in Figures 3.8 and 3.9 for the years 2001 and 2011, respectively. "Hot-spots", which are municipalities of high mortality incidences that are surrounded by municipalities with high mortality incidences, are indicated by a "High-High" (H-H) key on the map, while the "cold-spots", which are municipalities of low mortality incidences that are surrounded by municipalities with low mortality incidences, are indicated by a "Low-Low" (L-L) key. In addition, there are outliers indicated by "High-Low" (H-L), which are municipalities of high mortality incidences that are surrounded by municipalities with low mortality incidences, and "Low-High" (L-H), which are municipalities of low mortality incidences that are surrounded by municipalities with high mortality incidences. Municipalities whose clustering is not significant are denoted by "Not Significant" (N-S) key and have a white shade. The "hot-spots" are of major concern as they represent clusters of high risk of mortality due to the CVDs and have a black shade in the map.

Adjusted rates in Figure 3.8 have noticeable and well defined clusters as compared to raw and smoothed rates. Generally, clusters are found in the

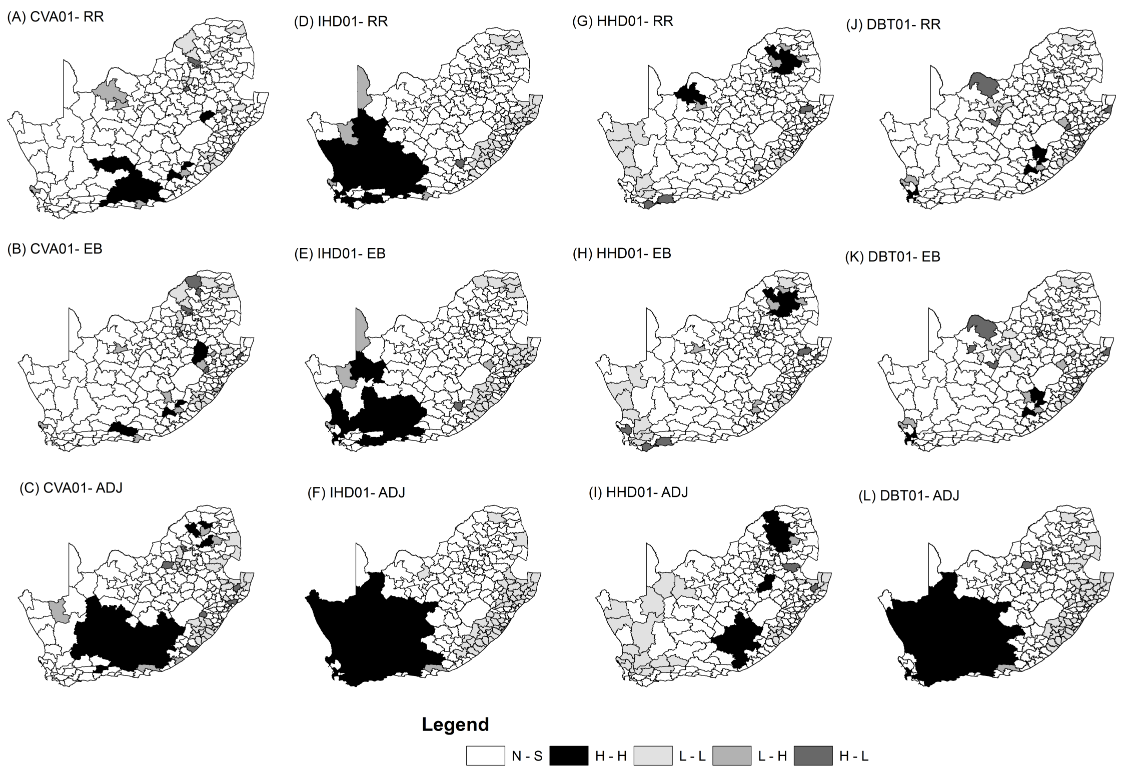


Figure 3.8: Univariate Moran's I LISA maps showing the distribution of clusters of raw, smoothed and adjusted mortality rates for the year 2001.

south west part of the country, except for HHD which has clusters in the south and north-east part of the country. The clusters for CVA and DBT seem to have reduced in size over the ten-year period under review. In Figure 3.8 C, for example, CVA01 LISA derived clusters comprise of 31 municipalities, but these have been more than halved to 16 municipalities in 2011 (see Figure 3.9 C). The disappearance or movement of the cluster from the south-east maybe due to intervention programmes aimed at alleviating the problem in the area. However, further investigation may help to explain what is truly happening, especially with DBT whose data suggest that mortality due to this disease has increased over the ten-year period under review.

The adjusted rates based clusters for IHD (see Figure 3.8 F and Figure 3.9

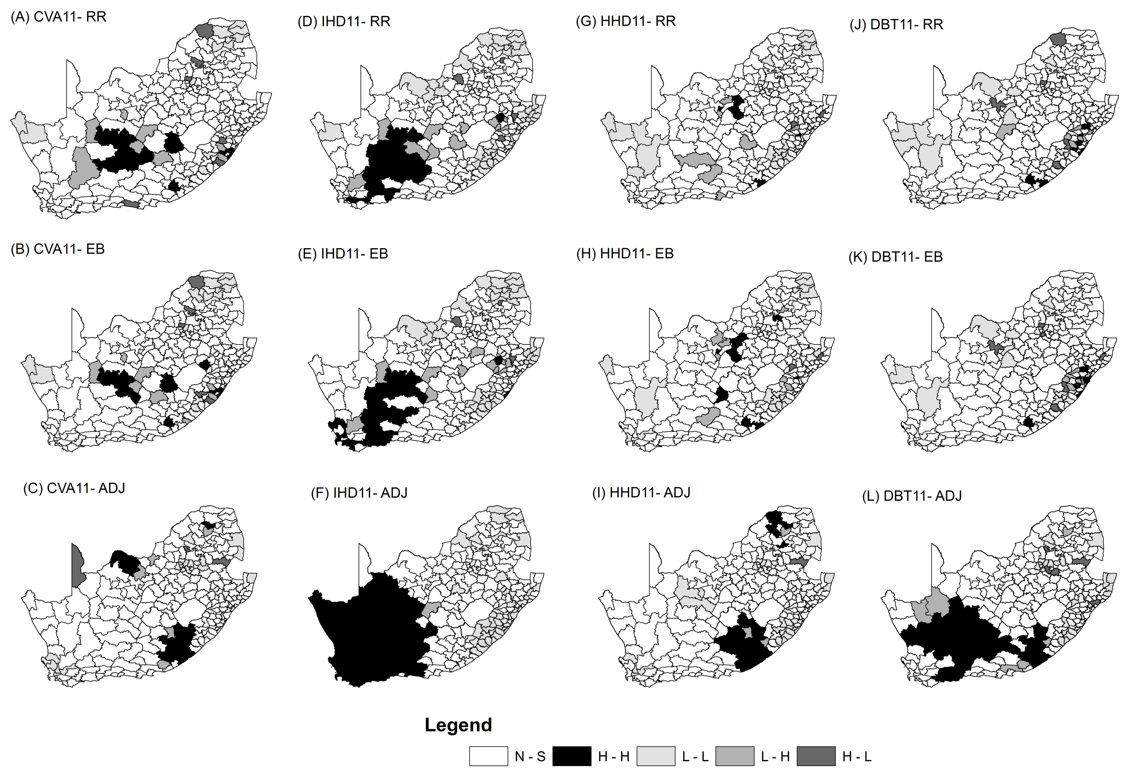


Figure 3.9: Univariate Moran's *I* LISA maps showing the distribution of clusters of raw, smoothed and adjusted mortality rates for the year 2011.

F) and HHD (see Figure 3.8 I and Figure 3.9 I) have not changed much over the period. This shows that the spatial dynamics of IHD and HHD are stable over the period under study, with IHD "hot-spots" located in the centre and spanning all the way to the south-west coast of the country. The LISA analysis for HHD (Figure 3.8 I and Figure 3.9 I) reveal two clusters in the south and north-east part of the country for both the years 2001 and 2011.

3.4 Application using a linear regression-based Moran's index

3.4.1 Analysis using distance based spatial weights

Choice of spatial weights to use

Chen (2013) proposed the use of distance based matrix, namely exponential matrix and inverse based matrix. The choice of the weight matrix is not obvious with both weight matrices dependent on the value of α. Usually α takes values 1 or 2. Application was done to the South African mortality rate data due to ischaemic heart condition for the year 2011 derived in section 3.3.2 based on the Poisson model. These data have been shown in section 3.3.3 to have a significant spatial pattern using the traditional Moran's I index.

Table 3.8: The results of Chen's regression approach to Moran's index when applied to mortality rates due to ischaemic heart conditions in South Africa: 2011.

	Exponential weights			**Inverse weights**		
α	Moran's I	P-value	s_f	Moran's I	P-value	s_f
1.00	0.012	<0,001	0.082	0.016	0.022	0.126
2.00	0.021	<0,05	0.136	0.040	0.106	0.328

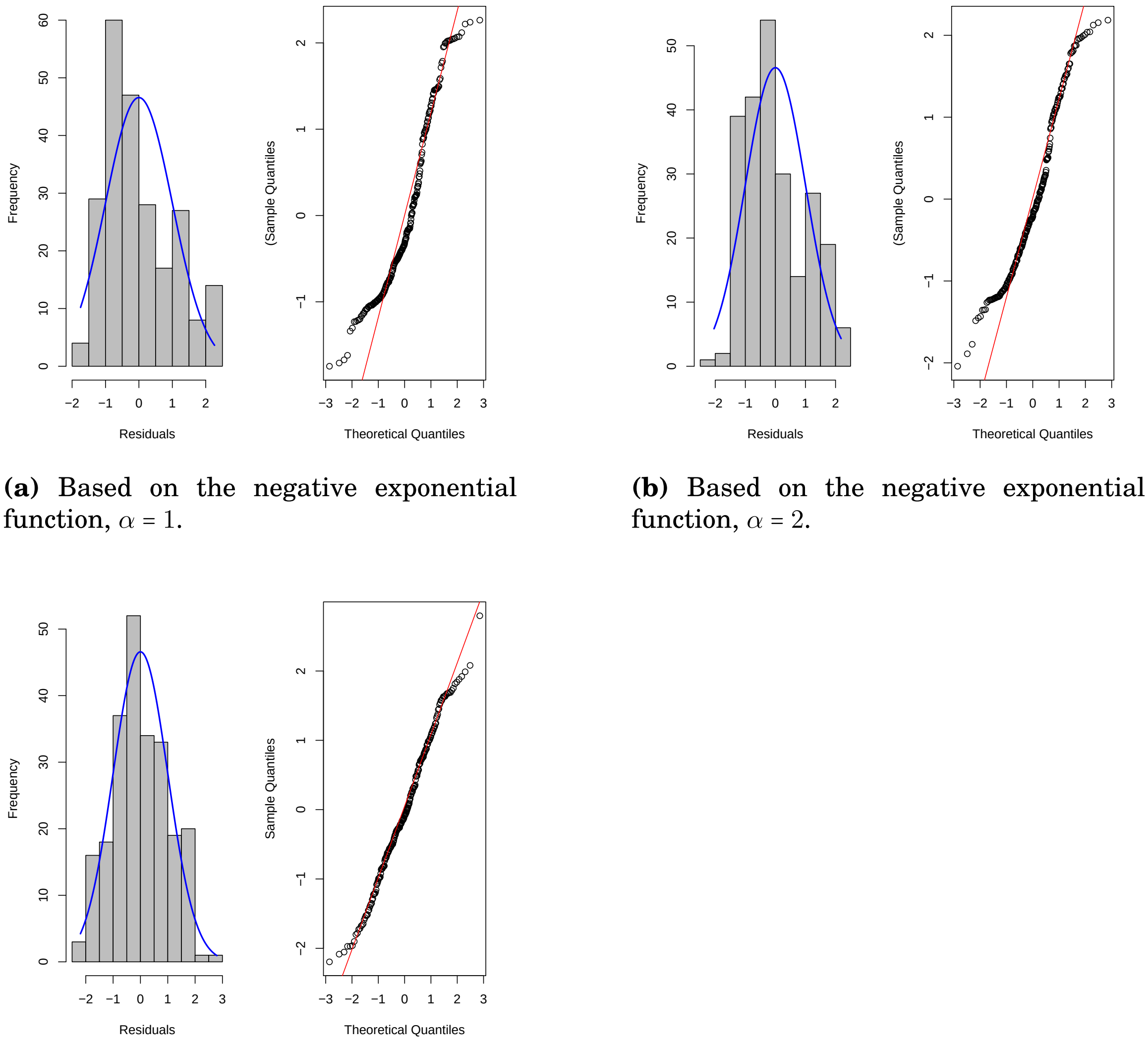
In Table 3.8, the data for the adjusted rates of mortality due to ischaemic heart diseases for the year 2011 were analysed to determine the which spatial weight to use when α is allowed to take values 1 or 2. The idea is to get an index which is statistically significant and at the same time with a low standard error. The Moran's index in Table 3.8 increases with an increase in the α values. There is evidence for spatial clustering for the negative exponential were weights for $\alpha = 1$ and $\alpha = 2$ and for the inverse function based spatial weights only $\alpha = 1$. These spatial weights corresponding to the significant Moran's indexes were used in the analyses for the remainder of this subsection.

Diagnostic checks

Having determined the spatial weights matrices to use, Chen (2013) requires that the residuals of the regression model, e_f, used to estimate the Moran's index follow a normal distribution. The residuals of the model were analysed for diagnostics checks using a histogram and Q-Q plots. Figure 3.10 shows the histogram and the corresponding Q-Q normal plots to determine if the residuals of the regression model used to determine Moran's I based on the spatial weights derived from a negative exponential function with $\alpha = 1$ (Figure 3.10 a), $\alpha = 1$ (Figure 3.10 b) and from an inverse power function with $\alpha = 1$ (Figure 3.10 c).

The residual plots shown in Figure 3.10 c, based on the inverse power function, seem to suggest that the distribution of the residuals follow an approximate normal distribution with almost all plotted points lying on the straight line of the Q-Q plots. It can be seen that the histogram of the residuals in Figure 3.10 c is almost symmetrical about the zero-mean. From this visual illustration one can assume with some high level of confidence that the residuals based on the inverse power function spatial weights follow a normal distribution. On the other hand, it can be seen in the histograms of Figure 3.10 a-b that those data are positively skewed and do not exhibit normality. In addition, the points on the Q-Q plot are generally not lying on the straight line with the plotted points curving away from the straight line at the ends of the line. Thus, the residual plots in Figure 2.10 a-b, based on the negative exponential functions, seem to suggest that the distributions of the residuals do not follow a normal distribution.

Statistical tests for the deviation from normality such as the Anderson-Darling test and the Shapiro-Wilk's W, may also be used to complement the visual illustrations. In this study, the Shapiro-Wilk's statistical test for normality was



(a) Based on the negative exponential function, $\alpha = 1$.

(b) Based on the negative exponential function, $\alpha = 2$.

(c) Based on the inverse power function, $\alpha = 1$.

Figure 3.10: Comparison of histogram and normal Q-Q plots of the residuals of the Moran's regression model using Poisson based ischaemic mortality adjusted rates in South Africa, 2011.

conducted, as graphical assessment alone would not be sufficient. The results for the statistical tests for normality are shown in Table 3.9. The test confirm that the residuals based on the inverse power function with $\alpha = 1$ are indeed normal with a Shapiro-Wilk's W value of 0.99 and a corresponding p-value os 0.102 which is greater than 0.05. This is in line with the visual assessment in

Figure 3.10.

Table 3.9: The Shapiro-Wilk's W test of normality on the residuals of ischaemic heart disease adjusted mortality rates for negative exponential and inverse power function weight matrices.

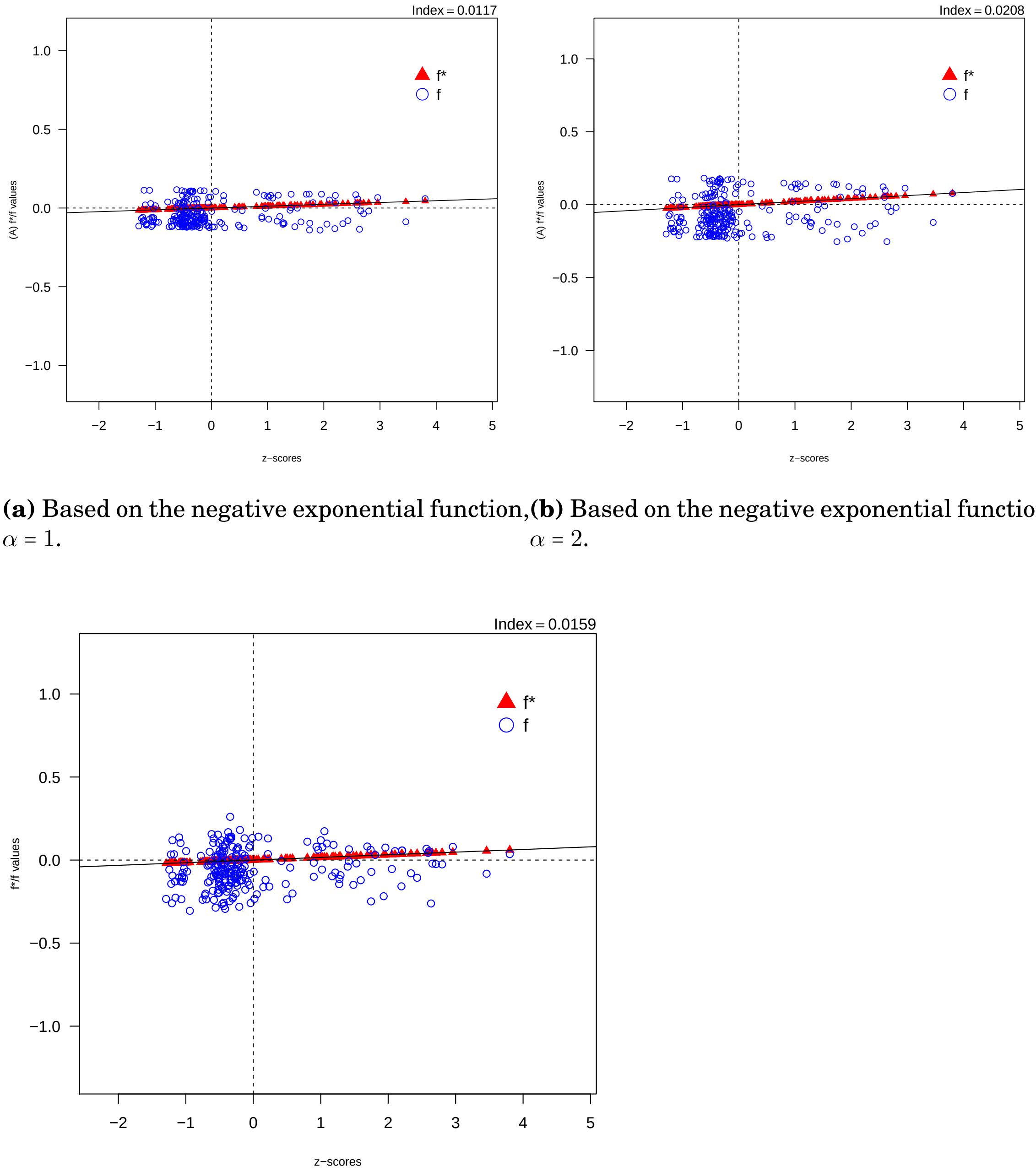
Spatial weights matrix	Shapiro W	P-value
Exponential (alpha=1)	0.922	<0.001
Exponential (alpha=2)	0.949	<0.001
Inverse (alpha=1)	0.990	0.105

Using Table 3.9 and Figure 3.10 it has been deduced that the appropriate spatial weight matrix to use with the adjusted ischaemic mortality rates da ta is that one based on the inverse power function with α value of 1. In the next section we use the regression approximation of the Moran's index using Chen's approach to determine local clusters for the three spatial weight matrices, irrespective of whether the residuals were normal or not, and then compare the results.

Cluster analysis

Two vectors, f^* and f, defined in Equation 2.14 and Equation 2.15, respectively, are central to Chen (2013)'s new regression approach in determining the global Moran's index and the corresponding local clusters from a scatter plot. This is because the line obtained by plotting f^* values against z values represent the spatial autocorrelation trend whose gradient provides the Moran's index. On the other hand, the relationship between f and z forms the autocorrelation pattern or scatter points on a two-dimensional plot. A superimposition of the two plots gives a revised scatter plot for the Moran's index, similar to the one done for the original Moran's index (Anselin, 1995).

The f^*'s were regressed on the z-scores to obtain the Moran's index for our data. The revised Moran's scatter plots of the data are illustrated in Figure



(a) Based on the negative exponential function, $\alpha = 1$.

(b) Based on the negative exponential function, $\alpha = 2$.

(c) Based on the inverse power function, $\alpha = 1$.

Figure 3.11: The scatter plots of the Moran's index using Poisson based ischaemic mortality adjusted rates in South Africa, 2011.

3.11. On the vertical-axis are values of f and f^* while on the horizontal-axis we have values of z. The solid line comprises of the fitted values of the couple $(z,\ f^*)$, while the plotted scatter points comprise of the observed values of the pair (z, f). The gradient of the fitted line provides the value of the Moran's index which is shown at the top right corner of the plot.

The scatter plots in Figure 3.11 can be used to determine the spatial autocorrelation patterns at South Africa's municipality level. This is because the quadrants in which the scatter points (z, f) are located gives an indication of the nature of the local spatial autocorrelation. Scatter points in the first quadrant of the plots (top right) have a "high-high" correlation at local level. In the second quadrant we have the "low-high" spatial correlation, third quadrant we have "low-low" correlation and the forth quadrant we have the "high-low" correlation at local level. Each (z, f) coordinate point represent a municipality and the nature of correlation was determined from these scatter points for each municipality in South Africa.

The ideal spatial weight matrix, M^*, was calculated using Equation 2.13. Elements of the diagonals of M^* provided the Moran's I based local indicators of spatial autocorrelation, I_i, for each municipality. An R code was developed to test the significance of these local Moran's index of spatial autocorrelation for each municipality. Appendix A.1 gives the R code used in the analyses of this section. Tables A2-A4 in Appendix A gives a summary of the cluster classifications due to the different weight matrices used.

The resulting univariate LISA cluster maps are shown in Figure 3.12. Three following categories with their usual meaning were plotted: "Not Significant"; "High-High"; and "Low-Low". The plots are almost similar with "hot spots" clusters in the south west part of the country while "cold spots" were in the north east and south east of the country. The exponential based weight matrices produced more "cold spots" than the inverse based spatial weight matrix.

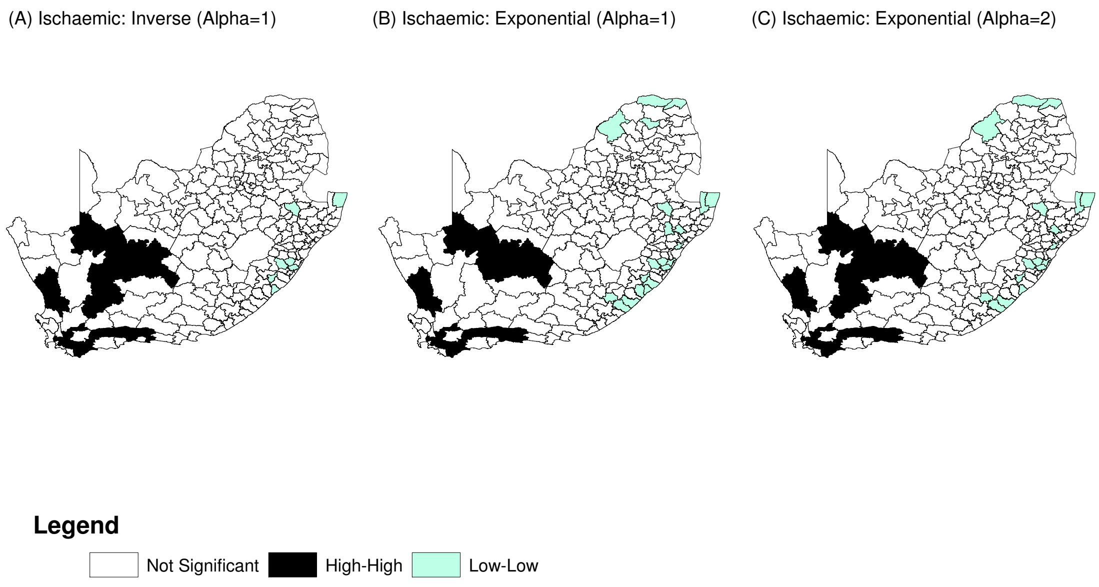

(A) Ischaemic: Inverse (Alpha=1)
(B) Ischaemic: Exponential (Alpha=1)
(C) Ischaemic: Exponential (Alpha=2)
Legend
Not Significant
High-High
Low-Low

Figure 3.12: The univariate local Moran's spatial clusters for ishaemic mortality rates based on Chen's regression approach, for the inverse and exponential spatial weights.

Concordance and inconsistency analysis

The objective of the concordance analysis was to to reveal any similarities in the diagonals or elements that are closer to diagonals in the classification contingency tables of 3 by 3 quantiles presented in Table 3.10. Each classification table shows the results of clustering using the local indicators of spatial autocorrelation. Cells in the diagonals represent the number of clusters that are common to both cluster maps due to their respective spatial weights. Off-diagonal celss represent clusters that are not common to the maps by the spatial weights under consideration.

In Table 3.10 a, 193 out of 194 municipalities classified as insignificant when inverse and exponential spatial weights both with $\alpha = 1$ are used. The

Table 3.10: Classification Tables for cluster categories.

		Exponential, $\alpha = 1$		
		Insignificant	High-High	Low-Low
Inverse $\alpha = 1$	Insignificant	193	2	14
	High-High	1	17	0
	Low-Low	0	0	7

(a) Inverse, $\alpha = 1$ vs. Exponential, $\alpha = 1$.

		Exponential, $\alpha = 2$		
		Insignificant	High-High	Low-Low
Inverse $\alpha = 1$	Insignificant	197	2	10
	High-High	0	18	0
	Low-Low	0	0	7

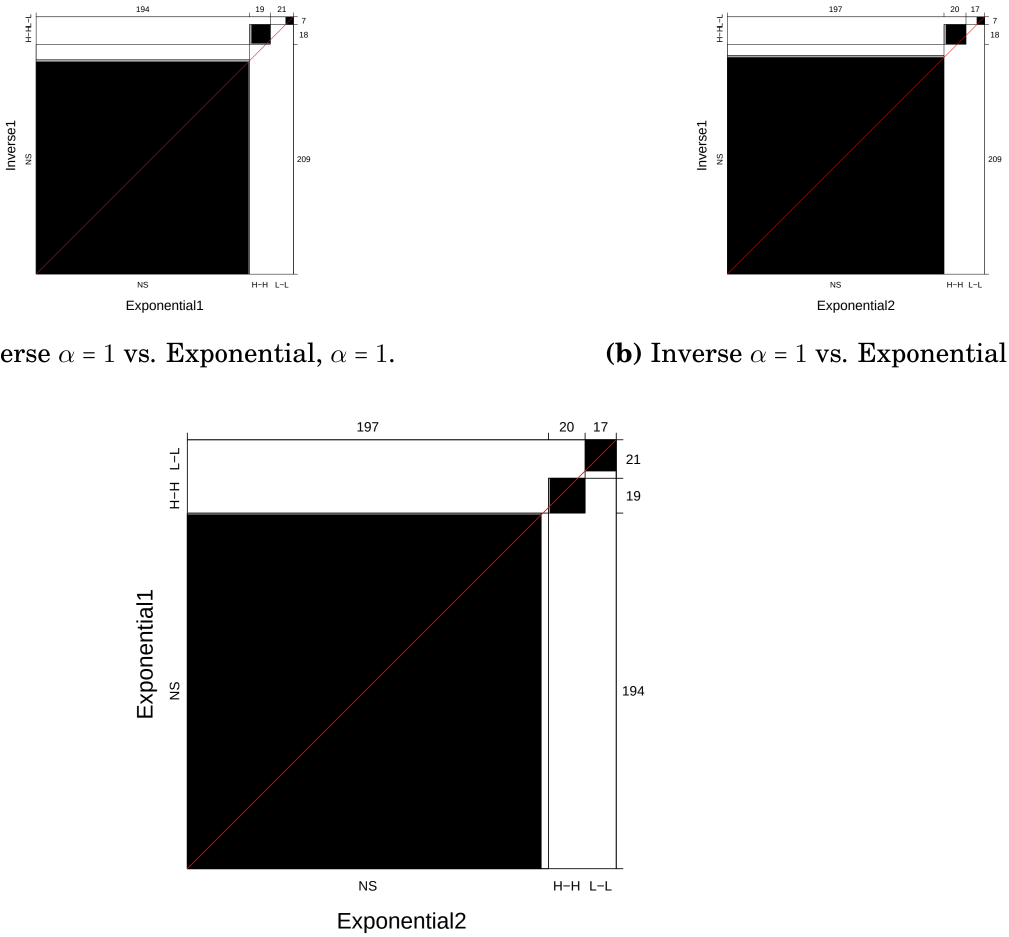
(b) Inverse, $\alpha = 1$ vs. Exponential, $\alpha = 2$

		Exponential $\alpha = 2$		
		Insignificant	High-High	Low-Low
Exponential $\alpha = 1$	Insignificant	193	1	0
	High-High	0	19	0
	Low-Low	4	0	17

(c) Exponential, $\alpha = 1$ vs. Exponential, $\alpha = 2$.

cluster map by the inverse spatial weight matrix had 18 municipalities forming "High-High" "hot spot" clusters, with 17 and all 18 of those municipalities identical to that for maps derived from exponential spatial weights with $\alpha = 1$ and $\alpha = 2$, respectively. There is just a difference in clusters by one or two municipalities. The main difference is with the "Low-Low" municipalities. Fourteen out of 21 "cold spot" municipalities for the cluster map of exponential weight matrix with $\alpha = 1$ were classified as "Insignificant" for the cluster map of the inverse weight matrix with $\alpha = 1$. Ten municipalities that were classified as "Low-Low" for the exponential weight matrix with $\alpha = 2$ were classified as "Insignificant" for the cluster map of the inverse weight matrix with $\alpha = 1$.

The cluster maps of based on the two exponential spatial weights are almost identical.The observer agreement charts for the 3 by 3 contingency tables in Table 3.10 are shown in Figure 3.13.



(a) Inverse $\alpha = 1$ vs. Exponential, $\alpha = 1$.

(b) Inverse $\alpha = 1$ vs. Exponential, $\alpha = 2$.

(c) Exponential $\alpha = 1$ vs. Exponential, $\alpha = 2$.

Figure 3.13: The agreement charts for comparing ischaemic heart disease mortality LISA cluster map categories by three spatial weight matrices.

There concordance exhibited by the agreement charts in Figure 3.13 is very high because the shade of the rectangles within the unit square are predominantly dark and the rectangles are almost filled up with the dark shade. In the charts, NS stands for “Not significant”, HH stands for “High-High” and “L-L” stands for “Low-Low”. It is observe in both Figures 3.13 a-b that inverse spatial weight matrix has a bias towards detecting “High-High” clusters than the exponential weight matrices. This is because the bulk of the rectangles for “High-High” are over the diagonal line in the two charts. The diagonal line is the line of no bias. The inconsistencies are shown by the white triangles for “NS” and “L-L” above the diagonal line of bias in Figures 3.13 a-b. We also note that there is hardly any ‘drift’ bias between the cluster maps of exponential spatial weights because diagonal line of no bias bisects the rectangles in Figure 3.13 c.

Table 3.11: The concordance strength Bangdiwala test statistics.

	Bangdiwala	Weighted Bangdiwala
Inverse ($\alpha = 1$) vs. Exponential ($\alpha = 1$)	0.916	0.928
Inverse ($\alpha = 1$) vs. Exponential ($\alpha = 2$)	0.941	0.948
Exponential ($\alpha = 1$) vs. Exponential ($\alpha = 2$)	0.973	0.977

Table 3.11 shows the Bangdiwala agreement strength statistics for the concordance analysis in Figure 3.13. Figure 3.13 c has a Bangdiwala’s B-statistic of 0.977, which happens to be highest. It is not a surprise that it is highest as the we would expect similar results from two spatial weights that are both based on the negative exponential function. A Bangdiwala’s B-statistic of 0.977 indicates that 97.7% of the 234 municipalities clusters of ischaemic mortality in the same categories from the two exponential based spatial weights. The Bangdiwala’s B-statistic for Figure 3.13a was the lowest at 0.928, yet reflects an almost perfect concordance of 92.8% for the local clusters due to the inverse based spatial weight and the exponential spatial weight with $\alpha = 1$. Munoz & Bangdiwala (1997) provides the interpretation for the Bangdiwala’s B-statistic as follows: Poor level of agreement (0.000 - 0.200); Weak level of agreement (0.201 - 0.400); Moderate level of agreement (0.401 - 0.600); Good level of agreement (0.601 - 0.800); and Excellent level of agreement (0.801 or greater). Thus the level of agreements in our analyses are generally at excellent levels.

3.4.2 Sensitivity Analysis

The choice of α to use with the negative exponential function or inverse power function based spatial weights is usually 1 or 2. However, the models derived using these α values may not be significant or the residuals violate the assumption of normality. If this is the case, then other values of α need to be explored. Sensitivity analysis was conducted to establish the effect of changing the spatial weights matrix, by adjusting α on the significance test (p-value) and on s_f values. While Chen (2013) recommended that s_f be less than 0.15 we only used this as a guideline. This analysis was done on two variables that are known to exhibit global spatial autocorrelation, namely, cerebrovascular mortality rates and ischaemic mortality rates for the years 2007 and 2011 combined. Two random variables that do not exhibit spatial autocorrelation patterns were also considered. The results are shown in Table 3.12.

Table 3.12: Sensitivity Analysis of the effect of changing the inverse spatial weights matrix on the p-value in significance testing of Moran's index and s_f when applying Chen's regression approach.

	Ischaemic			**Cerebrovascular**			**Random variable 1**			**Random variable 2**		
Alpha	MI	p-value	s_f	MI	p-value	s_f	MI	p-value	s_f	MI	p-value	s_f
0.01	-0.003	0.999	0.002	-0.004	0.999	0.001	-0.004	0.664	0.000	-0.004	0.852	0.001
0.02	-0.002	0.999	0.004	-0.004	0.999	0.002	-0.004	0.643	0.001	-0.004	0.717	0.001
0.04	0.001	0.999	0.008	-0.003	0.999	0.003	-0.004	0.624	0.001	-0.004	0.723	0.002
0.06	0.004	0.701	0.011	-0.003	0.999	0.005	-0.004	0.604	0.002	-0.004	0.709	0.003
0.08	0.007	<0.001	0.015	-0.002	0.999	0.007	-0.004	0.630	0.003	-0.004	0.629	0.004
0.10	0.010	<0.001	0.019	-0.002	0.999	0.009	-0.004	0.675	0.004	-0.004	0.675	0.006
0.20	0.025	<0.001	0.038	0.001	0.996	0.018	-0.004	0.641	0.008	-0.004	0.664	0.011
0.40	0.055	<0.001	0.075	0.007	0.046	0.036	-0.004	0.666	0.017	-0.004	0.558	0.023
0.60	0.086	<0.001	0.112	0.014	<0.001	0.057	-0.003	0.694	0.029	-0.005	0.451	0.036
0.80	0.120	<0.001	0.149	0.022	0.001	0.080	0.003	0.737	0.064	-0.007	0.379	0.051
1.00	0.154	<0.001	0.187	0.032	<0.001	0.109	-0.003	0.762	0.081	0.009	0.265	0.068
1.20	0.189	<0.001	0.225	0.044	<0.001	0.145	-0.003	0.812	0.088	-0.013	0.202	0.089
1.50	0.241	<0.001	0.286	0.066	<0.001	0.219	-0.004	0.818	0.132	-0.021	0.123	0.129
2.00	0.314	<0.001	0.394	0.111	<0.001	0.414	-0.010	0.732	0.229	-0.043	0.090	0.230
3.00	0.381	<0.001	0.636	0.233	0.004	1.142	-0.038	0.551	0.475	-0.101	0.138	0.552
4.00	0.372	<0.001	1.044	0.408	0.008	2.397	-0.060	0.591	0.788	-0.143	0.257	0.997
5.00	0.344	<0.001	1.752	0.633	0.004	4.179	-0.056	0.741	1.220	-0.154	0.438	1.539

In Table 3.12 it is observed that the s_f and Moran's index values increase with changes in alpha. Ischaemic and Cerebrovascular mortality are known to be spatially autocorrelated based on the original Moran's index. The Monte

Carlo simulation test confirms this with p-values<0.05 except at low values of alpha ($\alpha \leq 0.06$ for ischaemic mortality and $\alpha \leq 0.20$ for cerebrovascular mortality). Ischaemic mortality is more spatially autocorrelated than cerebrovascular mortality which explains why spatial autocorrelation for the former is detected at lower alpha values than the latter. At higher values of alpha spatial autocorrelation is easily detected. Generally, the Monte Carlo simulation test is robust to changes in the α values (with the exception at very low levels that are not normally used). Usually alpha values of 1 or 2 are used in spatial analysis.

In the case of randomly generated data, Table 3.12 shows that the data are not significant ($p - value > 0.05$) at all alpha values which is what we were expecting. It shows once again that the Monte Carlo simulation test is robust to changes in the α values. This is despite s_f values of lower than 0.15 for alpha values less than or equal to 1.50. It is clear from this analysis that it is not sufficient to base spatial autocorrelation conclusions on s_f being less than 0.15 alone. A balance has to be found between the p-values and s_f values in obtaining an optimal α value for given data. In the case of South African mortality data such a balance is achieved with an alpha value of close to 1 or just using $\alpha = 1$ in the inverse spatial weights matrix. Further studies, beyond the scope of this study, are needed to find the exact optimal α value should it be necessary, otherwise an approximate value should suffice.

3.4.3 Analysis using contiguity spatial weights

The analyses in this section are as per the formulations in Chen (2013) where distance based weight matrices were used. This does not mean contiguity based weight matrices cannot be applied. It will be interesting to see how the queen's spatial weight matrix will fare with the new method and if it can give similar results as the original Moran's index. In this subsection we give a brief of the results when Chen's regression approach is used with the queen's contiguity weight matrix.

Table 3.13: Comparison of the global univariate Moran's index of spatial autocorrelation between Chen's approach and the original approach using a Queen weight matrix.

