The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Geographically Weighted Regression and an extension

Karen M. Miller

Submitted to the Department of Statistical Sciences
in partial fulfillment of the requirements for the degree of

Master of Science in Statistical Sciences
at the
UNIVERSITY OF CAPE TOWN

January 5, 2008

The author hereby grants to the University of Cape Town permission to reproduce and to distribute copies of this document in whole or in part

Supervisors: Professor Linda Haines and Professor Christien Thiart

University of Cape Town

Abstract

It is often the case with spatial data that the relationships between variables vary over space, a situation referred to as spatial non-stationarity. As a result, global models with parameters that are assumed constant over space cannot adequately explain the relationship that exists between some set of variables. Various techniques have been developed to model relationships locally and a more recent development has been the method of Geographically Weighted Regression (GWR). GWR accounts for spatial non-stationarity by permitting parameters of a regression model to vary locally. In this dissertation GWR is described and an extension to the methodology is proposed. The extended model, termed Local Linear Geographically Weighted Regression (LLGWR) is developed through the expansion of the regression coefficients of the GWR model using a first order linear approximation. The aim of the expansion of the GWR model is to increase parameter flexibility and to capture more of the variability in a spatial data set. In the present study a small data set taken from soil science and a large data set taken from geology, are analysed using global regression, GWR, LLGWR and kriging models and results compared. The results produced from the two sets of analyses showed that both GWR and LLGWR models are superior to the global model. The results however are inconclusive as to whether the LLGWR model provides an improvement over the GWR approach, with the small data set showing some evidence of an improvement in the LLGWR model over the GWR model but the large data set showing no improvement.

University of Cape Town

Acknowledgements

I would like to express my gratitude to my supervisors, Prof. Linda Haines and Prof. Christien Thiart, for their support, guidance and patience throughout this process.

I would like to thank all my friends, colleagues and staff of the Statistical Sciences Department, who have all played a role in making my experience as a Masters student in the department an enjoyable and rewarding one.

I would like to thank the National Research Foundation (NRF) for providing me with a scholarship which has been of great financial assistance.

Last but not least, I would also like to thank my family and friends for their encouragement and support.

University of Cape Town

I hereby grant the University of Cape Town permission to copy and disseminate this work, or any part thereof, for the purposes of study and research.

Plagiarism Declaration

1. This dissertation is my own work. It has not been submitted before for any degree or examination to any other University.

2. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

3. Each significant contribution to, and quotation in, this dissertation from the work of other people has been cited and referenced.

_____

Signature

_____

Date

University of Cape Town

Contents

University of Cape Town

University of Cape Town

Chapter 1

Introduction

University of Cape Town

The increased availability of Geographic Information Systems (GIS) technology and software packages, which allow spatial information to be stored and managed, has resulted in increasing interest in the analysis of spatial data. The main aim of spatial analysis is to summarise and make inferences about properties and relationships between variables taking into account the location in space of the phenomenon under study. It is often the case with spatial data that the relationships between variables vary over space, a situation referred to as spatial non-stationarity. As a result, global models with parameters that are assumed constant over space cannot adequately explain the relationship that exists between a set of explanatory variables and a response variable. The development and application of spatial techniques which attempt to model relationships locally has received considerable attention in the literature. More recent research includes the development of Geographically Weighted Regression (GWR) (Brunsdon, Fotheringham and Charlton, 1996). GWR is a method that addresses the issue of non-stationarity by allowing parameters to be estimated locally for each location in space thus allowing different relationships between variables to exist at different points in space. The broad aims of this thesis are to describe GWR and to propose and investigate an extension to the GWR methodology.

1.1 Objectives

The specific research objectives of this dissertation are as follows:

- To explore the background to spatial regression modelling and its development.
- To provide a statistical description of Geographically Weighted Regression (GWR).
- To propose and develop an extension to the GWR model, termed Local Linear Geographically Weighted Regression (LLGWR), which is based on

the expansion of the regression coefficients using a first order linear approximation.

- To analyse a small data set through the implementation of ordinary least squares (OLS) regression, GWR, LLGWR and kriging models and to write prgrams in R to implement GWR and LLGWR.
- To compare the results from the various models, and in particular to investigate whether or not LLGWR provides an improvement over the GWR approach.
- To extend the investigation to a large data set.

1.2 Outline of dissertation

The thesis is organised as follows. In Chapter 2, a background to spatial regression modelling and its development is presented. In particular, a statistical description of GWR is given and the estimation of the model parameters discussed. In Chapter 3, an extension to the GWR model is proposed and the development of this extended model, termed LLGWR, is presented. Chapters 4 and 5 are application chapters in which the analyses of a small data set taken from soil science and a large data set taken from geology respectively are presented . The data sets are analysed using OLS, GWR, LLGWR and kriging models and the results obtained from the analyses based on these four models are compared. Finally conclusions are drawn and recommendations for future research are given in Chapter 6.

University of Cape Town

Chapter 2

Geographically Weighted Regression

University of Cape Town

In this chapter a background to spatial regression modelling is presented. In Section 2.1 the use of multiple regression and the problems associated with that technique are discussed and various approaches to addressing these problems are noted. The idea of Geographically Weighted Regression (GWR) and some examples of its applications are also given. In Section 2.2 a statistical description of GWR is presented. The estimation of the model parameters is discussed in Sections 2.3 and 2.4. Bandwidth selection for the weighting scheme and issues regarding spatial non-stationarity are considered in Sections 2.5 and 2.6 respectively.

2.1 Background

One of the main aims of the analysis of spatial data is to summarise and make inferences about properties and relationships between variables taking into account the location in space of the phenomenon under study. The advent of Geographic Information Systems (GIS) technology and software packages such as ArcGIS, which allows spatial information to be easily stored and managed, has resulted in increased availability of large and complex data sets.

Multiple regression has been a commonly applied method for analysing spatial data across various fields including social, environmental and geographical sciences. For example, it has been used in the investigation of rainfall across Southern Africa (Dent, Lynch and Schulze, 1988), the aim of the analysis being to predict the Mean Annual Precipitation from a number of explanatory variables such as altitude, sea roughness and distance from the sea. The multiple regression approach has been used in remote sensing, as for example to describe the relationship between an environmental variable measured at the Earth's

surface such as biomass and some measure of its remotely sensed image such as a vegetation index, and to make predictions of the environmental variable at other sites from the remotely sensed images (Foody, 2003).

Location in space generally plays no role in the modelling process when using regression models. However there are instances where latitude and longitude are included in the multiple regression model as independent variables. For example, Lobo and Martin-Piera (2002) constructed a multiple regression model to predict the species-richness distribution of dung beetles and found latitude to be a significant predictor. Specific problems with the use of the multiple regression model for modelling spatial data are that parameters are assumed constant over space and error terms are assumed to be independent. A multiple regression model incorporates a single regression equation and a single parameter associated with each explanatory variable that is assumed to apply globally over the entire study region. However, it is often the case with spatial data that the relationships between variables vary over space i.e. are spatially non-stationary. Thus a single set of parameters which are assumed constant over space cannot adequately explain the relationship that exists between some set of explanatory variables and the response variable. Furthermore, independence is rarely the case with spatial data where observations under study are often spatially dependent. Values of a variable at specific locations in space depend on those in the neighbouring locations and indeed according to Tobler's first law of geography 'everything is related to everything else, but near things are more related than distant things' (Tobler, 1970). As a result of spatial dependency, the fitting of regression models to spatial data will produce residuals that exhibit spatial autocorrelation. Spatial regression models have been developed that allow for spatial autocorrelation of the errors (Cressie, 1993; Anselin, 1993) but these models assume a global autocorrelation function for the errors. Such models incorporate global parameters but with local relationships introduced into the model through the covariance structure of the error terms. For example kriging, an interpolation technique that allows values of a variable to be predicted at locations where no measurements have been taken by using a weighted average of values at neighbouring locations, incorporates correlations between observations. The kriging technique has been a popular alternative to multiple regression and has been widely used in the analysis of spatial data (Cressie, 1993).

Various methods have been developed in an attempt to model relationships locally and a large body of literature that deals with these methods exists. The expansion method (Casetti, 1972) is one of the earliest attempts at modelling local relationships. In this framework parameters of a basic model are expanded by making them functions of other variables. In the spatial context, global models may be expanded by expressing parameters as functions of spatial locations, thus allowing parameters to vary over geographical space.

University of Cape Town

Trends in parameters over space can then be assessed (Eldridge and Jones, 1991). The method of spatial adaptive filtering (Foster and Gorr, 1986; Gorr and Olligschaeler, 1994), the spatial analogue of exponential smoothing in time series, has been proposed to model spatial variations in coefficients. This technique has limited applicability however as the parameter estimates produced cannot be tested statistically. In the random coefficient model approach (Swamy, 1971; Aikten, 1996) the parameters are assumed to be random variables. Specifically, the parameters of the model are assumed to vary from point to point and are drawn from some random distribution. Inferences about local parameters are obtained using Bayes' Theorem.

A more recent attempt at modelling spatial relationships locally within a multiple regression framework is the method of Geographically Weighted Regression (GWR) (Brunsdon, Fotheringham and Charlton, 1996). GWR is an extension of the traditional regression model for spatial data which takes into account location in space. Specifically, GWR attempts to account for spatial non-stationarity in the regression coefficients by permitting parameters to vary locally. Parameters of the model estimated by fitting the model to spatial data are specific to a given location and thus the relationship between variables differs at different locations. The model is fitted using a weighted regression approach whereby data are weighted according to their proximity to a point of interest termed a regression point, with data from nearby locations being given more weight than data from locations farther away. Brunsdon, Fotheringham and Charlton (1996) consider various choices of spatial weighting functions and Fotheringham, Brunsdon and Charlton (2002) consider spatially adaptive weighting functions which vary depending on the density of data points around the regression point.

The origins of GWR can be found in local regression (Cleveland, 1979) which, instead of fitting a global model to the entire data set, fits many models to localised sub-samples of the data around specific points in space. There are also parallels between GWR and kernel regression (Wand and Jones, 1995). In kernel regression, the dependent variable is modelled as a non-linear function of the independent variables by weighting data in attribute space rather than geographic space (Loader, 1999).

The main output of GWR is a set of local parameter estimates that can be mapped to show the parameter variations over space, and that can be used to describe the degree of spatial non-stationarity in a relationship. Developments of the GWR technique include a method to test for the stationarity of parameters based on a Monte Carlo approach which is described by Brunsdon, Fotheringham and Charlton (1998). Mixed GWR which extends the basic GWR model by allowing some parameters to be fixed globally and some to vary spatially are

University of Cape Town

discussed by Brunsdon, Fotheringham and Charlton (1999). Further extensions to the basic GWR model include Geographically Weighted Generalised Linear Models for situations where the distribution of the dependent variable is a member of the exponential family of distributions. Two examples, one using a Poisson model, and the other using a binomial model are given in Fotheringham, Brunsdon and Charlton (2002).

Fotheringham, Brunsdon and Charlton (2002) developed their own software, GWR software version 3.0, to perform GWR analyses. It is a windows-based application consisting of drop-down menus and tick boxes that allows users to specify and fit a GWR model. The program produces output files which contain location-specific parameter estimates and diagnostics that can be imported into a spreadsheet or mapping package. The program allows Gaussian, Poisson and binary logit regression models to be fitted. Fotheringham, Brunsdon and Charlton have also written code in the R language to perform GWR. The code is not available commercially but a beta version is obtainable from Brunsdon (2006). Furthermore, an implementation of GWR is readily available as an R package called `spgwr` written by Bivand and Yu (2007).

University of Cape Town

Since its development, there has been a number of applications of GWR across various fields reported in the literature. For example:

- Nelson (2000) developed a GWR model to examine the link between agricultural labour productivity and natural resource, socio-economic and farming system variables at the national level in Honduras. The results revealed that the GWR model described the data better than a global regression model and that the regression parameters vary over space.

- Foody (2003) used GWR to model the relationship between the normalised vegetation index (NDVI) and rainfall over North Africa and the Middle East. The analysis revealed the relationship between NDVI and rainfall to be spatially non-stationary and highlighted areas of local variation.

- Kam, Hossain, Bose and Villano (2005) applied GWR techniques to determine whether or not spatial differences occur between poverty indices and welfare-influencing factors in Bangladesh. Results indicated spatial differences in the relative importance of various poverty-influencing factors.