Condition	Anselin	Chen
Ischaemic	0.821**	0.752**
Cerebrovascular	0.297**	0.218**
Diabetes	0.316**	0.269**
Hypertension	0.329**	0.258**

An application was made to the Poisson adjusted cardiovascular mortality 2011 data; for ischaemic heart disease, cerebrovascular heart disease, hypertensive heart disease and diabetes. The distance based spatial weights of the inverse and exponential decay functions did not give significant Moran's indexes when used with Chen's index for diabetes, cerebrovascular and hypertensive mortality. But when the queen contiguity weight matrix was used the results of the global univariate Moran's indexes are shown in Table 3.13. Both methods were able to detect the presence of spatial clustering in all four rates of mortality. The local indicators of spatial autocorrelation using the Chen's regression approach was done and the cluster maps are shown in Figure 3.14.

Local indicators of spatial autocorrelation cluster maps observed when using the original Moran's index are shown in Figure 3.15 for comparison purposes. It can be seen in Figure 3.15 B and D that the clusters of cerebrovascular and hypertensive mortality are similar to the clusters in Figure 3.14 B and D, respectively. The clusters of ischaemic and diabetes mortality in Figure 3.14 A and C, respectively span across more municipalities than those in Figure 3.15 A and C, though "hot spots" clusters in all instances are located in the south western part of the country.

It is not possible to do a direct comparison between the clusters of Chen's

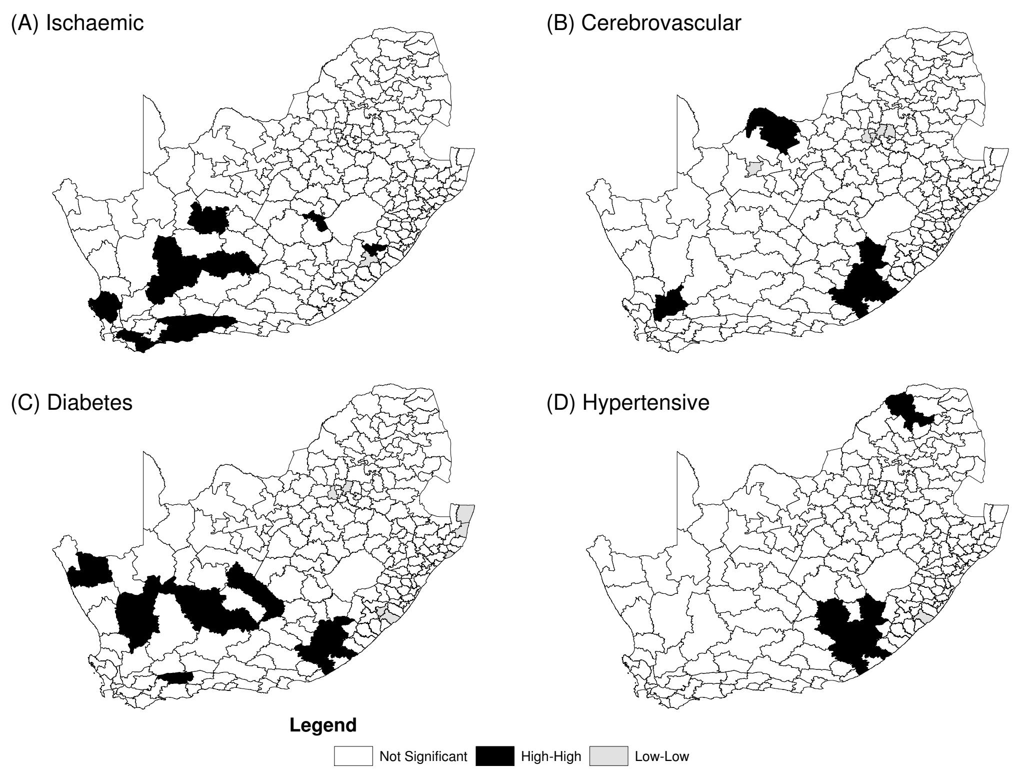


Figure 3.14: The univariate local Moran's spatial clusters for ishaemic mortality rates based on Chen's regression approach, for the Queen's spatial weights.

approach and those of the original method because the cluster categories differ as seen in the legends of Figures 3.14 and 2.15. However, the cluster map of ischaemic mortality rates in Fig 2.15 A has similarities with cluster maps based on distance based weight matrices in Figure 3.12. A concordance test between the cluster map of Fig 2.14 A and that of the inverse weight matrix with $\alpha = 1$ in Figure 3.12 A gives Bangdiwala statistic of 0.874, which suggest a high level of agreement.

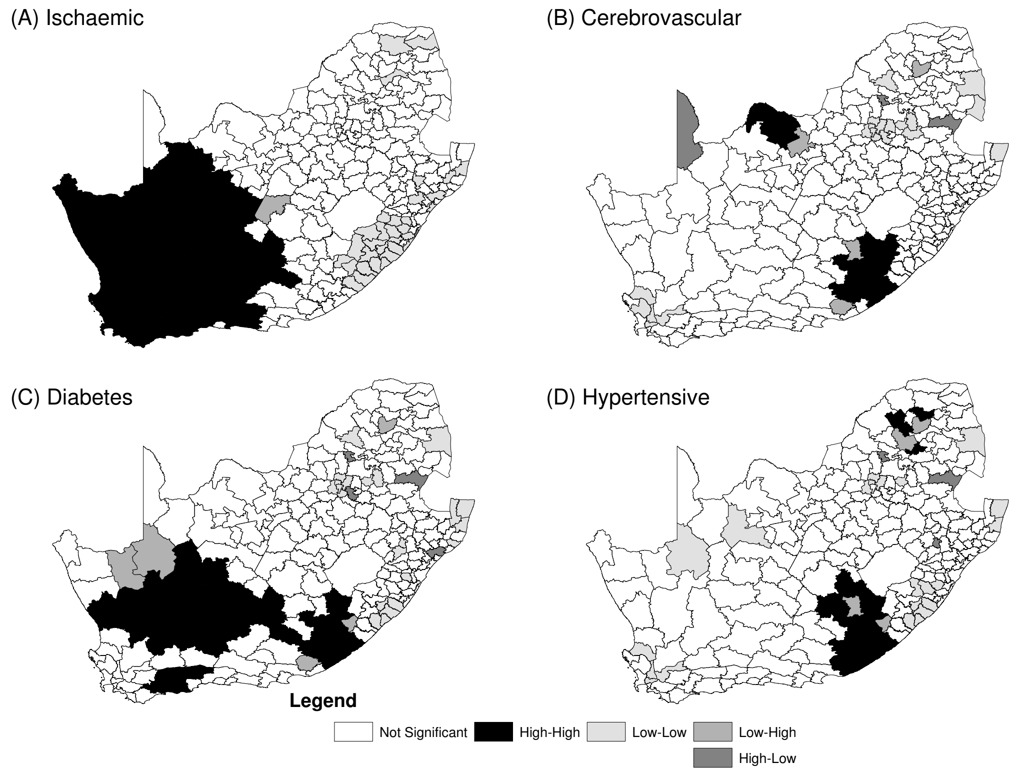


Figure 3.15: The univariate local Moran's spatial clusters for ishaemic mortality rates based on the original Moran's index, for the Queen's spatial weights.

3.5 Summary of the chapter

In this chapter the Moran's index and an alternative construct of the Moran's index based on linear regression were applied to two sets of CVD data in South Africa. A comparison of the results of Chen's regression approach was done between the spatial autocorrelation based on the inverse power distance function (with $\alpha = 1$) and that based on a negative exponential distance function (with $\alpha = 1; 2$). An R software programme was developed to determine global and local spatial clusters as well as ascertain their spatial significance. This is an innovative action which supplies research on the test of Chen's regression approach to Moran's I.

The first application was to cardiovascular prevalence data from SADHS of 2016. In this study we found significantly positive univariate spatial clustering for stroke (Moran's index = 0.128), smoking (0.606) hypertension (0.236) and high blood cholesterol (0.385). The second application of the chapter concerns the quantification of univariate spatial autocorrelations for CVD-related mortality in South Africa using the Moran's I. The study used mortality attributable to diabetes, cerebrovascular, ischaemic heart failure and hypertension captured by the country's Department of Home Affairs for the years 2001 and 2011. Univariate spatial clustering measures were derived using observed, empirical Bayes smoothed and empirical Bayes smoothed rates adjusted for age, race and poverty. Significant clustering was found in all the data except for diabetes which was only significant after adjusting for covariates using Poisson regression to estimate mortality rates. Clusters of CVD mortality were generally more pronounced in the south-west part of the country.

The Chen's Moran's indexes of spatial autocorrelation were found to be significant for the spatial weights based on the inverse power distance function with $\alpha = 1$ and that based on a negative exponential distance function $\alpha = 1$ and $\alpha = 2$. Only the residuals of the Moran's index based on the inverse power distance function with $\alpha = 1$ were normal. However, the local spatial clusters for all three spatial weight structures were similar indicating "hot-spots" of ischaemic mortality in the south western municipalities of the country. Sensitivity analysis to ascertain the effect of changing α in the inverse spatial weights matrix on ability to detect spatial autocorrelation showed that, on one hand, Monte Carlo simulation for the significance testing of the Moran's index was generally robust to changes in α. On the other hand, s_f values were found to increase with an increase in α with $s_f < 0.15$ in cases where data were known not to be spatially autocorrelated. The requirement that standard errors be less than 0.15 may not suffice. Instead the coefficient of variation may be used.

But dividing s_f by the mean of f's will make the coefficient of variation very big since the mean of f's will almost always be close or equal to zero.

Chapter 4

Review and application of bivariate spatial clustering methods.

4.1 Introduction

This chapter presents a review and an application of the bivariate spatial autocorrelation measures to two different South African health outcomes datasets recorded at municipality level that were discussed in Chapter 3. First, the original bivariate Moran's index is applied to prevalence rates of cardiovascular conditions and their risk factors. Then an application is made to cardiovascular mortality rates data using all three bivariate spatial autocorrelation measures, namely, Moran's I, Lee's L and Dray's H. Smoothing techniques and adjustments were made to the data before analysis in Chapter 3 as a way of mitigating against biases inherent in the data recorded at small areas.

Bivariate spatial autocorrelation was used to detect spatial co-clustering in two ways. Firstly, in both applications, it was used to test if the existence of high risk of one cardiovascular condition outcome in an area gives rise to a high risk outcome on nearby areas for a related disease or risk factor. The hypothesis being tested here is that interrelated health outcomes co-cluster. Identifying co-clusters of CVDs is important if a unified approach is to be employed in the prioritisation, prevention of the spread, diagnosis and cure of the related diseases.

Secondly, bivariate spatial autocorrelation was used to test if the spatial patterns of a CVD mortality differ between two time points. It is important to do this second test in order to establish if the spatial dynamics of a disease are changing with time. If the spatial dynamics of a disease are stable, it is much easier to predict the spatial patterns of the disease over time for planning and monitoring purposes.

4.1.1 Bivariate spatial autocorrelation measures

Bivariate spatial association measures were used in this PhD study to test spatial dependence between two diseases as well as to test if there is a difference in the spatial distribution of a disease over two time points. The bivariate methods applied in this study are variants of the formulation by Wartenberg (1985) and were derived using the popular Moran's I univariate spatial autocorrelation measure to detect clustering for one disease. The method suggested by Wartenberg (1985) for extending Moran's I to multivariate spatial analysis involves the derivation of a matrix of bivariate spatial autocorrelations. This matrix is, in turn, analysed using spatial principal component analysis, resulting in a set of spatial factors that represent the total spatial pattern. While it is preferable to use the row-sum standardised weights in the formulation of Wartenberg (1985), it was found to be problematic as it leads to an asymmetric matrix of bivariate spatial association measures to be diagonalised, which is complex to solve, as finding eigenvalues of such a matrix is difficult (Lee, 2001).

Lee (2001) gave conditions that must be satisfied by a bivariate spatial autocorrelation measure to be used for diagonalisation: a bivariate spatial autocorrelation measure must be a function of the respective individual univariate spatial autocorrelations and the "point to point" correlation of some sort between the two variables as measured by Pearson's correlation coefficient. Lee (2001) used the idea of a first order spatial lag (the weighted mean values for the immediate neighbours j of an area i) given by $Lx_i = \widetilde{x}_i = \sum_{j=1}^{n} w_{ij}x_j$, to show that the Moran's I in Equation 2.3 in Chapter 2, when applied with a row-sum standardised matrix, can be rewritten as:

$$I_X = \frac{\sum_{i=1}^{n}(\widetilde{x}_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{4.1}$$

The Pearson's correlation between variable X and its spatial lag $\widetilde{X}$ is given by

$$r_{X,\widetilde{X}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\widetilde{x}_i - \bar{\widetilde{x}})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(\widetilde{x}_i - \bar{\widetilde{x}})^2}}. \tag{4.2}$$

Dividing Equation 4.1 by Equation 4.2 and making the Moran's index the subject of the formular leads to:

$$I_X = \sqrt{\frac{\sum_{i=1}^{n}(\widetilde{x}_i - \bar{\widetilde{x}})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdot r_{X,\widetilde{X}}. \tag{4.3}$$

When the dispersion ratios in Equation 4.3 are further decomposed the following factorisation is realised:

$$I_X = \sqrt{\frac{\sum_{i=1}^{n}(\widetilde{x}_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdot \underbrace{\sqrt{\frac{\sum_{i=1}^{n}(\widetilde{x}_i - \bar{\widetilde{x}})^2}{\sum_{i=1}^{n}(\widetilde{x}_i - \bar{x})^2}}}_{\cong 1} \cdot r_{X,\widetilde{X}} \cong \sqrt{SSS_X} \cdot r_{X,\widetilde{X}}, \tag{4.4}$$

where SSS_X is a spatial smoothing scalar for variable X. This implies that the Moran's I is a product of a spatial smoothing scalar, SSS_X, and the correlation of a variable and its spatial lag, and can be written as $I_X = I_{X,\widetilde{X}}$. Deductively, the bivariate Moran's I between two variables X and Y was shown to be

$$I_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\widetilde{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \cong \sqrt{SSS_Y} \cdot r_{X,\widetilde{Y}}, \tag{4.5}$$

Equation 4.5 is a product of a spatial smoothing scalar (SSS) of a variable and the correlation of the variable and the spatial lag of the other variable. Clearly, the bivariate Moran's I does not satisfy the conditions set out by Lee (2001) as it is a function of only one univariate spatial association measure and a "point to point" association of two variables. Thus, Lee (2001) concluded that Wartenberg (1985) formulations are inadequate and should not be used in multivariate analysis. Lee (2001) went on to derive a bivariate spatial autocorrelation measure for use as a basis for multivariate spatial analysis:

$$\begin{aligned} L_{X,Y} &= \frac{n}{\sum_{i=1}^{n}(\sum_{j=1}^{n} v_{ij})^2} \cdot \frac{\sum_{i=1}^{n}[(\sum_{j=1}^{n} v_{ij}(x_j - \bar{x}))(\sum_{j=1}^{n} v_{ij}(y_j - \bar{y}))]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \sqrt{SSS_X} \cdot \sqrt{SSS_Y} \cdot r_{\widetilde{X},\widetilde{Y}} \\ &= \sqrt{BSSS_X} \cdot r_{\widetilde{X},\widetilde{Y}} \end{aligned} \tag{4.6}$$

,

where $BSSS$ denotes the bivariate spatial smoothing scalar and $[v_{ij}]_{n\times n}$ is a row-standardised weight matrix. Equation 4.6 by Lee (2001), known as Lee's L, is not only in line with his conditions for a bivariate spatial autocorrelation measure, but also produces a symmetric bivariate spatial autocorrelation matrix to be used for deriving total multivariate spatial autocorrelations. The Pearson's correlation part of Lee's derivation, $r_{\widetilde{X},\widetilde{Y}}$, is between the spatial lags of the two variables that will be considered. Additionally, Lee (2001) showed that if one puts X=Y, then

$$L_{X,X} = \sqrt{SSS_X} \cdot \sqrt{SSS_X} \cdot \underbrace{r_{\widetilde{X},\widetilde{X}}}_{=1} = SSS_X = S_X. \tag{4.7}$$

Equation 4.7 is often referred to as Lee's S and can be used to measure univariate spatial autocorrelation just like univariate Moran's I (Lee, 2001).

Despite the criticism of the ideas of Wartenberg (1985), the approach has remained popular, with Anselin *et al.* (2002) expanding the formulation to visual analysis of bivariate Moran's I spatial association measure. This expansion was done for both global and local indexes using a standardised weight matrix W. The Moran's bivariate measure does not meet the conditions set out by Lee (2001) as noted earlier. In order to overcome this difficulty, Dray *et al.* (2008) cautioned that instead of using W in his formulations, Lee (2001) should have used $\frac{\mathbf{W}+\mathbf{W}^T}{2}$ as originally suggested by de Jong *et al.* (1984). Dray *et al.* (2008) then proceeded to use the transformation by de Jong *et al.* (1984) to develop a bivariate spatial association measure:

$$H_{X,Y} = \frac{1}{2}\left[\sqrt{SSS_X} \cdot r_{X,\widetilde{Y}} + \sqrt{SSS_Y} \cdot r_{Y,\widetilde{X}}\right]. \tag{4.8}$$

The bivariate spatial association measure in Equation 4.8 is not only symmetric, but satisfies the conditions of Lee (2001). In addition, the measure is a function of the correlation of one variable and the lag of the second variable, thus indirectly connecting it to the regression formulations by Anselin *et al.* (2002).

4.2 Application to cardiovascular prevalence data

Co-clustering using raw data for all participants

The global bivariate spatial autocorrelation indexes for the association between the prevalence of CVDs and identified risk factors for all participants, irrespective of gender or age, were calculated and are shown in Table 4.1. In the diagonal are global univariate Moran's index values for detecting global spatial clustering that were calculated in Chapter 3. All, except heart attack, are exhibiting presence of spatial clustering at 5% significance level. Additionally, the following exhibit spatial dependency: stroke and smoking; stroke and HBC; smoking and HBC; smoking and hypertension. Although the distribution of heart attack did not show any spatial patterns, we still tested its dependency on the other variables and was found to be spatially dependent on smoking at 5% significance level.

Table 4.1: Global bivariate spatial autocorrelation association between the prevalence of CVDs and identified risk factors for all participants.

	Stroke	Heart attack	Smoking	HBC	Hypertension
Stroke	0.203**	0.031†	0.267**	0.305**	-0.011†
Heart attack		-0.013†	0.181**	0.108†	0.017†
Smoking			0.662**	0.426**	0.243**
HBC				0.503**	-0.002†
Hypertension					0.329**

Key: HBC, high blood cholesterol;†, insignificant at 5% level; **, significant at 5% level.

Figure 4.1 shows the clusters for CVDs and their risk factors that exhibit significant spatial dependents at district level in South Africa, disregarding age and gender. The key shows "hot-spots " (High-High) in the black colour and "cold-spots " (Low-Low) in the light grey colour. It can be observed in Figures 4.1 E-F that the joint "hot-spot " cluster of stroke and its risk factors of smoking and HBC, when all data is used, is found in the south western part of the country and comprises of City of Cape Town, Cape Winelands, Overberg and

Eden Districts. Smoking and HBC have a joint “hot-spot ” cluster comprising of the same four districts (Figure 4.1 G). This probably explains why the region is a “hot-spot ” for stroke. The biggest joint “hot-spot ” cluster is for smoking and hypertension, which spans a wider area from the south east, across the central part, to the north east part of the country (Figure 4.1 H). This cluster comprises of nine districts: Cape Winelands; Overberg; Eden; Cacadu; Lejweleputswa; Pixley ka Seme; Z F Mjcawu; Frances Baard; and Central Karoo.

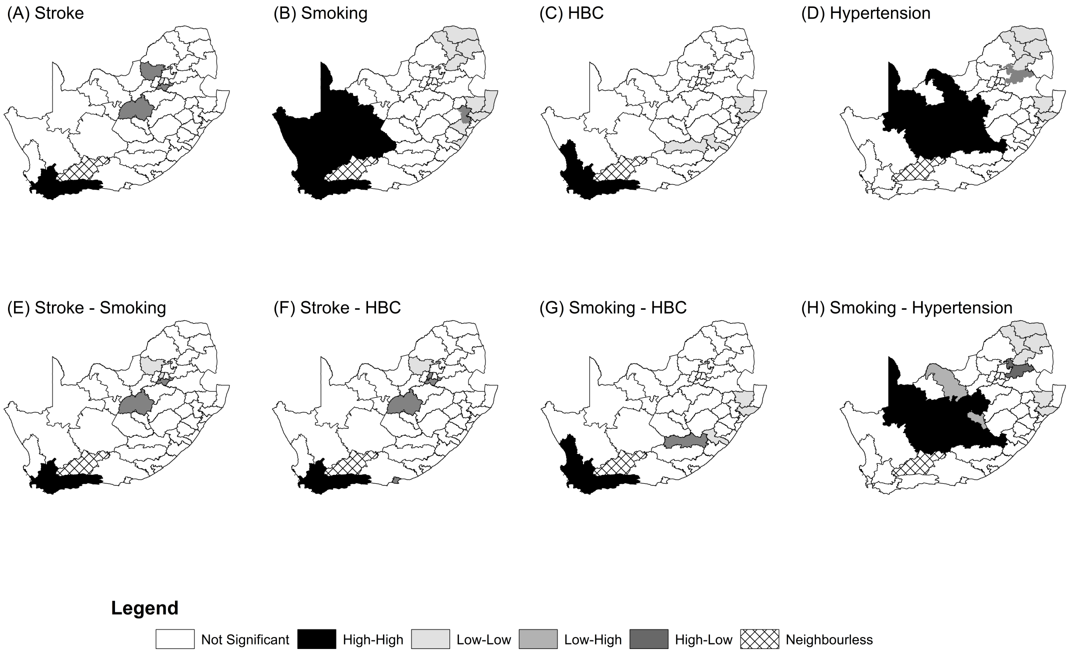


Figure 4.1: Univariate and joint spatial clusters of CVDs and their risk factors with significant association for all participants using raw rates.

4.2.1 Age-gender standardised joint spatial clustering analysis

In the previous section, our analyses used the raw prevalence of the two cardiovascular diseases and the three associated risk factors for the whole sample. As discussed in chapter, analysing spatial clustering using raw prevalence may be misleading due to confounding effects of covariates such as age and gender. It was then recommended to make use of standardised incidence ratios (SIR). We use SIRs here for the main bivariate spatial autocorrelation analyses.

The calculated values of univariate and bivariate measures of spatial clustering are presented in Table 4.2, where the diagonal values are the univariate global Moran's I values for the SIRs of CVDs and identified risk factors as calculated in Chapter 2. The off-diagonals (or upper triangle) are the global bivariate spatial autocorrelation indexes for the association between the SIRs of CVDs and identified risk factors for all participants.

Table 4.2: Global univariate and bivariate spatial autocorrelation association between the age-sex standardised incidence rates of CVDs and identified risk factors for all participants.

	Stroke	Heart Attack	Smoking	HBC	Hypertension
Stroke	0,128*	-0,019†	0,218**	0,184**	-0,075†
Heart Attack		-0,015†	-0,099†	-0,021†	-0,008†
Smoking			0,606***	0,366***	0,149†
HBC				0,355***	-0,077†
Hypertension					0,236**

Key: HBC, high blood cholesterol; †, insignificant at 5% level; **, significant at 5% level; *, significant at 10% level..

The SIRs for heart attack do not show any spatial patterns with non-significant univariate Moran's index. However, stroke (at 10%) and the three risk factors of smoking, HBC and hypertension, are exhibiting spatial significance at 5% significance level. It can also be seen that there is no evidence of spatial dependence between heart attack and all the three risk factors of CVDs at 5% significance level. Evidence is such that stroke is significantly spatially associated with smoking and HBC. In addition, there is also high spatial dependence between smoking and HBC (p-value less than 0.001).

The bivariate local indicators of spatial autocorrelations (LISA) for the five CVDs and risk factors were estimated and Figure 4.2 shows local joint cluster for different pairwise CVD and risk factors. Joint stroke-smoking “hot-spots ” district clusters (comprising West Coast, City of Cape Town, Cape Winelands, Overberg and Eden) were found in the south western part of the country. Similar joint “hot-spots ” clusters were found for stroke and HBC, and for smoking and HBC (Figure 4.2 B-C). Joint “hot-spots ” clusters of smoking and HBC are also concentrated in the Western Cape Province and is comprised of West Coast, City of Cape Town, Cape Winelands, Overberg and Eden districts.

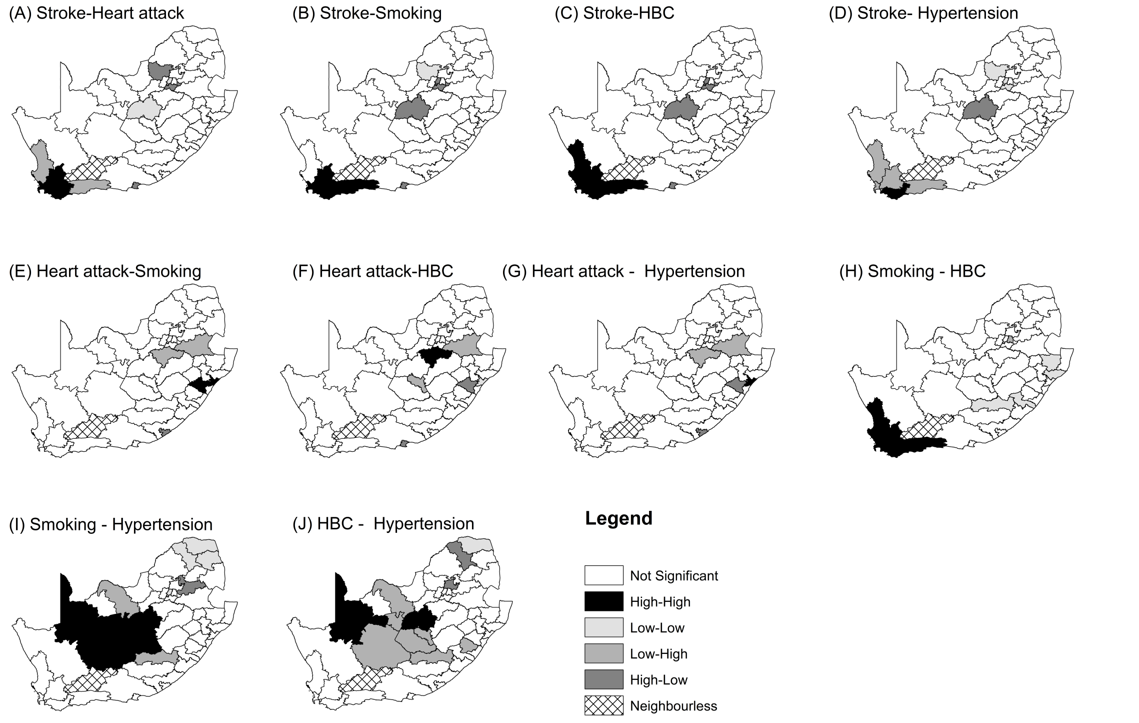


Figure 4.2: Joint Spatial Clusters of CVDs and their risk factors with significant association for all participants, adjusted for the national age-sex distribution of the sample.

The following “cold-spots ” were observed for significant associations: stroke and smoking, in Bojanala (in rural North West Province); stroke and HBC, in Sedibeng, West Rand (in urban and rural Gauteng Province), and Lejweleputswa (in rural Free State Province); and smoking and HBC, in Alfred Nzo and Joe Gqabi (in rural Eastern Cape Province), and Zululand and UThungulu (rural KZN Province). There were bivariate associations that were not significant: heart attack and stroke; heart attack and HBC; heart attack and hypertension; heart attack and smoking; smoking and hypertension; and HBC and hypertension.

4.3 Application to cardiovascular mortality rates

4.3.1 Bivariate association of individual CVD maps over time

It was shown in Chapter 2 that the geographical variation of mortality due to each CVD is significant over the years and the univariate clusters have been identified. What has not been shown, however, is whether or not the variation in the distribution of each CVD risk is the same over the years. In this section, bivariate spatial autocorrelation measures are used to determine if there is a difference in spatial distribution of mortality risk due to each of the CVDs between the two time points. This will shed some light on whether or not the geographical distribution of individual CVDs is changing over time.

The stability of spatial dynamics in the distribution of a disease is important if there is to be some semblance of predictability. This may be helpful when deciding the course of action to be taken when faced with an epidemic as intervention programmes are conceived. Changing spatial dynamics in the distribution of the disease may make it quite complex to contain the disease. Generally, one would not expect spatial dynamics of CVDs to change much within a short period of time as the factors and habits contributing to the emergence of these diseases take time to control. Thus, one would expect the bivariate spatial dependency of the distribution of each CVD over two time periods to be significantly positive.

The dependence of CVD rates in space for each of the conditions studied was tested for two different periods using bivariate Lee's L, Moran's I and Dray's H. The analysis was conducted on each CVD for the following comparative periods: 2001 versus 2011. Results are provided in Table 4.3.

Table 4.3: Bivariate global spatial autocorrelations measuring spatial dependence of individual CVD rates for raw, smoothed and adjusted mortality rates between the years 2001 and 2011.

Association (X-Y)		**Lee 2001** ($L_{X,Y}$)			**Anselin 2002** ($I_{X,Y}$)			**Dray 2008** ($H_{X,Y}$)		
X	**Y**	**RR**	**EB**	**Adj**	**RR**	**EB**	**Adj**	**RR**	**EB**	**Adj**
CVA01	CVA11	0.115**	0.125**	0,222***	0.060**	0,075**	0,179***	0.065**	0,075**	0,185***
IHD01	IHD11	0.250***	0,259***	0,811***	0.239**	0,243**	0,819***	0.254***	0,252***	0,821***
DBT01	DBT11	0,065†	0,076**	0,367***	0,065**	0,079**	0,359***	0,058**	0,071**	0,379***
HHD01	HHD11	0.084**	0,097**	0,331***	0.082**	0,094**	0,322***	0.084**	0,085**	0,325***

Key: *** = p-values <0,001; **= p-values < 0,05; † = Insignificant p-values.

Generally, the three indicators of bivariate spatial autocorrelation show that there is significant spatial dependency on how each disease is spatially distributed between the two time periods. It can, thus, be concluded that the spatial distribution of the risk of mortality due to each CVD has not significantly changed over the course of the ten-year period under review. Note that the bivariate Moran's I and Dray's H show similar results. This is not surprising as the methods are based on the same derivation. The bivariate LISA analyses of the combinations in Table 4.3 derived from the raw, smoothed and adjusted rates are presented in Figure 4.3.

In Figure 4.3 F, as an example, observed "hot-spots" are areas of high mortality of IHD in 2001 whose neighbourhood in 2011 also exhibits high mortality of IHD to form a co-cluster of high mortality for the two-time points

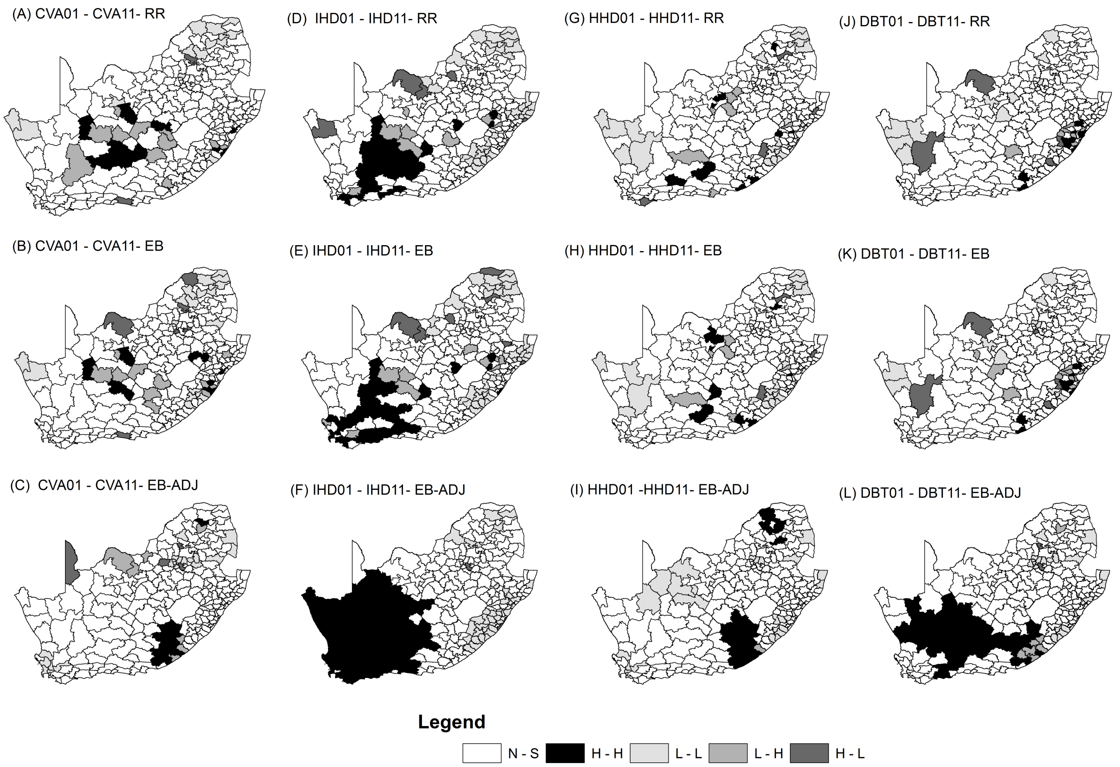


Figure 4.3: The raw, smoothed and adjusted mortality rates based Bivariate Moran's I LISA maps between same CVDs for the year 2001 and 2011.

in the south-western part of the country. The CVA and HHD co-clusters are similar and are found in the south and north-east part of the country.

4.3.2 Bivariate spatial association between two CVDs at a point in time

We also looked at determining spatial dependency between two different CVDs at a cross-section. One can hypothesise that CVDs should co-cluster or show spatial dependency at a point in time as they share risk factors. Table 4.4 presents the bivariate association measure values calculated for the possible combinations of the three CVDs for the years 2001 and 2011 to determine spatial dependence based on raw, smoothed and adjusted rates data.

Table 4.4: Bivariate global spatial autocorrelations measuring spatial dependence of individual CVD rates between two time periods for raw, smoothed and adjusted mortality rates between the years 2001 and 2011.

Association (X-Y)		**Lee 2001** ($L_{X,Y}$)			**Anselin 2002** ($I_{X,Y}$)			**Dray 2008** ($H_{X,Y}$)		
X	**Y**	**RR**	**EB**	**Adj**	**RR**	**EB**	**Adj**	**RR**	**EB**	**Adj**
CVA01	IHD01	0.159†	0,165**	0,338***	0.014†	-0000†	0,344***	0,018†	0,003†	0,359***
CVA11	IHD11	0.173†	0.166†	-0,084†	0.067**	0.051†	-0,105†	0.066**	0.048†	-0,113†
CVA01	HHD01	0.046†	0.046†	0,295***	-0.070**	-0.075**	0,279***	-0,066†	-0,073†	0,281***
CVA11	HHD11	0.135†	0.139†	0,350***	0.017†	0.012†	0,294***	0,011†	0.007†	0,274***
CVA01	DBT01	0.154†	0.176†	0,436***	-0.040†	-0.037†	0,432***	-0.033†	-0.032†	0,442***
CVA11	DBT11	0.187†	0.196†	0,269***	0.009†	-0.042†	0,196***	0.007†	0.013†	0,196***
DBT01	IHD01	0.192†	0,203**	0,713***	0.002†	0.003†	0,719***	0.002†	0.002†	0,725***
DBT11	IHD11	0.108†	0.117†	0,302***	0.008†	0.014†	0,289***	0.009†	0.015†	0,312***
HHD01	IHD01	-0.035†	-0.018†	-0.116†	-0.100**	-0.100**	-0.121†	-0,107**	-0.095†	-0.128†
HHD11	IHD11	0.105†	0.097†	-0.089†	0.034†	0.018†	-0.110†	0,033†	0.018†	-0.110†
HHD01	DBT01	0.042†	0.044†	0,065†	-0.069**	-0.067†	0.057†	-0,069†	-0.069**	0,058†
HHD11	DBT11	0.131†	0.134†	0,263***	-0,012†	-0,012†	0,203***	-0,011†	-0,012†	0,205***

Key: *** = p-values <0,001; **= p-values < 0,05; † = Insignificant p-values.

The bivariate Moran's I and Dray's H once again showed similar results. All three methods generally agree, based on the raw and smoothed rates, that there is no evidence of spatial dependence between all the associations tested. However, adjusted rates based tests revealed significant spatial dependence between the following maps: CVA01 and IHD01; CVA01 and HHD01; and CVA11 and HHD11. Importantly, DBT was found to have a significant association with all the three CVDs. This is in line with expectation as DBT is a well-known biomarker for CVDs. The other associations were either insignificant or their association was purely random with a negative Moran's index. The significant joint local "hot-spots " of the CVD associations, based on adjusted rate data, are shown in Figure 4.4.

Focusing on the most recent 2011 data, it can be seen that the "hot-spots " of DBT and the three CVDs in Figures 3.4 F-H are located in the south west part of the country. The joint clusters of CVA and HHD for the year 2011 are in the south and north-west of the country. The joint clusters of CVA-DBT

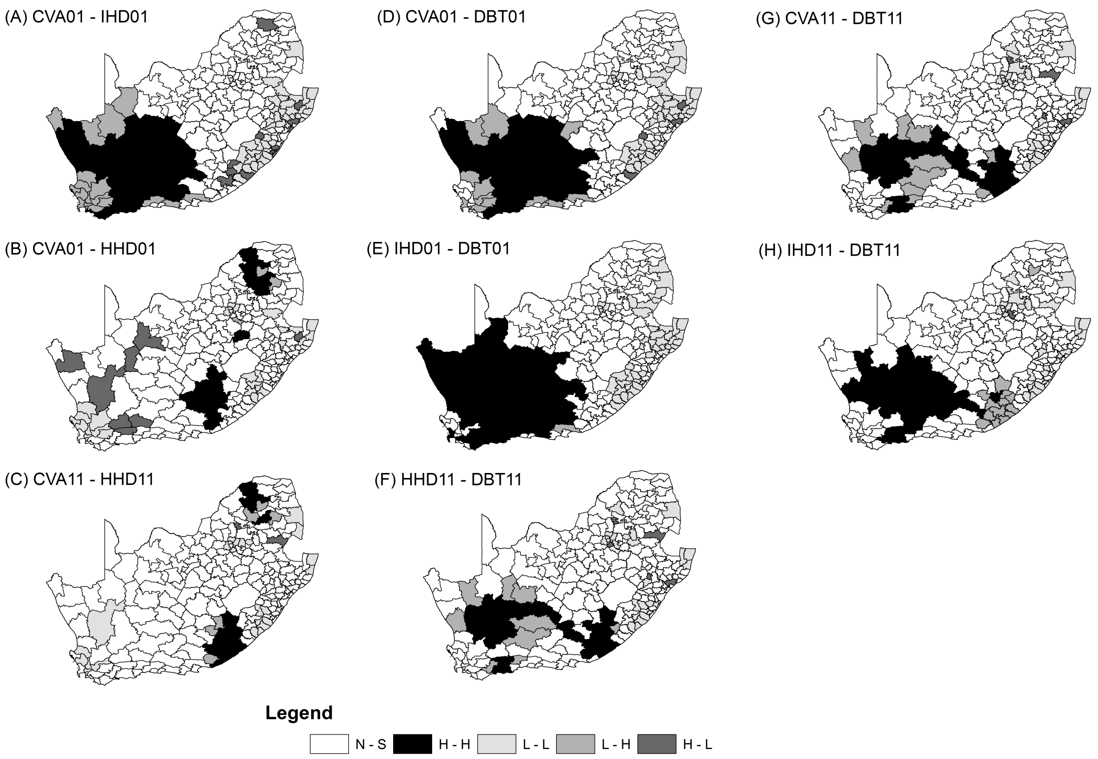


Figure 4.4: The significant adjusted mortality rates based Bivariate Moran's I LISA map between two CVDs at a point in time, 2001 and 2011.

and IHD-DBT have reduced in size over the period under review. This may be attributable to intervention programmes. However, joint clusters of HHD and DBT that were not in existence in 2001, have formed over the period under review as both the deaths and crude national rates attributable to the two diseases have increased over the period as was shown in Tables 2.5 and 2.6.

4.4 Chapter summary

In this chapter we set out to apply bivariate spatial autocorrelation measures to cardiovascular health outcomes. The first application was to cardiovascular prevalence data from SADHS of 2016. In this study we found significantly positive univariate spatial clustering for stroke (Moran; s Index = 0.128), smoking (0.606) hypertension (0.236) and high blood cholesterol (0.385). Smoking and high blood cholesterol (0.366), smoking and stroke (0.218) and stroke and high blood cholesterol (0.184) were the only bivariate outcomes with significant bivariate clustering. There was a joint stroke-smoking local “hot spots” cluster among four districts in the urban western part of the country (City of Cape Town; Cape Winelands; Overberg and Eden) and a joint “cold spots” cluster in the rural north-western part of the country. Similar joint “hot spots” clustering was found for stroke and high blood cholesterol, which also had “cold spots” cluster in the rural east-central part of the country. Smoking and high blood cholesterol had a “hot spots” cluster among five districts in the urban western part of the country (City of Cape Town; Cape Winelands; Overberg; Eden, and West Coast) and “cold spots” around the rural districts in east-southern parts of the country.

The second application of the chapter concerns the derivation and quantification of bivariate spatial autocorrelations for CVD-related mortality in South Africa using the three spatial autocorrelation methods of Moran, Lee and Dray. The study used mortality attributable to diabetes, cerebrovascular, ischaemic heart failure and hypertension captured by the country’s Department of Home Affairs for the years 2001 and 2011. Both univariate and pairwise spatial clustering measures were derived using observed, empirical Bayes smoothed and empirical Bayes smoothed rates adjusted for age, race and poverty. Cerebrovascular and ischaemic heart co-clustering was significant for the year 2011. Dray’s H and the Moran’s I gave identical results. All three methods generally had similar results. Data of adjusted rates revealed significant spatial dependence between the following mortality rates: CVA01 and IHD01; CVA01 and HHD01; and CVA11 and HHD11. Importantly, DBT was found to have a significant association with all the three CVDs. This is in line with expectation as DBT is a well-known biomarker for CVDs. Cerebrovascular and hypertension co-clustering was not significant and so were hypertension and ischaemic heart co-clustering. Co-clusters of cerebrovascular-ischaemic heart disease are the most profound and are located in the south-west part of the country. It was successfully demonstrated that bivariate spatial autocorrelations can be derived for spatially dependent mortality rates as exemplified by mortality rates attributed to the three cardiovascular conditions.

Chapter 5

The proposed measure of statistical multivariate spatial autocorrelation

5.1 Introduction

This chapter presents a new multivariate spatial autocorrelation statistic for detecting joint "hot-spots" for more than two outcomes that are spatially related. The proposed new multivariate spatial clustering method is based on canonical correlation. There were more than two cardiovascular-related health outcome data that were analysed in chapters 3 and 4. Unfortunately, we could not analyse all of them simultaneously as the available spatial autocorrelation methods can only cater for up to only two health outcomes. More data for related health outcomes that are geographically referenced are becoming readily available. This calls for need to develop multivariate spatial autocorrelation methods that caters for more than two health outcomes. Thus, the bivariate methods of Chapters 4 need to be extended to analyses of at least three related health outcomes, as more than two cardiovascular health outcomes were related.

Wartenberg (1985) proposed the use of principal component analysis to extend bivariate Cross Moran I to multivariate spatial autocorrelation. The diagonalisation method he proposed for the extension is complicated and not easy to apply. As a result, the proposal by Wartenberg (1985) has not been popular. Studies have been undertaken to develop and apply methods that employ regression analysis to determine and interpret spatial autocorrelation measures which are easily understood by most researchers (Anselin, 1995; Smouse *et al.*, 1986; Chen, 2013). But current methods which employs regression approach have not gone beyond bivariate spatial autocorrelation analysis (Anselin, 1995; Lee, 2001; Dray *et al.*, 2008; Chen, 2015). This chapter proposes an extension of the Moran's index of spatial autocorrelation to measures of multivariate spatial autocorrelation that cater for more than two variables.

Research has previously been undertaken in which principal component analysis (Jombart *et al.*, 2008; Montano & Jombart, 2017) and cross-correlations Eckardt & Mateu (2021) has been used to develop techniques to analyse areal data. The problem with these techniques is that they can only measure partial or semi-partial spatial autocorrelations. This is opposed to the traditional Moran's index which measures a direct simple correlation. Thus there is a need for the development of a multivariate method that can make use of a direct simple correlation between multiple variables as opposed to partial correlations.

There are three members of the family correlations that may be considered in the process of developing a multivariate spatial autocorrelation measure. Firstly, one may consider the Pearson's product moment correlation coefficient.

It is easy to apply, hence the most popular of the family of correlations. The univariate and bivariate measures of spatial autocorrelation indexes are derived from the Pearson's product moment correlation coefficient (Mantel, 1967; Anselin, 1995; Lee, 2001; Anselin *et al.*, 2002). It is used for pairwise analysis, and for this reason we cannot employ it to analyse more than two outcomes when considering multivariate associations. The extension of Moran's index to multivariate analysis requires higher order correlation methods. Secondly, one may consider a higher correlation method, namely, multiple linear regression analysis. This is the approach used in the calculation of Mantel tests and the partial Mantel test (Smouse *et al.*, 1986; Legendre, 2000). Although it is a higher order correlation method, it cannot be applied here as it measures partial correlations between the dependent variable and each of the independent variables involved. Besides that, the partial Mantel test is basically a bivariate spatial autocorrelation measure in which co-variates are being controlled for one of the variables involved (Legendre, 2000). Lastly, one may consider the canonical correlation coefficient, which is the most generalised form of the members of the family of correlation analysis (Clark, 1975). It is a higher order correlation method that assesses correlations across sets of data. Hence it is an ideal approach when considering multivariate associations. It is the aim of this chapter to extend the Moran's index of spatial autocorrelation to measures of multivariate spatial autocorrelation using canonical correlation analysis.

5.2 Methods

5.2.1 Standardised coefficients for simple linear regression equations

Linear regression coefficients can either be unstandardised or standardised. The standardisation of regression coefficients is necessitated by the need to compare the importance of predictor variables when predicting the criterion variable. Most often than not, predictor variables are measured in different metric units (metres, units, proportions, etc.), and are referred to as metric variables. The realised “metric coefficients” or unstandardised coefficients of a regression model depends on the units used to measure the predictor variables. Thus, when the units of the predictor variables are different, then the values of the calculated metric coefficients have no real meaning and are not comparable. The solution to this problem is to standardise both the criterion and predictor variables and make them “metric free”.

Standardised coefficients may be useful in comparing the importance of predictor variables in determining a criterion variable, but are generally not recommended in prediction analysis. In order to predict the criterion variables, one needs to employ the metric coefficients. A way is available to calculate the unstandardised coefficients from the standardised coefficients using relationships with variances, covariances and correlations. The simple regression equation of a criterion variable $\widetilde{\mathbf{Y}}$ versus the prediction variable $\mathbf{Y}$ is given by

$$\widetilde{\mathbf{Y}} = a + b\mathbf{Y}, \tag{5.1}$$

while the equation when the variables are standardised such that they have mean zero and variance 1, is given by:

$$\widetilde{\mathbf{Y}}^s = \beta \mathbf{Y}^s, \tag{5.2}$$

where a is a regression constant, while the unstandardised and standardised coefficients are represented by b and β, respectively. Note that the regression constant is zero when standardised variables are used, as shown in Equation 5.2. Also note that the standardised coefficient of $\mathbf{Y}$ is simply the correlation between $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}^s$. That is:

$$\rho_{y\widetilde{\mathbf{y}}} = \beta. \tag{5.3}$$

It can be shown that the relationship between the unstandardised and standardised coefficients of $\mathbf{Y}$ is given by:

$$\begin{gathered} b = \beta \times \frac{SD(\widetilde{\mathbf{Y}})}{SD(\mathbf{Y})} \text{ or} \\ b = \rho_{y\widetilde{y}} \times \frac{SD(\widetilde{\mathbf{Y}})}{SD(\mathbf{Y})}. \end{gathered} \tag{5.4}$$

The formula in Equation 5.4 shows that one can easily determine the metric coefficients from the standardised coefficients. The converse is also true. Thus, if any of the coefficients is known, there is no need to run a new regression equation to determine the other. We can simply apply Equation 5.4.

5.2.2 The Moran's index as linear regression and the notion of a multivariate Moran's index

Anselin (1995) has shown that the Moran's index can be obtained from a linear regression equation between values of a disease outcome in an area, say y_1, and the spatial-lagged values $\widetilde{y}_1$ formed by averaging all the observations in neighbouring areas. The univariate Moran's index, I_{y_1}, will be the slope in the regression equation

$$\widetilde{y}_1 = a + I_{y_1} y_1 + \varepsilon, \tag{5.5}$$

where a is a regression constant and ε is the error term. In other words, it measures the influence of the outcomes of one disease in an area on the prevalence of the same disease in the neighbourhood of a given area. When y_1 is standardised, then

$$\widetilde{y}_1^s = \rho_{y_1,\widetilde{y}_1} y_1^s + \varepsilon, \tag{5.6}$$

where $\rho_{y_1,\widetilde{y}_1}$ is the correlation between y_1 and $\widetilde{y}_1$, while $\widetilde{y}_1^s$ represents the standardised values of $\widetilde{y}_1$. This is because the regression coefficient, resulting from regressing two standardised variables on each other, is just the simple correlation coefficient between them. Thus, the univariate Moran's index, I_{y_1}, can be obtained from the regression coefficient of Equation 5.6 using the following formula:

$$I_{y_1} = \rho_{y_1,\widetilde{y}_1} \frac{SD(\widetilde{y}_1)}{SD(y_1)}. \tag{5.7}$$

The result in Equation 5.6 is obtained using Equation 5.4. It can also be established from the spatial smoothing scalar derivations of Lee (2001). These derivations are important in determining the relationship between the correlation and the bivariate Moran's index and are, therefore, discussed here.

Suppose $y_1, y_2, ..., y_n$ are the realisations of a geo-referenced variable Y. Define a spatial weight matrix as an $n \times n$ proximity matrix whose elements $\{w_{ij}\}_{i,j=1}^n$ define the "strength" of the neighbourhood relationship between two areas i and j. The global Moran's I using the standardised spatial weights, w_{ij}, is given by:

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} \cdot (y_i - \bar{y}) \cdot (y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (y_i - \bar{y}) \sum_{j=1}^n w_{ij} \cdot (y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Let $\widetilde{y}_i = \sum_{j=1}^n w_{ij} y_j$ be the spatial lag of observation y_i in area i. Using this spatial lag, Lee (2001)) rewrote the Moran's I for variable Y as

$$I_y = \frac{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{5.8}$$

and the Pearson's correlation between variable Y and its spatial lag $\widetilde{Y}$ as:

$$r_{y,\widetilde{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\widetilde{y}_i - \bar{\widetilde{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{\widetilde{y}})^2}}. \tag{5.9}$$

Equations 5.8 and 5.9 were combined and gave the following factorisation of the global Moran's index:

$$I_y = \sqrt{\frac{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{\widetilde{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \cdot r_{y,\widetilde{y}}. \tag{5.10}$$

The ratios of variances in Equation 5.10 were further decomposed to give

$$I_y = \sqrt{\frac{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \cdot \underbrace{\sqrt{\frac{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{\widetilde{y}})^2}{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{y})^2}}}_{\cong 1} \cdot r_{y,\widetilde{y}} \cong \sqrt{SSS_y} \cdot r_{y,\widetilde{y}}, \tag{5.11}$$

where SSS_y is a smoothing scalar and I_y can be written as $I_{y,\widetilde{y}}$. It follows that

$$I_{y,\widetilde{y}} = \sqrt{\frac{\sum_{i=1}^{n}(\widetilde{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \cdot r_{y,\widetilde{y}} = \frac{SD(\tilde{y})}{SD(y)} \cdot r_{y,\widetilde{y}}. \tag{5.12}$$

Equation 5.12 is equivalent to Equation 5.7. Based on Equation 5.8, it can be shown that the bivariate Moran's index between two geo-referenced variables Y_1 and Y_2 is defined as

$$I_{y_1,y_2} = \frac{\sum_{i=1}^{n}(y_2 i - \bar{y_2})(\widetilde{y_1 i} - \bar{y_1})}{\sqrt{\sum_{i=1}^{n}(y_2 i - \bar{y_2})^2}\sqrt{\sum_{i=1}^{n}(y_1 i - \bar{y_1})^2}} \cong \sqrt{SSS_{y_1}} \cdot r_{y_2,\widetilde{y_1}}, \tag{5.13}$$

where the spatial smoothing scalar SSS_{y_1} is given by

$$SSS_{y_1} = \frac{\sum_{i=1}^{n}(\widetilde{y}_{1i} - \bar{y}_1)^2}{\sum_{i=1}^{n}(y_{1i} - \bar{y}_1)^2}. \tag{5.14}$$

It follows that the bivariate Moran's index can be written as

$$I_{y_1,y_2} = \frac{SD(\tilde{y_1})}{SD(y_1)} \times r_{\widetilde{y_1},y_2}. \tag{5.15}$$

Based on Equation 5.15, bivariate Moran's index can be obtained from a linear regression equation between values y_2 of a disease outcome in an area and the spatial-lagged values of another spatially dependent disease $\widetilde{y}_1$. The bivariate Moran's index, I_{y_1,y_2}, will be the slope in the regression equation

$$\widetilde{y}_1 = I_{y_1,y_2} y_2 + \varepsilon, \tag{5.16}$$

where a is a regression constant and ε is the error term. It measures the influence of the outcomes of one disease in an area on the prevalence of another related disease in the neighbourhood of that area. When both $\widetilde{y}_1$ and y_2 are standardised, then

$$\widetilde{y}_1^s = \rho_{y_2,\widetilde{y}_1} y_2^s + \varepsilon, \tag{5.17}$$

where $\rho_{y_2,\widetilde{y}_1}$ is the correlation between y_2 and $\widetilde{y}_1$, while $\widetilde{y}_1^s$ and y_2 represent the standardised values of the spatial lagged values of $\widetilde{y}_1$ and that of y_2, respectively.

The bivariate Moran's index, I_{y_1,y_2}, can be calculated from the regression coefficient in Equation 5.9 as follows:

$$I_{y_1,y_2} = \rho_{y_2,\widetilde{y}_1} \frac{SD(\widetilde{y}_1)}{SD(y_1)}. \tag{5.18}$$

In both the univariate and bivariate Moran's indexes the assumption is that the prevalence of a disease in an area is influenced by either the outcome of that disease or another single disease in the neighbourhood, but nothing else. But it is possible that we can have three or more spatially dependent variables acting on each other. In the case of three disease outcomes, say y_1, y_2 and y_3, we may be interested in determining the influence of both y_2 and y_3 disease outcomes on the prevalence of y_1. This influence can be measured by determining the slope of the following regression equation

$$\widetilde{y_1} = I_{y_1,v}v + \varepsilon, \tag{5.19}$$

where $v = w_1y_1 + w_2y_2$ is the weighted average of the two disease outcomes occurring in the neighbouring areas. When y_1, y_2 and y_3 are all standardised, then we have

$$\widetilde{y_1^s} = \rho_{\widetilde{y_1},v}v^s + \varepsilon, \tag{5.20}$$

where $\rho_{\widetilde{y_1},v}$ is the correlation between $\widetilde{y_1}$ and v, while v^s represents the weighted average of standardised y_2 and y_3. If one can establish the correlation $\rho_{\widetilde{y_1},v}$, then $I_{y_1,v}$ can easily be calculated, through induction, by the following formula:

$$I_{y_1,v} = \rho_{\widetilde{y_1},v}\frac{SD(\widetilde{y_1})}{SD(y_1)}. \tag{5.21}$$

Thus, if we know the correlation $\rho_{\widetilde{y_1},v}$ then we should be able to calculate the multivariate Moran's index, $I_{y_1,v}$, using Equation 5.13.

5.2.3 Canonical correlation analysis

The canonical correlation analysis (CCA) was used to extend the global Moran's index. In CCA we have two random variables, $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$, that are correlated, but not necessarily in the same space. The two random variables are assumed to have a joint normal multivariate distribution. The objective of CCA is to find pairs of random scalars (u, v) that represent each instance of $(\widetilde{\mathbf{Y}}, \mathbf{Y})$ by preserving the correlation between $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$ as much as possible. The scalars are linear transformations of $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$ and are of the form: $\mathbf{u} = \mathbf{a}^T\widetilde{\mathbf{Y}}$ and $\mathrm{v} = \mathbf{b}^T\mathbf{Y}$, where b and a are coefficient matrices. Preserving the correlation between $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$ as much as possible is akin to maximising the correlation between u and v.

Suppose $\widetilde{\mathbf{Y}} = \left[\widetilde{Y}_1, \widetilde{Y}_2, ..., \widetilde{Y}_p\right]^T$ is a $p{\times}1$ random variable consisting of sub-vectors $\widetilde{Y}_1, \widetilde{Y}_2, ..., \widetilde{Y}_p$ and $\mathbf{Y} = [Y_1, Y_2, ..., Y_q]^T$ is a $q \times 1$ be random variable consisting of sub-vectors $Y_1, Y_2, ..., Y_q$, where $p \leq q$. Canonical variates are linear combinations of the variables in either $\widetilde{\mathbf{Y}}$ or $\mathbf{Y}$ and are defined as:

$$\begin{aligned} \mathbf{u}_i &= \mathbf{a}_i^T\widetilde{\mathbf{Y}} \\ \mathbf{v}_i &= \mathbf{b}_i^T\mathbf{Y}. \end{aligned} \tag{5.22}$$

The i^{th} canonical correlation is the one determined between corresponding canonical variates u_i and v_i. Each step of canonical correlation analysis involves maximising the correlation between the two canonical variates, that is maximise

$$corr(\mathbf{u}_i, \mathbf{v}_i) = C_i = \frac{\mathbf{a}^T\Sigma_{y\widetilde{y}}\mathbf{b}}{\sqrt{\mathbf{a}^T\Sigma_{\widetilde{y}\widetilde{y}}\mathbf{a}}\sqrt{\mathbf{b}^T\Sigma_{yy}\mathbf{b}}}, \tag{5.23}$$

where $i = 1, 2, ..., k$; $k \leqslant$ min(p, q); $\Sigma_{\widetilde{y}\widetilde{y}}$ and Σ_{yy} are the covariance of $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$, respectively and $\Sigma_{\widetilde{y}y}$ is covariance between $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$. Ultimately, this maximisation problem is reduced to the following standard eigenvalue problem:

$$\Sigma^{-1}_{\widetilde{y}\widetilde{y}}\Sigma_{\widetilde{y}y}\Sigma^{-1}_{yy}\Sigma_{y\widetilde{y}}\mathbf{b} = \lambda\mathbf{b}. \tag{5.24}$$

$$\Sigma^{-1}_{yy}\Sigma_{y\widetilde{y}}\Sigma^{-1}_{\widetilde{y}\widetilde{y}}\Sigma_{\widetilde{y}y}\mathbf{a} = \lambda\mathbf{a}. \tag{5.25}$$

This standard eigenvalue problem can then be solved using the following characteristic equations:

$$\left|\Sigma^{-1}_{\widetilde{y}\widetilde{y}}\Sigma_{y\widetilde{y}}\Sigma^{-1}_{yy}\Sigma_{\widetilde{y}y} - \lambda\mathbf{I}\right| = 0, \tag{5.26}$$

$$\left|\Sigma^{-1}_{yy}\Sigma_{\widetilde{y}y}\Sigma^{-1}_{\widetilde{y}\widetilde{y}}\Sigma_{\widetilde{y}y} - \lambda\mathbf{I}\right| = 0, \tag{5.27}$$

where $\lambda = C_i^2$ or $C_i = \sqrt{\lambda}$. It means the canonical correlations are obtained by taking the square roots of the eigenvalues of either Equation 5.18 or Equation 5.19. It can be inferred from these equations that a and b are eigenvectors that correspond to an eigenvalue λ in the equations 5.16 and 5.17, respectively.

5.2.4 Significance tests for canonical correlation

The statistical significance test of the canonical correlation between two canonical variate pairs, (U_i, V_i), can be done using the Wilk's lambda (Everitt & Rencher, 1996). The Wilks' lambda is a generalised equivalent of the R-Squared in the context of the multivariate analysis, but its interpretation is the reverse of the R-Squared (Everitt & Rencher, 1996). Low values (close to zero) of the Wilks' lambda statistic indicate high correlation, while high values (close to one) are an indication of low correlation. The null hypothesis to be tested is $H_0 : C_1 = C_2 = C_3 = ... = C_k = 0$ versus the alternative hypothesis that states that at least one of the canonical correlations is not equal to zero.

The F statistic is estimated to determine the statistical significance of C_i^2, and is given by

$$F = \frac{1 - \lambda_1^{\frac{1}{t}}}{\lambda_1^{\frac{1}{t}}} \times \frac{DF1}{DF2}\, F_{DF1,\, DF2,\, \alpha}, \tag{5.28}$$

where $\lambda_1 = \prod_{i=1}^{k}(1 - r_i^2); k = \min(p, q); DF1 = pq; DF2 = wt - \frac{1}{2}pq + 1;$

$W = n - \frac{1}{2}(p + q + 3)$; and $t = \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}}$. Here n is the sample size, p is the number of criterion variables in the $\widetilde{\mathbf{Y}}$ set and q is the number of independent variables in the $\mathbf{Y}$ set and $r_i^2 = C_i^2$ are the squared canonical correlations.

5.2.5 Multiple regression equivalence of canonical correlation

One of the major drawbacks of canonical correlation is the difficulty associated with interpreting its results. As a result, regression equivalences have been formulated to provide more information. This section presents a generelisation of the regression formulation originated by Johansson & Sheth (1974). Consider $\widetilde{\mathbf{Y}}$ as containing p sets of criterion (outcome) or dependent variables, and $\mathbf{Y}$ as comprised of q sets of explanatory (independent) variables or predictors. Additionally, suppose that the sets of variables in both $\widetilde{\mathbf{Y}}$ and $\mathbf{Y}$ are standardised with zero mean and variance equal to 1. The canonical variates u_i and v_i are a linear combination of standardised variables, hence they also have zero mean. In addition, they have an imposed variance of 1. It, therefore, follows that they are also standardised with mean zero and variance 1. It is well known that the regression coefficient resulting from regressing two standardised variables on each other is just the simple correlation coefficient between them (Johansson & Sheth, 1974). Thus, $\mathbf{u}_i = C_i\mathbf{v}_i$ where C_i is the i^{th} canonical correlation. It follows that the correlations of the canonical variates above may be rewritten as a system of linear equations as follows:

$$\begin{aligned}
a_{11}\widetilde{Y}_1 + a_{21}\widetilde{Y}_2 + \ldots + a_{p1}\widetilde{Y}_p &= C_1\left(b_{11}Y_1 + b_{21}Y_2 + \ldots + b_{q1}Y_q\right) + \epsilon_1 \\
a_{12}\widetilde{Y}_1 + a_{22}\widetilde{Y}_2 + \ldots + a_{p2}\widetilde{Y}_p &= C_2\left(b_{12}Y_1 + b_{22}Y_2 + \ldots + b_{q2}Y_q\right) + \epsilon_2 \\
&\vdots \\
a_{1k}\widetilde{Y}_1 + a_{2k}\widetilde{Y}_2 + \ldots + a_{pk}\widetilde{Y}_p &= C_k\left(b_{1k}Y_1 + b_{2k}Y_2 + \ldots + b_{qk}Y_q\right) + \epsilon_k.
\end{aligned} \tag{5.29}$$

Now if we define the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{Y}$, $\mathbf{Y}$ and ϵ are defined as

$$\mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{p1} \\ a_{12} & a_{22} & \dots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1k} & a_{2k} & \dots & a_{pk} \end{pmatrix}, \mathbf{B}^T = \begin{pmatrix} b_{11} & b_{21} & \dots & b_{q1} \\ b_{12} & b_{22} & \dots & b_{q2} \\ \vdots & \vdots & \vdots & \vdots \\ b_{1k} & b_{2k} & \dots & b_{qk} \end{pmatrix}, \mathbf{C} = \begin{pmatrix} C_1 & 0 & \dots & 0 \\ 0 & C_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & C_k \end{pmatrix}$$

$$\mathbf{Y}^T = \begin{pmatrix} \widetilde{Y}_1 \\ \widetilde{Y}_2 \\ \vdots \\ \widetilde{Y}_k \end{pmatrix}, \mathbf{Y}^T = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix}, \text{and } \epsilon^T = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k. \end{pmatrix} \tag{5.30}$$

then the system of linear equations in Equation 5.29 can be written in matrix form as

$$\widetilde{\mathbf{Y}}\mathbf{A} = \mathbf{YBC} + \epsilon, \tag{5.31}$$

Then Equation 5.32 can be written as a system of linear equations as follows:

$$\begin{aligned} \widetilde{Y}_1 = & \left(a^{11}C_1b_{11} + a^{12}C_2b_{12} + \dots + a^{1k}C_kb_{1k}\right)Y_1 + \left(a^{11}C_1b_{21} + a^{12}C_2b_{22} + \dots + a^{1k}C_kb_{2k}\right)Y_2 + \dots \\ & + \left(a^{11}C_1b_{q1} + a^{12}C_2b_{q2} + \dots + a^{1k}C_kb_{qk}\right)Y_q + e_1 \\ \widetilde{Y}_2 = & \left(a^{21}C_1b_{11} + a^{22}C_2b_{12} + \dots + a^{2k}C_kb_{1k}\right)Y_1 + \left(a^{21}C_1b_{21} + a^{22}C_2b_{22} + \dots + a^{2k}C_kb_{2k}\right)Y_2 + \dots \\ & + \left(a^{21}C_1b_{q1} + a^{22}C_2b_{q2} + \dots + a^{2k}C_kb_{qk}\right)Y_q + e_2 \\ \vdots = & \vdots \\ \widetilde{Y}_p = & \left(a^{p1}C_1b_{11} + a^{p2}C_2b_{12} + \dots + a^{pk}C_kb_{1k}\right)Y_1 + \left(a^{p1}C_1b_{21} + a^{p2}C_2b_{22} + \dots + a^{pk}C_kb_{2k}\right)Y_2 + \dots \\ & + \left(a^{p1}C_1b_{q1} + a^{p2}C_2b_{q2} + \dots + a^{pk}C_kb_{qk}\right)Y_q + e_k, \end{aligned} \tag{5.32}$$

where $e_i = \sum_{j=1}^{n} a^{ij}\epsilon_j$. The coefficients of the independent variables in Equation 5.34, which are in the form of summed products, measure the weight each Y_i

contributes to the canonical variates and the associated canonical root. Due to the fact that the Y_i's and the $\widetilde{Y}_i$'s are standardised, it follows that these coefficients are equivalent to the standardised beta coefficients of multivariate linear regression (Johansson & Sheth, 1974). The criterion variables, $\widetilde{Y}_i$'s, in this system of linear equations, may also be written as a function of the canonical variates composed of the predictor variables, Y_i's, as follows:

$$\begin{aligned}
\widetilde{Y}_1 &= a^{11}C_1\left(b_{11}Y_1 + b_{21}Y_2 + \dots + b_{q1}Y_q\right) + a^{12}C_2\left(b_{12}Y_1 + b_{22}Y_2 + \dots + b_{q2}Y_q\right) + \dots \\
&\quad + a^{1k}C_k\left(b_{1k}Y_1 + b_{2k}Y_2 + \dots + b_{qk}Y_q\right) + e_1 \\
\widetilde{Y}_2 &= a^{21}C_1\left(b_{11}Y_1 + b_{21}Y_2 + \dots + b_{q1}Y_q\right) + a^{22}C_2\left(b_{12}Y_1 + b_{22}Y_2 + \dots + b_{q2}Y_q\right) + \dots \\
&\quad + a^{2k}C_k\left(b_{1k}Y_1 + b_{2k}Y_2 + \dots + b_{qk}Y_q\right) + e_2 \\
\vdots &= \vdots \\
\widetilde{Y}_p &= a^{p1}C_1\left(b_{11}Y_1 + b_{21}Y_2 + \dots + b_{q1}Y_q\right) + a^{p2}C_2\left(b_{12}Y_1 + b_{22}Y_2 + \dots + b_{q2}Y_q\right) + \dots \\
&\quad + a^{pk}C_k\left(b_{1k}Y_1 + b_{2k}Y_2 + \dots + b_{qk}Y_q\right) + e_k,
\end{aligned} \tag{5.33}$$

which is similar to say

$$\begin{aligned}
\widetilde{Y}_1 &= \left(a^{11}C_1\right)\mathbf{v}_1 + \left(a^{12}C_2\right)\mathbf{v}_2 + \dots + \left(a^{1k}C_k\right)\mathbf{v}_k + e_1 \\
\widetilde{Y}_2 &= \left(a^{21}C_1\right)\mathbf{v}_1 + \left(a^{22}C_2\right)\mathbf{v}_2 + \dots + \left(a^{2k}C_k\right)\mathbf{v}_k + e_2 \\
\vdots &= \vdots \\
\widetilde{Y}_p &= \left(a^{p1}C_1\right)\mathbf{v}_1 + \left(a^{p2}C_2\right)\mathbf{v}_2 + \dots + \left(a^{pk}C_k\right)\mathbf{v}_k + e_k.
\end{aligned} \tag{5.34}$$

The coefficients of the canonical variates involving the predictor variables in Equation 5.36 measure the partial correlation between the $\widetilde{Y}_i$'s and each of the canonical variates v_i. This is as a result of both the $\widetilde{Y}_i$'s and the v_i's being standardised with mean zero and variance of 1. Additionally, Johansson & Sheth (1974) notes that these partial correlations are actually simple correlations of first order between the $\widetilde{Y}_i$'s and the v_i's since each v_i is orthogonal to the canonical variate preceding it. Thus, Equation 5.36 can be written in terms of correlations, as follows:

$$
\begin{aligned}
\widetilde{Y_1} &= \left(\rho_{y_1,v_1}\right)\mathbf{v}_1 + \left(\rho_{y_1,v_2}\right)\mathbf{v}_2 + \ldots + \left(\rho_{y_1,v_k}\right)\mathbf{v}_k + e_1\\
\widetilde{Y_2} &= \left(\rho_{y_2,v_1}\right)\mathbf{v}_1 + \left(\rho_{y_2,v_2}\right)\mathbf{v}_2 + \ldots + \left(\rho_{y_2,v_k}\right)\mathbf{v}_k + e_2\\
\vdots &= \vdots\\
\widetilde{Y_p} &= \left(\rho_{y_p,v_1}\right)\mathbf{v}_1 + \left(\rho_{y_p,v_2}\right)\mathbf{v}_2 + \ldots + \left(\rho_{y_p,v_k}\right)\mathbf{v}_k + e_k.
\end{aligned}
\tag{5.35}
$$

5.2.6 Using CCA to extend Moran's index to multivariate case

Case 1: Univariate Moran's index

In the univariate Moran's index case, consider the predictor matrix $\mathbf{Y}$ to contain one variable $\mathbf{Y}_1$, while the criterion matrix has only one variable $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_1$, the lag values of $\mathbf{Y}_1$. In this case, the canonical correlation problem reduces to a regression equation between $\mathbf{Y}_1$ and $\tilde{\mathbf{Y}}_1$. Since both $\mathbf{Y}_1$ and $\tilde{\mathbf{Y}}_1$ are standardised, then the standardised regression equation is given by:

$$\widetilde{\mathbf{Y}_1} = \rho_{y_1,\tilde{y}_1}\mathbf{Y}_1, \tag{5.36}$$

where $\rho_{y_1,\tilde{y}_1}$ is a standardised coefficient. The Moran's index will be the unstandardised coefficient of Equation 5.38 and is given by:

$$I_{y_1} = \frac{SD(\widetilde{\mathbf{Y}_1})}{SD(\mathbf{Y}_1)} \times \rho_{y_1,\tilde{y}_1}. \tag{5.37}$$

Case 2: Bivariate Moran's index

The bivariate Moran's index case considers the predictor matrix $\mathbf{Y} = \mathbf{Y}_2$, while the criterion matrix has only one variable $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_1$, the lag values of $\mathbf{Y}_1$. In this case the canonical correlation problem reduces to a regression equation between $\mathbf{Y}_2$ and $\tilde{\mathbf{Y}}_1$. Since both $\mathbf{Y}_2$ and $\tilde{\mathbf{Y}}_1$ are standardised, then the standardised regression equation is given by:

$$\widetilde{\mathbf{Y}}_1 = \rho_{y_2,\tilde{y}_1}\mathbf{Y}_2, \tag{5.38}$$

where $\rho_{y_2,\tilde{y}_1}$ is a standardised coefficient. The bivariate Moran's index will be the unstandardised coefficient of Equation 5.40, and is given by:

$$I_{y_1,y_2} = \frac{SD(\widetilde{\mathbf{Y}_1})}{SD(\mathbf{Y}_1)} \times \rho_{y_2,\tilde{y}_1}. \tag{5.39}$$

Case 3: Extension to mutivariate Moran's index

Consider the equation, for example,

$$\mathbf{Y}_i = \left(\rho_{y_i,v_1}\right)\mathbf{v}_1 + \left(\rho_{y_i,v_2}\right)\mathbf{v}_2 + \ldots + \left(\rho_{y_i,v_k}\right)\mathbf{v}_k + e_i. \tag{5.40}$$

Pre-multiplying Equation 5.42 by v_1^T and using the fact that $v_i^T v_j = 0 \; \forall i \neq j$ due to orthogonality, we get:

$$v_1^T\mathbf{Y}_i = v_1^T\left(\rho_{y_i,v_1}\right)\mathbf{v}_1 + v_1^T e_i. \tag{5.41}$$

Since

$$v_1^T\mathbf{Y}_i = v_1^T\left(\rho_{y_i,v_1}\mathbf{v}_1 + e_i\right), \tag{5.42}$$

it follows that:

$$\mathbf{Y}_i = \left(\rho_{\widetilde{y}_i,v_1}\right)\mathbf{v}_1 + e_i. \tag{5.43}$$

The first canonical is the most important and contributes the most variation in the predictor variable and can be considered for analysis. Since both $\mathbf{Y}_i$ and v_1 are standardised, then the multivariate Moran's index will be the unstandardised coefficient of Equation 5.45, and is given by:

$$I_{y_i,v_1} = \frac{SD(\widetilde{\mathbf{Y}_i})}{SD(\mathbf{Y}_i)} \times \rho_{\widetilde{y}_i,v_1}. \tag{5.44}$$

5.2.7 Multivariate LISA clusters

Having determined the multivariate global spatial autocorrelation the next step is to determine the multivariate LISA map. At this stage we will be having two variables: Y_i and the derived first canonical variate v_1. The bivariate LISA approach was then applied to these two variables (Y_i and v_1.) to determine the multivariate spatial clusters.

5.3 Hypothetical example

For illustrative purposes, hypothetical spatial data were generated on a spatial space comprising of 61 hexagons. This is along the same lines as the experiment by Lee (2001) who generated spatial data from 37 hexagons. A variable Y_1 was generated as a spatial random normal variable using a variance-covariance structure based on the Euclidean distance between the centroids of the hexagons. A variable Y_2 which is positively correlated to Y_1 ($r = 0.564$), was derived as a normal variable conditional on Y_1. Another variable, Y_3, was also derived as a normal variable conditional on both Y_1 and Y_2, with a positive correlation of r = 0.550 between Y_1 and Y_3. Values of 5, 3 and 1 were then assigned to the upper third, second third, and lower third of the corresponding quantile map for each respective variable. The variables Y_1, Y_2 and Y_3 are spatially dependent on each other. Each of these variables can be used as a criterion variable and the other two as predictor variables to establish positive spatial association for the three variables. Two spatially dispersed variables, Y_4 and Y_5 were also generated so that they both have negative correlation with Y_1. These two variables were employed to establish joint negative association between them and variables Y_1, Y_2 and Y_3. The spatial distribution of the variables Y_1, Y_2, Y_3, Y_4 and Y_5 is shown in Figure 5.1.

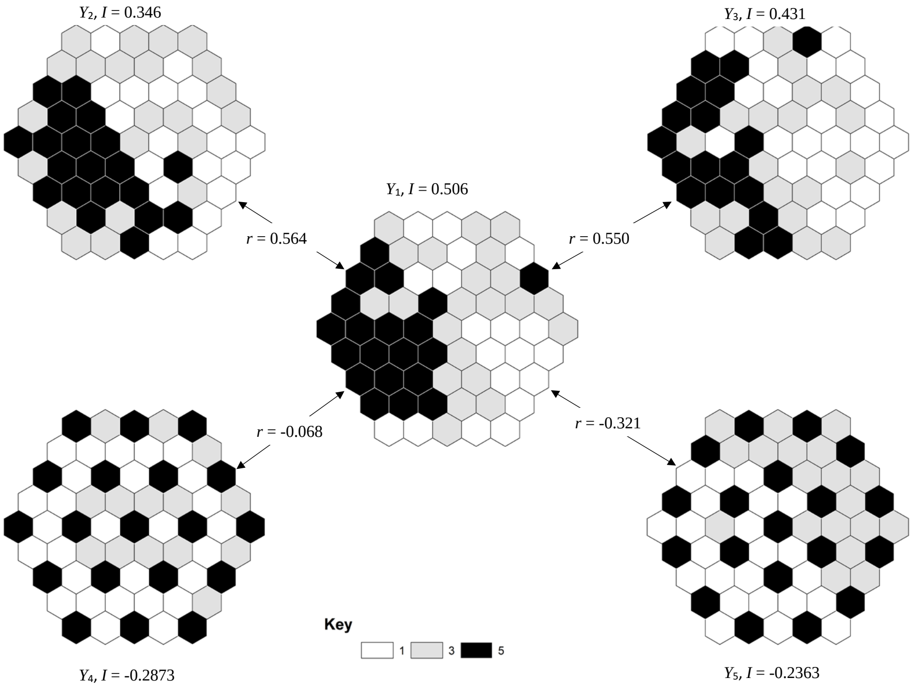


Figure 5.1: Hypothetical spatial data.