- Nakaya, Fotheringham, Brunsdon and Charlton (2005) used the Geographically Weighted Poisson Regression approach to examine the relationship between mortality rates and socio-economic factors across the metropolitan area of Tokyo. The results indicate that there are significant spatial variations in the relationships between working-age mortality and occupational segregation and between working-age mortality and unemployment throughout the Tokyo metropolitan area.

- Zhao, Chow, Li and Liu (2005) developed a GWR model to predict public transit use from a number of predictors, including demographic, socioeconomic, land use and pedestrian environment variables, for Broward County, Florida. The findings indicate that GWR can help improve transit demand analysis and identify areas where transit service may need improvement.

2.2 Statistical description of the GWR model

Consider the usual global multiple regression model

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + e_i \quad i = 1, ..., n \tag{2.1}$$

where y_i is the i^{th} observation of the dependent variable, $x_{i1}, ..., x_{ip}$ are p independent explanatory variables associated with that observation, n is the number of observations, β_k are unknown parameters and e_i are independent and identically distributed error terms with zero mean and variance σ^2. Then model (2.1) can be expressed as

$$y_i = \underline{x}_i^T \underline{\beta} + e_i \quad i = 1, ..., n$$

where $\underline{x}_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$ and $\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ and can be assembled into matrix form as

$$\underline{y} = X\underline{\beta} + \underline{e}$$

where $\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\underline{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \sim rv(\underline{0}, \sigma^2 I)$ and $X = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}$,

and where $\underline{0}$ is an $n \times 1$ vector of zeroes and I the identity matrix of order n.

In GWR, the regression coefficients depend on location and model (2.1) is extended to

$$y_i = \beta_0(s_i) + \sum_{k=1}^{p} \beta_k(s_i) x_{ik} + e_i \quad i = 1, ..., n \tag{2.2}$$

University of Cape Town

where y_i is the observed value of the dependent variable at location s_i where $s_i = (u_i, v_i)$ and u_i, v_i are the longitude and latitude coordinates respectively, $x_{i1}, ..., x_{ip}$ are the associated explanatory variables, $\beta_1(s_i) \ldots \beta_p(s_i)$ are the unknown parameters at location s_i, n is the number of observations and e_i are independent and identically distributed error terms with zero mean and variance σ^2. The observed values y_i may be expressed as $z(s_i)$, the notation frequently used in a spatial context, but since the emphasis here is on the regression framework, the use of the y_i notation is retained. Then

$$y_i = \underline{x}_i^T \underline{\beta}(s_i) + e_i \quad i = 1, ..., n$$

where $\underline{\beta}(s_i) = \begin{bmatrix} \beta_0(s_i) \\ \beta_1(s_i) \\ \vdots \\ \beta_p(s_i) \end{bmatrix}$. Model (2.2) may be assembled into matrix form as

$$\underline{y} = \begin{bmatrix} \underline{x}_1^T & \underline{0}^T & \ldots & \underline{0}^T \\ \underline{0}^T & \underline{x}_2^T & \ldots & \underline{0}^T \\ \vdots & \vdots & \vdots & \vdots \\ \underline{0}^T & \underline{0}^T & \ldots & \underline{x}_n^T \end{bmatrix} \begin{bmatrix} \underline{\beta}(s_1) \\ \underline{\beta}(s_2) \\ \vdots \\ \underline{\beta}(s_n) \end{bmatrix} + \underline{e}$$

This model is a varying-coefficient model (Hastie and Tibshirani, 1993) in which the coefficients vary as functions of location. Specifically, for each observation y_i there exists a unique set of parameters which are functions of location s_i and thus the model has $n \times (p+1)$ regression parameters. Model (2.1) is a special case of model (2.2), in which the parameters are assumed to be constant over space.

2.3 Estimation of the regression parameters

2.3.1 Weighted Regression

Model (2.2) allows the regression parameters to vary over the region of interest. However, it has more unknown parameters than observed responses and thus there are immediate problems in the estimation of the parameters of this model. Fotheringham, Brunsdon and Charlton (2002) proposed that a solution to estimating the parameters at a given location s_0 is to assume that the parameters consistent with model (2.2) are constant within the vicinity of the location s_0 and to perform a weighted regression using a subset of observations that are close to location s_0. Points at which parameters are to be estimated are called regression points. There are two possibilities for a regression point. The parameters may either be estimated at the observation point locations only, i.e. where data

University of Cape Town

are observed, allowing comparisons with the data to be made. Alternatively an arbitrary set of grid points may be defined over the region of interest and parameters estimated at the grid point locations, which may or may not be observation points, thereby allowing continuous surfaces for each regression coefficient to be approximated over the region of interest. Specifically, consider a fixed regression point at location s_0 which may or may not be an observation point. Then the weighted regression model for an observation at s_i is

$$\begin{aligned} y_i &= \beta_0(s_0) + \sum_{k=1}^{p} \beta_k(s_0) x_{ik} + e_i \quad i = 1, ..., n \\ &= \underline{x}_i^T \underline{\beta}(s_0) + e_i \end{aligned} \tag{2.3}$$

where $e_i \sim rv(0, \frac{\sigma^2}{w_{0i}})$ with w_{0i} denoting the geographical weighting of the observed data point at s_i which is a function of the Euclidean distance

$$d_{0i} = \sqrt{(u_0 - u_i)^2 + (v_0 - v_i)^2}.$$

Model (2.3) may be assembled into matrix form as

$$\underline{y} = X\underline{\beta}(s_0) + \underline{e}$$

where $\underline{y}$ is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times (p+1)$ matrix of independent variables, $\underline{\beta}(s_0)$ is a $(p+1) \times 1$ vector of parameters to be estimated, $\underline{e} \sim rv(\underline{0}, \sigma^2 W^{-1}(s_0))$, and $W(s_0)$ is an $n \times n$ diagonal spatial weighting matrix expressed in matrix form as

$$W(s_0) = \begin{bmatrix} w_{01} & 0 & \dots & 0 \\ 0 & w_{02} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & w_{0n} \end{bmatrix}$$

where w_{0i} is a function of the distance d_{0i}. Observations are weighted according to their distance from location s_0. It is assumed that observations closer to s_0 have similar regression coefficients to those at s_0 and thus a weighting system is used so that observations closer to s_0 have more influence on the parameter estimate at s_0. The estimate of $\underline{\beta}(s_0)$, the parameter vector at the regression point at location s_0 is

$$\hat{\underline{\beta}}(s_0) = (X^T W(s_0) X)^{-1} X^T W(s_0) \underline{y} \tag{2.4}$$

2.3.2 Spatial weighting functions

Fotheringham, Brunsdon and Charlton (2002) provide a number of possible models for the spatial weighting function which will be referred to as kernel functions

University of Cape Town

(Hastie and Tibshirani, 1990). The simplest model assigns equal weighting to all observations within a certain distance from the regression point and observations lying beyond the specified distance are excluded. This uniform kernel is specified as

$$\begin{aligned} w_{0i} &= 1 \text{ if } d_{0i} < d \qquad i = 1, ..., n \\ &= 0 \text{ otherwise} \end{aligned} \tag{2.5}$$

where d_{0i} is the distance between the data point at s_i and regression point at s_0 where s_i represents the location of a specific point in space at which data are observed and s_0 represents the location of any point in space for which parameters are to be estimated, that is a regression point. A specified threshold distance is given by d, and w_{0i} is the assigned weight to the observation at s_i. The advantage of this model is that it simplifies the computational procedure by excluding all data points further than a specified distance from the regression point. However, this model has the problem of imposing discontinuities which lead to estimated parameters changing sharply over the region of interest.

Two commonly used weighting functions are the Gaussian and bi-square functions. The Gaussian function is defined as

$$w_{0i} = \exp\left[-\frac{1}{2}\left(\frac{d_{0i}}{b}\right)^2\right] \tag{2.6}$$

and the bi-square function is defined as

$$\begin{aligned} w_{0i} &= [1 - (d_{0i}/b)^2]^2 \text{ if } d_{0i} < b \\ &= 0 \text{ otherwise} \end{aligned} \tag{2.7}$$

where b is a distance referred to as the bandwidth. According to the Gaussian function, if data are observed at the regression point, this point will be given a weighting of 1 and the weighting of the other data points will decrease according to a Gaussian curve as the distance between them and the regression point increases. This ensures that observations closer to the regression point will have more influence on the parameter estimates than observations further away. The bi-square function provides a continuous near-Gaussian weighting up to a distance b from the regression point, and data points at and beyond the distance b are weighted as zero. This technique has the advantage of simplifying the computational procedure by excluding observations that are further than a certain distance from the regression point while still maintaining the property of continuity. Fotheringham, Brunsdon and Charlton (2000) provide several other weighting functions which are not that commonly used and are therefore not discussed here.

University of Cape Town

The spatial kernels discussed above are fixed in shape and size over space. A problem that may arise with fixed spatial kernels is that for some regression points around which data are sparse, models might be fitted with very few data points and hence the estimated parameters will have large standard errors. To reduce this problem, spatial kernels can be constructed which vary their bandwidth in accordance with the density of data around the regression point so that the bandwidth is greater where data points are sparse than where data are dense. Fotheringham, Brunsdon and Charlton (2002) discuss various methods of producing such spatially adaptive kernels. One method is to rank the data points in terms of their distance from the regression point. The closest data point has a weight of 1 and the weights decrease as the rank increases according to some continuous function. Another method is to ensure that the weights for performing the regression for any point of interest sum to some constant C. In areas where data are sparse, the kernel will have to expand to ensure that the weights sum to C, whereas in areas where data are dense, the kernel will have to contract. The practitioner could simply choose a value of C arbitrarily or try to find an optimal value.

2.4 Estimation of σ^2

The estimation of the unknown variance parameter σ^2 in GWR may be done either globally or locally. The estimate is used in the estimation of the variance of the parameter estimates $\hat{\underline{\beta}}(s_0)$.

2.4.1 Global estimation of σ^2

In the basic GWR model σ^2 is assumed constant over the study region. The estimate of the vector of parameters at the location s_0 can be written as

$$\hat{\underline{\beta}}(s_0) = A(s_0)\underline{y} \tag{2.8}$$

where $A(s_0) = (X^T W(s_0) X)^{-1} X^T W(s_0)$. The fitted value at observation point i with that point acting as the regression point in the weighted regression is given by

$$\begin{aligned} \hat{y}_i &= \underline{x}_i^T \hat{\underline{\beta}}(s_i) \\ &= \underline{x}_i^T A(s_i)\underline{y} \\ &= \underline{l}_i^T \underline{y} \end{aligned} \tag{2.9}$$

where $\underline{l}_i^T = \underline{x}_i^T A(s_i)$ for $i = 1, ..., n$. Since the observation is acting as the regression point, it is the case that location $s_0 = s_i$.

University of Cape Town

The fitted values $\hat{y}_i$ defined in equation (2.9) can thus be assembled as

$$\begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \underline{\hat{y}} = L\underline{y} \tag{2.10}$$

where L is the matrix with rows $\underline{l}_i^T$ for $i = 1, ..., n$ which maps the data $\underline{y}$ to the fitted values $\underline{\hat{y}}$. Then, based on equation (2.10), the residual sums of squared errors (RSS) is obtained in the usual way as follows

$$RSS = \underline{\hat{e}}^T\underline{\hat{e}} \tag{2.11}$$

where

$$\begin{aligned} \underline{\hat{e}} &= \underline{y} - \underline{\hat{y}} \\ &= \underline{y} - L\underline{y} \\ &= (I - L)\underline{y} \end{aligned}$$

Thus,

$$\begin{aligned} \underline{\hat{e}}^T\underline{\hat{e}} &= \underline{y}^T(I - L)^T(I - L)\underline{y} \\ &= \underline{y}^T[I - L^T - L + L^TL]\underline{y}. \end{aligned} \tag{2.12}$$

Following usual practice in non-parametric smoothing, the effective degrees of freedom for error are given by the trace of the matrix $[I - L^T - L + L^TL]$, and thus by $n - 2tr(L) + tr(L^TL)$ (Hastie and Tibshirani, 1990). The estimate of σ^2, which is a global estimate as σ^2 is assumed constant across the study region, is then defined as

$$s^2 = \frac{\underline{\hat{e}}^T\underline{\hat{e}}}{(n - 2v_1 + v_2)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n - 2v_1 + v_2)} \tag{2.13}$$

where $v_1 = tr(L)$ and $v_2 = tr(L^TL)$. The technique of estimating σ^2 globally is the one implemented by Fotheringham, Brunsdon and Charlton (2002). The term $n - 2v_1 + v_2$ is known as the effective degrees of freedom of the residuals and the term $2v_1 - v_2$ is the effective number of parameters in the local GWR model (Fotheringham, Brunsdon and Charlton, 2002).