5.3.1 Verification of univariate and bivariate spatial autocorrelation results

In this section we show that the results of the new canonical approach give the same results as the original univariate and bivariate Moran's I. For univariate analysis using the canonical approach the criterion variable, $\widetilde{\mathbf{Y}} = [\widetilde{Y}_1]^T$, is a 1×1 random variable consisting of only one sub-vector $\widetilde{Y}_1$ the spatial lag variable, and the predictor variable $\mathbf{Y} = [Y_1]^T$ is a 1×1 random variable consisting of one sub-vectors Y_1, where $p = q = 1$. Similarly, in the case of bivariate analysis using the canonical approach the criterion variable,$\widetilde{\mathbf{Y}} = [\widetilde{Y}_1]^T$, is a 1×1 random variable consisting of only one sub-vector Y_1 the spatial-lagged variable of the criterion variable, and the predictor variable $\mathbf{Y} = [Y_2]^T$ is a 1×1 random variable consisting of one sub-vectors Y_2, where $p = q = 1$. The univariate analysis was done for the five variables Y_1, Y_2, Y_3, Y_4 and Y_5 based on Equation 5.7 using the R code given in Appendix B.1. Note that in this analysis $a^{11} = 1$ and the canonical correlation is simply equal to the correlation between the criterion and predictor variables. Table 5.1 provides the analyses results and a comparison with results of the original univariate Moran's index.

Table 5.1: Univariate Moran's index using the canonical approach.

Variable	Univariate Moran's I	Correlation	Standard deviations		SD ratio	New approach Moran's I
			SD1	SD2	SD1/SD2	MMI
Y1	0.5059**	0.7021	1.1615	1.6121	0.7205	0.5059**
Y2	0.3455**	0.4992	1.1157	1.6121	0.6920	0.3454**
Y3	0.4313**	0.6378	1.1234	1.6615	0.6762	0.4313**
Y4	-0.2873**	-0.7500	0.6683	1.7449	0.3830	-0.2873**
Y5	-0.2363**	-0.4274	0.9922	1.6681	0.5528	-0.2363**

Key: SD1, Standard deviation of spatial-lagged values of unstandardised criterion variable; SD2, Standard deviation of unstandardised criterion variable; MMI, Multivariate Moran's index.

Table 5.1 gives the variables whose spatial patterns are to be determined, the original univariate Moran's index, the standard deviation of the spatial-lagged values (SD1) of the variable and the standard deviation of the variable itself (SD2), the ratio of SD1 to SD2 (SD ratio) and the Moran's I from the canonical correlation approach (MMI). The canonical approach confirms the results of the original Moran's I. There is evidence of spatial clustering as indicated by significant positive Moran's I values for Y_1, Y_2 and Y_3, while Y_4 and Y_5 are significantly spatially dispersed with significant negative MMI values. The univariate clusters of Y_1, Y_2 and Y_3 are shown in Figures 5.2 A, B and C, respectively. Clusters are found in the western part of the polygon for all the three variables.

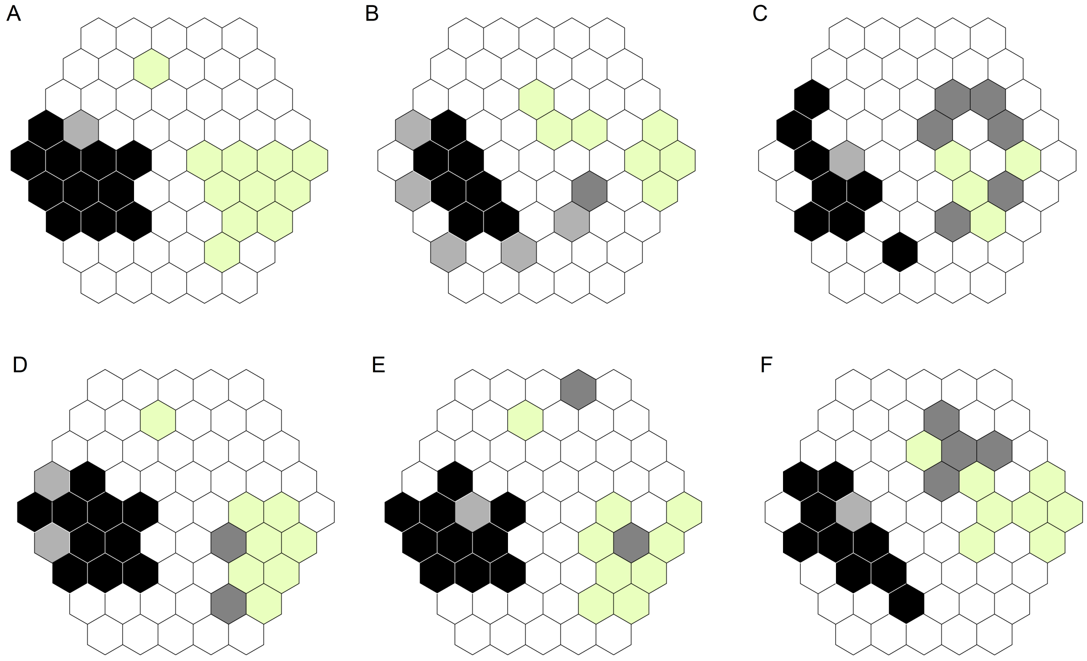


Figure 5.2: Univariate and bivariate LISA clusters for the hypothetical data.

Bivariate analysis was done to determine the pairwise association of the five variables Y_1, Y_2, Y_3, Y_4 and Y_5 based on Equation 5.41 using the same R code given in Appendix B.1. Table 5.2 provides the analysis results. The results are identical for the original Moran's I and the new canonical correlation approach.

There is evidence of significant positive spatial association between the following: Y_1 and Y_2; Y_1 and Y_3; Y_2 and Y_3; and Y_4 and Y_5. The bivariate joint clusters for Y_1 and Y_2, Y_1 and Y_3, and Y_2 and Y_3 are shown in Figures 5.2 D, E and F, respectively. The joint clusters in Figure 5.2 D-F and the univariate

Table 5.2: Bivariate spatial autocorrelation derived from the original Moran's index and the new multivariate approach.

Association	Bivariate Moran's I	**Correlation**	**Standard deviations**		**SD ratio**	**New approach Moran's** I
			SD1	**SD2**	**SD1/SD2**	**MMI**
Y1 - Y2	0.4524**	0.6279	1.1615	1.6151	0.7205	0.4524**
Y1 - Y3	0.4546**	0.6309	1.1615	1.6151	0.7205	0.4546**
Y1 - Y4	-0.0686	-0.0953	1.1615	1.6151	0.7205	-0.0686
Y1 - Y5	-0.1272**	-0.1765	1.1615	1.6151	0.7205	-0.1272**
Y2 - Y3	0.4695**	0.6784	1.1157	1.6116	0.6920	0.4695**
Y2 - Y4	-0.1214**	-0.1754	1.1157	1.6116	0.6920	-0.1214**
Y2 - Y5	-0.1579**	-0.2281	1.1157	1.6116	0.6920	-0.1579**
Y3 - Y4	-0.0478	-0.0709	1.1234	1.6615	0.6762	-0.0480
Y3 - Y5	-0.1533**	-0.2267	1.1234	1.6615	0.6762	-0.1533**
Y4 - Y5	0.1714**	0.4475	0.6683	1.7449	0.3830	0.1714**

clusters in Figures A-C, are found in the western parts of the hexagon.

5.3.2 Multivariate spatial autocorrelation analysis

In this section, we extend the univariate and bivariate Moran's analysis to multivariate (three or more variables) analysis using the canonical approach. The criterion vector $\mathbf{Y} = [\widetilde{Y}_i, \widetilde{Y}_{ic}]^T$ is a $2{\times}1$ random variable consisting of sub-vectors $\widetilde{Y}_i$ and $\widetilde{Y}_{ic}$ with the latter being a variable derived conditional on the distribution of the former. Assume that Z_1 and Z_2 are two independent standard normal random variables and define

$$\widetilde{Y}_i = Z_1 \tag{5.45}$$

$$\widetilde{Y}_{ic} = rZ_1 + Z_2\sqrt{1 - r^2} \tag{5.46}$$

where r is the specified correlation between variables $\widetilde{Y}_i$ and $\widetilde{Y}_{ic}$. Then, $\widetilde{Y}_i$ and $\widetilde{Y}_{ic}$ are assumed to be bivariate normal. These two standardised variables constituted the criterion vector in the analysis for each given criterion variable Y_i, where i takes values 1, 2 and 3. Using $\widetilde{Y}_i$ and $\widetilde{Y}_{ic}$ ensures that the two criterion variables are positively spatially associated and both variables will have similar associations with the predictor variables.

The predictor variable given by $\mathbf{Y} = [Y_j, Y_l]^T$ is a 2×1 random variable consisting of standardised sub-vectors Y_j and Y_l where $p = q = 2$. The indexes j and l take values 1, 2, 3, 4 and 5, with $i \neq j \neq l$. Only positively associated variables are to be used as predictor variables. Negatively associated variables were also used to determine if the multivariate method can detect spatial dispersion. Table 5.3 shows the canonical correlation analysis for different combinations of the hypothetical data. An $r = 0.6$ was used to determine the conditional variable to be included as a second criterion variable as per Equation 5.48.

Table 5.3: Summary results for the canonical correlation analysis for the hypothetical spatial data.

Y_1	Y_2	Y_3	**Canonical Variate Pair**	**Canonical correlation**	**Wilks Lambda**	**F**	**DF1**	**DF2**	**P-value**
Y_1	Y_2	Y_3	u_1v_1	0.7443	0.4349	14.7119	4	114	<0.001
			u_2v_2	0.1563	0.9756	1.4522	1	58	0.2331
Y_2	Y_1	Y_3	u_1v_1	0.7543	0.4219	15.3766	4	114	<0.001
			u_2v_2	0.1449	0.9790	1.2441	1	58	0.2698
Y_3	Y_1	Y_2	u_1v_1	0.7515	0.4352	14.7029	4	114	<0.001
			u_2v_2	0.0112	0.9999	0.0007	1	58	0.9325
Y_1	Y_4	Y_5	u_1v_1	0.2242	0.9497	0.7445	4	114	0.5636
			u_2v_2	0.0029	0.9999	0.0005	1	58	0.9827
Y_2	Y_4	Y_5	u_1v_1	0.3142	0.9012	1.5212	4	114	0.2006
			u_2v_2	0.0107	0.9999	0.0066	1	58	0.9356
Y_3	Y_4	Y_5	u_1v_1	0.2579	0.9331	1.0036	4	114	0.4088
			u_2v_2	0.0190	0.9996	0.0209	1	58	0.8856

The Wilks Lambda test was used to determine the significance of the canonical correlations for each canonical variate pair. It can be seen in Table 5.3 that the correlation of the first canonical pair (u_1v_1) is statistically significant for the combinations involving the three variables that have positive bivariate spatial association, namely, Y_1, Y_2 and Y_3, while the second canonical pairs (u_2v_2) are not significant for all the combinations considered in Table 5.3. When Y_1, Y_2 and Y_3 are each analysed with Y_4 and Y_5, both the first and second canonical variate pairs are not statistically significant. We then used Equation 5.45 to determine the multivariate spatial autocorrelation.

Table 5.4 shows the multivariate Moran's index calculated for an arbitrarily specified r value of 0.6 using the R code in Appendix B.2. Indications in Table 5.4 are that the variables Y_1, Y_2 and Y_3 have a shared spatial cluster among them with significant positive MMI values for the spatial association for the following: Y_1 and Y_2 – Y_3; Y_2 and Y_1 – Y_3; and Y_3 and Y_1 – Y_2.

Table 5.4: Multivariate spatial autocorrelation analysis of the hypothetical spatial data, *r = 0.6*.

r = 0.6

Criterion variable	**Predictor variables**	ρ_{y_1,v_1}	**Standard deviations**			**Multivariate Moran's** I
			SD1	**SD2**	**SD ratio**	**MMI**
Y_1	Y_2 Y_3	0.6163	1.1615	1.6121	0.7205	0.4440**
Y_2	Y_1 Y_3	0.6311	1.1157	1.6121	0.6920	0.4367**
Y_3	Y_1 Y_2	0.7023	1.1234	1.6615	0.6762	0.4748**
Y_1	Y_4 Y_5	-0.1688	1.1615	1.6121	0.7205	-0.1216^{INS}
Y_2	Y_4 Y_5	-0.2740	1.1157	1.6121	0.6920	-0.1896**
Y_3	Y_4 Y_5	-0.2026	1.1234	1.6615	0.6762	-0.1370^{INS}

The choice of correlation, r, to use in deriving the conditional variable to employ as one of the criterion variables may have an influence on the technique's ability to detect spatial heterogeneity or homogeneity. Thus, the study investigated the effects different values of r will have on MMI and its significance. Shown in Tables B.1-B.5 of Appendix B are results obtained when r takes values 0.3, 0.5, 0.7 and 0.9. It can be seen that the correlation between y_1 and v_1 decreases with increases in r and so does the values of MMI, in turn. While the significance of higher values of MMI (>0.20) remains unchanged between the interval 0.0 to 1.0 for r, the significance of smaller absolute MMI values that are between 0.20 and 0.15 are likely to change within the given interval. In order to maximise detection of spatial heterogeneity, it is recommended to use a small r value, say between 0.1 and 0.6.

5.4 Chapter summary

A new multivariate spatial autocorrelation measure for detecting joint "hot-spots" for more than two outcomes that are spatially related has been developed and is based on canonical correlation. Using hypothetical data it was shown that the proposed statistics performed very well in detecting joint clusters.

Chapter 6

Simulation Study and Application of the Proposed Multivariate Spatial Clustering Statistics

6.1 Introduction

This chapter presents a simulation study to assess the performance of the method in detecting multivariate spatial autocorrelation local clusters as well as an application to cardiovascular mortality. A Monte Carlo simulation was conducted by randomising the simulated sub-vector $\widetilde{Y}_{ic}$ in the set of dependent variables as described in Equation 5.48 of the previous chapter. The local spatial clusters created by the randomised data were then compared with the local clusters prior to randomisation using agreement analysis by Bangdiwala

& Shankar (2013) as originally demonstrated in Chien *et al.* (2018).

An application to real life data is also made to the mortality rates data adjusted for covariates using Poisson regression as described in Chapter 3. The data used are mortality rates due to diabetes; and three cardiovascular conditions of ischaemic, cerebrovascular and hypertensive heart conditions in South Africa for the years 2001 and 2011.

6.2 Simulation study

In this study we hypothesised three spatially related variables X_1, X_2 and Y_1 such that they form a specified number of joint clusters. Different hypothetical data sets were chosen such that they form joint clusters of sizes 2,4,10,12, 18 and 20. Let $\mathbf{X} = [X_1, X_2]^T$ be a 2×1 set of "explanatory" random variables and $\mathbf{Y} = [\widetilde{Y}_1, \widetilde{Y}_2]^T$ be a 2×1 set of "response variables" where $\widetilde{Y}_1$ is the spatial lag of Y_1 and $p = q = 2$. In this case $\widetilde{Y}_2$ is derived as

$$\widetilde{Y}_2 = r\widetilde{Y}_1 + Z\sqrt{1 - r^2} \tag{6.1}$$

where r is a specified correlation between $\widetilde{Y}_1$ and $\widetilde{Y}_2$, while Z is a standard normal variable independent of $\widetilde{Y}_1$. It follows that $\widetilde{Y}_1$ and $\widetilde{Y}_2$ can be assumed to be bivariate normal.

An illustration is shown here for the hypothetical spatial data forming joint clusters of size 20 which are shown in Figures 6.1 A-C. In this case $Y1$ is used to derive Y while $X1$ and $X2$ form the X set of variables. Each hexagonal tessellation is comprised of 217 hexagon cells. Note that the assignment of values in Figure 6.1 B was done such that there are only 20 high values (assigned 5) and low values (assigned 1). In addition, the 20 cells of high values are made to coincide with the high value cells of the variables in Figures 6.1 A and C. This was done to ensure that the joint "hot-spots" involving $X2$ will be located in the cells of high values for $X2$.

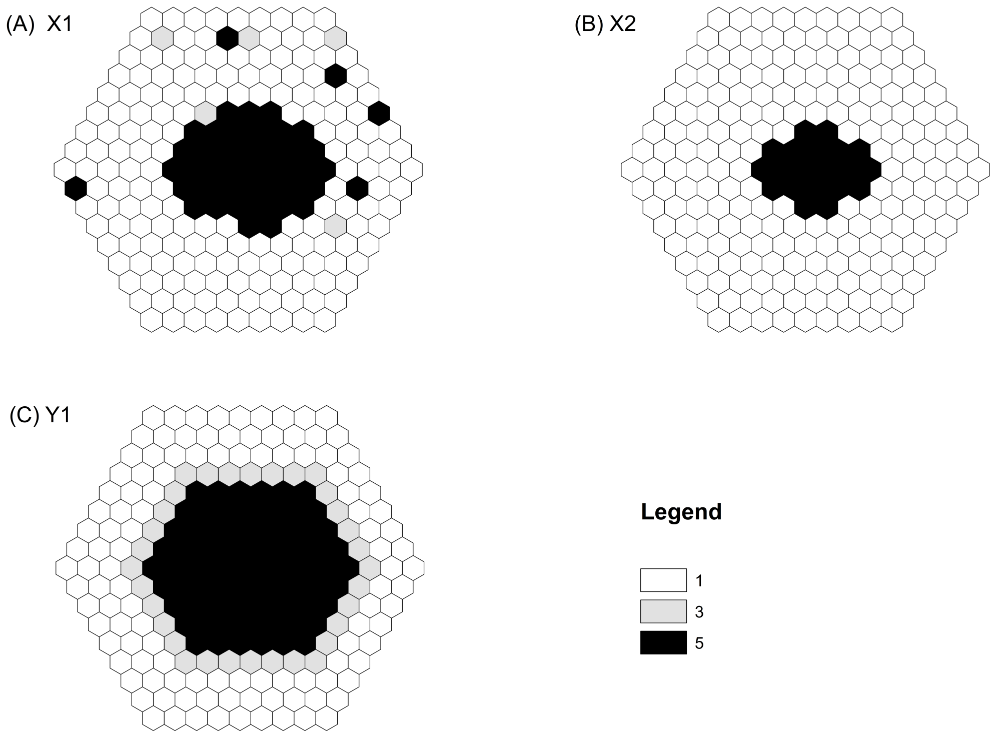


Figure 6.1: Unique value based spatial distribution of the hypothetical data.

The univariate LISA maps are shown in Figures 6.2 A-C for the spatial data of $Y1$, $X1$ and $X2$, respectively. Figures 6.2 D-F show the bivariate LISA maps between $Y1$ and $X1$, $Y1$ and $X2$ and between $X1$ and $X2$, respectively. It can be seen that the bivariate LISA patterns of $X2$ and each of $Y1$ and $X1$ have joint "hot-spots" of 20 cells similar to the univariate LISA "hot-spots" patterns for $X2$ in Figure 6.2 C.

The multivariate LISA map for the three variables are shown in Figure 6.3. It can be observed that the joint "hot-spots" for the three variables are the 20

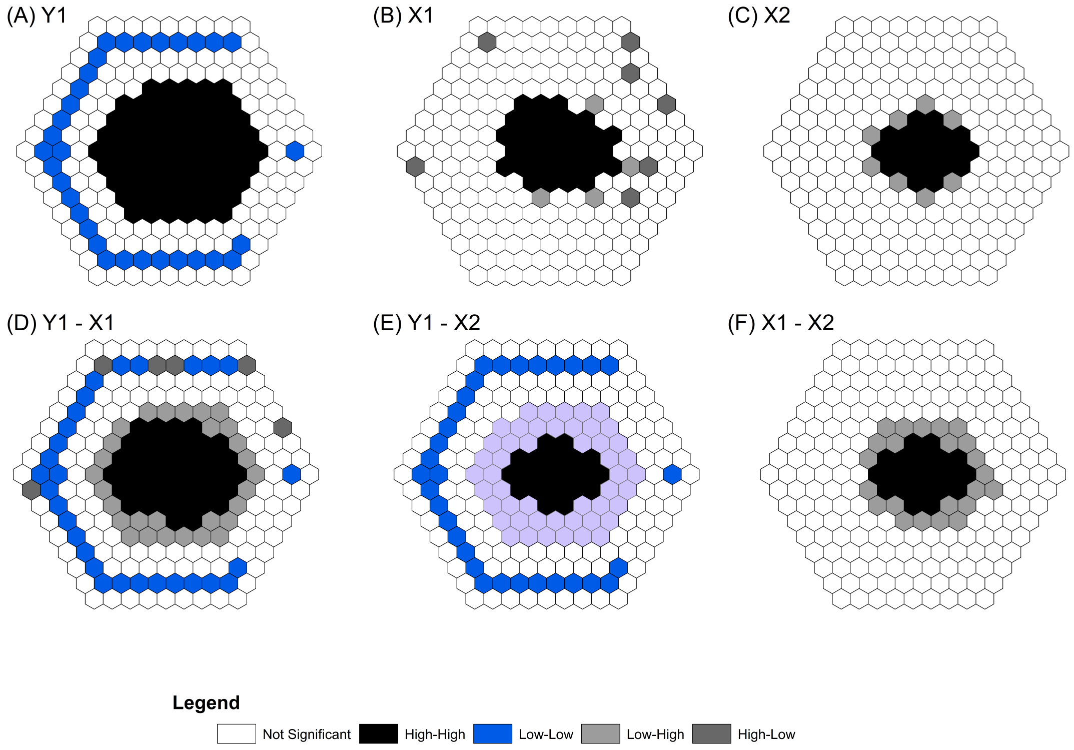


Figure 6.2: Univariate and bivariate LISA spatial patterns of the hypothetical data.

“hot-spots” similar to the univariate LISA “hot-spots” patterns for $X2$ in Figure 6.2 C. Thus, this multivariate spatial autocorrelation method is able to detect clusters attributed to three spatially related data.

Having established the ability of the method to detect multivariate spatial autocorrelation joint clusters we set out to determine if this can be done consistently through simulation. Categories cluster maps were compared with those obtained from simulated data after randomisation of the $\widetilde{Y}_2$ variable.

An example of the agreement table for the data with 20 “hot-spots” clusters and one simulation data is shown in Table 6.1. It can be observed that both the original data and the simulated data obtained the same cells of “High-High” “hot-spots”, but there are 12 cells categorised to be “Low-Low” by the original data which were classified to be “Not Significant” by the simulated data. This

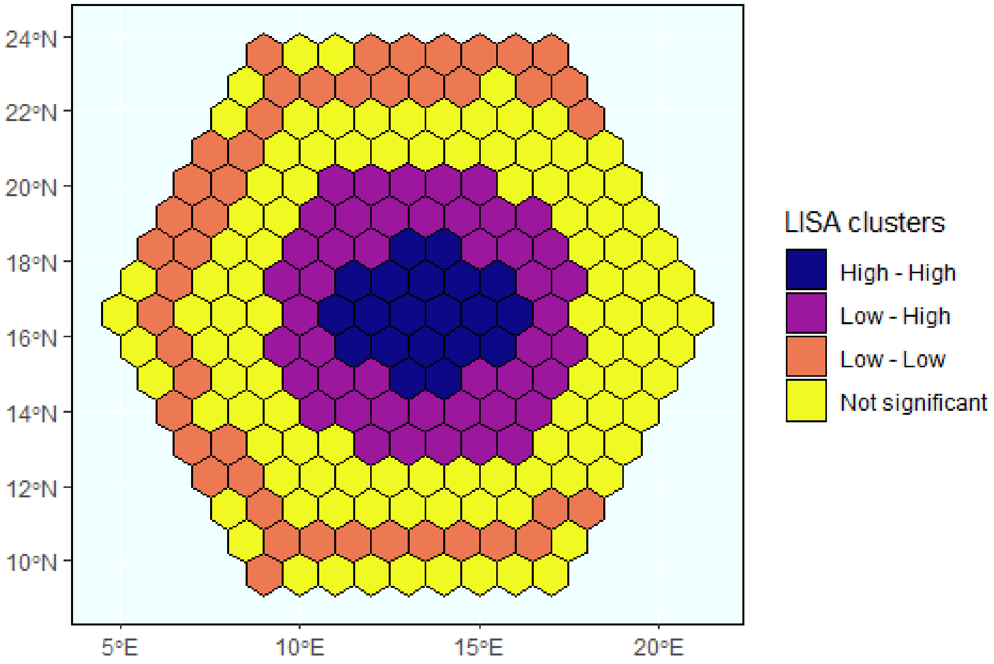


Figure 6.3: Multivariate LISA map for Y_1, X_1 and X_2 using the hypothetical data with 20 “hot-spots” clusters.

Table 6.1: Agreement table of the categories of multivariate joint clusters and simulated data when joint clusters equals to 20.

	High-High	High - Low	Low-High	Low-Low	Not Significant
High-High	20	0	0	0	0
High - Low	0	0	0	0	0
Low-High	0	0	44	0	2
Low-Low	0	0	0	34	12
Not Significant	0	0	1	16	88

agreement analysis had a weighted Bangdiwala statistic of 0.956. It shows that there is an excellent level of agreement between the multivariate LISA maps of the original and the simulated data as suggested by the interpretations in Munoz & Bangdiwala (1997).

Additionally, the simulation was done by changing the size of the joint clusters. This was done by specifying joint clusters of size 2, 4, 10, 12, 18 and 20. A total of 100 simulations were done for each set of hypothetical data by randomising $\widetilde{Y}_2$ variable and summary statistics calculated for the 100 weighted Bangdiwala statistics. The results are shown in Table 6.2.

Table 6.2: Simulation results

Number of "hot-spots"	Minimum	Median	Mean	Maximum
2	0.6653	0.7392	0.7419	0.8319
4	0.6532	0.7341	0.7335	0.8351
10	0.6632	0.7342	0.7348	0.8400
12	0.6455	0.7251	0.7453	0.9860
18	0.6592	0.9464	0.9373	0.9860
20	0.9202	0.9507	0.9490	0.9861

It can be seen in Table 6.2 that the mean level of agreement generally increases as the number of "hot-spots" increases. The mean agreement level ranges from 0.7335 (for 4 "hot-spots") to 0.9490 (for 20 "hot-spots"). In the case of data for two "hot-spots", the global joint clustering was not significant, but still the level of agreement can be described as good.

6.3 Application to cardiovascular mortality

The new multivariate Moran's index is hereby applied to Poisson regression estimated mortality rates data of Chapter 3. Diabetes and hypertensive heart diseases are known to be risk factors for other cardiovascular conditions. In this section, we illustrate the use of the new method by investigating how the average prevalence of both diabetes and hypertension mortality in a municipality may influence the prevalence of cerebrovascular mortality and ischaemic mortality in the neighbouring municipalities in South Africa. Table 6.3 reproduces the bivariate associations for the adjusted rates of the hypertensive heart disease (HHD), diabetes (DBT), ischaemic heart disease (IHD) and cerebrovascular heart disease (CVA) obtained using the original Moran's index in Chapter 3.

Cerebrovascular mortality is significantly spatially associated with both

Table 6.3: Bivariate spatial autocorrelation derived from the original Moran's index for Poisson regression adjusted cardiovascular mortality rates.

Association (X-Y)		**Moran's index**	**Significance**
X	**Y**	I	**Yes or No**
CVA01	HHD01	0,279	Yes
CVA01	DBT01	0,432	Yes
CVA11	HHD11	0,294	Yes
CVA11	DBT11	0,196	Yes
IHD01	HHD01	-0.121	No
IHD01	DBT01	0,719	Yes
IHD11	HHD11	-0.110	No
IHD11	DBT11	0,289	Yes

HHD and DBT for the years 2001 and 2011. But, the bivariate spatial association between ischaemic heart disease mortality rates and hypertensive heart disease is not significant with the Moran's index value of -0.121 and -0.110 for the years 2001 and 2011, respectively. So the extension to multivariate spatial association involving three variables will only be applied to the association between CVA, HHD and DBT. Two variable sets were used in this analysis. Firstly, the outcome criterion variable set, $\mathbf{Y} = [\widetilde{Y}_1, \widetilde{Y}_{1c}]^T$ is a 2×1 random variable consisting of the spatial-lagged values of CVA, $\widetilde{Y}_1$, and $\widetilde{Y}_{1c}$ is a sub-vector derived conditional on the distribution of spatial-lagged values of CVA with $r = 0.4$. Secondly, the independent variable set $\mathbf{Y} = [Y_2, Y_3]^T$, comprising of HHD and DBT adjusted mortality rates. Since each variable set is comprised of two sub-vectors, it implies that $p = q = k = 2$. Therefore, two canonical correlations were estimated to describe the interrelationship between the two variable sets. The summary results for the canonical correlation analysis are shown in Table 6.2.

Table 6.4 shows that only the first canonical correlation, $C_1 = \rho_{u_1,v_1}$, is statistically significant, for both the 2001 and 2011 data. So in both years, the regression equation reduces to: $\widetilde{Y}_1 = a^{11}C_1v_1 + e_1 = \rho_{Y_1,v_1}v_1 + e_1$. The estimates for a^{11} were estimated and found to be 0.9614307 and 1.042656 for the years 2001 and 2011,

Table 6.4: Summary results for the canonical correlation analysis.

Y_1	Y_2	Y_3	Canonical Variate Pair	Canonical Correlation	Wilks Lambda	F	DF1	DF2	P-value
CVA01	HHD01	DBT01	u_1v_1	-0.6796	0.5381	41.7745	4	460	<0.001
			u_2v_2	-0.0023	0.9999	0.0013	1	231	0.972
CVA11	HHD11	DBT11	u_1v_1	0.4372	0.8085	12.8930	4	460	<0.001
			u_2v_2	0.0204	0.9996	0.0964	1	231	0.756

respectively. Then the correlation between Y_1 and v_1, ρ_{Y_1,v_1}, were calculated to be $0.9614307 \times -0.6796 = -0.6534$ and $1.042656 \times 0.4372 = 0.4559$ for the years 2001 and 2011, respectively. Table 6.5 shows the remainder of the calculations estimating the multivariate Moran's indexes for the two periods under study.

Table 6.5: Summary results for the multivariate Moran's index estimation procedure for cardiovascular mortality in South Africa.

Y_1	Y_2	Y_3	ρ_{Y_1,v_1}	$SD(\widetilde{\mathbf{Y}}_1)$	$SD(\mathbf{Y_1})$	$\frac{SD(\widetilde{\mathbf{Y}}_1)}{SD(Y_1)}$	MMI	**P-value**
CVA01	HHD01	DBT01	-0.6534	12.6083	18.9554	0.6652	-0.4346	<0.001
CVA11	HHD11	DBT11	0.4559	10.8782	17.3337	0.6267	0.2861	<0.001

It can be seen in Table 6.5 that the multivariate Moran's indexes are highly statistically significant, with p-value less than 0.001, for both the data from 2001 (MMI=-0.4346) and 2011 (MMI=0.2861). There is evidence of positive multivariate spatial association between cerebrovascular mortality and the combined mortality rates due to hypertensive heart disease and diabetes for the year 2011 but not for the year 2001 which shows spatial dispersion (negative MMI). If we consider Y_1 and v_1 as two variables, we can determine the LISA clusters using the bivariate Moran's index used in Chapter 4. The LISA maps for the univariate, bivariate and multivariate approach for the year 2001 are shown in Figure 6.4.

As was shown in Chapter 3, the univariate Moran's indexes for CVA01 ($I = 0.422$), HHD01 ($I = 0.445$) and DBT ($I = 0.684$) are positive and statistically

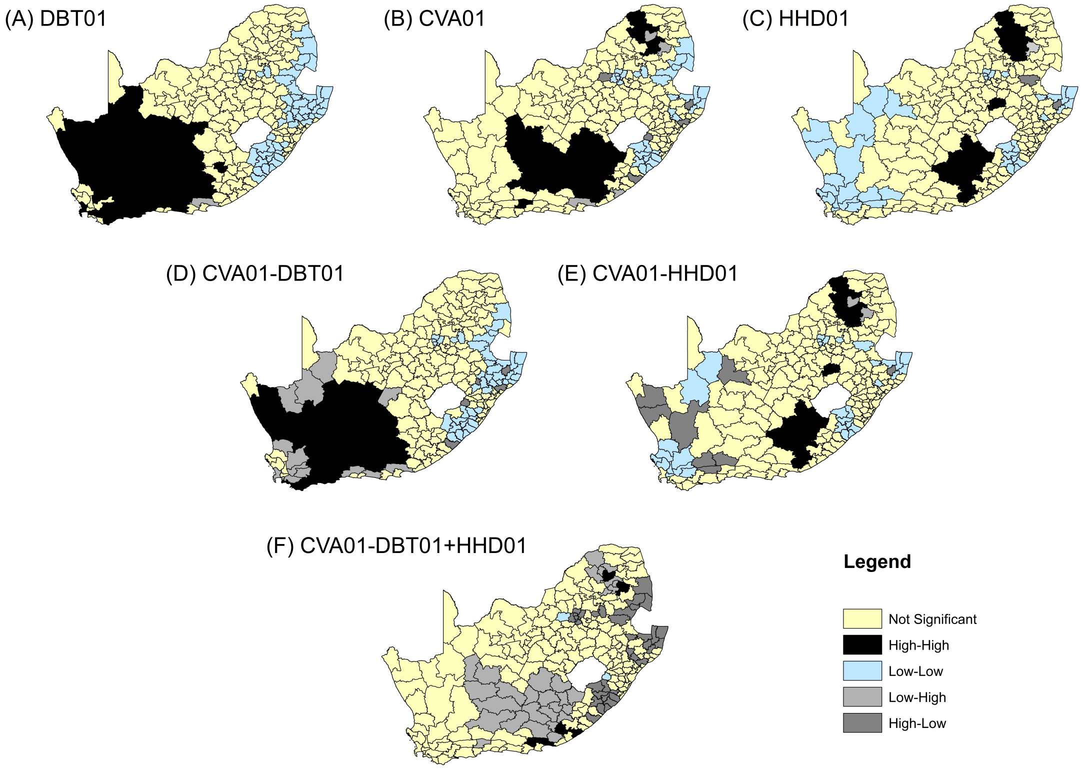


Figure 6.4: Univariate and bivariate LISA clusters for the hypothetical data.

significant. A mapping of the univariate LISA maps shows that DBT01 (Figure 6.4 A)and HHD01 (Figure 6.4 C) have only two "hot-spots" and a few "cold-spots" municipalities in common. The "cold-spots" of HHD01 in Figure 6.4 C are found in the western part of the country where the "hot-spots" municipalities for DBT01 are found in Figure 6.4 A. Some of the "hot-spots" for HHD01 are in the north-east of he country where we have "cold-spots" for diabetes. This helps to explain why the test for bivariate spatial dependency between CVA01 and DBT01 ($I = 0.057$) is not spatially significant in Table 6.3. There is significant positive bivariate spatial dependency between CVA01 and DBT01 and between CVA01 and HHD01 with the joint clusters are shown in Figures 6.4 D and E, respectively. However, there is a negative spatial association between CVA01 and weighted average of HHD01 and DBT01 ($I = -0.452$). This evidence of joint multivariate spatial dispersion is due to the differences in spatial distribution displayed by HHD01 and DBT01. Note that the bivariate Moran's index ($I = -0.452$) between Y_1 and v_1 is similar, if not identical, to MMI (-0.434) derived using the canonical approach. Discrepancies are due to random fluctuations since the canonical correlation between the canonical pair $u_1 v_1$ is not statistically significant.

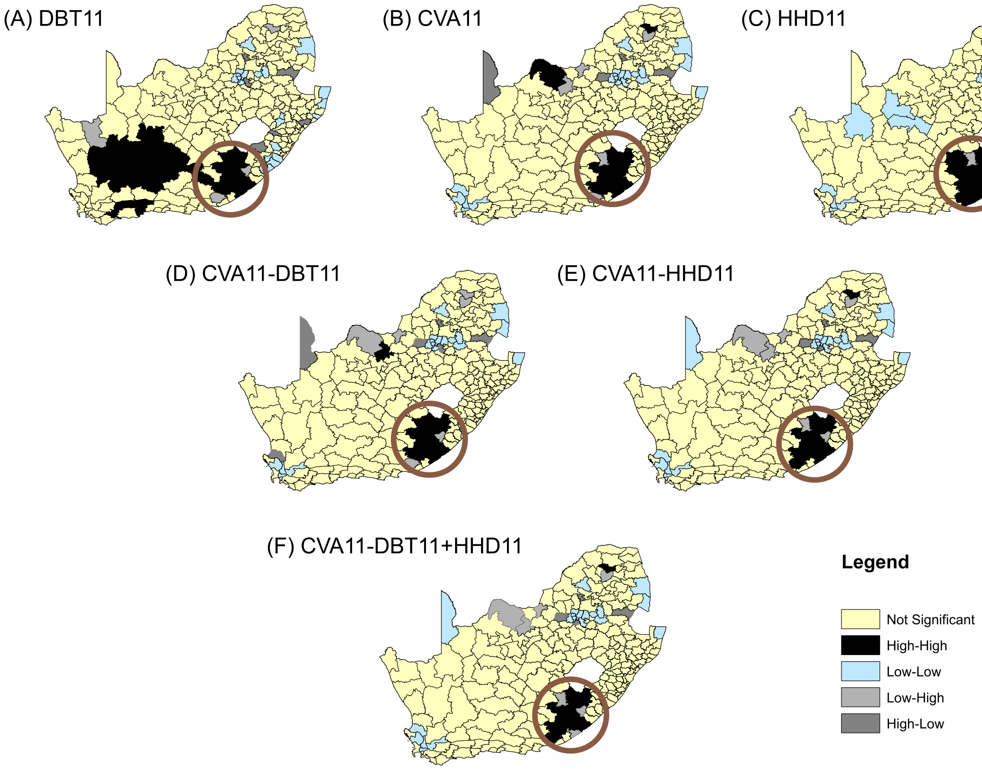


Figure 6.5: Univariate and bivariate LISA clusters for the hypothetical data.

The main difference between the results of the spatial data of 2001 and 2011 is that the spatial dependency between HHD and DBT is significant in the later year and not in the former year. A comparison of Figures 6.4 A-C and Figures 6.5 A-C shows that "cold-spots" of HHD mortality in the western part of the country and the "hot-spots" in the north-eastern part of the country has been disappearing over the 10 year period under review. The "hot-spots" clusters of DBT have also been reducing in the western part of the country and moving closer to the boundary of Lesotho (hollow patch on the map), while the DBT "cold-spots" have also reduced greatly on the eastern side of he country. There are more similar spatial patterns in the south and eastern part of the country for HHD11 and DBT11. This helps to explain why the bivariate dependency between the two is statistically significant. The bivariate spatial associations for all three possible combinations involving CVA11, HHD11 and DBT11 were statistically significant and the co-clusters are shown in Figures 6.5 D-F. It is clear that the circled joint "hot-spot" clusters are common to the individual univariate "hot-spot" clusters circled in Figures 6.4 A-C.

There is evidence that municipalities with low or high average mortality rates of HHD and DBT are surrounded by municipalities with low or high mortality rates due to cerebrovascular heart diseases for the year 2011, when Y_1 is regressed against v_1 with a significant bivariate Moran's index of 0.274.The bivariate Moran's index of 0.274 for Y_1 and v_1 is similar to the one obtained using the CCA approach (MMI=0.286), the difference being due to random fluctuations since the canonical correlation between the canonical pair u_1v_1 is not statistically significant. The multivariate joint clusters were validated using simulation approach described in Section 4.4 by randomising $\widetilde{Y}_{1c}$. A mean Bangdiwala weighted B-statistic of 99.87% shows that the method consistently detects the multivariate joint clusters with an excellent level of agreement.

6.4 Chapter summary

Simulation studies were done to determine the ability of the proposed multivariate spatial autocorrelation measure in detecting joint "hot-spots" for more than two outcomes that are spatially related. These studies were done for clusters of size two up to size 20. Simulation studies showed the proposed statistics performed very well in detecting joint clusters with agreement ranging from

73% for clusters of size 2 to 95% for clusters of size 20. It was shown that the level of agreement increased the more pronounced the spatial patterns.

The proposed method was applied to the spatial analysis of the possibly related mortality rates due to diabetes; and three cardiovascular conditions of ischaemic, cerebrovascular and hypertensive heart conditions in South Africa. There was significant multivariate spatial association between cerebrovascular and both hypertensive and diabetes.

Chapter 7

Discussion and conclusions

7.1 Introduction

The overall purpose of this study was to develop a single measure of multivariate spatial autocorrelation measures for detecting joint clustering for interrelated health outcomes and their risk factors. To achieve the objectives of the study spatial autocorrelation statistics were applied alongside other statistical methods for improving small area estimation of rates of health outcomes of cardiovascular conditions and their associated risk factors. This discussion will blend the findings of the thesis in such a way that it brings out the underlying concepts that led to the attainment of the main aim of this PhD study.

In this chapter, the discussion will begin with a summary of the main findings of the thesis. This will then be followed by limitations of the study and the contributions being made by this study towards the body of spatial statistics knowledge. Lastly, the direction of future studies will be presented before some concluding remarks are made.

7.2 Summary of the findings

A review of the spatial autocorrelation methods that are currently in use done in Chapters 2 and 4 showed that they can only analyse a maximum of two health outcomes simultaneously. The proposal by Wartenberg (1985) of extending the Moran's index beyond two variables is complex to the extent of making it practically impossible to implement. It has meant that the Moran's index has been restricted to analysis of not more than two variables. This is a gap in the literature of spatial statistics that this PhD study set to fill by developing a multivariate spatial autocorrelation measure that will cater for more than two health outcomes. Before developing the new measure, we investigated the feasibility of using the bivariate spatial autocorrelation methods in determining pairwise spatial associations and co-clustering patterns of CVD-related health outcomes in South Africa. This was done in Chapter 4.

The first application of bivariate spatial autocorrelation method in Chapter 4 was done to two cardiovascular diseases namely stroke and heart attack and three cardiovascular risk factors, namely tobacco smoking, hypertension and high blood cholesterol in South Africa. Globally, there was evidence of spatial dependency between stroke and smoking; stroke and high blood cholesterol; and between smoking and high blood cholesterol. This revealed that there is a tendency of nearby districts to have high or low joint stroke-smoking, stroke-high blood cholesterol and smoking-high blood cholesterol indexes of spatial autocorrelation. The study established joint local high-high clusters of stroke-smoking; stroke-high blood cholesterol; and smoking-high blood cholesterol in the urban districts in the western part of the country (City of Cape Town; Cape Winelands; Overberg and Eden). However, the same bivariate outcomes exhibited low-low clusters in rural north-western and east-southern parts of the country, respectively.

Thus, this study suggests that spatial clustering of CVDs and risk factors differ according to urbanisation or rurality locations, with urban districts having high-high district clusters and rural areas having low-low district clusters of CVDs and the risk factors. Differentials in urban and rural clustering of CVDs or their risk factors, based on the values of the rates, have been reported elsewhere (Penney *et al.*, 2014; Fabiyi & Garuba, 2015; Paquet *et al.*, 2016; Rajabi *et al.*, 2010).

In the more developed countries, for example, Sweden (Rajabi *et al.*, 2010) and Canada (Penney *et al.*, 2014), the "high-high" clustering areas, of CVD or their risk factors were found in rural areas while "low-low" clustering areas were found in urban areas. Thus, the process is more diffusing in rural areas in the western world, suggesting that risk factors such as physical inactivity, unhealthy dietary patterns and excessive alcohol drinking and smoking are yet under control or mitigated. The same processes could be driving high-high clustering in urban South Africa. For example, urban residents in South Africa take high fat and sugar content diets that are low in carbohydrates and fibres, while rural populations follow a traditional diet which is high in carbohydrates and fibre content, but low in fats and sugars (Steyn *et al.*, 2006).

Over the years, a transition from rural to urban life has seen the urban majority transiting to an urban life and diets (Manning *et al.*, 1974; Steyn *et al.*, 2006). Evidence has shown that a higher proportion of urban Black population with low economic status are heavily depended on fast food (Van Zyl *et al.*, 2010). Thus, dietary patterns and lifestyles may help to explain the disparities in the spatial co-clusters of CVDs and their risk factors across the districts in South Africa. There is a need for modification of the dietary patterns of the urban population in order to have adequate nutrient intake to prevent increased incidence of CVDs and their risk factors.

The presence of spatial clustering in CVDs and their risk factors has also been found in different countries such as Nigeria (Fabiyi & Garuba, 2015), Sweden (Rajabi *et al.*, 2010), France and Australia [(Paquet *et al.*, 2016), and the USA (Ford & Highfield, 2016). However, our modeling approach has allowed us to measure co-clustering of CVDs and risk factors. We have found that stroke, high blood cholesterol and smoking, co-cluster in space, which supports the notion that stroke, tobacco smoking and blood cholesterol are positively associated (Cappuccio & Miller, 2016; Ahmed *et al.*, 2019; Steyn *et al.*, 2006). Similar findings have been found in Ghana, where it was shown that raised cholesterol and smoking were associated with tobacco use (Vuvor *et al.*, 2016).

Our findings are generally in agreement with earlier studies in South Africa that have used spatial statistics methods to analyse CVDs and their risk factors. For example, Kandala *et al.* (2013), using a Bayesian geo-additive mixed model, found high levels of hypertension prevalence in north-central-western parts of the country and low prevalence in the north-eastern part of the country. Wandai *et al.* (2019) also found significantly above average prevalence of hypertension in the districts of the north-central-western parts of the country, as also revealed by this study.

In the second application in chapter 4, an examination of bivariate spatial pairwise co-clusters of cardiovascular mortality at local municipality level was done using Empirical Bayesian spatial and Poisson regression models. Cause of death data for the year 2001 and 2011 were used for this study. Mortality rates due to cerebrovascular heart disease, ischaemic heart disease, hypertensive heart disease and diabetes were analysed for the years 2001 and 2011. There were four objectives in this study. The first objective was to show how the

Empirical Bayesian approach and the Poisson regression approach may be applied to give reliable estimates of cardiovascular mortality rates. The second objective was to determine local municipalities of high and low CVD mortality risk as well as spatial dependence between two CVD mortality rates. In the case of the Poisson regression approach, the model further furnishes us with possible risk factors for CVD mortality that are measured at local municipality-level. The third objective was to detect bivariate spatial dependence between CVD rates for a given CVD at two time points (2001 and 2011). For a given period, the hypothesis being tested was that high risk of mortality of one CVD in a given municipality is associated with high risk of mortality of a related disease in the neighbouring municipalities. In other words, the hypothesis being tested here is that interrelated diseases co-cluster. At two-time points, the hypothesis being tested is that the high risk of mortality of a given CVD in the year 2001 in a given municipality is associated with high risk of mortality of the same CVD in the neighbouring municipalities of the given municipality in the year 2011. The co-clusters obtained were visually compared with those obtained using raw-rates. The fourth objective was to compare the results of three bivariate spatial autocorrelation methods in terms of detection of CVD mortality spatial clusters. These methods are the established bivariate Moran's index, the recently developed Lee's index and a variant of the Moran's index.

The three bivariate spatial autocorrelation techniques of the original Moran's I, the recently developed Lee's L and the Moran's I variant by Dray gave almost similar results in this study. These methods can be used to complement each other. After adjusting for known municipal-level covariates of age, race and poverty, the Poisson regression approach managed to detect co-clustering where the Empirical Bayes smoothed rates and raw rates could not. The bivariate analyses for CVDs between two time-points were significant for all four health outcomes under study. This showed the spatial distribution of mortality risk attributed to the mortality rates of cerebrovascular heart disease, ischaemic heart disease, hypertensive heart disease and diabetes have been stable over the years with minimal changes. In terms of intervention, it becomes easier to formulate programmes where there are minimal changes in identified clusters over time. Bivariate analysis of two CVD mortality rates in a given year showed that there was significant co-clustering between diabetes and the three CVDs. These clusters are located in the south west part of the country. Co-clustering was also significant between cerebrovascular heart disease and hypertensive heart disease. The joint clusters of cerebrovascular heart disease and hypertensive heart disease for the year 2011 are in the south and north west parts of the country. There was no evidence of co-clustering between ischaemic heart disease and both cerebrovascular heart disease and hypertensive heart disease for the 2011 data.

In Chapter 3, we explored the use of the recently developed alternative approach to Moran's I univariate spatial autocorrelation by Chen (2013). In the empirical analysis, the study is also devoted to checking the normal distribution of the residuals. Data for mortality rates due to cerebrovascular heart disease adjusted using the Poisson regression model were used. Improvements on the work done by Chen (2013) included a randomised statistical significance test that was developed and implemented in R programme. This enabled both the detection of clusters at national level and LISA cluster maps at municipal level. In addition, concordance and inconsistency of the LISA cluster maps was assessed among three different spatial weights constructs in South Africa. The three spatial weights based on the inverse power distance function with $\alpha = 1$ and that based on a negative exponential distance function $\alpha = 1$ and $\alpha = 2$ were chosen as they had significant new global Moran's index.

The results of the LISA maps show that all three spatial weight constructs displayed elevated cerebrovascular heart disease mortality risk in the south-western part of the country. The three spatial weight matrices may be different and the residuals of the exponential weight matrices were not normally distributed, but the three methods' "hot-spot" and "cold-spot" clustering are comparatively congruous based on Bangdiwala's B-statistic. This suggests that the outputs based on each of the three spatial weights would provide for similar policy making based on the identification of "hot-spot" and "cold-spot" clustering municipalities. Similar results were also visually observed when the original Moran's index was used. So the three spatial weight constructs would give similar guidelines on interventions and policy.

In Chapter 5, we were concerned with developing a multivariate spatial autocorrelation extension of the Moran's Index using the canonical correlation approach. The application of canonical correlation analysis to make inferences about multivariate spatial autocorrelation from at least three variables, has been demonstrated in order to underline the possible usefulness of this new procedural approach to exploratory spatial analysis. Having decided on the criterion variable, one can derive a variable conditional on the criterion variable to form a criterion variable set of two sub-vectors. The independent variable set is chosen such that it is comprised of at least two variables that have a positive bivariate spatial association with the criterion variable, and between themselves.

Canonical correlation analysis is then used to develop a simple linear regression between the criterion variable and the weighted average of the sub-vector in the independent variable set based on the first canonical variate associated with the independent variable set. Canonical analysis would have established the significance of the first canonical correlation for the first canonical variate pair. The first canonical is always the highest and most important while the second canonical is usually small. In the case of the South African cardiovascular data, it was shown to be not significant. In cases where the second canonical correlation is significant, but way smaller than the first canonical correlation, one can make the simplifying assumption that the weighted average of the sub-vectors in the independent variable set based on the second canonical variate do not disturb the spatial autocorrelation between the criterion variable and the weighted average of the sub-vectors in the independent variable set based on the first canonical variate patterns. But this will need further investigation.

The new method was applied to some hypothetical spatial data in Chapter 5 and real-life cardiovascular mortality data in Chapter 6, and the results show the potential utility of the method in detecting the presence of spatial autocorrelation patterns. In the case of the univariate and bivariate spatial autocorrelations, the results derived by using the multiple regression equivalent of canonical correlation are identical to those obtained using the original univariate and bivariate Moran's spatial autocorrelation approaches. This, unlike the regression approach by Chen (2013) which has different Moran's index values from the original method, validates the new method for the univariate and bivariate cases. The new method was validated in Chapter 6 via a simulation study and agreement analysis. The level of agreement was good to excellent in terms of the a set of given spatially correlated data in comparison with the simulated data

There is no known multivariate Moran's index that caters for more than two variables. A multivariate regression analysis could have been used for three or more variables, but this will only give coefficients that are partial correlations in which there is a blind “control” for other “relevant” variables (Blalock, 1961; Legendre, 1993, 2000). But in the use of the multiple regression equivalent of the canonical correlation procedure, a direct correlation between the criterion variable and a linear combination of the “relevant” variables is estimated which can then be used to derive a multivariate Moran’s index. This has the advantage of determining the spatial correlations of all “relevant” variables simultaneously instead of focusing on one variable at a time, as is in the univariate case, or just two variables at a time , as is in the bivariate case.

7.3 Limitations

Most of the limitations have already been discussed in full in the different chapters, but here, we highlight some before giving future direction of study. In Chapter 3, age-gender standardised mortality ratios were used to try to control for the two confounders. By using age-sex standardised incidence rates, our study removed the effects of age and gender. However, we still find pockets of high risks of CVDs and their risk factors, a finding that suggests other risk factors could be affecting the spatial variations in CVD incidence rates. It shows that differences in observed clustering that we have observed, even after accounting for differences in age and gender distribution across the districts, could be due to differences in other factors, but more data would be needed for more informed analysis. Chapter 3 introduces a method that can be used to address this problem. This was done through the introduction of a Poisson regression approach that adjusts for area-level covariates to estimate the mortality rates. However, covariates to explain the CVD mortality rate patterns are limited and more data will be needed.

The data on high blood cholesterol, smoking, stroke and heart attack described in Chapter 3 were self-reported. Newell *et al.* (1999) noted that inaccurate self-reporting could result in the misclassification or overestimation or underestimation of the incidence or prevalence of disease outcomes and their risk. Biomarkers can be used to redress the problem but unfortunately, there were not available. Without supporting data for validation, the results of the present study need to be treated with caution. However, self-reported values and directly measured values tend to be highly correlated even in the presence of bias (Celis-Morales *et al.*, 2012; Thomas *et al.*, 2016). It is our conviction that, even in the presence of bias in the self-reported values, the spatial autocorrelation patterns obtained in this study would not change much when measured values were to be used.

Our analyses were done at the municipal or district level, which is the level at which primary health is provided in South Africa. Aggregation of the results has the effect of introducing ecological fallacy and large geographical units of analysis may mask some information of interest (Newell *et al.*, 1999). Results and efficiency may be improved by having smaller units of analysis (Newell *et al.*, 1999). According to Paquet *et al.* (2016), when conducting spatial epidemiology, the administrative unit to use in the analysis goes beyond just the size of the unit of analysis and will need to be studied for each given setting. Our study excludes adults older than 64 years old. This was done to focus on the spatial patterns attributable to the productive age group of 15-64 years which overlaps with the age range in which premature mortality occurs (less than 70 years). However, it is hereby acknowledged that this limits the ability to evaluate patterns in the age groups that are at the highest risk of cardiovascular disease (65 years and greater).

In developing the multivariate spatial Moran's index in Chapter 5, writing the system of equations as in Equation 5.37 has the added advantage of seeing the consequence of excluding any pair of the canonical variates $(\mathrm{u}_i, \mathrm{v}_i)$ from the analysis. It is clear that one consequence is the reduction in the explanation of the dependent variables $\widetilde{Y}$ (Johansson & Sheth, 1974). It is then desirable for at least the second canonical pair to be statistically not significant for the

MMI to have more accurate meaning. The second canonical correlations for the data used in this study were all not statistically significant but this may not all ways be the case. So an investigation into the effects of statistically significant second canonical pair on MMI results will need to be investigated.

The new multivariate spatial autocorrelation method is subject to certain practical limitations such as linearity assumptions, differences in level of measurement (solved with standardisation of variables), choice of spatial weights and sample size. But the same limitations also apply to the univariate and bivariate Moran's indexes of spatial autocorrelation as well as the multivariate spatial autocorrelations based on partial autocorrelations such as the Mantel test (Legendre, 2000).

The distribution properties of the new multivariate Moran's index has not yet been ascertained. This will provide equations for the distribution properties such as the mean and variance for the new multivariate spatial association measure, MMI. Such distribution properties can be used to evaluate simulations done using this new approach. Lee (2004) has provided a method that can be used as a basis for future development of the properties.

7.4 Strength

The strength of our study has been the novel application of multivariate spatial autocorrelation modeling approach to measure clustering and local co-clusters of CVDs and their risk factors. Studies by Fabiyi & Garuba (2015), Penney *et al.* (2014), Rajabi *et al.* (2010) and Paquet *et al.* (2016) employed univariate spatial clustering methods. Their approaches could be limited as CVDs and risk factors tend to co-occur at both individual and ecological levels (Ford & Highfield, 2016; Penney *et al.*, 2014). Kandala *et al.* (2013), noted that CVDs and their risk factors have similar aetiology such that analysing them independently would be less efficient. In addition, estimating joint “hot-spot ” and low cluster of districts for two or more CVDs will provide more evidence for an integrated intervention approach that targets all the modelled diseases, instead of targeting only one CVD.

Additionally, by using age-sex standardised incidence rates, our study removed the effects of age and gender, two of the major determinants of health. However, we still find pockets of high risks of CVDs and their risk factors, a finding that suggests other risk factors, could be affecting the spatial variations in CVD incidence rates. As alluded to in Mena *et al.* (2018) and Elmadfa & Meyer (2010), accessibility to health services, socio-economic factors, level of urbanity, educational level, food composition and intake of nutrients, water quality, temperature and other environmental factors could also impact on geographical variations in CVDs and risk factors. Thus, differences in observed clustering that we have observed, even after accounting for differences in age and gender distribution across the districts, could be due to differences in these other factors, but more data would be needed to confirm this assertion.

7.5 Contribution to knowledge

In this thesis a review was made of the relevant literature pertaining to the Moran’s spatial autocorrelation measure. This review revealed that there is a lack of multivariate spatial autocorrelation that extent the Moran’s index to analyse more than two variables. Thus, this study fills that gap in literature by extending the Moran’s index to cater for more than two health outcomes.

7.6 Future direction

The direction to be taken in any future work will be guided by the limitations discussed in the subsection above. More co-variates explaining the health outcomes will need to be measured and incorporated in future analyses to improve results. Excluding higher canonical pairs from the multivariate Moran's index lessens the explanation on the outcome variable by the predictor variables. An investigation on the extent of the reduction of the explanation ought to be instituted and quantified. In addition, the distribution properties of the new multivariate Moran's index are unknown. They need to be developed and be used to evaluate the approach in future research, using Lee (2004) as a basis. A comparative analysis with other multivariate techniques developed for areal data by Jombart *et al.* (2008), Montano & Jombart (2017) and Eckardt & Mateu (2021) will need to be instituted. Despite some of the limitations inherent in the method, we are convinced that if this method is to be used and interpreted properly, it should demonstrate to be a powerful and useful tool in the theoretical advancement of spatial autocorrelations, particularly where multiple variables and complex spatial associations are involved.

7.7 Conclusion

Using novel spatial clustering statistical techniques, the study has identified joint spatial association and locations of similar rates of CVDs and their risk factors among adults in South Africa. Even though the findings of the study are mostly confirmatory, they are nonetheless important in supporting the identification of priority areas for public health interventions. The finding that districts tend to co-cluster in the urban areas and have higher rates of CVDs and risk factors than district that co-cluster in rural areas suggest that there are more contagious and spatial diffusion processes among interdependent districts in urban districts than in the rural areas. Urbanisation or rurality of locations need to be considered when intervention initiatives are implemented. Evidence of co-clustering may point to having an integrated intervention programme targeting several CVDs and associated risk factors simultaneously, mainly in these urban districts and might be more effective and less costly.

References

Ahmed, S. H., Meyer, H. E., Kjollesdal, M. K., Marjerrison, N., Mdala, I., Htet, A. S., Bjertness, E. & Madar, A. A., 2019 The prevalence of selected risk factors for non-communicable diseases in Hargeisa, Somaliland: a cross-sectional study. *BMC Public Health* **19**, 878. 120

Alberti, K., Zimmet, P. & Shaw, J., 2005 The metabolic syndrome — a new worldwide definition. *The Lancet* pp. 1059–1062. 5, 7

Alberts, M., Urdal, P., Steyn, K., Stensvold, I., Tverdal, A., Nel, J. H. & Steyn, N. P., 2005 Prevalence of cardiovascular diseases and associated risk factors in a rural black population of South Africa. *European journal of cardiovascular prevention and rehabilitation* **12**, 347–54. 5

Anselin, L., 1995 Local indicators of spatial association (LISA). *Geographical Analysis* **27**, 93–115. 1, 10, 51, 82, 83, 85

Anselin, L., Syabri, I. & Smirov, O., 2002 . 1, 8, 69, 83

Bangdiwala, S. I. & Shankar, V., 2013 The agreement chart. *BMC medical research methodology* **13**, 97–97. 105

Birnbaum, J., Murray, C & Lozano, R., 2011 Exposing misclassified hiv/aids deaths in south africa. *Bull World Health Organ* **89**, 278–285. 37

Blalock, H. M., 1961 Correlation and causality: The multivariate case. *Social Forces* **39**, 246–251. 124

Bradshaw, D., Bourne, D., Schneider, M. & Sayed, R., 1995 Mortality patterns of chronic diseases of lifestyle in South Africa. In *Review of research and identification of essential health research priorities* (eds. J. Fourie & K. Steyn). Cape Town: South Africa Medical Research Council. 6, 8

Bradshaw, D., Schneider, M., Norman & Bourne, D., 2006 Mortality patterns of chronic diseases of lifestyle in South Africa. In *Chronic Diseases of Lifestyle in South Africa: 1995 - 2005* (eds. K. Steyn, J. Fourie & N. Temple). Cape Town: South Africa Medical Research Council, Burden of Disease Research Unit. 7, 8

Cappuccio, F. & Miller, M., 2016 Cardiovascular disease and hypertension in Sub-Saharan Africa: burden, risk and interventions. *Internal and Emergency Medicine* **11**, 299–305. 7, 120

Celis-Morales, C. A., Perez-Bravo, F., Ibañez, L., Salas, C., Bailey, M. E. S. & Gill, J. M. R., 2012 Objective vs. self-reported physical activity and sedentary time: effects of measurement method on relationships with risk biomarkers. *PloS One* **7**, e36345–e36345. 126

Chen, Y., 2013 New approaches for calculating moran's index of spatial autocorrelation. *PloS One* **7**, 1–14. 2, 14, 17, 18, 19, 20, 21, 48, 49, 51, 58, 59, 82, 122, 124

Chen, Y., 2015 A new methodology of spatial cross-correlation analysis. *PloS One* **10**, 1–20. 2, 82

Chien, L., Lin, G., Li, X. & Zhang, X., 2018 Disparity of imputed data from small area estimate approaches–a case study on diabetes prevalence at the county level in the us. *Data Science Journal* **17**, 1–11. 106

Clark, D., 1975 *Concepts and Techniques in Modern Geography (CATMOG 3): Understanding Canonical Correlation Analysis*. Norwich: Geo Abstracts, University of Anglia. 83

Clayton, D. & Kaldor, J., 1987 Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* . 37

Cliff, A. D. & Ord, J. K., 1969 *The Problem of Spatial Autocorrelation*, p. 25–55. London Papers in Regional Science. London: Pion. 10

Day, C., Groenewald, P., Laubscher, R., Chaudhry, S., van Schaik, N. & Bradshaw, D., 2014 Monitoring of non-communicable diseases such as hypertension in South Africa: Challenges for the post-2015 global development agenda. *South African Medical Journal* **104**, 680–687. 35

de Jong, P., Sprenger, C. & van Veen, F., 1984 On extreme values of Moran's I and Geary's c. *Geographical Analysis* **16**, 17–24. 69

Dray, S., Said, S. & Debias, F., 2008 Spatial ordination of vegetation data using a generalization of wartenberg's multivariate spatial correlation. *Ecological Modelling Journal of Vegetation Science* **19**, 45–56. 69, 82

Eckardt, M. & Mateu, J., 2021 Partial and semi-partial statistics of spatial associations for multivariate areal data. *Geographical Analysis* . 82, 129

Elmadfa, I. & Meyer, A. L., 2010 Importance of food composition data to nutrition and public health. *European Journal of Clinical Nutrition* **64 Suppl 3**, S4–7. 128

Everitt, B. & Rencher, A., 1996 Methods of multivariate analysis. *The Statistician* **45**, 535. 91

Fabiyi, O. O. & Garuba, O. E., 2015 Geo-spatial analysis of cardiovascular disease and biomedical risk factors in ibadan, South-Western Nigeria. *Journal of Settlements and Spatial Planning* **6**, 61–69. URL https://search.proquest.com/docview/1695792143?accountid=16460. 119, 120, 127

Ford, M. M. & Highfield, L. D., 2016 Exploring the spatial association between social deprivation and cardiovascular disease mortality at the neighbourhood level. *PloS One* **11**, 1–17. 2, 6, 7, 120, 127

Gaziano, T., 2007 Reducing the growing burden of cardiovascular disease in the developing world. *Health affairs* **26**, 13–24. 4

GBD Collaborators, 2017 Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *Lancet* **390**, 1151–1210. 3, 4

Griffith, D. & Chun, Y., 2014 Spatial autocorrelation and eigenvector spatial filtering. In *Handbook of Regional Science* (eds. M. Fischer & P. Nijkamp). Berlin Heidelberg: Springer-Verlag. 45

Griffith, D. A., 1987 Spatial autocorrelation: A primer. *Association of American Geographers, Resource Publications in Geography* . 14

Groenewald, P., Bradshaw, D., Day, C. & Laubscher, R., 2014 Burden of disease. In *District Health Barometer: 2013-14* (eds. N. Massyn, C. Day, N. Peer, A. Padarath, P. Barron & R. English). Durban: Health Systems Trust. 8, 35

Johansson, J. K. & Sheth, J. N., 1974 Canonical correlation, multiple regression, and simultaneous systems; some equivalences and their implications. *Faculty working papers number 150*, University of Illinois at Urbana-Champaign, College of Commerce and Business Administration, Chicago. 92, 94, 126

Jombart, T., Devillard, S., A-B., D. & Pontier, D., 2008 Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**, 92–103. 82, 129

Joubert, J., Rao, C., Bradshaw, D., Vos, T. & Lopez, A., 2013 Evaluating the quality of national mortality statistics from civil registration in South Africa, 1997–2007. *PloS One* **8**, e64592. 35, 37

Kandala, N.-B., Mandad, S. O., Tigbea, William W.and Mwambi, H. & Stranges, S., 2014 Geographic distribution of cardiovascular comorbidities in South Africa: a national cross-sectional analysis. *Journal of Applied Statistics* **41**, 1203–1216. 8

Kandala, N.-B., Tigbe, W., Manda, S. O. M. & Stranges, S., 2013 Geographic variation of hypertension in Sub-Saharan Africa: A case study of South Africa. *American Journal of Hypertension* **26**, 382–391. 2, 120, 127

Kosfeld, R., 2010 Spatial econometrics. URL `http://www.uni-kassel.de/~rkosfeld/lehre/spatial.html`. [Online; posted 27-May-2012]. 11, 12, 13, 19

Lee, S.-I., 2001 Developing a bivariate spatial association measure: An integration of Pearson's r and Moran's I. *Journal of Geographical Systems* **3**, 369–385. 1, 2, 10, 67, 68, 69, 82, 83, 86, 87, 97

Lee, S.-I., 2004 A generalized significance testing method for global measures of spatial association: an extension of the mantel test. *Environment and Planning* **36**, 1687 – 1703. 2, 19, 127, 129

Legendre, P., 1993 Spatial autocorrelation: Trouble or new paradigm? *Ecology* **74**, 1659–1673. 124

Legendre, P., 2000 Comparison of permutation methods for the partial correlation and partial mantel tests. *Journal of Statistical Computation and Simulation* **67**, 37–73. 83, 124, 127

Leyland, A. H. & Davies, C. A., 2005 Empirical bayes methods for disease mapping. *Statistical Methods in Medical Research* **14**, 17–34. 39

Manda, S. O., Lombard, C. J. & Mosala, T., 2012 Divergent spatial patterns in the prevalence of the human immunodeficiency virus (HIV) and syphilis in South African pregnant women. *Geospatial Health* pp. 221–231. 8

Manning, E., Mann, J., Sophangisa, E. & Truswell, A., 1974 Dietary patterns in urbanised blacks. *South African Mededical Journal* **48**, 485–498. 7, 119

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220. 19, 83

Marshall, R. J., 1991 Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society* . 39

Matsha, T. E., Hassan, M. S., Kidd, M. & Erasmus, R. T., 2012 The 30-year cardiovascular risk profile of South Africans with diagnosed diabetes, undiagnosed diabetes, pre-diabetes or normoglycaemia: the Bellville, South Africa pilot study. *Cardiovascular journal of Africa* **23**, 5–11. 5

Mena, C., Sepulveda, C., Fuentes, E., Ormazabal, Y. & Palomo, I., 2018 Spatial analysis for the epidemiological study of cardiovascular diseases: A systematic literature search. *Geospat Health* **13**, 587. 128

Montano, V. & Jombart, T., 2017 An eigenvalue test for spatial principal component analysis. *BMC Bioinformatics* **18**, 562. 82, 129

Moran, P. A. P., 1950 Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23. 1, 10

Munoz, S. & Bangdiwala, S., 1997 Interpretation of kappa and b statistics measures of agreement. *Journal of Applied Statistics* **24**, 105–112. 57, 109

NDoH, 2013 Strategic plan for the prevention and control of non-communicable diseases 2013-17. *Tech. Rep. RP06/2013*, National Department of Health. 7

Neupane, S., Prakash, K. C. & Doku, D. T., 2016 Overweight and obesity among women: analysis of demographic and health survey data from 32 Sub-Saharan African countries. *BMC public health* **16**, 30–30. 5

Newell, S. A., Girgis, A., Sanson-Fisher, R. W. & Savolainen, N. J., 1999 The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population 1: A critical review. *American Journal of Preventive Medicine* **17**, 211–229. 125, 126

Njelekela, M. A., Mpembeni, R., Muhihi, A., Mligiliche, N. L., Spiegelman, D., Hertzmark, E., Liu, E., Finkelstein, J. L., Fawzi, W. W., Willett, W. C. & Mtabaji, J., 2009 Gender-related differences in the prevalence of cardiovascular disease risk factors and their correlates in urban Tanzania. *BMC Cardiovascular Disorders* **9**, 30. 5

Noubiap, J. J. N., Nansseu, J. R. N., Bigna, J. J. R., Jingi, A. M. & Kengne, A. P., 2015 Prevalence and incidence of dyslipidaemia among adults in Africa: a systematic review and meta-analysis protocol. *BMJ open* **5**, e007404. 3

Olawuyi, A. T. & Adeoye, I. A., 2018 The prevalence and associated factors of non-communicable disease risk factors among civil servants in Ibadan, Nigeria. *PloS One* **13**, e0203587–e0203587. 5

Ord, J. K. & Getis, A., 1995 Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* **27**, 286–306. URL http://dx.doi.org/10.1111/j.1538-4632.1995.tb00912.x. 16

Paquet, C., Chaix, B., Howard, N. J., Coffee, N. T., Adams, R. J., Taylor, A. W., Thomas, F. & Daniel, M., 2016 Geographic clustering of cardiometabolic risk factors in metropolitan centres in France and Australia. *International journal of environmental research and public health* **13**, 519. 119, 120, 126, 127

Peer, N., Steyn, K., Lombard, C., Lambert, E. V., Vythilingum, B. & Levitt, N. S., 2012 Rising diabetes prevalence among urban-dwelling Black South Africans. *PloS One* **7**, e43336. 4, 5

Pelzom, D., Isaakidis, P., Oo, M. M., Gurung, M. S. & Yangchen, P., 2017 Alarming prevalence and clustering of modifiable noncommunicable disease risk factors among adults in Bhutan: a nationwide cross-sectional community survey. *BMC public health* **17**, 975. 5

Penney, T. L., Rainham, D. G., Dummer, T. J. & Kirk, S. F., 2014 A spatial analysis of community level overweight and obesity. *J Hum Nutr Diet* **27 Suppl 2**, 65–74. 119, 127

Pestana, J. A. X., Steyn, K., Leiman, A. & Hartzenberg, G., 1996 The direct and indirect costs of cardiovascular disease in South Africa in 1991. *S Atr Med J* **86**, 679–684. 4, 7

Petoumenos, K., Reiss, P., Ryom, L., Rickenbach, M., Sabin, C., El-Sadr, W., d'Arminio Monforte, A., Phillips, A. N., De Wit, S., Kirk, O., Dabis, F., Pradier, C., Lundgren, J. D., Law, M. G. & study group, D., 2014 Increased risk of cardiovascular disease (CVD) with age in HIV-positive men: a comparison of the D:A:D CVD risk equation and general population CVD risk equations. *HIV Medicine* **15**, 595–603. 22

Pillay-van Wyk, V., Bradshaw, D., Groenewald, P. & Laubscher, R., 2011 Improving the quality of medical certification of cause of death: The time is now! *South African Medical Journal* **101**, 171–177. 37

Rajabi, M., Mansourian, A., Pilesjö, P., Åström, D. O., Cederin, K. & Sundquist, K., 2010 Exploring spatial patterns of cardiovascular disease in Sweden between 2000 and 2010. *Scandinavian journal of public health* **46**, 647–658. 119, 120, 127

Sarndal, C. E., 1984 Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association* **79**, 624–631. 37

Sartorius, B., Kahn, K., Collinson, M. A., Vounatsou, P. & Tollman, S. M., 2010 Space and time clustering of mortality in rural South Africa (Agincourt HDSS), 1992-2007. *Global Health Action Supplement 1* pp. 50–58. 8

Sartorius, B., Sartorius, K., Chirwa, T. & Fonn, S., 2011 Infant mortality in South Africa - distribution, associations and policy implications, 2007: an ecological spatial analysis. *International Journal of Health Geographics* **10**, 61. 8

Schutte, A. E., 2018 Urgency for South Africa to prioritise cardiovascular disease management. *Lancet* **7**, e177. 7

Smouse, P. E., Long, J. C. & Sokal, R. R., 1986 Multiple regression and correlation extensions of the mantel test of matrix correspondence. *Systematic Zoology* **35**, 627–632. 82, 83

Statistics South Africa, 2014 The South African MPI: Creating a multidimensional poverty index using census data. *Report no.: 03-10-08*, Statistics South Africa, Pretoria. 39

Steyn, N., Bradshaw, D., Norman, R., Joubert, J. D., Schneider, M. & Steyn, K., 2006 Dietary changes and the health transition in South Africa: implications for health policy. 7, 119, 120

Tanser, F., Barnighausen, T. & Cooke, G., 2009 Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. *International Journal of Epidemiology* **38**, 1008–1016. 8

Thomas, J., r., Paulet, M. & Rajpura, J. R., 2016 Consistency between self-reported and recorded values for clinical measures. *Cardiol Res Pract* **2016**, 4364761. 126

Tsai, P. J., Lin, M. L., Chu, C. M. & Perng, C. H., 2009 Spatial autocorrelation analysis of health care hotspots in taiwan in 2006. *BMC Public Health* **9**, 1–13. 6

van Rheenen, S. M., 2015 A Spatial Epidemiological Analysis of Stroke in Alberta, Canada, Using GIS. Ph.D. thesis. 6

Van Zyl, M., Steyn, N. & Marais, M., 2010 Characteristics and factors influencing fast food intake of young adult consumers in Johannesburg South Africa. *South African Journal of Clinical Nutrition* **23**, 124–130. 119

Vuvor, F., Steiner-Asiedu, M., Saalia, K. & Owusu, W., 2016 Predictors of hypertension, hypercholesterolemia, and dyslipidemia of men living in a periurban community in Ghana. *Journal of Health Research and Reviews* **3**, 66–71. 120

Waller, L. A. & Gotway, C. A., 2004 *Applied spatial statistics for public health data*. New Jersey: Wiley. 1, 8, 12, 13, 14, 20

Wandai, M. E., Norris, S. A., Aagaard-Hansen, J. & Manda, S. O., 2019 Geographical influence on the distribution of the prevalence of hypertension in South Africa: a multilevel analysis. *Cardiovasc J Afr* **30**, 1–8. 120

Wartenberg, D., 1985 Multivariate spatial correlation: A method for exploratory geographical analysis. *Geographical Analysis* **17**, 263–283. 1, 66, 67, 68, 69, 82, 118

Wesonga, R., Guwatudde, D., Bahendeka, S. K., Mutungi, G., Nabugoomu, F. & Muwonge, J., 2016 Burden of cumulative risk factors associated with non-communicable diseases among adults in Uganda: evidence from a national baseline survey. *International journal for equity in health* **15**, 195–195. 4, 5

WHO, 2013 Draft comprehensive global monitoring framework and targets for the prevention and control of noncommunicable diseases A66/8. *Technical report*, WHO, Geneva. 7

WHO, 2015 Noncommunicable diseases. In *Health in 2015: from MDGS to SDGs* (ed. WHO). Geneva: WHO,. 3

WHO, 2018 World health statistics 2018: Monitoring health for the SDGs, sustainable development goals. *Technical report*, WHO, Geneva. 3, 4

World Health Organization, 2004 ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth revision. *Tech. rep.*, W. 35

Yaya, S., Ekholuenetale, M. & Bishwajit, G., 2018 Differentials in prevalence and correlates of metabolic risk factors of non-communicable diseases among women in sub-Saharan Africa: evidence from 33 countries. *BMC public health* **18**, 1168–1168. 4, 5

Appendix A

Appendix A for Chapter 3

Table A.1: Correlation analysis between CVDs and their risk factors for the whole sample, by age, gender and age-gender combinations.