The variance of $\underline{\hat{\beta}}(s_0)$, the parameter estimates at any regression point s_0 using the global estimate of σ^2 is then given by

$$Var[\underline{\hat{\beta}}(s_0)] = A(s_0)W^{-1}(s_0)A(s_0)^Ts^2 \tag{2.14}$$

This expression holds for all regression points.

University of Cape Town

2.4.2 Local estimation of σ^2

An alternative to the global approach for the estimation of σ^2 discussed in the previous section is to obtain a locally based estimate of the unknown variance σ^2 which is based on the assumption that σ^2 varies across the study region. The local estimate of σ^2 is that obtained by fitting the weighted regression model given by equation (2.3) and is defined as

$$s^2(s_0) = \frac{\sum_{i=1}^{n} w_{0i}(y_i - \hat{y}_i^{w_0})^2}{(n - p - 1)} \tag{2.15}$$

where $\hat{y}_i^{w_0}$ is the fitted value at s_i in the weighted regression centered at s_0 and w_{0i} is the geographical weighting of the observed data point relative to the regression point with the weightings w_{0i} summing to n. This technique for estimating σ^2 is the technique implemented in the `spgwr` R package (Bivand and Yu, 2007).

The variance of $\hat{\underline{\beta}}(s_0)$, the parameter estimates at any regression point using the local estimate of σ^2 is then given by

$$Var[\hat{\underline{\beta}}(s_0)] = (X^T W(s_0) X)^{-1} s^2(s_0) \tag{2.16}$$

2.5 Choice of bandwidth

Estimated parameters from GWR are dependent on the choice of weighting function and on the choice of bandwidth. Numerous weighting functions exist. However Brunsdon, Fotheringham and Charlton (1999) state that if the weighting function is continuous, with the weights decreasing as distance from the regression point increases, the choice of an appropriate bandwidth is far more important than the choice of weighting function. Larger bandwidths produce parameter estimates that are similar in value across the study area and have high bias. In other words as the bandwidth increases, the parameters estimated by GWR become closer to those estimated by a global model. Smaller bandwidths, on the other hand, produce parameter estimates with increased variance, since fewer points are included that carry weight. As a result of this 'bias-variance' dilemma, the selection of an appropriate bandwidth is very important. There are various options that can be used for bandwidth selection. For example:

- Expert opinion: The choice of bandwidth may be made subjectively, using an expert opinion, if one has prior beliefs about the value of bandwidth in a particular situation based on a sound theoretical understanding of that situation (Silverman, 1986).

- Cross-validation and Akaike Information Criterion (AIC): Cross-validation and AIC may be used to select bandwidth. The AIC criterion requires the

University of Cape Town

assumption of normality of the data to hold whereas the cross-validation technique does not require this assumption.

- k-nearest neighbours: Several rules of thumb have been suggested for estimating bandwidth selection based on the distances of the k-nearest neighbours from the point of interest (Bailey and Gatrell, 1995).

2.5.1 Cross-validation

Cross-validation, a technique widely used in Statistics with nonparametric modelling, involves the refitting of the model to predict each data point, with that data point being omitted from the fitting process (Hastie, Tibshirani and Friedman, 2001). The optimal bandwidth is that which minimises the sum of squares

$$CV(b) = \sum_{i=1}^{n} [y_i - \hat{y}_{(-i)}(b)]^2 \tag{2.17}$$

where $\hat{y}_{(-i)}(b)$ is the predicted value of y_i using a bandwidth b with the data point at i used as the regression point but omitted from the computations in the weighted regression model. $CV(b)$ is termed the CV score. The model may be refitted repeatedly with various values for bandwidth and the associated cross validation score calculated. By plotting the CV scores against bandwidths, guidance on selection of an appropriate bandwidth may be provided (Fotheringham, Brunsdon and Charlton, 2002). Clearly an optimal bandwidth is one which minimises the CV score. If a bandwidth which minimises the CV score is identified graphically, a more accurate value for this bandwidth may then be obtained by using an optimisation routine.

University of Cape Town

2.5.2 Akaike's Information Criterion

Akaike's Information Criterion (AIC) is widely used in Statistics to compare different models (Hastie, Tibshirani and Friedman, 2001). The general definition of AIC is

$$AIC = -2\ ln(L) + 2m \tag{2.18}$$

where L is the estimated likelihood function and m is the number of parameters in the model of interest. It is a measure of goodness of fit of a model that penalises the log-likelihood by the number of parameters. The model with the smallest AIC is chosen as the 'best' fitting model. A version of AIC for regression models, adapted to correct for bias, was presented by Hurvich and Tsai (1989), and is defined as

$$AIC_c = 2\ n\ ln(s) + n\ ln(2\pi) + n\left\{\frac{n+p+1}{n-3-p}\right\} \tag{2.19}$$

where n is the sample size, s is the estimated standard deviation of the error term, and p is the number of independent variables in the model. Following Hurvich and Tsai (1989), Fotheringham, Brunsdon and Charlton (2002) define an adapted AIC for GWR by replacing $p+1$ with $\mathrm{tr}(L)$ in expression (2.19), the trace of the matrix L which maps $\underline{\hat{y}}$ on to $\underline{y}$ as shown in equation (2.12). The adapted AIC used in GWR is thus defined as

$$AIC_c = 2\ n\ ln(s) + n\ ln(2\pi) + n\left\{\frac{n+tr(L)}{n-2-tr(L)}\right\} \tag{2.20}$$

2.6 Spatial Non-stationarity

An important question to answer is whether a particular set of local parameter estimates obtained from fitting a GWR model exhibit significant spatial variation over the region of interest, and hence whether the use of a spatially varying regression model is justified. It is necessary to determine whether the observed pattern has arisen due to true spatial trend in the data or simply due to random variation. To do so, the stationarity of parameters based on their variability over space when estimated using GWR can be tested as follows.
Consider n data points within a region and consider a particular regression coefficient $\hat{\beta}_k$. A GWR estimate of the regression coefficient, $\hat{\beta}_k(s_i)$, is taken at each data point $i = 1, \ldots, n$, and the variance of the n parameter estimates can be computed. The variance provides a useful statistic to measure the variability of the parameter estimates over space, and is defined as

$$s^2(\hat{\beta}_k) = \frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_k(s_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_k(s_i)]^2. \tag{2.21}$$

To test the hypothesis that the parameter is globally constant, the sampling distribution of $s^2(\hat{\beta}_k)$ under the null hypothesis that the global model holds, has to be determined. To do so, a Monte Carlo approach is adopted, since this distribution is analytically intractable (Fotheringham, Brunsdon and Charlton, 2002). Under the null hypothesis, it is assumed that the parameters do not vary over space and thus any permutation of the regression variables against their locations should be equally likely to occur. Thus, the data can be repeatedly rearranged in space by randomly allocating the observed data pairs $(y_i, \underline{x}_i)$ for $i = 1, ..., n$ across locations without replacement. The GWR model is then fitted to the randomly allocated pairs and the estimated variance defined in equation (2.21) calculated. This procedure is repeated a large number of times and the

University of Cape Town

actual value of $s^2(\hat{\beta}_k)$ for the data at the correct locations can be compared to those obtained from the randomised distributions to obtain an experimental significance level. The proportion of the randomised values of the variance $s^2(\hat{\beta}_k)$ exceeding the actual $s^2(\hat{\beta}_k)$ is calculated and used as a p-value in the test of the hypothesis

$$H_0 : \beta_k \text{ is stationary across the region of interest vs.}$$
$$H_A : \beta_k \text{ is non-stationary across the region of interest}$$

This procedure is repeated for all parameters β_k for $k = 0, \ldots, p$.

University of Cape Town

Chapter 3

A Proposed Extension to the GWR Model

University of Cape Town

In this chapter, an extension to the GWR model is proposed. The extended model is called Local Linear GWR (LLGWR) and is essentially an application of the expansion method (Casetti, 1972) to the GWR model. The general expansion model is presented in Section 3.1 and the development of the model is discussed in Section 3.2.

3.1 The Expansion Method

Casetti (1972) defines an expansion method for the construction and modification of a given model and illustrates the method by a number of examples. The method is demonstrated on a number of models concerned with population growth. The expansion method is a procedure whereby a more complex model is generated from a simpler initial model by redefining at least some of the parameters of the initial model as functions of relevant variables. The expanded parameters are then substituted back into the initial model to produce a new model. The expansion method builds upon an initial model and may be used to construct new models that meet requirements that an initial model does not satisfy, to improve predictability or to remove inadequacies of an initial model. The expansion method is especially suited to cases in which the parameters of initial models appear to vary in a trend-like manner. Several models in the literature have been developed by applying the expansion method. Hastie and Tibshirani (1993) describe the varying-coefficients model which is a particular case of the expansion method and which expands regression models with the aim of increasing the flexibility of the models by allowing the regression coefficients to vary as functions of other variables. Although most applications of the expansion method relate to non-spatial settings, there have been applications for which parameters have been expanded based on the coordinates of location in space (Elridge and

Jones, 1991; McMillen, 1996).

3.2 Development of the LLGWR model

The Local Linear GWR model is developed through the expansion of the GWR model. The aim of the expansion of the GWR model is to increase parameter flexibility and to capture more of the variability in a spatial data set.

For a fixed regression point at location s_0, where $s_0 = (u_0, v_0)$ with u_0 and v_0 the longitude and latitude respectively, the GWR model for an observation at location s_i as defined in equation (2.3) is given by

$$\begin{aligned} y_i &= \beta_0(s_0) + \sum_{k=1}^{p} \beta_k(s_0) x_{ik} + e_i \\ &= \underline{x}_i^T \underline{\beta}(s_0) + e_i \end{aligned}$$

where y_i is the observed value of the dependent variable at location (u_i, v_i), x_{ik} are observed values of the independent variables for $i = 1, ..., n$ and $\beta_k(s_0)$ are unknown parameters at location (u_0, v_0). The error terms are random variables with zero mean and variance $(\frac{\sigma^2}{w_{0i}})$, where w_{0i} denotes the weight given to the observed point at s_i which is a function of the distance of that point from the regression point at s_0.

Consider a specific regression point at location s_0. Then model (2.3) may be extended by expanding the regression coefficients in terms of the latitude u_i and longitude v_i of an observed point relative to the latitude u_0 and longitude v_0 of the the regression point. Specifically consider

$$y_i = \beta_0^*(s_0) + \sum_{k=1}^{p} \beta_k^*(s_0) x_{ik} + e_i \tag{3.1}$$

where the coefficients are expanded as

$$\beta_k^*(s_0) = \beta_k(s_0) + \beta_k^u(s_0)(u_i - u_0) + \beta_k^v(s_0)(v_i - v_0) \tag{3.2}$$

for $k = 0, 1, \ldots, p$ and for each observation $i = 1, ..., n$. Equation (3.2) expands the regression coefficients in model (3.1) using a first order linear approximation. The use of first order approximations has been shown to be effective in the literature on local regression (Loader, 1999). For estimation, the parameters $\beta_k(s_0)$, $\beta_k^u(s_0)$ and $\beta_k^v(s_0)$ are assumed constant in the neighbourhood of the regression point, as is the case of GWR.

University of Cape Town

To fix ideas, the technique is illustrated for simple regression. The GWR model for a fixed regression point at s_0 is

$$y_i = \beta_0(s_0) + \beta_1(s_0)x_i + e_i \qquad i = 1, ..., n \tag{3.3}$$

where $e_i \sim rv(0, \frac{\sigma^2}{w_{0i}})$. The model is extended through the expansion of the coefficients as defined in equation (3.2) to yield the LLGWR model given by

$$\begin{aligned} y_i = \quad & \beta_0^*(s_0) + \beta_1^* x_i + e_i \\ = \quad & \beta_0(s_0) + \beta_0^u(s_0)(u_i - u_0) + \beta_0^v(s_0)(v_i - v_0) + \\ & \beta_1(s_0)x_i + \beta_1^u(s_0)(u_i - u_0)x_i + \beta_1^v(s_0)(v_i - v_0)x_i + e_i \qquad i = 1, ..., n \end{aligned} \tag{3.4}$$

Model (3.4) may be written in matrix form as

$$\underline{y} = X^*(s_0)\underline{\beta}^*(s_0) + \underline{e}$$

where $\underline{y}$ is an $n \times 1$ vector of observations on the dependent variable, $X^*(s_0)$ is an $n \times 6$ augmented matrix of independent variables, $\underline{\beta}^*(s_0)$ is a 6×1 vector of the parameters to be estimated, $\underline{e} \sim rv(\underline{0}, \sigma^2 W^{-1}(s_0))$, and $W(s_0)$ is an $n \times n$ diagonal spatial weighting matrix. The augmented matrix $X^*(s_0)$ is assembled as

$$X^*(s_0) = \begin{bmatrix} 1 & (u_1 - u_0) & (v_1 - v_0) & x_1 & (u_1 - u_0)x_1 & (v_1 - v_0)x_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (u_i - u_0) & (v_i - v_0) & x_i & (u_i - u_0)x_i & (v_i - v_0)x_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (u_n - u_0) & (v_n - v_0) & x_n & (u_n - u_0)x_n & (v_n - v_0)x_n \end{bmatrix}$$

and the vector of parameters is assembled as $\underline{\beta}^*(s_0) = \begin{bmatrix} \beta_0(s_0) \\ \beta_0^u(s_0) \\ \beta_0^v(s_0) \\ \beta_1(s_0) \\ \beta_1^u(s_0) \\ \beta_1^v(s_0) \end{bmatrix}$.

More generally the LLGWR model for p explanatory variables at a regression point $s_0 = (u_0, v_0)$, can be summarised as

$$\begin{aligned} y_i \quad = \quad & \beta_0(s_0) + \beta_0^u(s_0)(u_i - u_0) + \beta_0^v(s_0)(v_i - v_0) \\ & + \sum_{k=1}^{p} [\beta_k(s_0)x_{ik} + \beta_k^u(s_0)(u_i - u_0)x_{ik} + \beta_k^v(s_0)(v_i - v_0)x_{ik}] + e_i \\ = \quad & (\underline{x}_i^*)^T \underline{\beta}^*(s_0) + e_i \qquad i = 1, \ldots, n \end{aligned} \tag{3.5}$$

University of Cape Town

where $(\underline{x}_i^*)^T$ are row vectors of the $n \times 3(p+1)$ augmented matrix of independent variables $X^*(s_0)$, which includes additional variables through which the linear extension is incorporated, $\underline{\beta}^*(s_0)$ is a $3(p+1) \times 1$ vector of parameters to be estimated and the error terms are random variables with zero mean and variance σ^2/w_{0i}.

The LLGWR model has too many parameters i.e. $n \times 3(p+1)$ since there are only n observations. It is implemented by assuming that the paramters are locally constant as is the case with GWR and is thus fitted using weighted regression on the expanded model. The estimate of $\underline{\beta}^*(s_0)$, the parameter vector at the regression point at location s_0, is given by

$$\underline{\hat{\beta}}^*(s_0) = (X^*(s_0)^T W(s_0) X(^*s_0))^{-1} X^*(s_0)^T W(s_0) \underline{y} \tag{3.6}$$

and the variance of $\hat{\beta}^*(s_0)$ is given by

$$Var[\underline{\hat{\beta}}^*(s_0)] = A(s_0) W^{-1}(s_0) A(s_0)^T s^2$$

where $A(s_0) = (X^*(s_0)^T W(s_0) X^*(s_0))^{-1} X^*(s_0) W(s_0)$ and s^2 is the estimate of the unknown variance parameter σ^2 defined in equation 2.13. The model can be fitted using weighted regression centered on observation points at which predictions can be made and can also be fitted at arbitrarily defined grid points allowing continuous surfaces for each regression coefficient to be approximated over the region of interest. The spatial variability of the additional parameters may also be examined.

In the GWR model, the spatial variability of the regression coefficients is accommodated by invoking weighted regression centered at a point of interest and with weights decreasing as the distance of observations from that point increases. In the LLGWR model, the spatial variability in the regression coefficients is accommodated by including a weighting function as well as by expanding the regression coefficients, thus allowing for more flexibility in the model. At the same time however the LLGWR model produces a large number of parameters, 3 times as many parameters as the GWR model, and thus could become unwieldy. The aim of the dissertation is to implement LLGWR and to investigate whether or not it provides an improvement over the GWR approach.

University of Cape Town

Chapter 4

A Small Data Set taken from Soil Science

University of Cape Town

The analysis of a data set taken from soil science is presented in this chapter. A description of the data is given in Section 4.1. The data are summarised by means of some simple graphing techniques and descriptive statistics in Section 4.2 and a global analysis of the data using ordinary least squares regression is presented in Section 4.3. This is followed by the applications of local analyses, namely Geographically Weighted Regression (GWR), Local Linear GWR and kriging presented in Sections 4.4, 4.5 and 4.6 respectively. A summary and comparison of the results obtained from the analyses based on these four models are given in Section 4.7.

4.1 Data

A data set was taken from a bulletin by Clarke and Dane (1991) written at Auburn University, Alabama, USA. The data set comprised records of water and clay content recorded at 60 locations in a rectangular field of 50m $\times$ 100m at the Alabama Agricultural Experiment Station. The data are presented in Appendix A. Measurements of water content in cm^3 of water per cm^3 of soil and clay content as a percentage of unit weight of soil were taken at a depth of 80 cm at each location. The data also include latitude and longitude as distances along the x and y axes from the origin $(0, 0)$ respectively. Two of the locations had missing data and were excluded from the present analysis. The analysis was thus performed on the 58 locations for which data were recorded. The aim of the analysis was to establish a relationship between water content as the dependent variable and clay content as the independent variable.

4.2 Exploratory Data Analysis

The initial step in the analysis of a spatial data set is to summarize the data through some simple graphing techniques and descriptive statistics. The spatial distributions of water and clay content are plotted in Figure 4.1. From this figure it can be seen that water and clay content have similar distributions across the field suggesting a positive relationship between the two variables. The concentrations of both water and clay are highest at the north end of the field and decrease toward the south. Along the south edge of the field is a band of observation points containing the lowest water and clay concentrations. Observations in the northern half of the field comprise mostly of high water and clay concentrations but there are 2 locations in the north-eastern corner where the concentrations of water and clay are low. There are also high concentrations of clay located in the mid-west of the field.

A scatterplot of water against clay content is presented in Figure 4.2. From this graph it is apparent that a positive relationship exists between the two variables. Descriptive statistics of water and clay content are given in Table 4.1. Clay content has a higher coefficient of variation which measures the amount of variation relative to the mean, than water content, indicating greater relative variation in clay content than in water content. The distribution of water content is negatively skewed as is indicated by the skewness coefficient and as can be seen in the histogram of water content presented in Figure 4.3. The histogram of clay content also shown in Figure 4.3 appears to be reasonably symmetrical and this is confirmed by the small positive skewness coefficient.

	Water content (cm^3/cm^3)	Clay content (%)
Minimum	0.204	13.400
Median	0.270	22.750
Maximum	0.314	32.900
Mean	0.269	22.717
Std deviation	0.027	4.386
Coefficient of variation	0.102	0.193
Skewness	-0.280	0.106

Table 4.1: Descriptive statistics of water and clay content

University of Cape Town

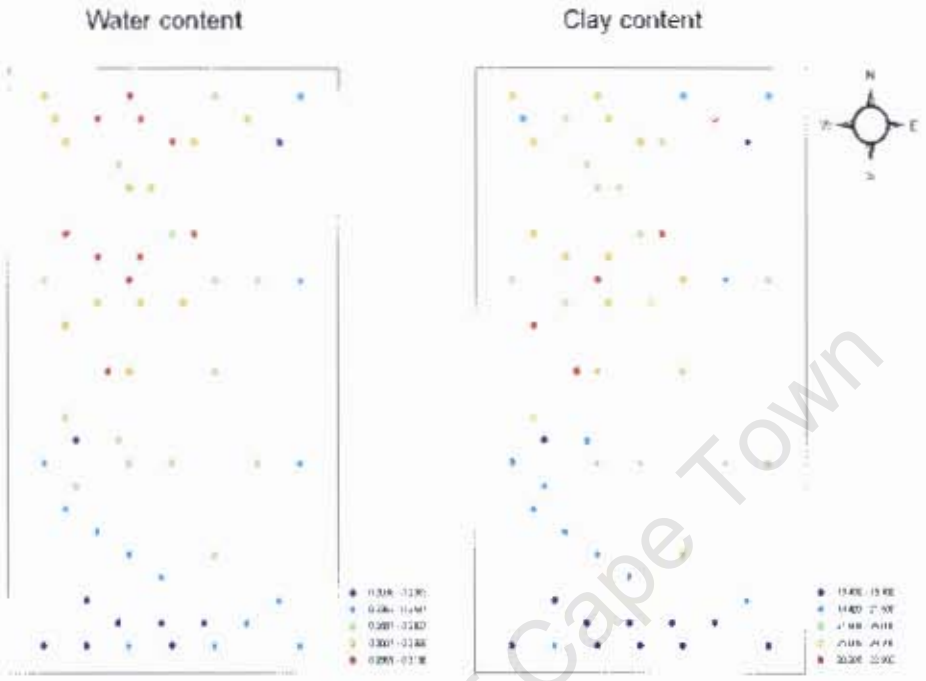


Figure 4.1: Spatial distribution of water content (cm^3/cm^3) and clay content(%)

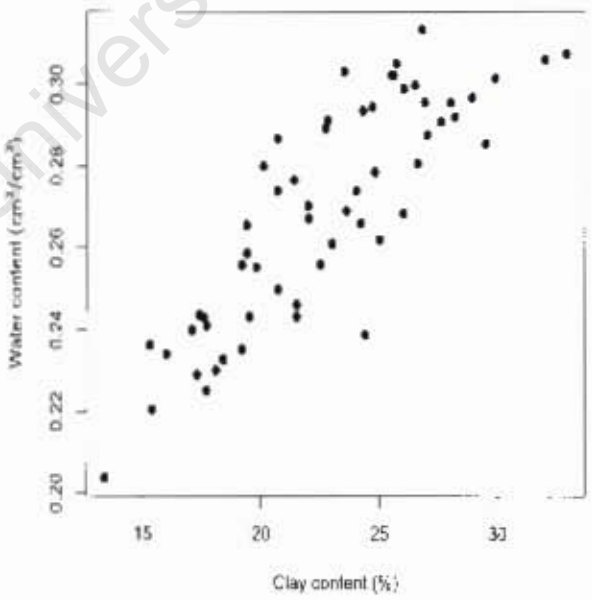


Figure 4.2: Scatterplot of water against clay

University of Cape Town

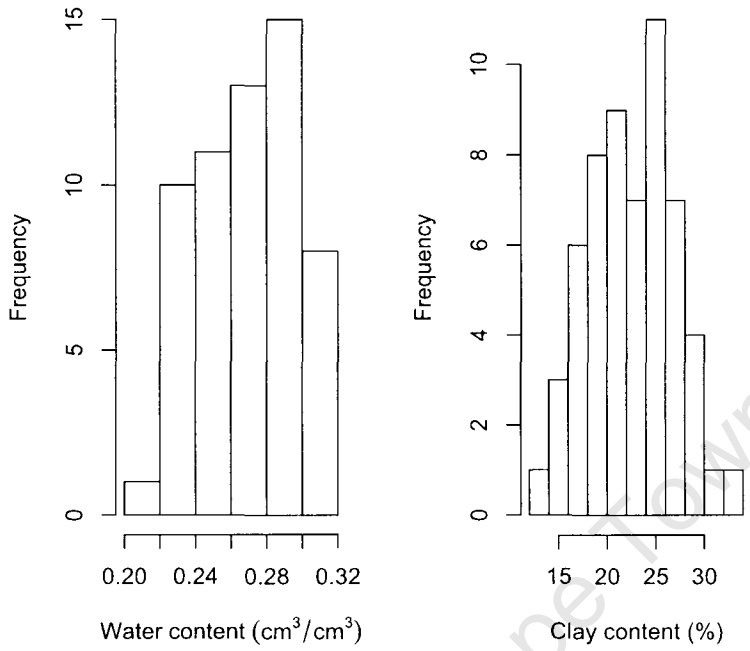


Figure 4.3: Histograms of water and clay content

University of Cape Town

4.3 Global analysis

4.3.1 Global model

The simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \ldots, 58 \tag{4.1}$$

where y_i is the i^{th} observation of water content, x_i the clay content and e_i an independent random error term was fitted to the data using ordinary least squares regression. This was implemented using the package R. Transformations of the clay variable were considered but they provided no better fit to the data than the use of the untransformed clay variable. The parameter estimates obtained by fitting model (4.1) to the data, their standard errors and their p-values from t-tests of the hypotheses,

$$H_0 : \beta_k = 0 \text{ vs. } H_A : \beta_k \neq 0 \text{ for k} = 0,1$$

are reported in Table 4.2. The t-tests are based on the assumptions that the error terms are normally distributed and have constant variance however, t-tests are robust to deviations from those assumptions. According to the results of the t-tests, the intercept and clay coefficients are both significantly different to zero.