	(A) Overall Sample										
	Stroke	Heart attack	Smoking	HBC	Hypertension						
Stroke	1.00	0.85	0.47	0.82	0.41						
Heart attack		1.00	0.38	0.71	0.41						
Smoking			1.00	0.55	0.73						
HBC				1.00	0.53						
Hypertension					1.00						
	(B) Male						**(C) Female**				
	Stroke	Heart attack	Smoking	HBC	Hypertension		Stroke	Heart attack	Smoking	HBC	Hypertension
Stroke	1.00	0.81	0.15	0.66	0.45	Stroke	1.00	0.22	0.00	0.29	0.04
Heart attack		1.00	0.14	0.64	0.50	Heart attack			-0.06	0.24	0.16
Smoke			1.00	0.16	0.61	Smoking				0.12	0.03
HBC				1.00	0.55	HBC					0.40
Hypertension					1.00	Hypertension					1.00
	(D) 15-39 years						**(E) 40-64 years**				
	Stroke	Heart attack	Smoking	HBC	Hypertension		Stroke	Heart attack	Smoking	HBC	Hypertension
Stroke	1.00	0.10	0.14	-0.03	-0.04	Stroke	1.00	0.61	0.40	0.68	0.01
Heart attack		1.00	0.09	-0.11	0.05	Heart attack		1.00	0.25	0.49	0.14
Smoking			1.00	0.46	0.73	Smoking			1.00	0.41	0.32
HBC				1.00	0.50	HBC				1.00	0.11
Hypertension					1.00	Hypertension					1.00
	(F) Males aged 15-39 years						**(G) Females aged 15-39 years**				
	Stroke	Heart attack	Smoking	HBC	Hypertension		Stroke	Heart attack	Smoking	HBC	Hypertension
Stroke	1.00	0.36	0.05	0.03	0.08	Stroke	1.00	-0.01	-0.05	0.00	-0.18
Heart attack		1.00	0.08	-0.03	-0.02	Heart attack		1.00	-0.03	0.00	0.14
Smoking			1.00	0.30	0.63	Smoking			1.00	0.10	0.53
HBC				1.00	0.54	HBC				1.00	0.26
Hypertension					1.00	Hypertension					1.00
	(H) Males aged 40-64 years						**(I) Females aged 40-64 years**				
	Stroke	Heart attack	Smoking	HBC	Hypertension		Stroke	Heart attack	Smoking	HBC	Hypertension
Stroke	1.00	0.45	0.24	0.48	0.08	Stroke	1.00	0.53	0.10	0.81	0.12
Heart attack		1.00	0.16	0.49	0.35	Heart attack		1.00	-0.02	0.50	0.03
Smoking			1.00	0.22	0.08	Smoking			1.00	0.14	-0.11
HBC				1.00	0.17	HBC				1.00	0.28
Hypertension					1.00	Hypertension					1.00

Key: CVD, cardiovascular disease; HBC, high blood cholesterol.

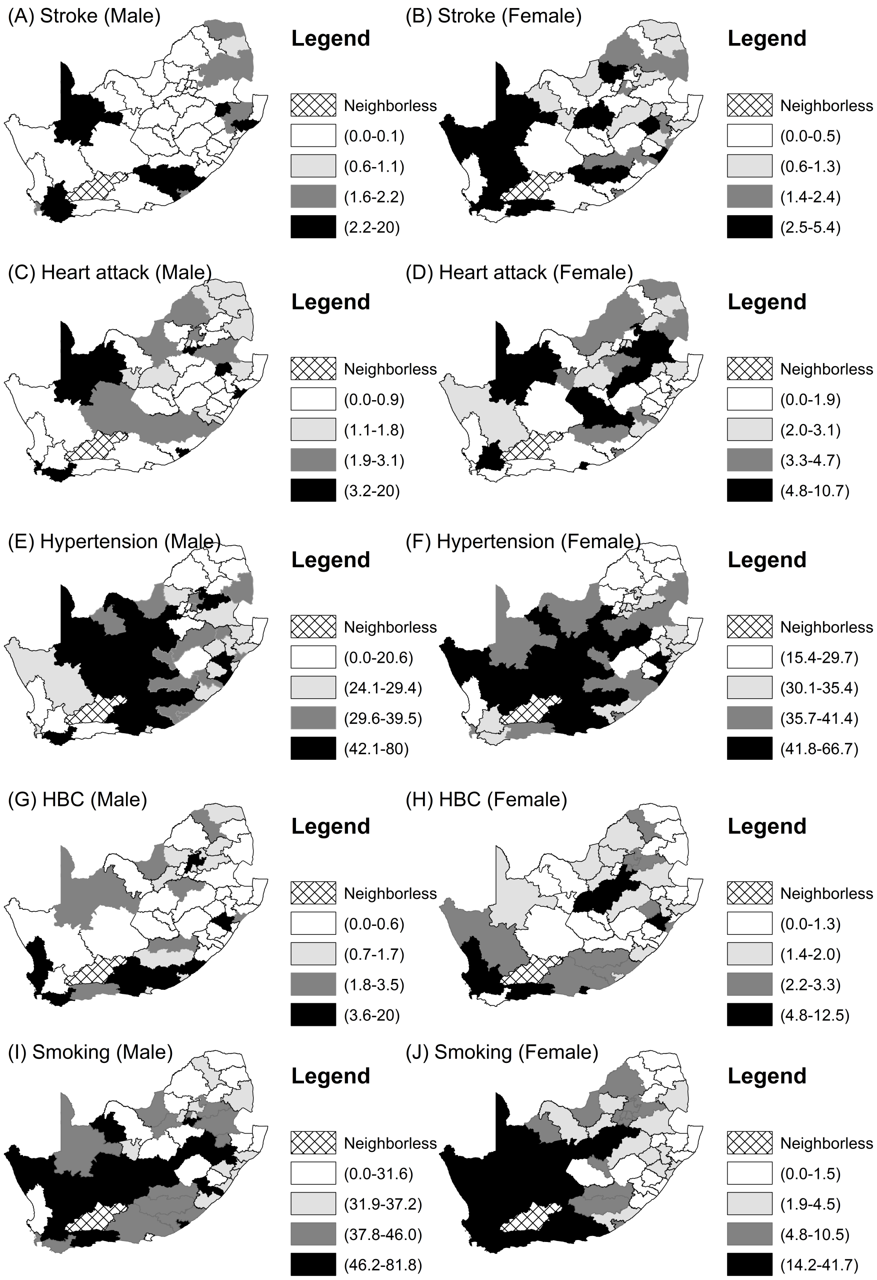


Figure A.1: Quantile maps of prevalence rates of the CVDs and their related risk factors by gender.

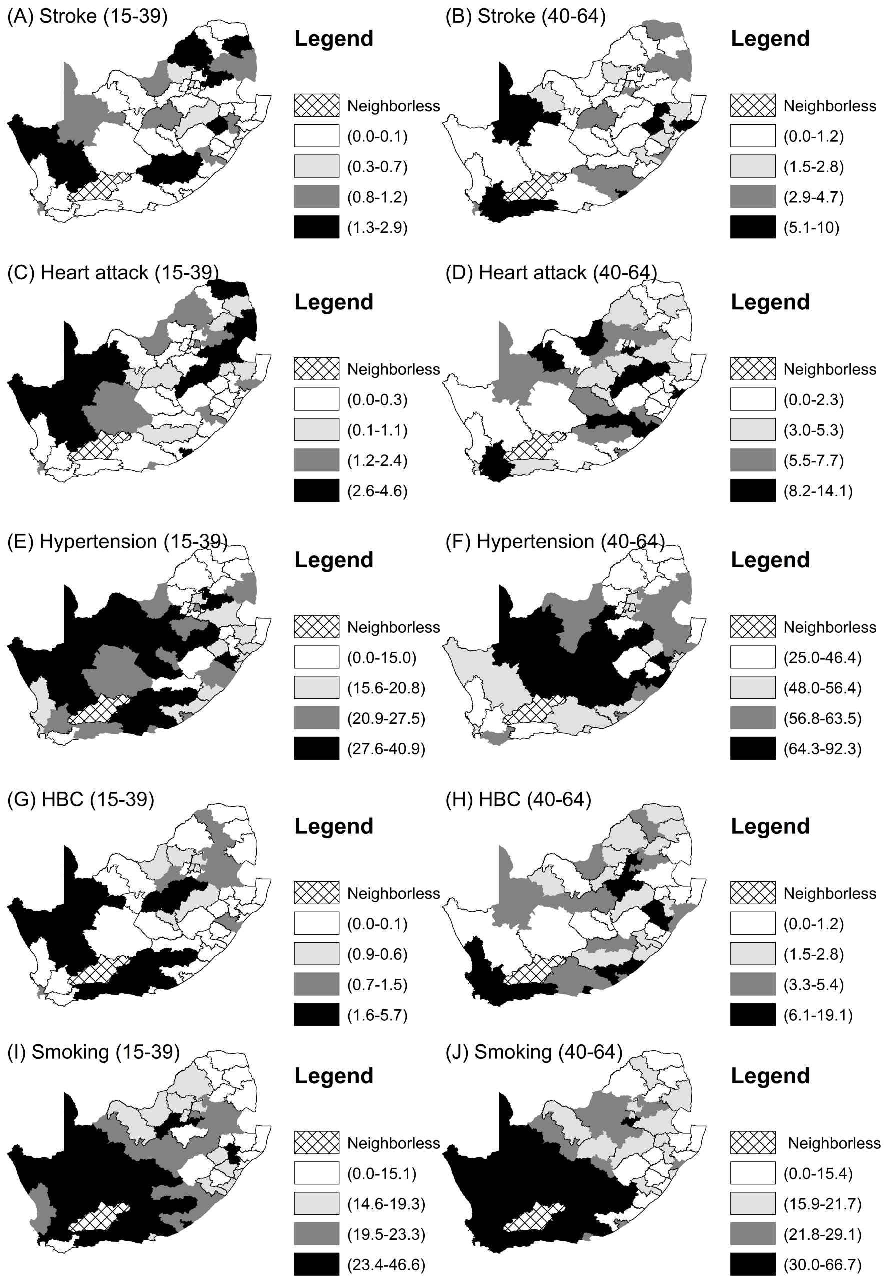


Figure A.2: Quantile maps of prevalence rates of the CVDs and their related risk factors by age.

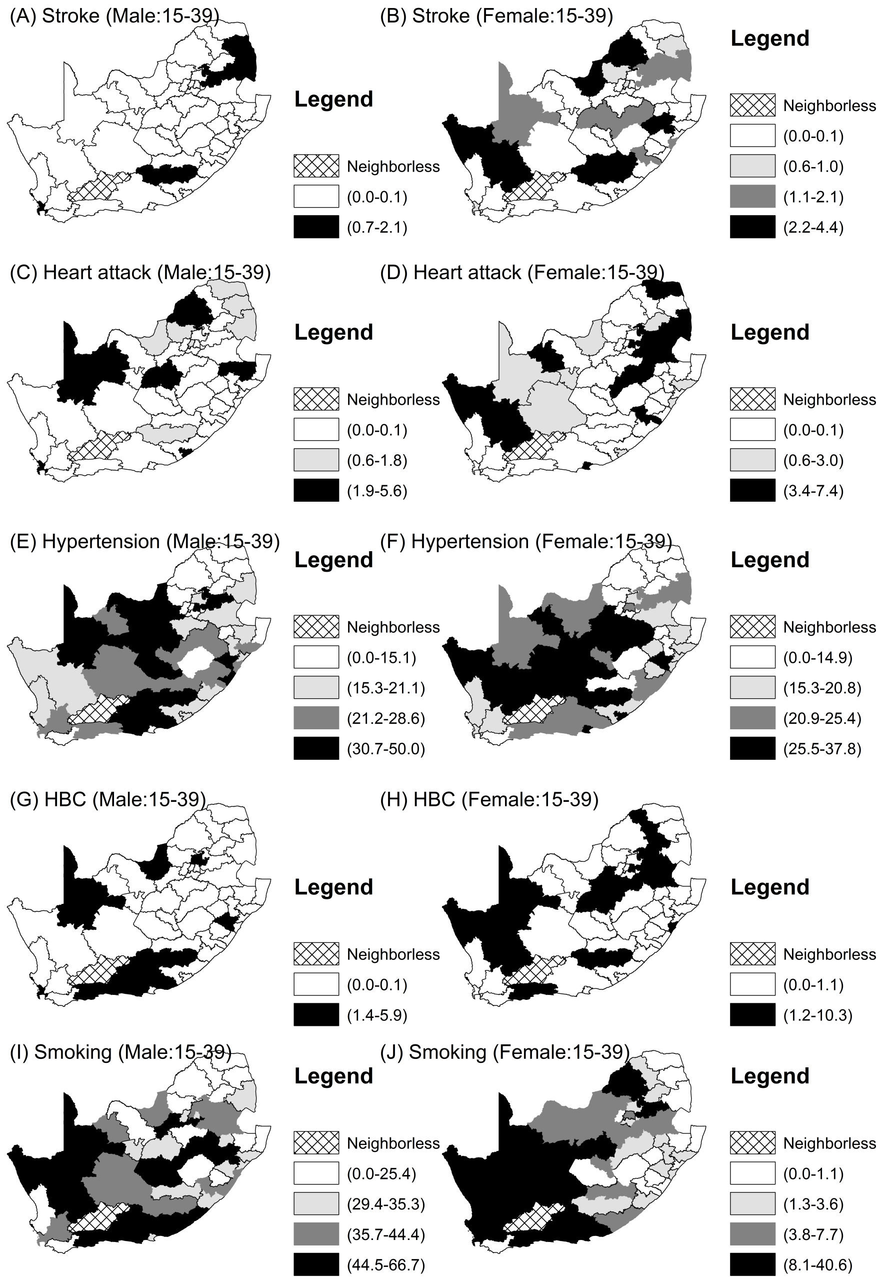


Figure A.3: Quantile maps of prevalence rates of the CVDs and their related risk factors by gender for ages 15-39 years.

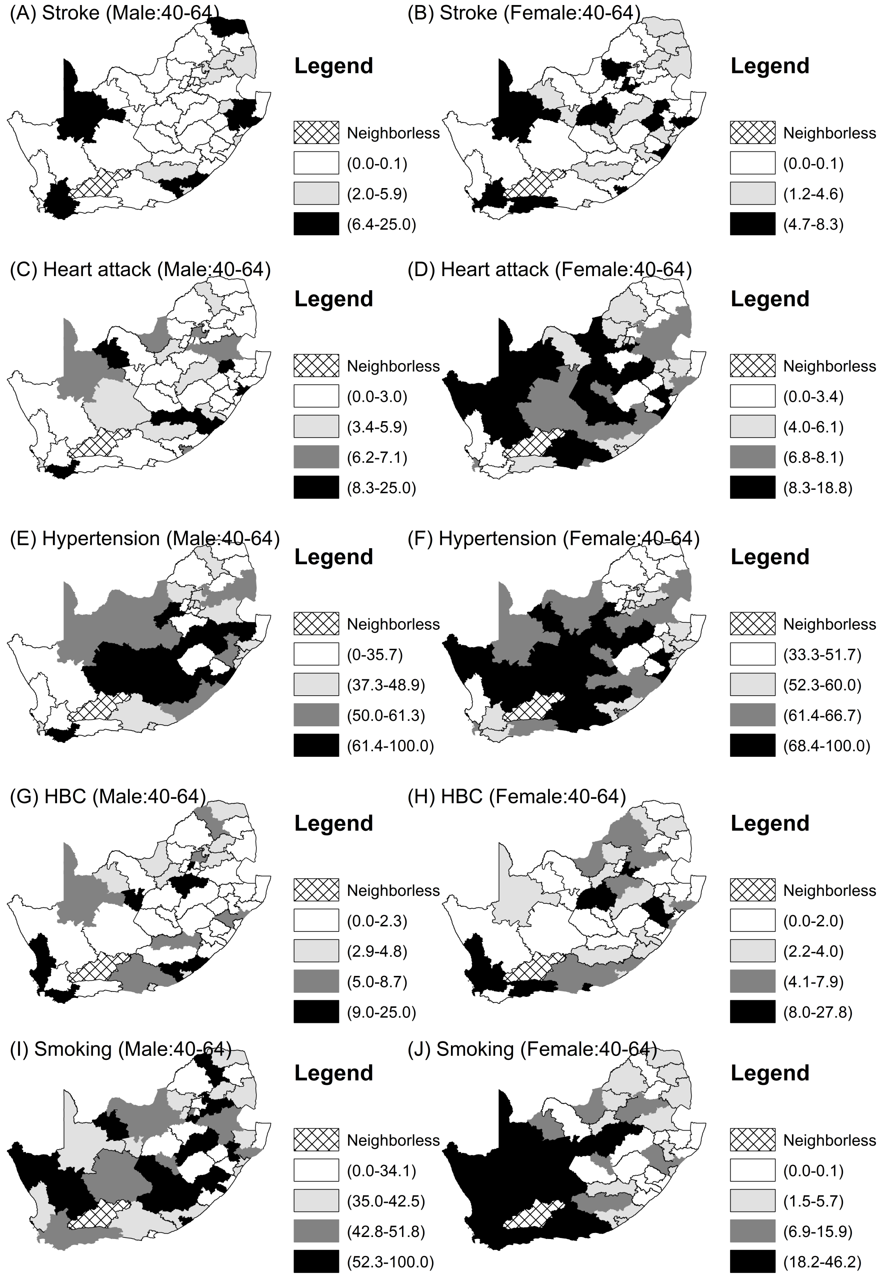


Figure A.4: Quantile maps of prevalence rates of the CVDs and their related risk factors by gender for ages 40-64 years.

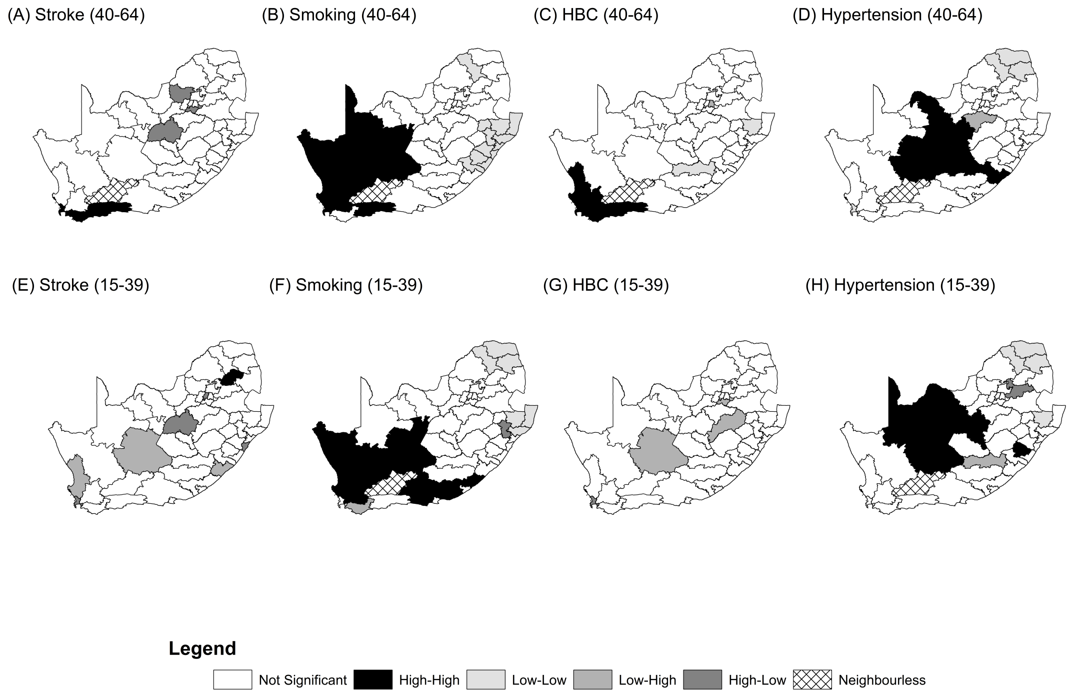


Figure A.5: Univariate spatial clusters of CVDs and their risk factors by age groups.

A.1 R Code for Chen’s regression approach

```
library(foreign)
#Create Weight Matrix from distance Matrix
wgts1<-read.csv("SA_dist_matrix.csv")
n<-dim(wgts1)[1]
tmp<-matrix(rbinom(n * n, 1, 1), ncol = n, nrow = n)
tmp<-as.matrix(tmp)
#Calculate an inverse power function weight matrix
wgts2<-tmp/wgts1
colnames(wgts2)<-1:n
rownames(wgts2)<-1:n
#Use the power alpha by changing value
alpha<-1.0
wgts3<-wgts2^(alpha)
## Make all diagonal elements zero
diag(wgts3)<-0
# Add all elements in the weights matrix
sum_wgts<-sum(wgts3)
#Divide all elements in the weights matrix by the totals
wgts<-(1/sum_wgts)*(wgts3)
#Check if sum equals 1
sum(wgts)

library(maptools) #Contains the overlay command ## Data management
SAshp <- readShapePoly("SAshpCVD(Bin,Poi).shp",
                      proj4string=CRS("+proj=longlat +datum=WGS84"))
#data
```

```
x <- SAshp$Poi_IHD11

y<-x-mean(x)

s<-sqrt(sum(y^2)/n)

z <- (y)/(s)

z<-as.matrix(z)


#The Ideal Spatial Weight Matrix (ISWM)- M_Star is given by

M_Star<-z%*%t(z)%*%(wgts)

# Calculate the Diagonal matrix to give LISA:

A<-diag(M_Star)

Local_MI<-as.matrix(A)

## Calculate the Moran’s I’

I = Global_MI=sum(A)

# Permutations for Moran’s I

nsims <-999

W=wgts

Local_moran <- function(x, W){

  n   = nrow(W)

  z <- (x - mean(x, na.rm=T))/sd(x, na.rm=T)

  z<-as.matrix(z)

  M_Star<-z%*%t(z)%*%(W)

  A<-diag(M_Star)

  local_moran<-as.matrix(A)

  list(local_moran  = local_moran)

}

Local_mor<-as.matrix(Local_moran(x, W))

local_sims  = matrix(NA, nrow = n, ncol=nsims)

nsims <-999
```

```
for(i in 1:nsims){
  x2<-sample(x)
  local_sims[[i]]  <-Local_moran(x2, W)
}

D<-as.matrix(local_sims)
D2<-lapply(D, na.omit)
#Combine lists into a dataframe:use a plyr function:
library(plyr)
df <- ldply(D2, data.frame)
# create matrix with elements
y <- matrix(df$local_moran, nrow=234, ncol=nsims )
y<-as.matrix(y)
Loc_I.count <- vector(mode="numeric", length=s)

for (i in 1:234){
  Loc_I.count [i]<- sum(abs(y[i,]) >as.numeric(Local_MI[i]))
}
Loc_I.count
p_value <- Loc_I.count/(nsims+1)
#Calculate f*:
f_Star=M_Star%*%z
#The Real Spatial Weight Matrix (RSWM)- M is given by
M<-n*(wgts)
#Calculate f:
f<-M%*%z
#Calculate the error terms  e_f
```

```
e_f<-(f-f_Star)


library(nortest)
#ad.test(e_f)
shapiro.test(e_f)
#Calculate z_Star
z_Star<- (1/Global_MI)*f
#Calculate the error terms  e_z
e_z<-(z-z_Star)
##Put all the variables in one file
dat<-as.data.frame(SAshp$MN_CODE)
dat$Poi_IHD11 <-SAshp$Poi_IHD11
dat$z_score <-z
dat$f <-f
dat$f_Star <-f_Star
dat$z_Star <-z_Star
dat$LISA <-A
dat$P_value <-as.vector(p_value)


library(dplyr)
dat<- dat %>%
  mutate(Level1 = if_else(z > 0, ’H’, ’L’))
dat<- dat %>%
  mutate(Level2 = if_else(f > 0, ’H’, ’L’))


#unite function from tidyr
require(tidyverse)
dat <- dat %>% mutate(
  Type = paste(Level1, Level2, sep = "-")
)

dat<- dat %>%
  mutate(Significant = if_else(P_value > 0.05, ’No’, ’Yes’))

#Remove columns by setting them to NULL
dat$Level1 <- NULL
dat$Level2 <- NULL

## DIAGNOSTIC CHECKS OF THE RESIDUALS
# Calculate standard error between f and f_Star, s_f:
s_f<-sqrt((1/n)*t(e_f)%*%(e_f))
# Calculate standard error between z and z_Star, s_z:
s_z<-sqrt((1/n)*t(e_z)%*%(e_z))

# Histogram of standardised residuals:
b<-e_f-mean(e_f)
se<-sqrt(sum(b^2)/n)
z_e <- (e_f-mean(e_f))/(se)
z_e<-as.matrix(z_e)

#x11(width = 12, height = 7, pointsize = 12)
x11()
par(mfrow=c(1,2))

library(rcompanion)
plotNormalHistogram(z_e,xlab="Residuals")
```

```
qqnorm(z_e,ylab="Sample Quantiles",main="")
qqline(z_e,col="red")

#Scatter plot
a<-max(dat$f_Star)
b<-max(dat$f)
c<-max(a,b)+1
a1<-min(dat$f_Star)
b1<-min(dat$f)
c1<-min(a1,b1)-1
a2<-max(dat$z_score)+1
b2<-min(dat$z_score)-1
x11()
library(rcompanion)
# Fit a regression line in which the intercept has been
# forced to be zero and display the line on the scattter
mC <- lm(dat$f_Star ~dat$z_score)
plot(dat$z_score,dat$f_Star,xlim=range(c(b2,a2)),ylim=range(c(c1,c)), pc
     ylab="f*/f values",col="red",cex.lab=0.75,lwd=2,las=1)
points(dat$z_score,dat$f,col="blue",pch=1)
abline(coef = coef(mC),col = "black",pch=4)
abline(v=0,col="black",lty=2)
abline(h=0,col="black",lty=2)
lm_coef <- round(coef(mC), 4) # extract coefficients
mtext(bquote(Index == .(lm_coef[2])),
      adj=1, padj=0) # display equation
# Add a legend
legend("topright",
       legend = c("f*", "f"),
       col = c("red","blue"),
       pch = c(17,1),
       bty = "n",
       pt.cex = 2,
       cex = 1.2,
       text.col = "black",
       horiz = F ,
       inset = c(0.1, 0.1))
#sf values
s_f
```

Table A.2: Classification of spatial autocorrelation based on inverse power spatial weight, $\alpha = 1$.

Municipality	**Ischaemic**	z	f	f^*	z^*	**LISA**	P value	**Type**	**Significant**
Matzikama	80.99	2.10	0.05	0.03	3.39	0.0005	0.0130	H-H	Yes
Cederberg	74.93	1.81	0.03	0.03	1.94	0.0002	0.0390	H-H	Yes
Bergrivier	77.52	1.93	-0.22	0.03	-13.69	-0.0018	0.9990	H-L	No
Saldanha Bay	61.7	1.17	-0.10	0.02	-6.12	-0.0005	0.9990	H-L	No
Swartland	68.23	1.48	-0.15	0.02	-9.38	-0.0009	0.9990	H-L	No
Witzenberg	57.4	0.96	0.06	0.02	3.93	0.0003	0.1980	H-H	No
Drakenstein	63.95	1.28	-0.11	0.02	-7.18	-0.0006	0.9990	H-L	No
Stellenbosch	58.22	1.00	0.12	0.02	7.51	0.0005	0.0300	H-H	Yes
Breede Valley	62.13	1.19	0.09	0.02	5.77	0.0005	0.0240	H-H	Yes
Langeberg	70.48	1.59	-0.12	0.03	-7.63	-0.0008	0.9990	H-L	No
Swellendam	78.05	1.96	0.08	0.03	4.71	0.0006	0.0300	H-H	Yes
Theewaterskloof	59.33	1.06	0.17	0.02	10.91	0.0008	0.0190	H-H	Yes
Overstrand	58.59	1.02	-0.06	0.02	-3.61	-0.0002	0.9990	H-L	No
Cape Agulhas	83.21	2.21	0.06	0.04	3.63	0.0005	0.0240	H-H	Yes
Kannaland	90.67	2.57	0.07	0.04	4.27	0.0007	0.0360	H-H	Yes
Hessequa	92.48	2.66	-0.02	0.04	-1.35	-0.0002	0.9990	H-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Mossel Bay	64.27	1.29	-0.09	0.02	-5.74	-0.0005	0.9990	H-L	No
George	60.08	1.09	0.10	0.02	6.27	0.0005	0.0260	H-H	Yes
Oudtshoorn	83.16	2.21	0.06	0.04	3.63	0.0005	0.0460	H-H	Yes
Bitou	54.09	0.80	0.11	0.01	6.96	0.0004	0.0470	H-H	Yes
Knysna	66.87	1.42	0.07	0.02	4.09	0.0004	0.1090	H-H	No
Laingsburg	85.9	2.34	-0.08	0.04	-5.01	-0.0008	0.9990	H-L	No
Prince Albert	87.8	2.43	-0.11	0.04	-6.72	-0.0011	0.9990	H-L	No
Beaufort West	73.75	1.75	-0.07	0.03	-4.50	-0.0005	0.9990	H-L	No
City of Cape Town	49.53	0.58	-0.20	0.01	-12.66	-0.0005	0.9990	H-L	No
Buffalo City	32.58	-0.24	-0.15	0.00	-9.68	0.0002	0.2480	L-L	No
Camdeboo	66.47	1.40	-0.04	0.02	-2.49	-0.0002	0.9990	H-L	No
Blue Crane Route	41.26	0.18	-0.12	0.00	-7.53	-0.0001	0.9990	H-L	No
Ikwezi	56.07	0.90	-0.02	0.01	-1.00	-0.0001	0.9990	H-L	No
Makana	29.58	-0.38	-0.05	-0.01	-3.01	0.0001	0.5860	L-L	No
Ndlambe	39.07	0.08	0.14	0.00	8.86	0.0000	0.6970	H-H	No
Sundays River Valley	30.49	-0.34	-0.12	-0.01	-7.66	0.0002	0.3000	L-L	No
Baviaans	80.05	2.06	-0.05	0.03	-3.39	-0.0005	0.9990	H-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Kouga	63.85	1.27	-0.15	0.02	-9.10	-0.0008	0.9990	H-L	No
Kou-Kamma	56.07	0.90	-0.10	0.01	-6.32	-0.0004	0.9990	H-L	No
Mbhashe	13.5	-1.16	-0.13	-0.02	-8.10	0.0006	0.0780	L-L	No
Mnquma	17.15	-0.98	-0.07	-0.02	-4.36	0.0003	0.1650	L-L	No
Great Kei	32.62	-0.24	0.07	0.00	4.36	-0.0001	0.9990	L-H	No
Amahlathi	31.81	-0.27	-0.08	0.00	-5.15	0.0001	0.5480	L-L	No
Ngqushwa	32.55	-0.24	-0.11	0.00	-6.74	0.0001	0.5680	L-L	No
Nkonkobe	32.34	-0.25	-0.14	0.00	-8.55	0.0001	0.4430	L-L	No
Nxuba	41.98	0.22	0.04	0.00	2.26	0.0000	0.6190	H-H	No
Inxuba Yethemba	42.41	0.24	-0.16	0.00	-10.05	-0.0002	0.9990	H-L	No
Tsolwana	34.11	-0.16	0.04	0.00	2.36	0.0000	0.9990	L-H	No
Inkwanca	26.59	-0.53	-0.07	-0.01	-4.09	0.0001	0.3890	L-L	No
Lukanji	31.15	-0.31	-0.23	-0.01	-14.59	0.0003	0.2480	L-L	No
Intsika Yethu	18.02	-0.94	-0.31	-0.02	-19.22	0.0012	0.0590	L-L	No
Emalahleni	40.57	0.15	-0.16	0.00	-10.23	-0.0001	0.9990	H-L	No
Engcobo	13.14	-1.18	0.04	-0.02	2.21	-0.0002	0.9990	L-H	No
Sakhisizwe	30.95	-0.32	0.02	-0.01	1.07	0.0000	0.9990	L-H	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Elundini	16.85	-1.00	0.05	-0.02	3.40	-0.0002	0.9990	L-H	No
Senqu	30.39	-0.34	0.26	-0.01	16.37	-0.0004	0.9990	L-H	No
Maletswai	31.88	-0.27	-0.10	0.00	-6.23	0.0001	0.4250	L-L	No
Gariep	32.78	-0.23	0.08	0.00	4.83	-0.0001	0.9990	L-H	No
Ngquza Hill	11.72	-1.24	-0.06	-0.02	-3.64	0.0003	0.1290	L-L	No
Port St Johns	12.29	-1.22	-0.14	-0.02	-9.01	0.0007	0.0440	L-L	Yes
Nyandeni	14.69	-1.10	-0.11	-0.02	-6.61	0.0005	0.0610	L-L	No
Mhlontlo	16.16	-1.03	-0.01	-0.02	-0.86	0.0001	0.5580	L-L	No
King Sabata Dalindyebo	27.01	-0.51	-0.11	-0.01	-6.80	0.0002	0.2130	L-L	No
Matatiele	16.25	-1.03	-0.11	-0.02	-6.91	0.0005	0.0830	L-L	No
Umzimvubu	12.67	-1.20	0.12	-0.02	7.49	-0.0006	0.9990	L-H	No
Mbizana	14.73	-1.10	0.14	-0.02	8.58	-0.0006	0.9990	L-H	No
Ntabankulu	12.52	-1.21	-0.26	-0.02	-16.35	0.0013	0.0400	L-L	Yes
Nelson Mandela Bay	38.57	0.05	-0.21	0.00	-12.97	0.0000	0.9990	H-L	No
Joe Morolong	37.17	-0.02	0.13	0.00	8.34	0.0000	0.9990	L-H	No
Ga-Segonyana	29.88	-0.37	-0.17	-0.01	-10.84	0.0003	0.2330	L-L	No
Gamagara	35.76	-0.08	-0.13	0.00	-8.22	0.0000	0.7550	L-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Richtersveld	73.65	1.75	-0.25	0.03	-15.65	-0.0019	0.9990	H-L	No
Nama Khoi	92.06	2.64	-0.26	0.04	-16.44	-0.0029	0.9990	H-L	No
Kamiesberg	95.46	2.80	-0.03	0.05	-1.64	-0.0003	0.9990	H-L	No
Hantam	93.57	2.71	-0.02	0.04	-1.52	-0.0003	0.9990	H-L	No
Karoo Hoogland	91.17	2.59	0.04	0.04	2.71	0.0005	0.0340	H-H	Yes
Khâi-Ma	66.83	1.42	-0.01	0.02	-0.46	0.0000	0.9990	H-L	No
Ubuntu	82.98	2.20	-0.16	0.04	-9.95	-0.0015	0.9990	H-L	No
Umsobomvu	46.11	0.42	-0.01	0.01	-0.36	0.0000	0.9990	H-L	No
Emthanjeni	58.7	1.02	0.08	0.02	4.91	0.0003	0.0130	H-H	Yes
Kareeberg	116.24	3.80	0.04	0.06	2.25	0.0006	0.0030	H-H	Yes
Renosterberg	73.52	1.74	0.06	0.03	3.87	0.0005	0.0660	H-H	No
Thembelihle	91.45	2.61	0.05	0.04	3.38	0.0006	0.0070	H-H	Yes
Siyathemba	72.5	1.69	0.08	0.03	5.14	0.0006	0.0060	H-H	Yes
Siyancuma	69.15	1.53	-0.02	0.02	-1.28	-0.0001	0.9990	H-L	No
Mier	109.13	3.46	-0.08	0.06	-5.16	-0.0012	0.9990	H-L	No
Kai !Garib	56.91	0.94	0.08	0.02	5.05	0.0003	0.0220	H-H	Yes
//Khara Hais	62.65	1.22	-0.08	0.02	-4.74	-0.0004	0.9990	H-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
!Kheis	98.76	2.96	0.08	0.05	5.01	0.0010	0.0180	H-H	Yes
Tsantsabane	48.82	0.55	-0.05	0.01	-2.87	-0.0001	0.9990	H-L	No
Kgatelopele	42	0.22	0.13	0.00	8.18	0.0001	0.3360	H-H	No
Sol Plaatjie	36.52	-0.05	0.09	0.00	5.50	0.0000	0.9990	L-H	No
Dikgatlong	47.44	0.48	-0.14	0.01	-9.06	-0.0003	0.9990	H-L	No
Magareng	36.68	-0.04	-0.26	0.00	-16.29	0.0000	0.8190	L-L	No
Phokwane	34.81	-0.13	0.13	0.00	8.24	-0.0001	0.9990	L-H	No
Letsemeng	33.4	-0.20	0.18	0.00	11.39	-0.0002	0.9990	L-H	No
Kopanong	33.34	-0.20	-0.12	0.00	-7.79	0.0001	0.4470	L-L	No
Mohokare	30.29	-0.35	-0.08	-0.01	-4.86	0.0001	0.4350	L-L	No
Naledi	24.63	-0.62	-0.10	-0.01	-6.30	0.0003	0.2630	L-L	No
Masilonyana	23.89	-0.66	-0.03	-0.01	-1.82	0.0001	0.3570	L-L	No
Tokologo	32.84	-0.22	-0.02	0.00	-1.33	0.0000	0.8720	L-L	No
Tswelopele	24.53	-0.63	-0.03	-0.01	-1.57	0.0001	0.4270	L-L	No
Matjhabeng	25.2	-0.59	0.10	-0.01	6.49	-0.0003	0.9990	L-H	No
Nala	24.72	-0.62	-0.01	-0.01	-0.49	0.0000	0.8460	L-L	No
Setsoto	29.42	-0.39	0.00	-0.01	-0.25	0.0000	0.9220	L-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Dihlabeng	30.66	-0.33	-0.14	-0.01	-8.66	0.0002	0.2220	L-L	No
Nketoana	30.67	-0.33	-0.03	-0.01	-2.08	0.0000	0.6720	L-L	No
Maluti a Phofung	27.15	-0.50	0.06	-0.01	3.43	-0.0001	0.9990	L-H	No
Phumelela	30.58	-0.33	-0.08	-0.01	-4.74	0.0001	0.3930	L-L	No
Mantsopa	25.35	-0.59	-0.10	-0.01	-6.00	0.0002	0.2570	L-L	No
Moqhaka	26.5	-0.53	0.02	-0.01	1.29	0.0000	0.9990	L-H	No
Ngwathe	27.47	-0.48	-0.15	-0.01	-9.67	0.0003	0.1280	L-L	No
Metsimaholo	25.54	-0.58	-0.06	-0.01	-3.76	0.0001	0.3360	L-L	No
Mafube	30.24	-0.35	-0.25	-0.01	-15.61	0.0004	0.2240	L-L	No
Mangaung	26.05	-0.55	-0.09	-0.01	-5.50	0.0002	0.1610	L-L	No
Umzumbe	15.61	-1.06	-0.10	-0.02	-6.24	0.0004	0.0460	L-L	Yes
UMuziwabantu	33.18	-0.21	-0.28	0.00	-17.68	0.0002	0.3500	L-L	No
Ezingoleni	28.26	-0.45	-0.05	-0.01	-3.31	0.0001	0.4170	L-L	No
Hibiscus Coast	34.65	-0.14	-0.09	0.00	-5.44	0.0001	0.6140	L-L	No
Emnambithi/Ladysmith	28.81	-0.42	-0.12	-0.01	-7.34	0.0002	0.1460	L-L	No
Newcastle	24.07	-0.65	-0.13	-0.01	-8.03	0.0004	0.1120	L-L	No
Emadlangeni	15.6	-1.06	-0.11	-0.02	-6.58	0.0005	0.0490	L-L	Yes