Coefficient	Estimate	Std Error	p-value
β_0	0.1494	0.01029	< 0.001
β_1	0.0053	0.0004	< 0.001

Table 4.2: Parameter estimates of the simple linear regression (global) model

The regression has an R^2 value of 71%, and thus the model provides a good fit to the data. The AIC for the model is -320.62, and the Residual Sums of Squares (RSS) value is 0.012.

4.3.2 Residuals

A normal Q-Q plot of residuals, defined as $e_i = y_i - \hat{y}_i$ for $i = 1, ..., 58$ where y_i is the i^{th} observed value of water content and $\hat{y}_i$ the corresponding fitted value from model (4.1), is shown in Figure 4.4 (a) and a plot of the residuals against the fitted values is shown in Figure 4.4 (b). From Figure 4.4 (a), the residuals appear to be fairly normally distributed. The points appear to be randomly spaced in the plot given by Figure 4.4(b). The spatial distribution of the residuals shown in Figure 4.5 however, appears to be non-random. Large positive residuals are located in the north-east of the map, and negative residuals are located in the south. The residuals were further investigated to identify potential outliers. Standardised residuals, calculated as

$$\frac{y_i - \hat{y}_i}{s}$$

where s is the standard error obtained from fitting the regression model, are useful for outlier detection. Generally standardised residuals which are greater than 2 in absolute value are considered to be potential outliers. A plot of the standardised residuals is presented in Figure 4.6. An examination of this plot revealed observations 38 and 52 as potential outliers. The impact of these observations on the regression analysis was investigated but deleting them made very little difference to the results. It was thus decided to retain these observations in the remainder of the study.

4.3.3 Global models fitted over quadrants

The field was divided into four quadrants as is shown in Figure 4.7 and simple linear regression models were fitted separately to the data for each quadrant. This allows a simple way of examining whether the relationship modelled between water and clay content is likely to be stationary over space. Results of the global models fitted separately for each quadrant are presented in Table 4.3 and scatterplots of water against clay content with associated fitted regression lines for each quadrant are presented in Figure 4.8.

University of Cape Town

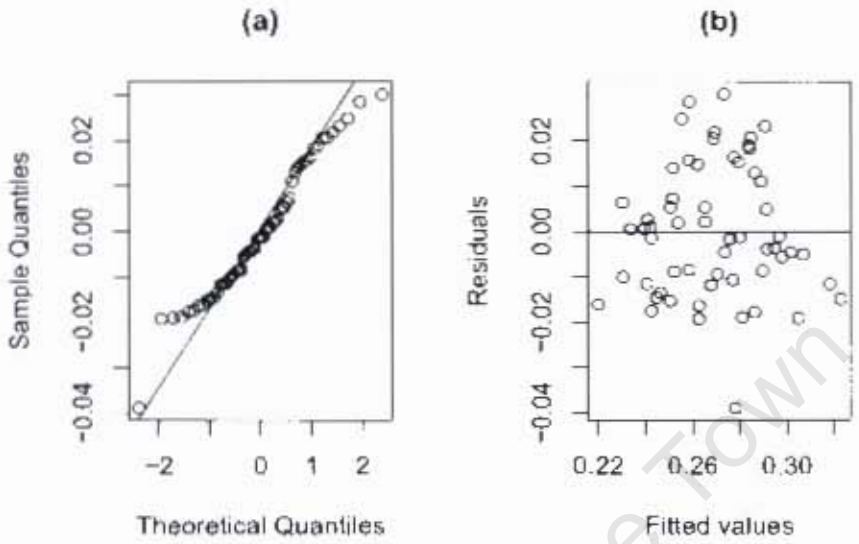


Figure 4.4: (a)Normal Q-Q plot of residuals from global model and (b)Plot of residuals against fitted values from the global model

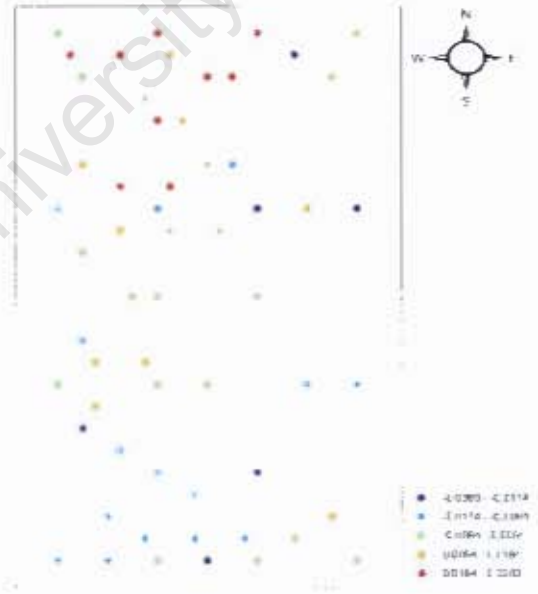


Figure 4.5: The distribution of the residuals for the global model

University of Cape Town

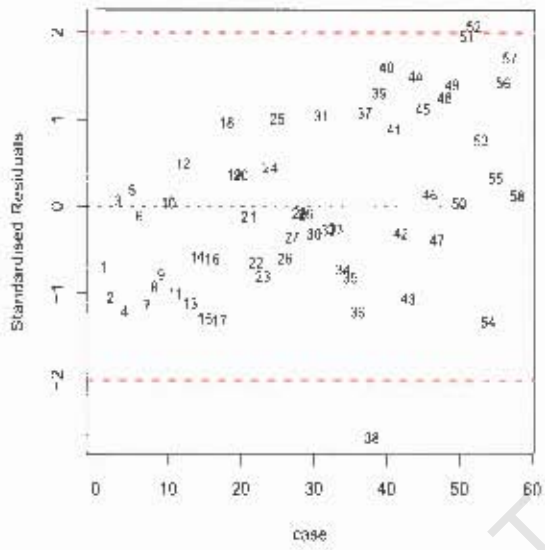


Figure 4.6: Plot of standardised residuals

		Intercept		Clay		
	n_j	$\hat{\beta}_0$	Std Error of $\hat{\beta}_0$	$\hat{\beta}_1$	Std Error of $\hat{\beta}_1$	R^2
Quadrant 1	20	0.1372	0.0122	0.0056	0.0006	0.84
Quadrant 2	19	0.2274	0.0273	0.0026	0.0011	0.26
Quadrant 3	10	0.1959	0.0287	0.0032	0.0012	0.48
Quadrant 4	9	0.1712	0.0168	0.0039	0.0008	0.77

Table 4.3: Results from separate regressions for each quadrant where n_j ($j = 1, \ldots, 4$) is the number of data points in each quadrant

Examination of these results reveals some differences in the parameter estimates between the quadrants suggesting non-stationarity. In particular, quadrant 1 characterised by areas of high and areas of very low water concentration has the lowest intercept and steepest gradient. Quadrant 2, characterised entirely by high water concentration, has the the highest intercept and flattest gradient. The models fitted to the data in quadrants 1 and 4 provided good fits with R^2 values of 84% and 77% respectively. The models fitted to the data in quadrants 2 and 3 exhibit poor fits, especially the model fitted in quadrant 2 which has a very low R^2 value of 26%.

University of Cape Town

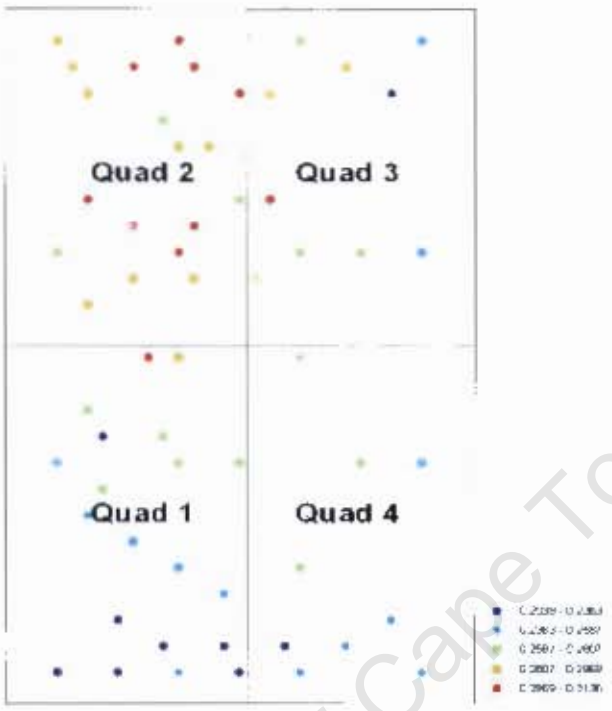


Figure 4.7: The distribution of observations of water content divided into 4 quadrants

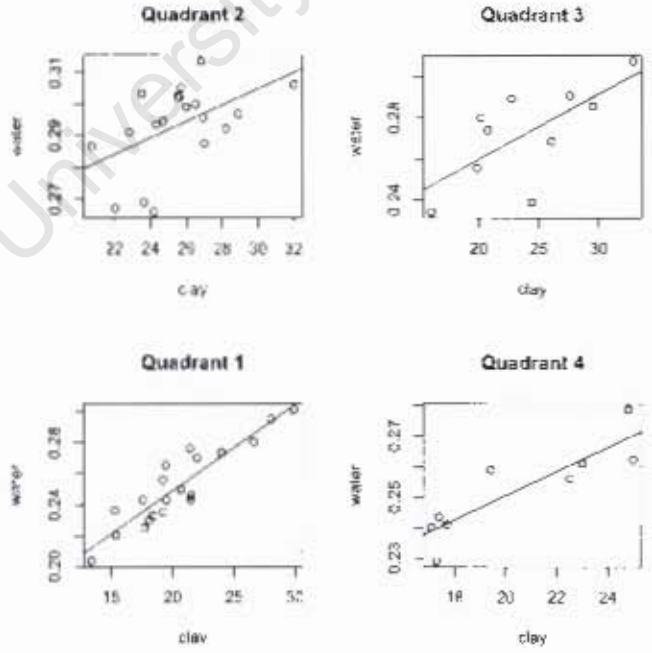


Figure 4.8: Scatterplots of water against clay content for each quadrant

University of Cape Town

4.4 Application of GWR

GWR analysis was performed using programs written in the language R (R Development Core Team, 2006). The code for the programs is provided in Appendix B. Results were crossed-checked with GWR 3 software as well as with the `spgwr` library in the package R. The weighted regression model centered at a regression point located at s_0 and for an observation at s_i as defined in equation (2.3) is given by

$$y_i = \beta_0(s_0) + \beta_1(s_0)x_i + e_i \quad i = 1, \ldots, 58 \tag{4.2}$$

where y_i is the observed value of water content at location s_i, x_i is the observed value of clay content, $\beta_0(s_0)$ and $\beta_1(s_0)$ are unknown parameters and $e_i \sim rv(0, \frac{\sigma^2}{w_{0i}})$, with w_{0i} denoting the weighting given to the observed point at s_i which is a function of the distance of that point from the regression point at s_0. Model (4.2) was fitted to the data using weighted regression centered on each observation point. A fixed Gaussian kernel was used to define the weights associated with each observation since the Gaussian kernel is one of the most commonly used weighting functions and according to Brunsdon, Fotheringham and Charlton (1999) as discussed in Section 2.5, the choice of bandwidth is far more important than the choice of weighting function. The cross validation criterion was invoked for bandwidth selection. The model was fitted with various possible values of bandwidth and the resulting CV scores calculated according to equation (2.17). The CV scores are plotted against bandwidth in Figure 4.9 to provide some guidance with bandwidth selection. The optimal value for bandwidth is that which minimises the CV score. From Figure 4.9, it may be seen that a minimum CV score exists for a bandwidth of approximately 17 m. The optimal value for bandwidth was calculated more exactly as 17.282 m using the routine `optim` in R which implements a form of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method of optimisation (Byrd, Lu, Nocedal, and Zhu, 1995).

Parameter estimates were obtained using observation points as the regression points and using the optimal bandwidth of 17.282 m. A five-number summary of the estimated model parameters at the 58 observation point locations is given in Table 4.4. The individual parameters may be tested for significance using a pseudo t-test (Fotheringham, Brunsdon and Charlton, 2002). The t-statistics are obtained by dividing each local estimate by the corresponding local standard error of the estimate. The usual critical value is used to assess the level of significance of the t-value even though it is not really applicable since many hypotheses are being tested. The hypothesis

$$H_0 : \beta_0 = 0 \text{ vs. } H_A : \beta_0 \neq 0$$

was rejected for all cases and the hypothesis

University of Cape Town

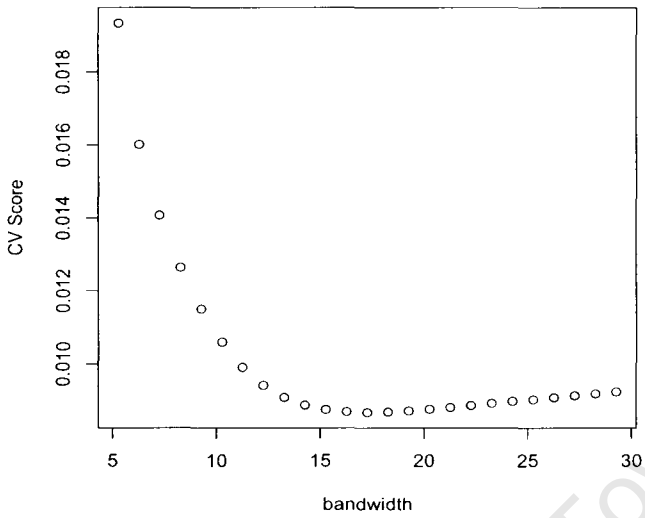


Figure 4.9: Variation in CV score with bandwidth using a Gaussian kernel. The minimum CV score exists for a bandwidth of approximately 17m.

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

was rejected for all except one case. The model has an AIC_c value of -342.471 and a Residual Sums of Squares (RSS) value of 0.006. The effective degrees of freedom for the model are 45.55.

	$\hat{\beta}_0$	$\hat{\beta}_1$
Minimum	0.01359	0.0020
Lower quartile	0.1535	0.0033
Median	0.1785	0.0040
Upper quartile	0.2035	0.0048
Maximum	0.2424	0.0055

Table 4.4: Five-number summary of parameter estimates for model (4.2) over the 58 locations.

Maps of the estimated parameters over the experimental region were produced to illustrate their spatial variation. A grid of 50×100 equally spaced points was defined over the study region. The grid was created using a program written in the language R by Brunsdon (2006). Model (4.2) was fitted to the data using weighted regression centered on each grid point. Parameters were thus estimated at each of the grid points producing 5000 estimates for each parameter over space. These estimates as well as their standard errors defined as $se(\hat{\beta}_k) = \sqrt{Var[\hat{\beta}_k]}$ for $k = 0, 1$, were mapped using ArcGIS software (ESRI, 2005) and are presented in Figures 4.10 and 4.11 respectively. Spatial variations

University of Cape Town

are evident in both parameter estimates. The intercept coefficient estimates, as can be seen in Figure 4.10 (a), show a clear pattern with higher values located in the north-west of the field and lower values located in the south. The standard errors of these estimates as can be seen in Figure 4.10 (b) are highest in the corners of the field. The estimated clay coefficients mapped in Figure 4.11 (a) have the highest values located in the south-west of the field, and lowest values in the north-west. The standard errors of these estimates are highest in the north-western corner as well as along the southern edge of the field. A comparison of Figure 4.10 (a) with Figure 4.11 (a) shows that high intercept values correspond to low values of the clay coefficient and low intercept values correspond to high values of the clay coefficient.

The Monte Carlo method described in Section 2.6 was used to determine whether or not the parameters displayed significant non-stationarity. Specifically the data were rearranged in space by randomly allocating the (y_i, x_i) pairs of data points without replacement across observed locations. The GWR methodology was implemented using the randomised observations and the variances of the parameter estimates, $s^2(\hat{\beta}_0)$ and $s^2(\hat{\beta}_1)$ as defined in equation (2.21) computed. This randomisation was repeated 1000 times and the proportions of the variances $s^2(\hat{\beta}_k)$ for $k = 0, 1$ exceeding the actual variance obtained from the data at the correct locations were calculated and found to be 0.001 and 0.022 respectively. These proportions provide a measure of the probability of observing variation in the local parameter estimates at least as extreme as that observed for the actual data if the parameter were globally constant. The hypothesis

$$H_0 : \beta_k \text{ is stationary across space vs.}$$
$$H_A : \beta_k \text{ is non-stationary across space for } k = 0, 1$$

is thus rejected with p-values of 0.001 and 0.022 for the intercept and clay coefficients respectively and it may be concluded that significant spatial non-stationarity exists in both coefficients.

University of Cape Town

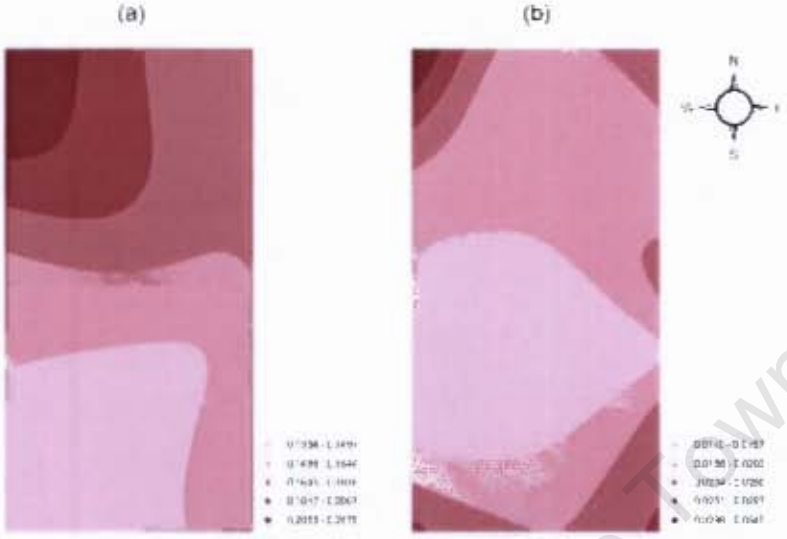


Figure 4.10: Maps of (a) $\hat{\beta}_0$ and (b) $se(\hat{\beta}_0)$ from the GWR analysis

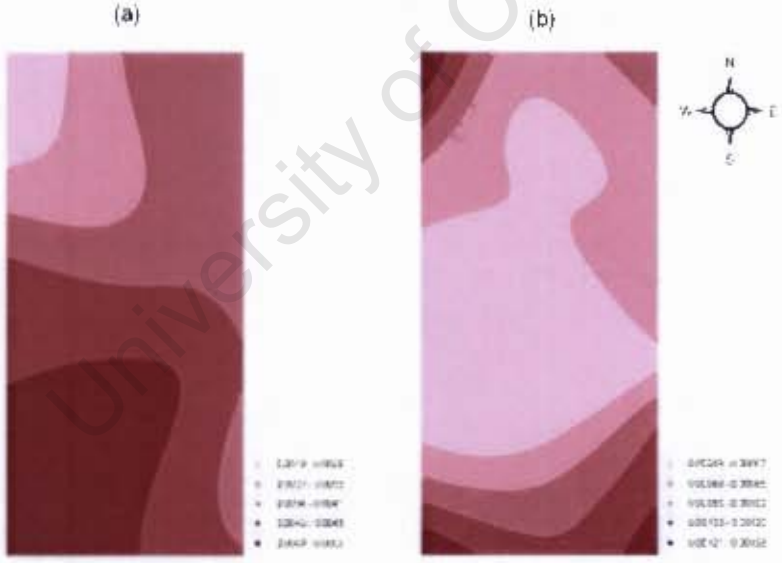


Figure 4.11: Maps of (a) $\hat{\beta}_1$ and (b) $se(\hat{\beta}_1)$ from the GWR analysis

4.5 Implementation of LLGWR

Local Linear GWR analysis was performed on the data using programs written in the language R. The code for the programs is provided in Appendix C. The weighted regression model for an observation at location s_i and centered at a

University of Cape Town

regression point at s_0 as defined in equation (3.6) is given by

$$\begin{aligned} y_i &= \beta_0(s_0) + \beta_0^u(s_0)(u_i - u_0) + \beta_0^v(s_0)(v_i - v_0) \\ &\quad + [\beta_1(s_0) + \beta_1^u(s_0)(u_i - u_0) + \beta_1^v(s_0)(v_i - v_0)]x_i + e_i \\ &\qquad\qquad i = 1, \ldots, 58 \end{aligned} \quad (4.3)$$

where y_i is the observed value of water content at location s_i, the location of all data points $i = 1, \ldots, 58$, x_i the clay content, $\beta_k(s_0)$, $\beta_k^u(s_0)$, $\beta_k^v(s_0)$ for $k = 0, 1$ are the unknown parameters at the regression point at location s_0, (u_i, v_i) denotes the location of the i-th point and $e_i \sim rv(0, \frac{\sigma^2}{w_{0i}})$. Model (4.3) was fitted to the data using weighted regression centered on each observation point. A fixed Gaussian kernel was used to define the weights associated with each observation and bandwidth was selected based on the cross validation criterion. The model was fitted for a range of bandwidths and the resulting CV scores calculated. The CV scores are plotted against bandwidth in Figure 4.12 to provide guidance with bandwidth selection. From this figure it may be seen that an optimal value for bandwidth exists, with a much clearer minimum CV score than in the case of GWR. The bandwidth corresponding to the minimum CV score was calculated as 30.69 m, again using the `optim` routine in R. This optimal bandwidth is much larger than the one calculated in the case of GWR and the reason for this cannot readily be explained.

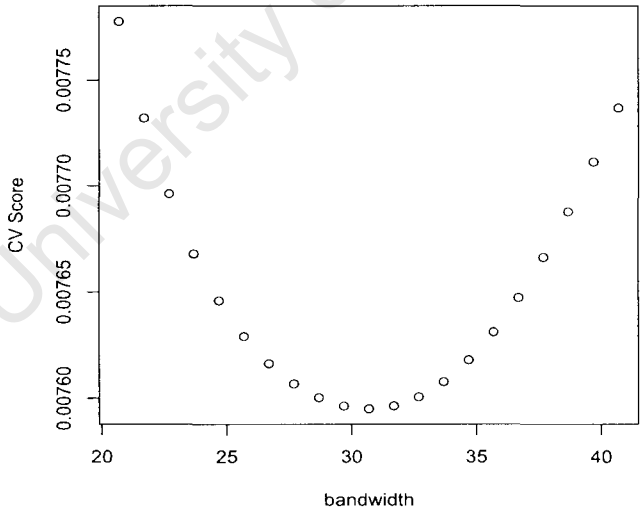


Figure 4.12: Variation in CV score with bandwidth for the LLGWR model using a Gaussian kernel. The minimum CV score exists for a bandwidth of approximately 30m.