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Dannhauser	29.41	-0.39	0.12	-0.01	7.33	-0.0002	0.9990	L-H	No
Abaqulusi	28.17	-0.45	0.02	-0.01	1.35	0.0000	0.9990	L-H	No
uMhlathuze	22.62	-0.72	-0.21	-0.01	-13.37	0.0007	0.0840	L-L	No
Nkandla	15.78	-1.05	-0.08	-0.02	-4.81	0.0003	0.1060	L-L	No
Maphumulo	16.25	-1.03	-0.05	-0.02	-3.25	0.0002	0.3440	L-L	No
Vulamehlo	12.67	-1.20	-0.09	-0.02	-5.90	0.0005	0.1110	L-L	No
Umdoni	37.85	0.02	-0.23	0.00	-14.71	0.0000	0.9990	H-L	No
uMshwathi	28.72	-0.42	-0.07	-0.01	-4.36	0.0001	0.4190	L-L	No
uMngeni	30.01	-0.36	-0.07	-0.01	-4.58	0.0001	0.5600	L-L	No
Mpofana	27.88	-0.46	0.04	-0.01	2.43	-0.0001	0.9990	L-H	No
Impendle	31.3	-0.30	-0.22	-0.01	-14.01	0.0003	0.3510	L-L	No
The Msunduzi	26.84	-0.51	-0.11	-0.01	-7.18	0.0003	0.1470	L-L	No
Mkhambathini	27.99	-0.46	0.05	-0.01	3.29	-0.0001	0.9990	L-H	No
Richmond	26.77	-0.52	-0.03	-0.01	-2.03	0.0001	0.5450	L-L	No
Indaka	34.83	-0.13	-0.04	0.00	-2.37	0.0000	0.8180	L-L	No
Umtshezi	35.13	-0.11	0.01	0.00	0.44	0.0000	0.9990	L-H	No
Okhahlamba	34.14	-0.16	-0.05	0.00	-3.42	0.0000	0.6320	L-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Imbabazane	33.4	-0.20	-0.08	0.00	-4.98	0.0001	0.5940	L-L	No
Endumeni	32.24	-0.25	-0.07	0.00	-4.23	0.0001	0.5580	L-L	No
Nqutu	33.98	-0.17	-0.12	0.00	-7.73	0.0001	0.4630	L-L	No
Msinga	12.2	-1.22	0.03	-0.02	2.13	-0.0002	0.9990	L-H	No
Umvoti	35.03	-0.12	0.01	0.00	0.63	0.0000	0.9990	L-H	No
eDumbe	27.73	-0.47	0.12	-0.01	7.49	-0.0002	0.9990	L-H	No
UPhongolo	26.58	-0.53	-0.20	-0.01	-12.55	0.0004	0.1770	L-L	No
Nongoma	27.91	-0.46	-0.17	-0.01	-10.41	0.0003	0.2120	L-L	No
Ulundi	27.14	-0.50	-0.09	-0.01	-5.74	0.0002	0.1960	L-L	No
Umhlabuyalingana	10.71	-1.29	-0.23	-0.02	-14.71	0.0013	0.0260	L-L	Yes
Jozini	13.45	-1.16	-0.13	-0.02	-8.07	0.0006	0.0820	L-L	No
The Big 5 False Bay	31.89	-0.27	-0.05	0.00	-3.25	0.0001	0.4670	L-L	No
Hlabisa	33.76	-0.18	-0.07	0.00	-4.51	0.0001	0.5100	L-L	No
Mtubatuba	26.33	-0.54	-0.09	-0.01	-5.83	0.0002	0.1800	L-L	No
Mfolozi	25.87	-0.56	-0.08	-0.01	-4.99	0.0002	0.1700	L-L	No
Ntambanana	33.56	-0.19	-0.05	0.00	-2.95	0.0000	0.6790	L-L	No
uMlalazi	28.78	-0.42	0.09	-0.01	5.90	-0.0002	0.9990	L-H	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Mthonjaneni	32.11	-0.26	0.03	0.00	1.83	0.0000	0.9990	L-H	No
Mandeni	25.4	-0.58	-0.24	-0.01	-15.09	0.0006	0.1210	L-L	No
KwaDukuza	32.28	-0.25	0.06	0.00	3.92	-0.0001	0.9990	L-H	No
Ndwedwe	15.83	-1.05	-0.13	-0.02	-8.23	0.0006	0.0820	L-L	No
Ingwe	15.15	-1.08	0.10	-0.02	6.45	-0.0005	0.9990	L-H	No
Kwa Sani	29.74	-0.37	-0.16	-0.01	-9.80	0.0002	0.2720	L-L	No
Greater Kokstad	27	-0.51	-0.13	-0.01	-8.30	0.0003	0.1500	L-L	No
Ubuhlebezwe	15.41	-1.07	-0.24	-0.02	-14.83	0.0011	0.0350	L-L	Yes
Umzimkhulu	15.21	-1.08	-0.13	-0.02	-8.16	0.0006	0.0240	L-L	Yes
eThekwini	36.43	-0.05	-0.08	0.00	-5.28	0.0000	0.7870	L-L	No
Moretele	28.65	-0.43	0.07	-0.01	4.44	-0.0001	0.9990	L-H	No
Madibeng	28.96	-0.41	-0.13	-0.01	-8.43	0.0002	0.2500	L-L	No
Rustenburg	26.81	-0.52	-0.18	-0.01	-11.22	0.0004	0.1630	L-L	No
Kgetlengrivier	35.04	-0.12	-0.18	0.00	-11.21	0.0001	0.6010	L-L	No
Moses Kotane	28.4	-0.44	-0.26	-0.01	-16.14	0.0005	0.1700	L-L	No
Ratlou	36.24	-0.06	0.08	0.00	4.79	0.0000	0.9990	L-H	No
Tswaing	30.82	-0.32	-0.06	-0.01	-3.44	0.0001	0.6220	L-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Mafikeng	26.88	-0.51	-0.16	-0.01	-9.80	0.0003	0.1350	L-L	No
Ditsobotla	30.71	-0.33	0.13	-0.01	8.35	-0.0002	0.9990	L-H	No
Ramotshere Moiloa	30.36	-0.34	0.13	-0.01	7.83	-0.0002	0.9990	L-H	No
Naledi	37.62	0.01	-0.07	0.00	-4.40	0.0000	0.9990	H-L	No
Mamusa	30.03	-0.36	-0.15	-0.01	-9.63	0.0002	0.1650	L-L	No
Greater Taung	36.28	-0.06	-0.15	0.00	-9.51	0.0000	0.7070	L-L	No
Lekwa-Teemane	28.14	-0.45	-0.08	-0.01	-5.27	0.0002	0.3250	L-L	No
Kagisano/Molopo	35.12	-0.11	-0.11	0.00	-7.01	0.0001	0.7160	L-L	No
Ventersdorp	30.51	-0.34	0.13	-0.01	8.03	-0.0002	0.9990	L-H	No
Tlokwe City Council	32.02	-0.26	0.00	0.00	0.15	0.0000	0.9990	L-H	No
City of Matlosana	27.57	-0.48	-0.02	-0.01	-1.38	0.0000	0.7940	L-L	No
Maquassi Hills	30.98	-0.31	-0.02	-0.01	-1.45	0.0000	0.6680	L-L	No
Emfuleni	25.88	-0.56	0.02	-0.01	1.15	0.0000	0.9990	L-H	No
Midvaal	47.88	0.50	-0.24	0.01	-14.86	-0.0005	0.9990	H-L	No
Lesedi	28.62	-0.43	0.05	-0.01	3.29	-0.0001	0.9990	L-H	No
Mogale City	34.84	-0.13	0.03	0.00	1.75	0.0000	0.9990	L-H	No
Randfontein	32.42	-0.24	0.03	0.00	1.72	0.0000	0.9990	L-H	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Westonaria	30.53	-0.34	0.13	-0.01	8.22	-0.0002	0.9990	L-H	No
Merafong City	28.44	-0.44	-0.02	-0.01	-1.53	0.0000	0.6450	L-L	No
Ekurhuleni	32.07	-0.26	-0.14	0.00	-8.62	0.0002	0.2150	L-L	No
City of Johannesburg	26.57	-0.53	0.03	-0.01	2.05	-0.0001	0.9990	L-H	No
City of Tshwane	28.07	-0.46	0.08	-0.01	5.15	-0.0002	0.9990	L-H	No
Albert Luthuli	27.93	-0.46	-0.01	-0.01	-0.72	0.0000	0.7590	L-L	No
Msukaligwa	29.78	-0.37	0.17	-0.01	10.55	-0.0003	0.9990	L-H	No
Mkhondo	32.73	-0.23	-0.09	0.00	-5.86	0.0001	0.5200	L-L	No
Pixley Ka Seme	31.17	-0.31	0.08	-0.01	4.98	-0.0001	0.9990	L-H	No
Lekwa	26.63	-0.52	0.15	-0.01	9.61	-0.0003	0.9990	L-H	No
Dipaleseng	30.01	-0.36	0.14	-0.01	8.63	-0.0002	0.9990	L-H	No
Govan Mbeki	26.53	-0.53	-0.11	-0.01	-6.60	0.0002	0.1860	L-L	No
Victor Khanye	32.5	-0.24	-0.03	0.00	-1.73	0.0000	0.6850	L-L	No
Emalahleni	31.7	-0.28	-0.11	0.00	-6.93	0.0001	0.3830	L-L	No
Steve Tshwete	29.34	-0.39	-0.19	-0.01	-12.00	0.0003	0.2880	L-L	No
Emakhazeni	30.72	-0.33	0.14	-0.01	8.90	-0.0002	0.9990	L-H	No
Thembisile	21.38	-0.78	0.03	-0.01	1.61	-0.0001	0.9990	L-H	No

Continued on next page

Table A.2 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Dr JS Moroka	23.42	-0.68	-0.03	-0.01	-2.13	0.0001	0.4990	L-L	No
Thaba Chweu	31.43	-0.29	-0.09	-0.01	-5.65	0.0001	0.3780	L-L	No
Mbombela	27.51	-0.48	-0.07	-0.01	-4.13	0.0001	0.4210	L-L	No
Umjindi	28.74	-0.42	-0.07	-0.01	-4.10	0.0001	0.3250	L-L	No
Nkomazi	24.09	-0.65	0.05	-0.01	3.13	-0.0001	0.9990	L-H	No
Bushbuckridge	25.31	-0.59	0.13	-0.01	8.19	-0.0003	0.9990	L-H	No
Greater Giyani	26.11	-0.55	-0.02	-0.01	-1.47	0.0001	0.5700	L-L	No
Greater Letaba	27.17	-0.50	-0.09	-0.01	-5.71	0.0002	0.2100	L-L	No
Greater Tzaneen	26.62	-0.53	-0.20	-0.01	-12.58	0.0004	0.1800	L-L	No
Ba-Phalaborwa	26.84	-0.51	0.10	-0.01	6.49	-0.0002	0.9990	L-H	No
Maruleng	27	-0.51	-0.20	-0.01	-12.37	0.0004	0.1200	L-L	No
Mutale	13.6	-1.15	-0.23	-0.02	-14.25	0.0011	0.0610	L-L	No
Thulamela	25.77	-0.57	-0.04	-0.01	-2.30	0.0001	0.4120	L-L	No
Musina	22.81	-0.71	-0.20	-0.01	-12.71	0.0006	0.0960	L-L	No
Makhado	27.12	-0.50	-0.09	-0.01	-5.70	0.0002	0.2220	L-L	No
Blouberg	28.47	-0.44	-0.28	-0.01	-17.40	0.0005	0.1510	L-L	No
Aganang	32.07	-0.26	-0.20	0.00	-12.79	0.0002	0.3910	L-L	No

Continued on next page

Table A.2 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Molemole	23.04	-0.70	-0.24	-0.01	-14.88	0.0007	0.0650	L-L	No
Polokwane	27.37	-0.49	0.04	-0.01	2.41	-0.0001	0.9990	L-H	No
Lepele-Nkumpi	28.29	-0.44	-0.17	-0.01	-10.87	0.0003	0.2320	L-L	No
Thabazimbi	27.99	-0.46	-0.26	-0.01	-16.56	0.0005	0.1550	L-L	No
Lephalale	21.92	-0.75	-0.24	-0.01	-15.04	0.0008	0.0790	L-L	No
Mookgopong	25.06	-0.60	-0.19	-0.01	-11.61	0.0005	0.1380	L-L	No
Modimolle	23.56	-0.67	-0.14	-0.01	-8.50	0.0004	0.1540	L-L	No
Bela-Bela	25.93	-0.56	-0.29	-0.01	-18.03	0.0007	0.1030	L-L	No
Mogalakwena	28.89	-0.42	-0.08	-0.01	-5.29	0.0001	0.3290	L-L	No
Ephraim Mogale	27.95	-0.46	0.00	-0.01	0.28	0.0000	0.9990	L-H	No
Elias Motsoaledi	28.63	-0.43	-0.02	-0.01	-1.33	0.0000	0.7750	L-L	No
Makhuduthamaga	28.84	-0.42	-0.29	-0.01	-18.48	0.0005	0.1980	L-L	No
Fetakgomo	27.93	-0.46	-0.07	-0.01	-4.30	0.0001	0.5850	L-L	No
Greater Tubatse	24.69	-0.62	0.16	-0.01	9.82	-0.0004	0.9990	L-H	No

Table A.3: Classification of spatial autocorrelation based on negative exponential spatial weight, $\alpha = 1$.

Municipality	**Ischaemic**	z	f	f^*	z^*	**LISA**	P value	**Type**	**Significant**
Matzikama	80.99	2.101	0.051	0.024	4.386	0.0005	0.006	H-H	Yes
Cederberg	74.93	1.808	0.034	0.021	2.942	0.0003	0.032	H-H	Yes
Bergrivier	77.52	1.933	-0.140	0.023	-12.035	-0.0012	0.999	H-L	No
Saldanha Bay	61.7	1.169	-0.084	0.014	-7.163	-0.0004	0.999	H-L	No
Swartland	68.23	1.484	-0.119	0.017	-10.172	-0.0008	0.999	H-L	No
Witzenberg	57.4	0.962	-0.002	0.011	-0.183	0.0000	0.999	H-L	No
Drakenstein	63.95	1.278	-0.093	0.015	-8.007	-0.0005	0.999	H-L	No
Stellenbosch	58.22	1.001	0.076	0.012	6.547	0.0003	0.035	H-H	Yes
Breede Valley	62.13	1.190	0.082	0.014	6.997	0.0004	0.014	H-H	Yes
Langeberg	70.48	1.593	-0.088	0.019	-7.574	-0.0006	0.999	H-L	No
Swellendam	78.05	1.959	0.088	0.023	7.552	0.0007	0.004	H-H	Yes
Theewaterskloof	59.33	1.055	0.081	0.012	6.926	0.0004	0.027	H-H	Yes
Overstrand	58.59	1.019	-0.071	0.012	-6.092	-0.0003	0.999	H-L	No
Cape Agulhas	83.21	2.208	0.033	0.026	2.804	0.0003	0.041	H-H	Yes
Kannaland	90.67	2.568	0.084	0.030	7.238	0.0009	0.001	H-H	Yes
Hessequa	92.48	2.655	-0.018	0.031	-1.555	-0.0002	0.999	H-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Mossel Bay	64.27	1.293	-0.099	0.015	-8.499	-0.0005	0.999	H-L	No
George	60.08	1.091	0.063	0.013	5.432	0.0003	0.044	H-H	Yes
Oudtshoorn	83.16	2.205	0.080	0.026	6.876	0.0008	0.002	H-H	Yes
Bitou	54.09	0.802	0.100	0.009	8.577	0.0003	0.021	H-H	Yes
Knysna	66.87	1.419	0.088	0.017	7.511	0.0005	0.010	H-H	Yes
Laingsburg	85.9	2.338	-0.100	0.027	-8.544	-0.0010	0.999	H-L	No
Prince Albert	87.8	2.430	-0.080	0.028	-6.877	-0.0008	0.999	H-L	No
Beaufort West	73.75	1.751	-0.096	0.020	-8.267	-0.0007	0.999	H-L	No
City of Cape Town	49.53	0.582	-0.127	0.007	-10.853	-0.0003	0.999	H-L	No
Buffalo City	32.58	-0.237	-0.097	-0.003	-8.339	0.0001	0.220	L-L	No
Camdeboo	66.47	1.399	-0.014	0.016	-1.222	-0.0001	0.999	H-L	No
Blue Crane Route	41.26	0.182	-0.098	0.002	-8.430	-0.0001	0.999	H-L	No
Ikwezi	56.07	0.897	-0.055	0.010	-4.758	-0.0002	0.999	H-L	No
Makana	29.58	-0.382	-0.078	-0.004	-6.653	0.0001	0.195	L-L	No
Ndlambe	39.07	0.077	0.106	0.001	9.061	0.0000	0.567	H-H	No
Sundays River Valley	30.49	-0.338	-0.115	-0.004	-9.902	0.0002	0.130	L-L	No
Baviaans	80.05	2.055	-0.102	0.024	-8.789	-0.0009	0.999	H-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Kouga	63.85	1.273	-0.104	0.015	-8.900	-0.0006	0.999	H-L	No
Kou-Kamma	56.07	0.897	-0.065	0.010	-5.610	-0.0003	0.999	H-L	No
Mbhashe	13.5	-1.158	-0.109	-0.014	-9.335	0.0005	0.009	L-L	Yes
Mnquma	17.15	-0.982	-0.089	-0.011	-7.669	0.0004	0.036	L-L	Yes
Great Kei	32.62	-0.235	-0.006	-0.003	-0.479	0.0000	0.879	L-L	No
Amahlathi	31.81	-0.274	-0.092	-0.003	-7.868	0.0001	0.241	L-L	No
Ngqushwa	32.55	-0.238	-0.079	-0.003	-6.748	0.0001	0.285	L-L	No
Nkonkobe	32.34	-0.248	-0.117	-0.003	-10.028	0.0001	0.173	L-L	No
Nxuba	41.98	0.217	0.046	0.003	3.972	0.0000	0.410	H-H	No
Inxuba Yethemba	42.41	0.238	-0.125	0.003	-10.735	-0.0001	0.999	H-L	No
Tsolwana	34.11	-0.163	-0.027	-0.002	-2.332	0.0000	0.716	L-L	No
Inkwanca	26.59	-0.526	-0.082	-0.006	-6.998	0.0002	0.118	L-L	No
Lukanji	31.15	-0.306	-0.116	-0.004	-9.985	0.0002	0.153	L-L	No
Intsika Yethu	18.02	-0.940	-0.091	-0.011	-7.772	0.0004	0.026	L-L	Yes
Emalahleni	40.57	0.149	-0.089	0.002	-7.671	-0.0001	0.999	H-L	No
Engcobo	13.14	-1.176	0.020	-0.014	1.706	-0.0001	0.999	L-H	No
Sakhisizwe	30.95	-0.316	0.024	-0.004	2.076	0.0000	0.999	L-H	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Elundini	16.85	-0.996	0.015	-0.012	1.260	-0.0001	0.999	L-H	No
Senqu	30.39	-0.343	0.100	-0.004	8.597	-0.0001	0.999	L-H	No
Maletswai	31.88	-0.271	-0.097	-0.003	-8.359	0.0001	0.205	L-L	No
Gariep	32.78	-0.227	-0.003	-0.003	-0.263	0.0000	0.930	L-L	No
Ngquza Hill	11.72	-1.244	-0.066	-0.015	-5.697	0.0004	0.033	L-L	Yes
Port St Johns	12.29	-1.216	-0.083	-0.014	-7.143	0.0004	0.031	L-L	Yes
Nyandeni	14.69	-1.100	-0.070	-0.013	-6.007	0.0003	0.049	L-L	Yes
Mhlontlo	16.16	-1.029	-0.004	-0.012	-0.353	0.0000	0.668	L-L	No
King Sabata Dalindyebo	27.01	-0.506	-0.087	-0.006	-7.442	0.0002	0.103	L-L	No
Matatiele	16.25	-1.025	-0.075	-0.012	-6.393	0.0003	0.054	L-L	No
Umzimvubu	12.67	-1.198	0.113	-0.014	9.689	-0.0006	0.999	L-H	No
Mbizana	14.73	-1.099	0.112	-0.013	9.638	-0.0005	0.999	L-H	No
Ntabankulu	12.52	-1.205	-0.088	-0.014	-7.530	0.0005	0.021	L-L	Yes
Nelson Mandela Bay	38.57	0.052	-0.120	0.001	-10.326	0.0000	0.999	H-L	No
Joe Morolong	37.17	-0.015	0.070	0.000	6.041	0.0000	0.999	L-H	No
Ga-Segonyana	29.88	-0.367	-0.112	-0.004	-9.566	0.0002	0.117	L-L	No
Gamagara	35.76	-0.083	-0.111	-0.001	-9.503	0.0000	0.519	L-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Richtersveld	73.65	1.746	-0.140	0.020	-11.993	-0.0010	0.999	H-L	No
Nama Khoi	92.06	2.635	-0.135	0.031	-11.567	-0.0015	0.999	H-L	No
Kamiesberg	95.46	2.799	-0.019	0.033	-1.648	-0.0002	0.999	H-L	No
Hantam	93.57	2.708	-0.037	0.032	-3.143	-0.0004	0.999	H-L	No
Karoo Hoogland	91.17	2.592	0.019	0.030	1.602	0.0002	0.059	H-H	No
Khâi-Ma	66.83	1.417	0.002	0.017	0.151	0.0000	0.763	H-H	No
Ubuntu	82.98	2.197	-0.131	0.026	-11.212	-0.0012	0.999	H-L	No
Umsobomvu	46.11	0.416	-0.009	0.005	-0.757	0.0000	0.999	H-L	No
Emthanjeni	58.7	1.024	0.069	0.012	5.918	0.0003	0.020	H-H	Yes
Kareeberg	116.24	3.803	0.060	0.044	5.111	0.0010	0.000	H-H	Yes
Renosterberg	73.52	1.740	0.088	0.020	7.591	0.0007	0.004	H-H	Yes
Thembelihle	91.45	2.605	0.057	0.030	4.903	0.0006	0.001	H-H	Yes
Siyathemba	72.5	1.691	0.088	0.020	7.524	0.0006	0.001	H-H	Yes
Siyancuma	69.15	1.529	-0.001	0.018	-0.073	0.0000	0.999	H-L	No
Mier	109.13	3.459	-0.087	0.040	-7.500	-0.0013	0.999	H-L	No
Kai !Garib	56.91	0.938	0.080	0.011	6.902	0.0003	0.018	H-H	Yes
//Khara Hais	62.65	1.215	-0.054	0.014	-4.606	-0.0003	0.999	H-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
!Kheis	98.76	2.959	0.066	0.034	5.693	0.0008	0.000	H-H	Yes
Tsantsabane	48.82	0.547	-0.017	0.006	-1.433	0.0000	0.999	H-L	No
Kgatelopele	42	0.218	0.079	0.003	6.782	0.0001	0.272	H-H	No
Sol Plaatjie	36.52	-0.047	0.068	-0.001	5.836	0.0000	0.999	L-H	No
Dikgatlong	47.44	0.481	-0.109	0.006	-9.320	-0.0002	0.999	H-L	No
Magareng	36.68	-0.039	-0.124	0.000	-10.620	0.0000	0.723	L-L	No
Phokwane	34.81	-0.129	0.110	-0.002	9.393	-0.0001	0.999	L-H	No
Letsemeng	33.4	-0.197	0.110	-0.002	9.457	-0.0001	0.999	L-H	No
Kopanong	33.34	-0.200	-0.105	-0.002	-8.983	0.0001	0.251	L-L	No
Mohokare	30.29	-0.347	-0.032	-0.004	-2.772	0.0000	0.381	L-L	No
Naledi	24.63	-0.621	-0.079	-0.007	-6.795	0.0002	0.096	L-L	No
Masilonyana	23.89	-0.656	-0.052	-0.008	-4.488	0.0001	0.131	L-L	No
Tokologo	32.84	-0.224	-0.087	-0.003	-7.492	0.0001	0.320	L-L	No
Tswelopele	24.53	-0.626	-0.002	-0.007	-0.158	0.0000	0.880	L-L	No
Matjhabeng	25.2	-0.593	0.015	-0.007	1.297	0.0000	0.999	L-H	No
Nala	24.72	-0.616	-0.061	-0.007	-5.263	0.0002	0.140	L-L	No
Setsoto	29.42	-0.390	-0.034	-0.005	-2.926	0.0001	0.373	L-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Dihlabeng	30.66	-0.329	-0.099	-0.004	-8.454	0.0001	0.151	L-L	No
Nketoana	30.67	-0.329	-0.014	-0.004	-1.216	0.0000	0.632	L-L	No
Maluti a Phofung	27.15	-0.499	0.039	-0.006	3.323	-0.0001	0.999	L-H	No
Phumelela	30.58	-0.333	-0.069	-0.004	-5.879	0.0001	0.269	L-L	No
Mantsopa	25.35	-0.586	-0.084	-0.007	-7.198	0.0002	0.102	L-L	No
Moqhaka	26.5	-0.530	-0.012	-0.006	-0.987	0.0000	0.610	L-L	No
Ngwathe	27.47	-0.483	-0.106	-0.006	-9.131	0.0002	0.071	L-L	No
Metsimaholo	25.54	-0.577	-0.029	-0.007	-2.529	0.0001	0.267	L-L	No
Mafube	30.24	-0.350	-0.122	-0.004	-10.430	0.0002	0.123	L-L	No
Mangaung	26.05	-0.552	-0.042	-0.006	-3.633	0.0001	0.177	L-L	No
Umzumbe	15.61	-1.056	-0.060	-0.012	-5.106	0.0003	0.035	L-L	Yes
UMuziwabantu	33.18	-0.208	-0.125	-0.002	-10.715	0.0001	0.215	L-L	No
Ezingoleni	28.26	-0.445	-0.047	-0.005	-4.000	0.0001	0.223	L-L	No
Hibiscus Coast	34.65	-0.137	-0.086	-0.002	-7.417	0.0001	0.478	L-L	No
Emnambithi/Ladysmith	28.81	-0.419	-0.057	-0.005	-4.923	0.0001	0.170	L-L	No
Newcastle	24.07	-0.648	-0.076	-0.008	-6.562	0.0002	0.086	L-L	No
Emadlangeni	15.6	-1.057	-0.062	-0.012	-5.309	0.0003	0.029	L-L	Yes

Continued on next page

Table A.3 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Dannhauser	29.41	-0.390	0.105	-0.005	9.044	-0.0002	0.999	L-H	No
Abaqulusi	28.17	-0.450	-0.026	-0.005	-2.229	0.0000	0.450	L-L	No
uMhlathuze	22.62	-0.718	-0.083	-0.008	-7.140	0.0003	0.054	L-L	No
Nkandla	15.78	-1.048	-0.087	-0.012	-7.479	0.0004	0.030	L-L	Yes
Maphumulo	16.25	-1.025	-0.061	-0.012	-5.265	0.0003	0.056	L-L	No
Vulamehlo	12.67	-1.198	-0.078	-0.014	-6.674	0.0004	0.024	L-L	Yes
Umdoni	37.85	0.017	-0.120	0.000	-10.320	0.0000	0.999	H-L	No
uMshwathi	28.72	-0.423	-0.056	-0.005	-4.825	0.0001	0.207	L-L	No
uMngeni	30.01	-0.361	-0.083	-0.004	-7.104	0.0001	0.177	L-L	No
Mpofana	27.88	-0.464	0.044	-0.005	3.792	-0.0001	0.999	L-H	No
Impendle	31.3	-0.299	-0.115	-0.003	-9.846	0.0001	0.162	L-L	No
The Msunduzi	26.84	-0.514	-0.109	-0.006	-9.377	0.0002	0.070	L-L	No
Mkhambathini	27.99	-0.458	-0.011	-0.005	-0.926	0.0000	0.700	L-L	No
Richmond	26.77	-0.517	-0.064	-0.006	-5.452	0.0001	0.164	L-L	No
Indaka	34.83	-0.128	-0.069	-0.001	-5.877	0.0000	0.547	L-L	No
Umtshezi	35.13	-0.114	-0.069	-0.001	-5.940	0.0000	0.589	L-L	No
Okhahlamba	34.14	-0.162	-0.032	-0.002	-2.788	0.0000	0.620	L-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Imbabazane	33.4	-0.197	-0.074	-0.002	-6.318	0.0001	0.394	L-L	No
Endumeni	32.24	-0.253	-0.076	-0.003	-6.500	0.0001	0.343	L-L	No
Nqutu	33.98	-0.169	-0.076	-0.002	-6.528	0.0001	0.376	L-L	No
Msinga	12.2	-1.221	-0.059	-0.014	-5.025	0.0003	0.050	L-L	Yes
Umvoti	35.03	-0.119	-0.051	-0.001	-4.363	0.0000	0.641	L-L	No
eDumbe	27.73	-0.471	0.104	-0.005	8.927	-0.0002	0.999	L-H	No
UPhongolo	26.58	-0.526	-0.115	-0.006	-9.861	0.0003	0.066	L-L	No
Nongoma	27.91	-0.462	-0.121	-0.005	-10.360	0.0002	0.071	L-L	No
Ulundi	27.14	-0.500	-0.103	-0.006	-8.813	0.0002	0.092	L-L	No
Umhlabuyalingana	10.71	-1.293	-0.114	-0.015	-9.807	0.0006	0.006	L-L	Yes
Jozini	13.45	-1.160	-0.105	-0.014	-9.020	0.0005	0.016	L-L	Yes
The Big 5 False Bay	31.89	-0.270	-0.061	-0.003	-5.267	0.0001	0.330	L-L	No
Hlabisa	33.76	-0.180	-0.067	-0.002	-5.757	0.0001	0.404	L-L	No
Mtubatuba	26.33	-0.539	-0.056	-0.006	-4.833	0.0001	0.128	L-L	No
Mfolozi	25.87	-0.561	-0.044	-0.007	-3.742	0.0001	0.168	L-L	No
Ntambanana	33.56	-0.190	-0.023	-0.002	-1.940	0.0000	0.669	L-L	No
uMlalazi	28.78	-0.420	0.087	-0.005	7.490	-0.0002	0.999	L-H	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Mthonjaneni	32.11	-0.260	0.031	-0.003	2.668	0.0000	0.999	L-H	No
Mandeni	25.4	-0.584	-0.119	-0.007	-10.166	0.0003	0.056	L-L	No
KwaDukuza	32.28	-0.251	0.057	-0.003	4.867	-0.0001	0.999	L-H	No
Ndwedwe	15.83	-1.046	-0.109	-0.012	-9.314	0.0005	0.011	L-L	Yes
Ingwe	15.15	-1.078	0.028	-0.013	2.444	-0.0001	0.999	L-H	No
Kwa Sani	29.74	-0.374	-0.074	-0.004	-6.381	0.0001	0.181	L-L	No
Greater Kokstad	27	-0.506	-0.083	-0.006	-7.091	0.0002	0.118	L-L	No
Ubuhlebezwe	15.41	-1.066	-0.118	-0.012	-10.083	0.0005	0.017	L-L	Yes
Umzimkhulu	15.21	-1.075	-0.089	-0.013	-7.640	0.0004	0.022	L-L	Yes
eThekwini	36.43	-0.051	-0.095	-0.001	-8.171	0.0000	0.687	L-L	No
Moretele	28.65	-0.427	0.027	-0.005	2.284	0.0000	0.999	L-H	No
Madibeng	28.96	-0.411	-0.115	-0.005	-9.850	0.0002	0.098	L-L	No
Rustenburg	26.81	-0.515	-0.121	-0.006	-10.337	0.0003	0.066	L-L	No
Kgetlengrivier	35.04	-0.118	-0.078	-0.001	-6.701	0.0000	0.517	L-L	No
Moses Kotane	28.4	-0.439	-0.120	-0.005	-10.318	0.0002	0.081	L-L	No
Ratlou	36.24	-0.060	0.033	-0.001	2.792	0.0000	0.999	L-H	No
Tswaing	30.82	-0.322	-0.053	-0.004	-4.539	0.0001	0.314	L-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Mafikeng	26.88	-0.512	-0.109	-0.006	-9.307	0.0002	0.069	L-L	No
Ditsobotla	30.71	-0.327	0.104	-0.004	8.916	-0.0001	0.999	L-H	No
Ramotshere Moiloa	30.36	-0.344	0.103	-0.004	8.867	-0.0002	0.999	L-H	No
Naledi	37.62	0.007	-0.022	0.000	-1.855	0.0000	0.999	H-L	No
Mamusa	30.03	-0.360	-0.109	-0.004	-9.355	0.0002	0.124	L-L	No
Greater Taung	36.28	-0.058	-0.114	-0.001	-9.753	0.0000	0.635	L-L	No
Lekwa-Teemane	28.14	-0.451	-0.076	-0.005	-6.477	0.0001	0.171	L-L	No
Kagisano/Molopo	35.12	-0.114	-0.064	-0.001	-5.517	0.0000	0.584	L-L	No
Ventersdorp	30.51	-0.337	0.105	-0.004	9.015	-0.0002	0.999	L-H	No
Tlokwe City Council	32.02	-0.264	-0.039	-0.003	-3.381	0.0000	0.436	L-L	No
City of Matlosana	27.57	-0.479	-0.067	-0.006	-5.732	0.0001	0.172	L-L	No
Maquassi Hills	30.98	-0.314	-0.026	-0.004	-2.219	0.0000	0.513	L-L	No
Emfuleni	25.88	-0.560	0.030	-0.007	2.605	-0.0001	0.999	L-H	No
Midvaal	47.88	0.502	-0.120	0.006	-10.278	-0.0003	0.999	H-L	No
Lesedi	28.62	-0.428	0.073	-0.005	6.236	-0.0001	0.999	L-H	No
Mogale City	34.84	-0.128	-0.060	-0.001	-5.175	0.0000	0.562	L-L	No
Randfontein	32.42	-0.244	-0.025	-0.003	-2.131	0.0000	0.631	L-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Westonaria	30.53	-0.336	0.106	-0.004	9.054	-0.0002	0.999	L-H	No
Merafong City	28.44	-0.436	-0.034	-0.005	-2.920	0.0001	0.320	L-L	No
Ekurhuleni	32.07	-0.261	-0.082	-0.003	-6.997	0.0001	0.196	L-L	No
City of Johannesburg	26.57	-0.527	0.035	-0.006	2.995	-0.0001	0.999	L-H	No
City of Tshwane	28.07	-0.455	0.064	-0.005	5.471	-0.0001	0.999	L-H	No
Albert Luthuli	27.93	-0.461	0.001	-0.005	0.087	0.0000	0.999	L-H	No
Msukaligwa	29.78	-0.372	0.107	-0.004	9.205	-0.0002	0.999	L-H	No
Mkhondo	32.73	-0.229	-0.095	-0.003	-8.136	0.0001	0.288	L-L	No
Pixley Ka Seme	31.17	-0.305	0.057	-0.004	4.907	-0.0001	0.999	L-H	No
Lekwa	26.63	-0.524	0.110	-0.006	9.456	-0.0002	0.999	L-H	No
Dipaleseng	30.01	-0.361	0.109	-0.004	9.331	-0.0002	0.999	L-H	No
Govan Mbeki	26.53	-0.529	-0.091	-0.006	-7.779	0.0002	0.099	L-L	No
Victor Khanye	32.5	-0.241	-0.053	-0.003	-4.553	0.0001	0.421	L-L	No
Emalahleni	31.7	-0.279	-0.109	-0.003	-9.317	0.0001	0.168	L-L	No
Steve Tshwete	29.34	-0.393	-0.114	-0.005	-9.808	0.0002	0.105	L-L	No
Emakhazeni	30.72	-0.326	0.107	-0.004	9.176	-0.0001	0.999	L-H	No
Thembisile	21.38	-0.777	0.039	-0.009	3.318	-0.0001	0.999	L-H	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Dr JS Moroka	23.42	-0.679	-0.079	-0.008	-6.814	0.0002	0.076	L-L	No
Thaba Chweu	31.43	-0.292	-0.071	-0.003	-6.062	0.0001	0.261	L-L	No
Mbombela	27.51	-0.482	-0.065	-0.006	-5.617	0.0001	0.138	L-L	No
Umjindi	28.74	-0.422	-0.026	-0.005	-2.245	0.0000	0.390	L-L	No
Nkomazi	24.09	-0.647	0.042	-0.008	3.561	-0.0001	0.999	L-H	No
Bushbuckridge	25.31	-0.588	0.031	-0.007	2.653	-0.0001	0.999	L-H	No
Greater Giyani	26.11	-0.549	-0.035	-0.006	-3.033	0.0001	0.224	L-L	No
Greater Letaba	27.17	-0.498	-0.044	-0.006	-3.773	0.0001	0.181	L-L	No
Greater Tzaneen	26.62	-0.525	-0.104	-0.006	-8.960	0.0002	0.088	L-L	No
Ba-Phalaborwa	26.84	-0.514	0.081	-0.006	6.938	-0.0002	0.999	L-H	No
Maruleng	27	-0.506	-0.120	-0.006	-10.286	0.0003	0.066	L-L	No
Mutale	13.6	-1.153	-0.099	-0.013	-8.530	0.0005	0.016	L-L	Yes
Thulamela	25.77	-0.566	-0.098	-0.007	-8.438	0.0002	0.068	L-L	No
Musina	22.81	-0.708	-0.115	-0.008	-9.876	0.0003	0.047	L-L	Yes
Makhado	27.12	-0.501	-0.097	-0.006	-8.288	0.0002	0.090	L-L	No
Blouberg	28.47	-0.435	-0.122	-0.005	-10.490	0.0002	0.076	L-L	No
Aganang	32.07	-0.262	-0.116	-0.003	-9.937	0.0001	0.176	L-L	No

Continued on next page

Table A.3 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Molemole	23.04	-0.697	-0.117	-0.008	-10.020	0.0003	0.048	L-L	Yes
Polokwane	27.37	-0.488	0.029	-0.006	2.524	-0.0001	0.999	L-H	No
Lepele-Nkumpi	28.29	-0.444	-0.117	-0.005	-9.998	0.0002	0.081	L-L	No
Thabazimbi	27.99	-0.458	-0.094	-0.005	-8.071	0.0002	0.104	L-L	No
Lephalale	21.92	-0.751	-0.119	-0.009	-10.228	0.0004	0.031	L-L	Yes
Mookgopong	25.06	-0.600	-0.096	-0.007	-8.273	0.0002	0.070	L-L	No
Modimolle	23.56	-0.672	-0.072	-0.008	-6.149	0.0002	0.086	L-L	No
Bela-Bela	25.93	-0.558	-0.101	-0.007	-8.689	0.0002	0.086	L-L	No
Mogalakwena	28.89	-0.415	-0.111	-0.005	-9.540	0.0002	0.109	L-L	No
Ephraim Mogale	27.95	-0.460	-0.056	-0.005	-4.766	0.0001	0.182	L-L	No
Elias Motsoaledi	28.63	-0.427	-0.085	-0.005	-7.317	0.0002	0.152	L-L	No
Makhuduthamaga	28.84	-0.417	-0.109	-0.005	-9.332	0.0002	0.090	L-L	No
Fetakgomo	27.93	-0.461	-0.071	-0.005	-6.086	0.0001	0.170	L-L	No
Greater Tubatse	24.69	-0.618	0.116	-0.007	9.985	-0.0003	0.999	L-H	No

Table A.4: Classification of spatial autocorrelation based negative exponential spatial weight, $\alpha = 2$.