Parameter estimates were obtained using observation points as the regression points and the using the optimal bandwidth of 30.69 m. A five-number summary

University of Cape Town

of the estimated model parameters at the 58 observation point locations are presented in Table 4.5. The individual parameters were tested for significance using a pseudo t-test. The proportions of cases for which each of the hypotheses

$$\begin{aligned} H_0 &: \beta_k = 0 \quad \text{vs.} \quad H_A : \beta_k \neq 0 \\ H_0 &: \beta_k^u = 0 \quad \text{vs.} \quad H_A : \beta_k^u \neq 0 \\ H_0 &: \beta_k^v = 0 \quad \text{vs.} \quad H_A : \beta_k^v \neq 0 \quad \text{for k} = 0, 1 \end{aligned}$$

are rejected at the 5% significance level are summarised in Table 4.6. The parameter β_0 was found to be significantly different from zero at all of the locations and β_1 was found to be significantly different from zero at 91% of the locations. The parameter β_1^v was found to be nonsignificant at all of the locations, β_0^u and β_1^u were found to be significant at approximately a third of the locations and β_0^v was found to be significant at just over half of the locations. It thus appears worthwhile to include the linear extension. The model has an AIC_c value of -348.12 and a Residual Sums of Squares (RSS) value of 0.005. The effective degrees of freedom for the model are 45.46.

Parameter	$\hat{\beta}_0$	$\hat{\beta}_0^u$	$\hat{\beta}_0^v$	$\hat{\beta}_1$	$\hat{\beta}_1^u$	$\hat{\beta}_1^v$
Minimum	0.133071	-0.001360	0.000438	0.001620	-0.000120	-0.000120
Lower quartile	0.170441	-0.000434	0.000699	0.003091	-0.000085	-0.000020
Median	0.193142	0.000746	0.000781	0.003459	-0.000052	-0.000016
Upper quartile	0.170441	0.000434	0.000699	0.003091	-0.000085	-0.000020
Maximum	0.247656	0.002440	0.001187	0.005392	0.000036	-0.000004

Table 4.5: Five-number summary of parameter estimates for model (4.3)

	β_0	β_0^u	β_0^v	β_1	β_1^u	β_1^v
Proportion of locations where the parameter is significantly different to zero	1.00	0.38	0.52	0.91	0.33	0.00

Table 4.6: Proportion of locations where the individual parameters in model (4.3) were found to be significantly different to zero

Maps of the estimates of the parameters were produced to illustrate their variation over space. The same grid defined for the GWR model was used and model (4.3) was fitted to the data using weighted regression centered on each grid point. Parameters were thus estimated at each of the grid points producing 5000 estimates for each parameter. The estimates of the coefficients as well as the standard errors of the estimates from the LLGWR analysis were mapped using ArcGIS software and are presented in Figures 4.13 to 4.18. Spatial patterns

University of Cape Town

are evident for some of the parameters. The Monte Carlo method, as detailed in Section 2.6, was used to determine whether or not the parameters displayed significant non-stationarity. 1000 randomisations of the data were performed and the results of the tests of the following hypotheses

$H_0 : \beta_k$ is stationary across the region of interest vs.
$H_A : \beta_k$ is non-stationary across the region of interest

$H_0 : \beta_k^u$ is stationary across the region of interest vs.
$H_A : \beta_k^u$ is non-stationary across the region of interest

$H_0 : \beta_k^v$ is stationary across the region of interest vs.
$H_A : \beta_k^v$ is non-stationary across the region of interest for k = 0, 1

are presented in Table 4.7. The hypothes were rejected for the parameters β_0 and β_1 but not for any of the additional parameters. Thus significant spatial non-stationarity exists only in the intercept and clay coefficients and not in the linear parameters introduced in the LLGWR model. From Figure 4.13 (a) it can be seen that the intercept coefficient has lower estimates located in the south-western corner of the field, and higher estimates located in the north-western corner, as well as along the south-eastern border of the field. From Figure 4.14 (a) it can be seen that the clay coefficient has high estimates located in the south-western corner of the field. The standard errors of the estimates of both the intercept and clay coefficients as can be seen in Figures 4.13 (b) and 4.14 (b) respectively are lowest in the center of the field and highest in the corners. Low estimates for the clay coefficient are located in the north-western corner, as well as along the south-eastern border of the field. The general trends shown in the maps of the estimates of the intercept and clay coefficients from the LLGWR analysis are similar to those from the GWR analysis but the LLGWR maps show more detailed variation in the estimated coefficients.

Based on the results of the significance tests of the individual parameters, the coefficient β_1^v which was found to be not significantly different from zero at all locations could be omitted from the model. Furthermore, based on the results of the tests for non-stationarity of the parameters, only the intercept and clay coefficients were found to be non-stationary over the study area and thus a mixed LLGWR model may be fitted whereby the stationary parameters are modelled globally and only the non-stationary parameters modelled locally.

University of Cape Town

Parameter	p-value
β_0	0.000
β_0^x	0.986
β_0^y	0.422
β_1	0.000
β_1^x	1.000
β_1^y	1.000

Table 4.7: Monte Carlo test for non-stationarity

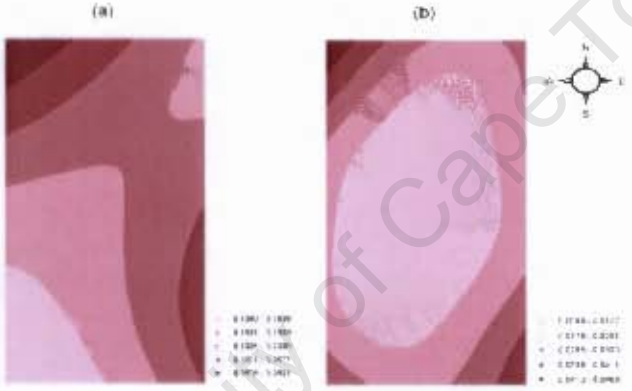


Figure 4.13: Maps of (a) β_0 and (b) $se(\hat{\beta}_0)$ from the LLGWR analysis

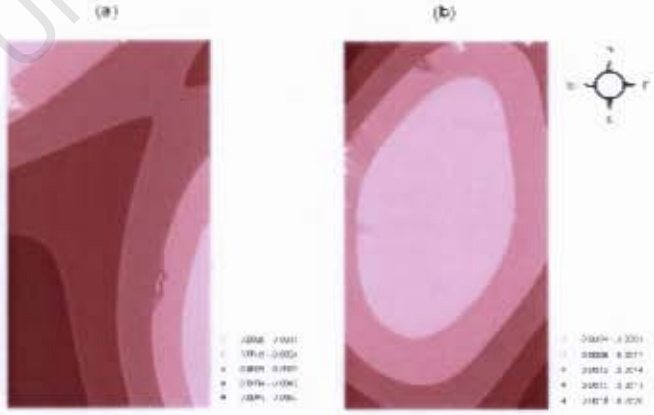


Figure 4.14: Maps of (a) β_1 and (b) $se(\hat{\beta}_1)$ from the LLGWR analysis

University of Cape Town

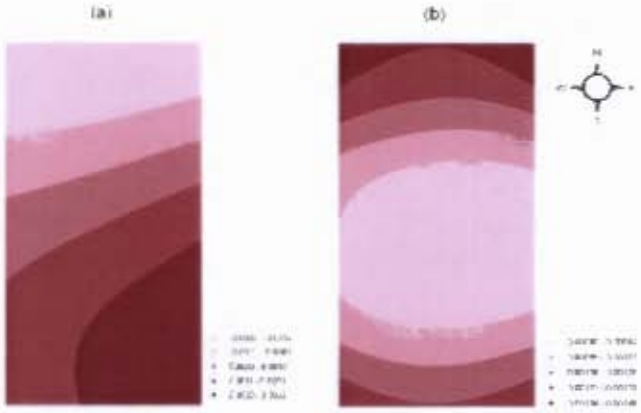


Figure 4.15: Maps of (a) $\hat{\beta}_2$ and (b) $se(\hat{\beta}_2)$ from the LLGWR analysis

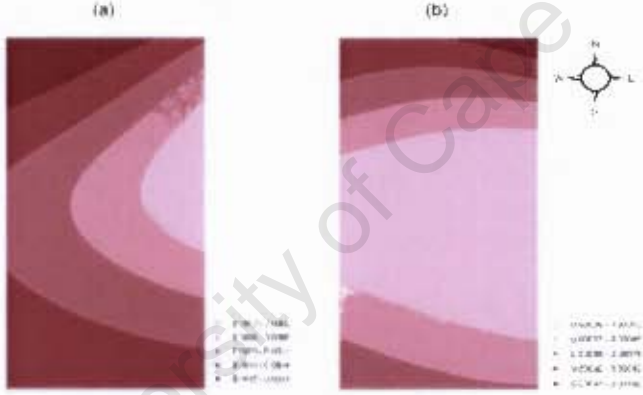


Figure 4.16: Maps of (a) $\hat{\beta}_3$ and (b) $se(\hat{\beta}_3)$ from the LLGWR analysis

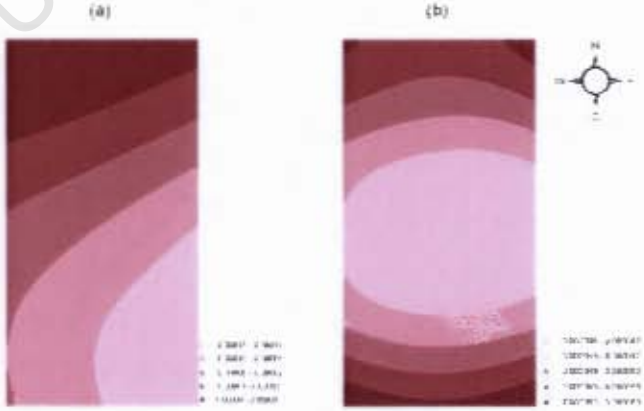


Figure 4.17: Maps of (a) $\hat{\beta}_4$ and (b) $se(\hat{\beta}_4)$ from the LLGWR analysis

University of Cape Town

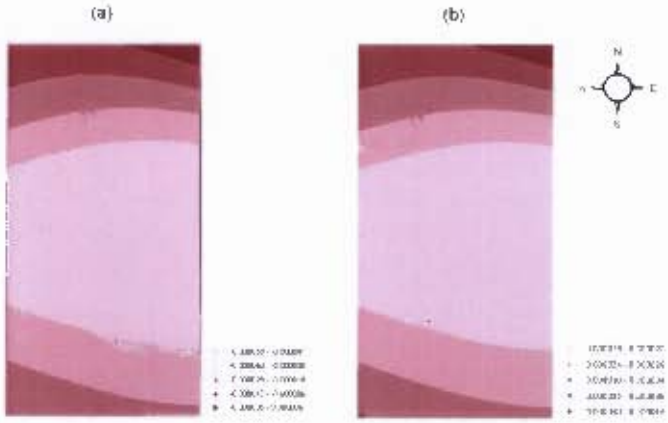


Figure 4.18: Maps of (a) $\hat{\beta}_5$ and (b) $se(\hat{\beta}_5)$ from the LLGWR analysis

4.6 Kriging application

Kriging is commonly used in the modelling and analysis of spatial data. ArcGIS Geostatistical Analyst was used to perform kriging on the current data set in order to predict water content at each of the observation point locations. Kriging on the current data set is used only for comparative purposes and mathematical details of the kriging model are not presented. The kriging model can be specified as

$$z(s_i) = \mu_i + \epsilon(s_i) \tag{4.4}$$

where $z(s_i)$ represents the value of water content at a spatial location s_i and is analogous to y_i at location s_i used in the regression framework, μ_i represents an unknown mean and $\epsilon(s_i)$ represents the random error term. The parameter μ_i can be taken to be constant as is the case of ordinary kriging or it can include trend terms as is the case with universal kriging. The error terms $\epsilon(s_i)$ are correlated and are described by a spatial correlation function called a semivariogram (Cressie, 1993). Several types of theoretical semivariograms were fitted to the data and following Clarke and Dane (1991), the spherical model was chosen. An elliptical search neighbourhood featuring 20 neighbouring points was used as suggested by Clarke and Dane (1991). Ordinary kriging and universal kriging with linear, quadratic and cubic trends were performed (Cressie, 1993). Cokriging, the multivariate extension of kriging, was also performed to incorporate clay content in the prediction of water content (Cressie, 1993). The various models fitted were compared on the basis of the root mean square prediction error (RMSPE) which is defined as the square root of the average squared differences between true and predicted values at observation point locations. The results of fitting the various models are presented in Table 4.8 and from these it can be seen that the ordinary cokriging model was the best in terms of having the smallest RMSPE. Predictions

University of Cape Town

	RMSPE
Ordinary kriging	0.1531
Universal kriging with:	
linear trend	0.1676
quadratic trend	0.1398
cubic trend	0.1404
Ordinary cokriging	**0.0127**
Universal cokriging with:	
linear trend	0.0160
quadratic trend	0.0139
cubic trend	0.0140

Table 4.8: RMSPE for various kriging models fitted to the data

obtained from this model as well as the standard errors of predictions are mapped in Figure 4.19. High values of water content are predicted in the north-western region of the field and low values along the southern border.