Municipality	**Ischaemic**	z	f	f^*	z^*	**LISA**	P value	**Type**	**Significant**
Matzikama	80.99	2.101	0.085	0.044	4.098	0.0008	0.008	H-H	Yes
Cederberg	74.93	1.808	0.053	0.038	2.547	0.0004	0.034	H-H	Yes
Bergrivier	77.52	1.933	-0.236	0.04	-11.335	-0.0019	0.999	H-L	No
Saldanha Bay	61.7	1.169	-0.109	0.024	-5.269	-0.0005	0.999	H-L	No
Swartland	68.23	1.484	-0.177	0.031	-8.541	-0.0011	0.999	H-L	No
Witzenberg	57.4	0.962	0.017	0.02	0.833	0.0001	0.531	H-H	No
Drakenstein	63.95	1.278	-0.121	0.027	-5.804	-0.0007	0.999	H-L	No
Stellenbosch	58.22	1.001	0.142	0.021	6.815	0.0006	0.023	H-H	Yes
Breede Valley	62.13	1.19	0.144	0.025	6.911	0.0007	0.013	H-H	Yes
Langeberg	70.48	1.593	-0.119	0.033	-5.744	-0.0008	0.999	H-L	No
Swellendam	78.05	1.959	0.124	0.041	5.964	0.0010	0.004	H-H	Yes
Theewaterskloof	59.33	1.055	0.143	0.022	6.883	0.0006	0.023	H-H	Yes
Overstrand	58.59	1.019	-0.084	0.021	-4.046	-0.0004	0.999	H-L	No
Cape Agulhas	83.21	2.208	0.074	0.046	3.562	0.0007	0.024	H-H	Yes
Kannaland	90.67	2.568	0.123	0.053	5.913	0.0013	0.003	H-H	Yes
Hessequa	92.48	2.655	-0.014	0.055	-0.696	-0.0002	0.999	H-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Mossel Bay	64.27	1.293	-0.127	0.027	-6.091	-0.0007	0.999	H-L	No
George	60.08	1.091	0.123	0.023	5.934	0.0006	0.02	H-H	Yes
Oudtshoorn	83.16	2.205	0.11	0.046	5.296	0.0010	0.007	H-H	Yes
Bitou	54.09	0.802	0.148	0.017	7.112	0.0005	0.037	H-H	Yes
Knysna	66.87	1.419	0.117	0.029	5.65	0.0007	0.024	H-H	Yes
Laingsburg	85.9	2.338	-0.147	0.049	-7.067	-0.0015	0.999	H-L	No
Prince Albert	87.8	2.43	-0.13	0.05	-6.253	-0.0013	0.999	H-L	No
Beaufort West	73.75	1.751	-0.135	0.036	-6.473	-0.0010	0.999	H-L	No
City of Cape Town	49.53	0.582	-0.223	0.012	-10.717	-0.0006	0.999	H-L	No
Buffalo City	32.58	-0.237	-0.182	-0.005	-8.758	0.0002	0.22	L-L	No
Camdeboo	66.47	1.399	-0.034	0.029	-1.644	-0.0002	0.999	H-L	No
Blue Crane Route	41.26	0.182	-0.158	0.004	-7.593	-0.0001	0.999	H-L	No
Ikwezi	56.07	0.897	-0.074	0.019	-3.561	-0.0003	0.999	H-L	No
Makana	29.58	-0.382	-0.1	-0.008	-4.835	0.0002	0.328	L-L	No
Ndlambe	39.07	0.077	0.156	0.002	7.509	0.0001	0.602	H-H	No
Sundays River Valley	30.49	-0.338	-0.187	-0.007	-8.982	0.0003	0.171	L-L	No
Baviaans	80.05	2.055	-0.151	0.043	-7.282	-0.0013	0.999	H-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Kouga	63.85	1.273	-0.15	0.026	-7.231	-0.0008	0.999	H-L	No
Kou-Kamma	56.07	0.897	-0.116	0.019	-5.582	-0.0004	0.999	H-L	No
Mbhashe	13.5	-1.158	-0.183	-0.024	-8.791	0.0009	0.024	L-L	Yes
Mnquma	17.15	-0.982	-0.125	-0.02	-5.993	0.0005	0.043	L-L	Yes
Great Kei	32.62	-0.235	0.007	-0.005	0.341	0.0000	0.999	L-H	No
Amahlathi	31.81	-0.274	-0.133	-0.006	-6.384	0.0002	0.318	L-L	No
Ngqushwa	32.55	-0.238	-0.122	-0.005	-5.847	0.0001	0.388	L-L	No
Nkonkobe	32.34	-0.248	-0.211	-0.005	-10.155	0.0002	0.214	L-L	No
Nxuba	41.98	0.217	0.077	0.005	3.727	0.0001	0.406	H-H	No
Inxuba Yethemba	42.41	0.238	-0.22	0.005	-10.596	-0.0002	0.999	H-L	No
Tsolwana	34.11	-0.163	-0.034	-0.003	-1.634	0.0000	0.807	L-L	No
Inkwanca	26.59	-0.526	-0.109	-0.011	-5.257	0.0002	0.19	L-L	No
Lukanji	31.15	-0.306	-0.217	-0.006	-10.442	0.0003	0.177	L-L	No
Intsika Yethu	18.02	-0.94	-0.175	-0.02	-8.415	0.0007	0.037	L-L	Yes
Emalahleni	40.57	0.149	-0.125	0.003	-6.013	-0.0001	0.999	H-L	No
Engcobo	13.14	-1.176	0.03	-0.024	1.422	-0.0001	0.999	L-H	No
Sakhisizwe	30.95	-0.316	0.037	-0.007	1.773	0.0000	0.999	L-H	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Elundini	16.85	-0.996	0.033	-0.021	1.573	-0.0001	0.999	L-H	No
Senqu	30.39	-0.343	0.177	-0.007	8.511	-0.0003	0.999	L-H	No
Maletswai	31.88	-0.271	-0.149	-0.006	-7.192	0.0002	0.266	L-L	No
Gariep	32.78	-0.227	0.015	-0.005	0.699	0.0000	0.999	L-H	No
Ngquza Hill	11.72	-1.244	-0.076	-0.026	-3.666	0.0004	0.076	L-L	No
Port St Johns	12.29	-1.216	-0.16	-0.025	-7.723	0.0008	0.019	L-L	Yes
Nyandeni	14.69	-1.1	-0.088	-0.023	-4.235	0.0004	0.089	L-L	No
Mhlontlo	16.16	-1.029	-0.014	-0.021	-0.672	0.0001	0.542	L-L	No
King Sabata Dalindyebo	27.01	-0.506	-0.118	-0.011	-5.682	0.0003	0.152	L-L	No
Matatiele	16.25	-1.025	-0.096	-0.021	-4.643	0.0004	0.084	L-L	No
Umzimvubu	12.67	-1.198	0.177	-0.025	8.511	-0.0009	0.999	L-H	No
Mbizana	14.73	-1.099	0.175	-0.023	8.431	-0.0008	0.999	L-H	No
Ntabankulu	12.52	-1.205	-0.168	-0.025	-8.07	0.0009	0.021	L-L	Yes
Nelson Mandela Bay	38.57	0.052	-0.196	0.001	-9.423	0.0000	0.999	H-L	No
Joe Morolong	37.17	-0.015	0.137	0	6.597	0.0000	0.999	L-H	No
Ga-Segonyana	29.88	-0.367	-0.208	-0.008	-10.031	0.0003	0.138	L-L	No
Gamagara	35.76	-0.083	-0.208	-0.002	-9.998	0.0001	0.581	L-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Richtersveld	73.65	1.746	-0.253	0.036	-12.198	-0.0019	0.999	H-L	No
Nama Khoi	92.06	2.635	-0.253	0.055	-12.191	-0.0029	0.999	H-L	No
Kamiesberg	95.46	2.799	-0.019	0.058	-0.929	-0.0002	0.999	H-L	No
Hantam	93.57	2.708	-0.046	0.056	-2.235	-0.0005	0.999	H-L	No
Karoo Hoogland	91.17	2.592	0.05	0.054	2.385	0.0005	0.029	H-H	Yes
Khâi-Ma	66.83	1.417	0.002	0.029	0.083	0.0000	0.852	H-H	No
Ubuntu	82.98	2.197	-0.195	0.046	-9.4	-0.0018	0.999	H-L	No
Umsobomvu	46.11	0.416	-0.011	0.009	-0.513	0.0000	0.999	H-L	No
Emthanjeni	58.7	1.024	0.11	0.021	5.279	0.0005	0.019	H-H	Yes
Kareeberg	116.24	3.803	0.077	0.079	3.684	0.0012	0	H-H	Yes
Renosterberg	73.52	1.74	0.137	0.036	6.617	0.0010	0.012	H-H	Yes
Thembelihle	91.45	2.605	0.097	0.054	4.682	0.0011	0	H-H	Yes
Siyathemba	72.5	1.691	0.142	0.035	6.839	0.0010	0.003	H-H	Yes
Siyancuma	69.15	1.529	-0.009	0.032	-0.444	-0.0001	0.999	H-L	No
Mier	109.13	3.459	-0.122	0.072	-5.855	-0.0018	0.999	H-L	No
Kai !Garib	56.91	0.938	0.12	0.019	5.796	0.0005	0.025	H-H	Yes
//Khara Hais	62.65	1.215	-0.086	0.025	-4.145	-0.0004	0.999	H-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
!Kheis	98.76	2.959	0.113	0.061	5.437	0.0014	0.001	H-H	Yes
Tsantsabane	48.82	0.547	-0.038	0.011	-1.829	-0.0001	0.999	H-L	No
Kgatelopele	42	0.218	0.142	0.005	6.827	0.0001	0.279	H-H	No
Sol Plaatjie	36.52	-0.047	0.122	-0.001	5.849	0.0000	0.999	L-H	No
Dikgatlong	47.44	0.481	-0.204	0.01	-9.829	-0.0004	0.999	H-L	No
Magareng	36.68	-0.039	-0.225	-0.001	-10.846	0.0000	0.732	L-L	No
Phokwane	34.81	-0.129	0.176	-0.003	8.466	-0.0001	0.999	L-H	No
Letsemeng	33.4	-0.197	0.171	-0.004	8.219	-0.0001	0.999	L-H	No
Kopanong	33.34	-0.2	-0.169	-0.004	-8.153	0.0001	0.322	L-L	No
Mohokare	30.29	-0.347	-0.066	-0.007	-3.164	0.0001	0.383	L-L	No
Naledi	24.63	-0.621	-0.106	-0.013	-5.099	0.0003	0.158	L-L	No
Masilonyana	23.89	-0.656	-0.055	-0.014	-2.633	0.0002	0.261	L-L	No
Tokologo	32.84	-0.224	-0.13	-0.005	-6.244	0.0001	0.393	L-L	No
Tswelopele	24.53	-0.626	-0.017	-0.013	-0.797	0.0000	0.624	L-L	No
Matjhabeng	25.2	-0.593	0.048	-0.012	2.293	-0.0001	0.999	L-H	No
Nala	24.72	-0.616	-0.071	-0.013	-3.424	0.0002	0.247	L-L	No
Setsoto	29.42	-0.39	-0.04	-0.008	-1.942	0.0001	0.51	L-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Dihlabeng	30.66	-0.329	-0.169	-0.007	-8.125	0.0002	0.182	L-L	No
Nketoana	30.67	-0.329	-0.036	-0.007	-1.74	0.0001	0.614	L-L	No
Maluti a Phofung	27.15	-0.499	0.08	-0.01	3.827	-0.0002	0.999	L-H	No
Phumelela	30.58	-0.333	-0.083	-0.007	-3.99	0.0001	0.352	L-L	No
Mantsopa	25.35	-0.586	-0.114	-0.012	-5.502	0.0003	0.16	L-L	No
Moqhaka	26.5	-0.53	-0.013	-0.011	-0.612	0.0000	0.734	L-L	No
Ngwathe	27.47	-0.483	-0.195	-0.01	-9.403	0.0004	0.086	L-L	No
Metsimaholo	25.54	-0.577	-0.056	-0.012	-2.676	0.0001	0.277	L-L	No
Mafube	30.24	-0.35	-0.218	-0.007	-10.471	0.0003	0.155	L-L	No
Mangaung	26.05	-0.552	-0.09	-0.011	-4.341	0.0002	0.141	L-L	No
Umzumbe	15.61	-1.056	-0.12	-0.022	-5.77	0.0005	0.031	L-L	Yes
UMuziwabantu	33.18	-0.208	-0.23	-0.004	-11.053	0.0002	0.268	L-L	No
Ezingoleni	28.26	-0.445	-0.082	-0.009	-3.931	0.0002	0.257	L-L	No
Hibiscus Coast	34.65	-0.137	-0.112	-0.003	-5.402	0.0001	0.554	L-L	No
Emnambithi/Ladysmith	28.81	-0.419	-0.117	-0.009	-5.648	0.0002	0.159	L-L	No
Newcastle	24.07	-0.648	-0.145	-0.013	-6.992	0.0004	0.084	L-L	No
Emadlangeni	15.6	-1.057	-0.12	-0.022	-5.778	0.0005	0.026	L-L	Yes

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Dannhauser	29.41	-0.39	0.156	-0.008	7.484	-0.0003	0.999	L-H	No
Abaqulusi	28.17	-0.45	-0.032	-0.009	-1.523	0.0001	0.588	L-L	No
uMhlathuze	22.62	-0.718	-0.159	-0.015	-7.661	0.0005	0.081	L-L	No
Nkandla	15.78	-1.048	-0.113	-0.022	-5.445	0.0005	0.039	L-L	Yes
Maphumulo	16.25	-1.025	-0.086	-0.021	-4.141	0.0004	0.116	L-L	No
Vulamehlo	12.67	-1.198	-0.126	-0.025	-6.083	0.0006	0.032	L-L	Yes
Umdoni	37.85	0.017	-0.199	0	-9.583	0.0000	0.999	H-L	No
uMshwathi	28.72	-0.423	-0.099	-0.009	-4.777	0.0002	0.262	L-L	No
uMngeni	30.01	-0.361	-0.127	-0.008	-6.111	0.0002	0.256	L-L	No
Mpofana	27.88	-0.464	0.06	-0.01	2.879	-0.0001	0.999	L-H	No
Impendle	31.3	-0.299	-0.217	-0.006	-10.419	0.0003	0.201	L-L	No
The Msunduzi	26.84	-0.514	-0.17	-0.011	-8.167	0.0004	0.122	L-L	No
Mkhambathini	27.99	-0.458	-0.001	-0.01	-0.026	0.0000	0.985	L-L	No
Richmond	26.77	-0.517	-0.074	-0.011	-3.584	0.0002	0.301	L-L	No
Indaka	34.83	-0.128	-0.082	-0.003	-3.955	0.0000	0.672	L-L	No
Umtshezi	35.13	-0.114	-0.09	-0.002	-4.313	0.0000	0.699	L-L	No
Okhahlamba	34.14	-0.162	-0.063	-0.003	-3.026	0.0000	0.619	L-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Imbabazane	33.4	-0.197	-0.093	-0.004	-4.485	0.0001	0.513	L-L	No
Endumeni	32.24	-0.253	-0.097	-0.005	-4.691	0.0001	0.446	L-L	No
Nqutu	33.98	-0.169	-0.137	-0.004	-6.589	0.0001	0.398	L-L	No
Msinga	12.2	-1.221	-0.065	-0.025	-3.115	0.0003	0.126	L-L	No
Umvoti	35.03	-0.119	-0.054	-0.002	-2.591	0.0000	0.787	L-L	No
eDumbe	27.73	-0.471	0.167	-0.01	8.023	-0.0003	0.999	L-H	No
UPhongolo	26.58	-0.526	-0.215	-0.011	-10.365	0.0005	0.091	L-L	No
Nongoma	27.91	-0.462	-0.219	-0.01	-10.526	0.0004	0.111	L-L	No
Ulundi	27.14	-0.5	-0.155	-0.01	-7.475	0.0003	0.117	L-L	No
Umhlabuyalingana	10.71	-1.293	-0.201	-0.027	-9.686	0.0011	0.009	L-L	Yes
Jozini	13.45	-1.16	-0.164	-0.024	-7.875	0.0008	0.016	L-L	Yes
The Big 5 False Bay	31.89	-0.27	-0.068	-0.006	-3.256	0.0001	0.45	L-L	No
Hlabisa	33.76	-0.18	-0.076	-0.004	-3.671	0.0001	0.519	L-L	No
Mtubatuba	26.33	-0.539	-0.105	-0.011	-5.061	0.0002	0.14	L-L	No
Mfolozi	25.87	-0.561	-0.089	-0.012	-4.273	0.0002	0.136	L-L	No
Ntambanana	33.56	-0.19	-0.04	-0.004	-1.941	0.0000	0.711	L-L	No
uMlalazi	28.78	-0.42	0.121	-0.009	5.804	-0.0002	0.999	L-H	No

Continued on next page

Table A.4 – continued from previous page

Municipality	Ischaemic	z	f	f^*	z^*	LISA	P value	Type	Significant
Mthonjaneni	32.11	-0.26	0.063	-0.005	3.043	-0.0001	0.999	L-H	No
Mandeni	25.4	-0.584	-0.22	-0.012	-10.568	0.0005	0.071	L-L	No
KwaDukuza	32.28	-0.251	0.103	-0.005	4.96	-0.0001	0.999	L-H	No
Ndwedwe	15.83	-1.046	-0.188	-0.022	-9.023	0.0008	0.022	L-L	Yes
Ingwe	15.15	-1.078	0.066	-0.022	3.166	-0.0003	0.999	L-H	No
Kwa Sani	29.74	-0.374	-0.138	-0.008	-6.634	0.0002	0.208	L-L	No
Greater Kokstad	27	-0.506	-0.144	-0.011	-6.938	0.0003	0.149	L-L	No
Ubuhlebezwe	15.41	-1.066	-0.212	-0.022	-10.212	0.0010	0.013	L-L	Yes
Umzimkhulu	15.21	-1.075	-0.163	-0.022	-7.834	0.0007	0.013	L-L	Yes
eThekwini	36.43	-0.051	-0.122	-0.001	-5.863	0.0000	0.734	L-L	No
Moretele	28.65	-0.427	0.064	-0.009	3.098	-0.0001	0.999	L-H	No
Madibeng	28.96	-0.411	-0.183	-0.009	-8.783	0.0003	0.137	L-L	No
Rustenburg	26.81	-0.515	-0.212	-0.011	-10.204	0.0005	0.084	L-L	No
Kgetlengrivier	35.04	-0.118	-0.147	-0.002	-7.08	0.0001	0.538	L-L	No
Moses Kotane	28.4	-0.439	-0.207	-0.009	-9.956	0.0004	0.109	L-L	No
Ratlou	36.24	-0.06	0.075	-0.001	3.632	0.0000	0.999	L-H	No
Tswaing	30.82	-0.322	-0.09	-0.007	-4.318	0.0001	0.377	L-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Mafikeng	26.88	-0.512	-0.201	-0.011	-9.683	0.0004	0.083	L-L	No
Ditsobotla	30.71	-0.327	0.174	-0.007	8.357	-0.0002	0.999	L-H	No
Ramotshere Moiloa	30.36	-0.344	0.148	-0.007	7.11	-0.0002	0.999	L-H	No
Naledi	37.62	0.007	-0.048	0	-2.313	0.0000	0.999	H-L	No
Mamusa	30.03	-0.36	-0.191	-0.007	-9.214	0.0003	0.148	L-L	No
Greater Taung	36.28	-0.058	-0.187	-0.001	-8.993	0.0000	0.668	L-L	No
Lekwa-Teemane	28.14	-0.451	-0.098	-0.009	-4.71	0.0002	0.236	L-L	No
Kagisano/Molopo	35.12	-0.114	-0.114	-0.002	-5.473	0.0001	0.624	L-L	No
Ventersdorp	30.51	-0.337	0.177	-0.007	8.541	-0.0003	0.999	L-H	No
Tlokwe City Council	32.02	-0.264	-0.04	-0.005	-1.923	0.0000	0.624	L-L	No
City of Matlosana	27.57	-0.479	-0.09	-0.01	-4.311	0.0002	0.274	L-L	No
Maquassi Hills	30.98	-0.314	-0.035	-0.007	-1.66	0.0000	0.586	L-L	No
Emfuleni	25.88	-0.56	0.046	-0.012	2.21	-0.0001	0.999	L-H	No
Midvaal	47.88	0.502	-0.227	0.01	-10.901	-0.0005	0.999	H-L	No
Lesedi	28.62	-0.428	0.085	-0.009	4.105	-0.0002	0.999	L-H	No
Mogale City	34.84	-0.128	-0.07	-0.003	-3.391	0.0000	0.723	L-L	No
Randfontein	32.42	-0.244	-0.031	-0.005	-1.511	0.0000	0.719	L-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Westonaria	30.53	-0.336	0.15	-0.007	7.207	-0.0002	0.999	L-H	No
Merafong City	28.44	-0.436	-0.057	-0.009	-2.724	0.0001	0.385	L-L	No
Ekurhuleni	32.07	-0.261	-0.158	-0.005	-7.59	0.0002	0.199	L-L	No
City of Johannesburg	26.57	-0.527	0.052	-0.011	2.501	-0.0001	0.999	L-H	No
City of Tshwane	28.07	-0.455	0.106	-0.009	5.12	-0.0002	0.999	L-H	No
Albert Luthuli	27.93	-0.461	-0.01	-0.01	-0.498	0.0000	0.787	L-L	No
Msukaligwa	29.78	-0.372	0.157	-0.008	7.542	-0.0002	0.999	L-H	No
Mkhondo	32.73	-0.229	-0.137	-0.005	-6.593	0.0001	0.356	L-L	No
Pixley Ka Seme	31.17	-0.305	0.108	-0.006	5.207	-0.0001	0.999	L-H	No
Lekwa	26.63	-0.524	0.162	-0.011	7.794	-0.0004	0.999	L-H	No
Dipaleseng	30.01	-0.361	0.167	-0.008	8.042	-0.0003	0.999	L-H	No
Govan Mbeki	26.53	-0.529	-0.124	-0.011	-5.958	0.0003	0.144	L-L	No
Victor Khanye	32.5	-0.241	-0.058	-0.005	-2.779	0.0001	0.562	L-L	No
Emalahleni	31.7	-0.279	-0.16	-0.006	-7.702	0.0002	0.243	L-L	No
Steve Tshwete	29.34	-0.393	-0.216	-0.008	-10.382	0.0004	0.134	L-L	No
Emakhazeni	30.72	-0.326	0.158	-0.007	7.597	-0.0002	0.999	L-H	No
Thembisile	21.38	-0.777	0.057	-0.016	2.766	-0.0002	0.999	L-H	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Dr JS Moroka	23.42	-0.679	-0.105	-0.014	-5.054	0.0003	0.153	L-L	No
Thaba Chweu	31.43	-0.292	-0.083	-0.006	-4.011	0.0001	0.359	L-L	No
Mbombela	27.51	-0.482	-0.101	-0.01	-4.863	0.0002	0.214	L-L	No
Umjindi	28.74	-0.422	-0.061	-0.009	-2.914	0.0001	0.364	L-L	No
Nkomazi	24.09	-0.647	0.067	-0.013	3.211	-0.0002	0.999	L-H	No
Bushbuckridge	25.31	-0.588	0.074	-0.012	3.567	-0.0002	0.999	L-H	No
Greater Giyani	26.11	-0.549	-0.053	-0.011	-2.56	0.0001	0.318	L-L	No
Greater Letaba	27.17	-0.498	-0.09	-0.01	-4.341	0.0002	0.187	L-L	No
Greater Tzaneen	26.62	-0.525	-0.2	-0.011	-9.615	0.0004	0.096	L-L	No
Ba-Phalaborwa	26.84	-0.514	0.142	-0.011	6.84	-0.0003	0.999	L-H	No
Maruleng	27	-0.506	-0.205	-0.011	-9.868	0.0004	0.093	L-L	No
Mutale	13.6	-1.153	-0.188	-0.024	-9.037	0.0009	0.018	L-L	Yes
Thulamela	25.77	-0.566	-0.13	-0.012	-6.241	0.0003	0.102	L-L	No
Musina	22.81	-0.708	-0.195	-0.015	-9.389	0.0006	0.047	L-L	Yes
Makhado	27.12	-0.501	-0.133	-0.01	-6.379	0.0003	0.126	L-L	No
Blouberg	28.47	-0.435	-0.219	-0.009	-10.52	0.0004	0.091	L-L	No
Aganang	32.07	-0.262	-0.218	-0.005	-10.493	0.0002	0.226	L-L	No

Continued on next page

Table A.4 – continued from previous page

Municipality	**Ischaemic**	z	f	f^*	z^*	LISA	P value	Type	Significant
Molemole	23.04	-0.697	-0.219	-0.014	-10.561	0.0007	0.057	L-L	No
Polokwane	27.37	-0.488	0.053	-0.01	2.553	-0.0001	0.999	L-H	No
Lepele-Nkumpi	28.29	-0.444	-0.216	-0.009	-10.376	0.0004	0.122	L-L	No
Thabazimbi	27.99	-0.458	-0.182	-0.01	-8.769	0.0004	0.105	L-L	No
Lephalale	21.92	-0.751	-0.221	-0.016	-10.655	0.0007	0.049	L-L	Yes
Mookgopong	25.06	-0.6	-0.186	-0.012	-8.928	0.0005	0.08	L-L	No
Modimolle	23.56	-0.672	-0.135	-0.014	-6.514	0.0004	0.092	L-L	No
Bela-Bela	25.93	-0.558	-0.195	-0.012	-9.366	0.0005	0.091	L-L	No
Mogalakwena	28.89	-0.415	-0.169	-0.009	-8.137	0.0003	0.132	L-L	No
Ephraim Mogale	27.95	-0.46	-0.073	-0.01	-3.528	0.0001	0.3	L-L	No
Elias Motsoaledi	28.63	-0.427	-0.12	-0.009	-5.798	0.0002	0.221	L-L	No
Makhuduthamaga	28.84	-0.417	-0.206	-0.009	-9.934	0.0004	0.114	L-L	No
Fetakgomo	27.93	-0.461	-0.102	-0.01	-4.9	0.0002	0.262	L-L	No
Greater Tubatse	24.69	-0.618	0.182	-0.013	8.741	-0.0005	0.999	L-H	No

Appendix B

Appendix B for Chapter 5

B.1 R code for univariate and bivariate Moran's I using canonical approach

```
HEXshp <- readShapePoly("HEX61_Man3.shp",
                        proj4string=CRS("+proj=longlat +datum=WGS84"))
# Queen based contiguity
library(spdep)
HEX.nb<-poly2nb(HEXshp, queen=T);
HEX.wt<-nb2listw(neighbours=HEX.nb, style="W")
library("CCA")
##Y values
Y <- HEXshp@data$Y3
# X values
X <- HEXshp@data$Y3
#A function to calculate Univariate and Bivariate Moran's I
mmi<- function(Y,X,r,HEX.nb)
```

```
  {
 HEXshp$Y<-Y
 Y0 <- lag.listw(nb2listw(HEX.nb),HEXshp$Y)
 Y1 <- scale(lag.listw(nb2listw(HEX.nb),HEXshp$Y))

 X1 <- X
 Y_vec <- cbind.data.frame(Y1)
 X_vec <- cbind.data.frame(X1)
Zx <- scale(X_vec); Zy <- (Y_vec);
Zx<-as.matrix(Zx);Zy<-as.matrix(Zy)
 n <- nrow(X_vec);p <- ncol(X_vec);
 q <- ncol(Y_vec);k<-min(p,q)
 #Calculate Covariances
 Sx<-cov(Zx);Sxy<-cov(Zx,Zy);Syx<-t(Sxy);Sy<-cov(Zy)
 #Eigenvectors
 Ex<-(solve(Sx)%*%Sxy%*%solve(Sy)%*%Syx);
 Ey<-(solve(Sy)%*%Syx%*%solve(Sx)%*%Sxy)
Ahat <- (as.matrix(eigen(Ey)$vectors[,1:k]));
Bhat <- (as.matrix(eigen(Ex)$vectors[,1:k]))

 #Calculate Canonical variables:

 #U <- Zy %*% Ahat
 #V <- Zx %*% Bhat

 u1 <- as.matrix(Zy) %*% as.matrix(eigen(Ey)$vectors[,1])
 v1 <- as.matrix(Zx) %*% as.matrix(eigen(Ex)$vectors[,1])
 # display the canonical correlations
```

```
  canon.corr <-cor(u1,v1)
  canon.corr
   # Calculate the CCA Coefficients
  Inv_A=solve(Ahat)
  rho_y1_v1<-Inv_A[1,1]*canon.corr[1]

  SD_Y_lag<-sqrt(var(Y0))
  SD_Y<-sqrt(var(HEXshp$Y))
  SD_Ratio<-SD_Y_lag/SD_Y
 mmi<-SD_Ratio*rho_y1_v1
  #return(mmi)
 (dmat <- cbind(rho=rho_y1_v1, SD1=SD_Y_lag,SD2=SD_Y,
  SDRatio=SD_Ratio, mm_i = mmi))
    }
dmat<-mmi(Y,X,r,HEX.nb)
dmat
mmi_global<-dmat[,5]
mmi_global
```

B.2 R code for multivariate Moran's I using canonical approach

```
HEXshp <- readShapePoly("hex61_Man_2.shp",
                        proj4string=CRS("+proj=longlat +datum=WGS84"))
# Queen based contiguity
library(spdep)
```

```
HEX.nb<-poly2nb(HEXshp, queen=T);
HEX.wt<-nb2listw(neighbours=HEX.nb, style="W")

library("CCA")

##Y values
Y <- HEXshp@data$Y3cY1cY2
#For calculating X values
X <- cbind.data.frame(HEXshp@data$Y2,HEXshp@data$Y3)
#A function to calculate Trivariate Moran’s I

mmi<- function(Y,X,r,HEX.nb)

   {
  HEXshp$Y<-Y
  Y0 <- lag.listw(nb2listw(HEX.nb),HEXshp$Y)
  Y1 <- scale(lag.listw(nb2listw(HEX.nb),HEXshp$Y))
    n<-length(Y1)
  set.seed(123)
  Y2 <- r*Y1 + rnorm(n, mean = 0,sd=sqrt(1-r ˆ2))
  X1 <- X[,1]
  X2 <- X[,2]
  Y_vec <- cbind.data.frame(Y1,Y2)
  X_vec <- cbind.data.frame(X1,X2)
  Zx <- scale(X_vec); Zy <- (Y_vec);
  Zx<-as.matrix(Zx);Zy<-as.matrix(Zy)
  n <- nrow(X_vec);p <- ncol(X_vec);
  q <- ncol(Y_vec);k<-min(p,q)
```

```
 #Calculate Covariances
 Sx<-cov(Zx);Sxy<-cov(Zx,Zy);Syx<-t(Sxy);Sy<-cov(Zy)

 #Eigenvectors
 Ex<-(solve(Sx)%*%Sxy%*%solve(Sy)%*%Syx);
 Ey<-(solve(Sy)%*%Syx%*%solve(Sx)%*%Sxy)

Ahat <- (as.matrix(eigen(Ey)$vectors[,1:k]));
Bhat <- (as.matrix(eigen(Ex)$vectors[,1:k]))

 u1 <- as.matrix(Zy) %*% as.matrix(eigen(Ey)$vectors[,1])
 v1 <- as.matrix(Zx) %*% as.matrix(eigen(Ex)$vectors[,1])

 # display the canonical correlations

 # canon.corr <- sqrt(eigen(Ey)$values) Note: Do not use this
 #because always positive; instead use
 canon.corr <-(-1)*cor(u1,v1)
 canon.corr

 # Calculate the CCA Coefficients
 Inv_A=solve(Ahat)

 rho_y1_v1<-Inv_A[1,1]*canon.corr[1]
 SD_Y_lag<-sqrt(var(Y0))
 SD_Y<-sqrt(var(HEXshp$Y))
 SD_Ratio<-SD_Y_lag/SD_Y
```

```
  mmi<-SD_Ratio*rho_y1_v1


  #return(mmi)


 (dmat <- cbind(rho=rho_y1_v1, SD1=SD_Y_lag,SD2=SD_Y,

  SDRatio=SD_Ratio, mm_i = mmi))


  }


r=0.6

dmat<-mmi(Y,X,r,HEX.nb)

dmat

mmi_global<-dmat[,5]

mmi_global
```

Appendix 5.3

Table B.1: Multivariate spatial autocorrelation analysis of the hypothetical spatial data, *r = 0.1*.

r = 0.1

Criterion variable	Predictor variables	ρ_{y_i,v_1}	Standard Deviations			Multivariate Moran's I
			SD1	SD2	SD ratio	MMI
Y1	Y2 Y3	0.6997	1.1615	1.6121	0.7205	0.5041**
Y2	Y1 Y3	0.7106	1.1157	1.6121	0.6920	0.4918**
Y3	Y1 Y2	0.7377	1.1234	1.6615	0.6762	0.4988**
Y1	Y4 Y5	-0.2018	1.1615	1.6121	0.7205	-0.1454^{INS}
Y2	Y4 Y5	-0.3006	1.1157	1.6121	0.6920	-0.2080**
Y3	Y4 Y5	-0.2379	1.1234	1.6615	0.6762	-0.1608**

Table B.2: Multivariate spatial autocorrelation analysis of the hypothetical spatial data, *r = 0.3*.

r = 0.3

Criterion variable	Predictor variables	ρ_{y_i,v_1}	Standard Deviations			Multivariate Moran's I
			SD1	SD2	SD ratio	MMI
Y1	Y2 Y3	0.6703	1.1615	1.6121	0.7205	0.4829**
Y2	Y1 Y3	0.6828	1.1157	1.6121	0.6920	0.4725**
Y3	Y1 Y2	0.7258	1.1234	1.6615	0.6762	0.4908**
Y1	Y4 Y5	-0.1893	1.1615	1.6121	0.7205	-0.1364^{INS}
Y2	Y4 Y5	-0.2915	1.1157	1.6121	0.6920	-0.2017**
Y3	Y4 Y5	-0.2250	1.1234	1.6615	0.6762	-0.1521**

Table B.3: Multivariate spatial autocorrelation analysis of the hypothetical spatial data, *r = 0.5*.

r = 0.5

Criterion variable	Predictor variables	ρ_{y_i,v_1}	Standard Deviations			Multivariate Moran's I
			SD1	SD2	SD ratio	MMI
Y1	Y2 Y3	0.6367	1.1615	1.6121	0.7205	0.4587**
Y2	Y1 Y3	0.6507	1.1157	1.6121	0.6920	0.4503**
Y3	Y1 Y2	0.7115	1.1234	1.6615	0.6762	0.4811**
Y1	Y4 Y5	-0.1761	1.1615	1.6121	0.7205	-0.1269^{INS}
Y2	Y4 Y5	-0.2807	1.1157	1.6121	0.6920	-0.1943**
Y3	Y4 Y5	0.2108	1.1234	1.6615	0.6762	-0.1425^{INS}

Table B.4: Multivariate spatial autocorrelation analysis of the hypothetical spatial data, *r = 0.7*.

r = 0.7

Criterion variable	Predictor variables	ρ_{y_i,v_1}	Standard Deviations			Multivariate Moran's I
			SD1	**SD2**	**SD ratio**	**MMI**
Y1	Y2 Y3	0.5913	1.1615	1.6121	0.7205	0.4260**
Y2	Y1 Y3	0.6068	1.1157	1.6121	0.6920	0.4199**
Y3	Y1 Y2	0.6902	1.1234	1.6615	0.6762	0.4667**
Y1	Y4 Y5	-0.1613	1.1615	1.6121	0.7205	-0.1162^{INS}
Y2	Y4 Y5	-0.2654	1.1157	1.6121	0.6920	-0.1837**
Y3	Y4 Y5	-0.1933	1.1234	1.6615	0.6762	-0.1307^{INS}

Table B.5: Multivariate spatial autocorrelation analysis of the hypothetical spatial data, *r = 0.9*.

r = 0.9

Criterion variable	Predictor variables	ρ_{y_i,v_1}	Standard Deviations			Multivariate Moran's I
			SD1	**SD2**	**SD ratio**	**MMI**
Y1	Y2 Y3	0.5194	1.1615	1.6121	0.7205	0.3742**
Y2	Y1 Y3	0.5305	1.1157	1.6121	0.6920	0.3671**
Y3	Y1 Y2	0.6376	1.1234	1.6615	0.6762	0.4311**
Y1	Y4 Y5	-0.1648	1.1615	1.6121	0.7205	0.1188^{INS}
Y2	Y4 Y5	-0.2333	1.1157	1.6121	0.6920	-0.1614^{INS}
Y3	Y4 Y5	-0.1818	1.1234	1.6615	0.6762	-0.1229^{INS}