4.7 Comparative results

A summary of the results on goodness of fit obtained from the analyses of the soil science data based on the four models, namely global regression, GWR, LLGWR and ordinary cokriging, are presented in Table 4.9. The global model, which has the lowest R^2 value and largest AIC_c and RSS values, is clearly inferior to the local models in modelling the relationship between water and clay content. The LLGWR model has the smallest RSS value, followed by GWR and then kriging. Boxplots of the residuals from the four models are presented in Figure 4.20. It may be seen from these plots that the GWR and LLGWR produced less extreme residuals compared to the global model indicating that GWR and LLGWR provide better fit to the data and are able to accommodate the spatial variation in the observations. The LLGWR model has a smaller AIC_c and RSS value, and higher R^2 value than the GWR model. It thus appears that the introduction of the linear extension provides some improvement to the GWR methodology for this particular example. The data set used in this example is however too small to draw many meaningful conclusions. It thus instructive to repeat the analyses on a larger data set and the results for such a data set are presented in the following the chapter.

University of Cape Town

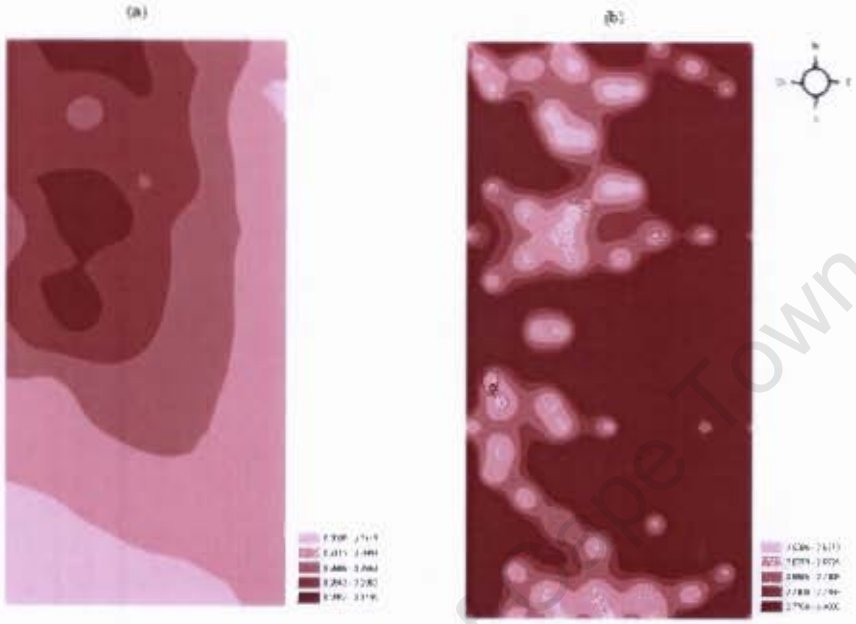


Figure 4.19: Maps of (a)Predicted water content and (b) Standard error of predictions from ordinary cokriging

University of Cape Town

Method	AIC_c	R^2	RSS
Global	-320.62	71.38	0.0719
GWR	-342.47	85.84	0.0060
LLGWR	-348.12	88.10	0.0050
Ordinary cokriging	-	-	0.0094

Table 4.9: Comparative results for the various models fitted to the soil science data. AIC_c and R^2 are not relevant for Ordinary cokriging.

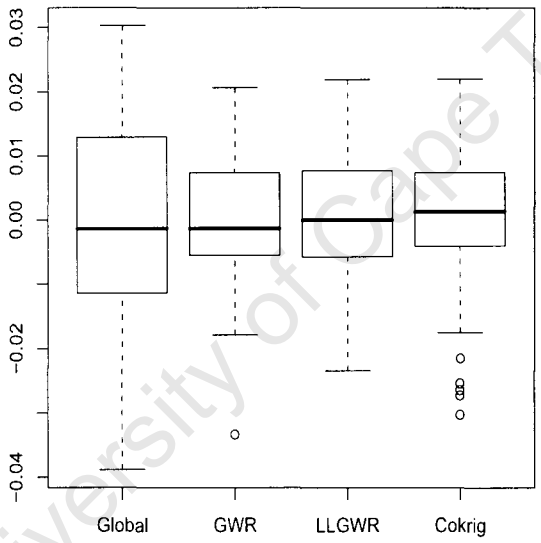


Figure 4.20: Boxplots of residuals from the various models fitted to the soil science data

University of Cape Town

Chapter 5

A Large Data Set taken from Geology

University of Cape Town

The analysis of a data set taken from geology is presented in this chapter. A description of the data is given in Section 5.1. The data are summarized by means of graphs and descriptive statistics in Section 5.2. and the global analysis of the data by means of ordinary least squares regression is discussed in Section 5.3. This is followed by local analyses of the data through the use of GWR, LLGWR and kriging techniques presented in Sections 5.4, 5.5 and 5.6 respectively. A summary and comparison of the results obtained from the analyses are given in Section 5.7 and comparative results based on a validation data set are presented in Section 5.8.

5.1 Data

The data set used in the present study, termed the Jura data set, was taken from the book by Goovaerts (1997). The topsoil of a 14.5 km^2 region near La Chaux-de-Fonds in the Jura Mountains, Switzerland, was surveyed by the Swiss Federal Institute of Technology. The soil was sampled at 359 locations and the concentrations of seven heavy metals, namely cadmium, cobalt, chromium, copper, nickel, lead and zinc, were measured at 25 cm depths at each location. The concentrations of the heavy metals are expressed in parts per million, i.e. milligrams of metal per kilogram of soil. In addition, rock type (Argovian, Kimmeridgian, Sequanium, Portlandian, Quaternary) and land use (forest, pasture, meadow, tillage) were recorded for each location. In the present study, 259 data points were randomly chosen from the entire data set and these comprise a training set used in the model building process. The remaining 100 data points formed the validation set. The locations at which data were recorded are shown in Figure 5.1. The aim of the analysis is to model the relationship between the concentration of chromium as the dependent variable, and the concentration of the other

heavy metals, the land use and the rock type as independent variables.

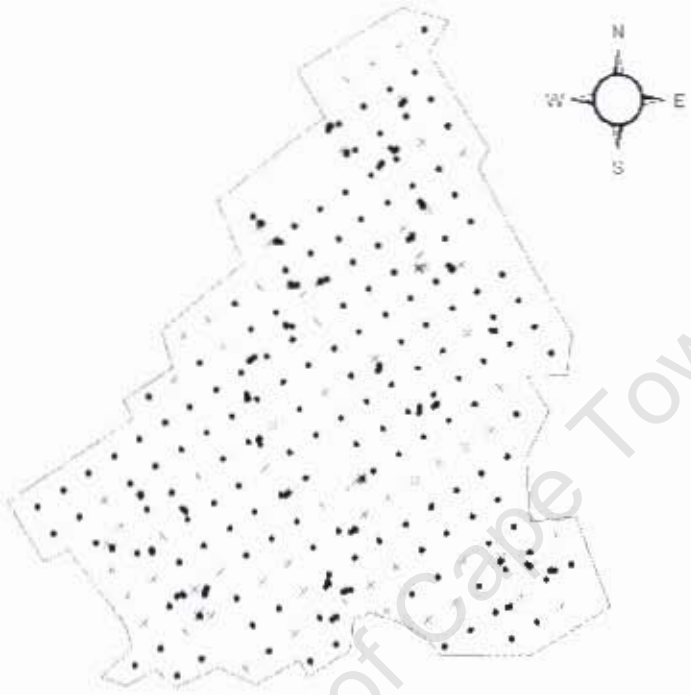


Figure 5.1: Map showing locations of the training data (circles) and validation data (crosses). The outline is the extent of the study area.

University of Cape Town

5.2 Exploratory Data Analysis

An exploratory data analysis was first carried out on the sampled training data set through some simple graphing techniques and descriptive statistics.

5.2.1 Categorical variables

Land use

The spatial distribution of land use is shown in Figure 5.2. From these maps it is apparent that most of the sampled locations are meadow and pasture land. Meadow and pasture land are fairly evenly distributed across the study region while sampled locations in the east of the study region are dominated by forest. The frequencies of occurrence of the different land uses throughout the study area are summarized in Table 5.1. Almost 60% of the sampled locations are meadow, 23.5% are pasture, only 2.7% are tillage and the remaining 14.7% are forest.

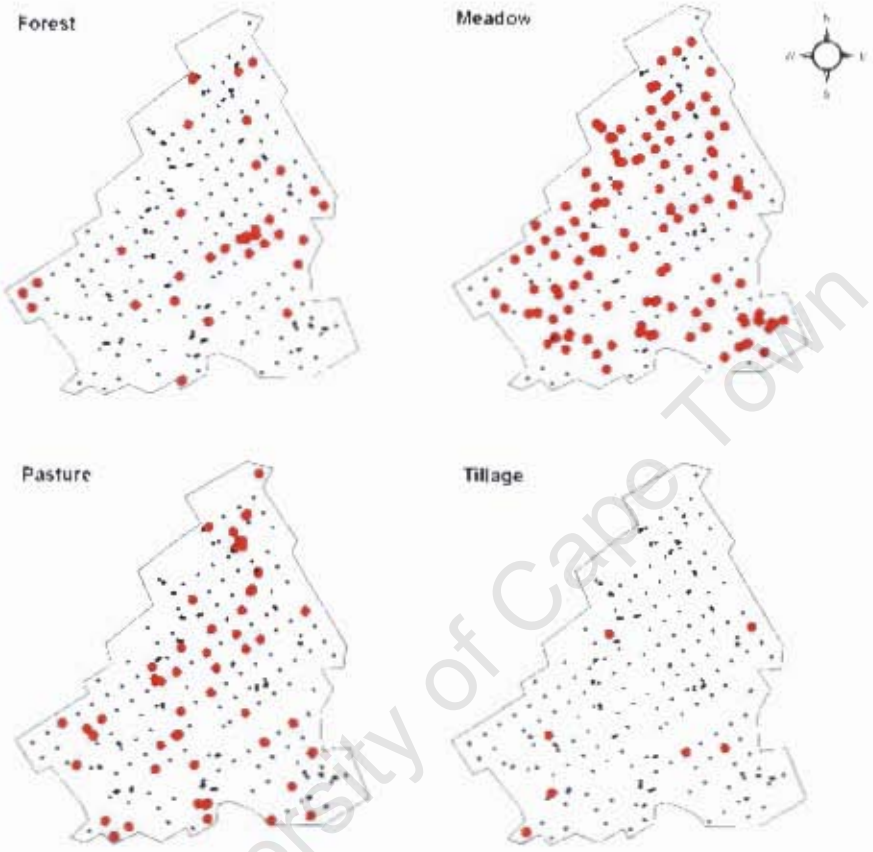


Figure 5.2: Maps showing the spatial distribution of land use for the training data set. The bold red dots indicate the locations at which each land type occurs.

University of Cape Town

Land use	Frequency	Frequency(%)
Forest	38	14.67
Pasture	61	23.55
Meadow	153	59.07
Tillage	7	2.70

Table 5.1: Frequencies of occurrence of land use

Rock type

The spatial distribution of rock type is shown in Figure 5.3. Argovian rock formations are found at locations furthest north and furthest south of the study region. Sequanian rock formations occur mostly in the west of the region and Quaternary rock formations occur mostly in the northern half of the region. Sampled locations in the middle of the study region are dominated by Kimmeridgian rock formations and there are only 4 sampled locations in total with Portlandian rock formations. The frequencies of occurrence of the different rock types are shown in Table 5.2. Over 60% of the locations consist of Kimmeridgian and Sequanian rock formations. Quaternary and Argovian each represent roughly 20% of the locations and Portlandian represents only 1.5% of the locations.

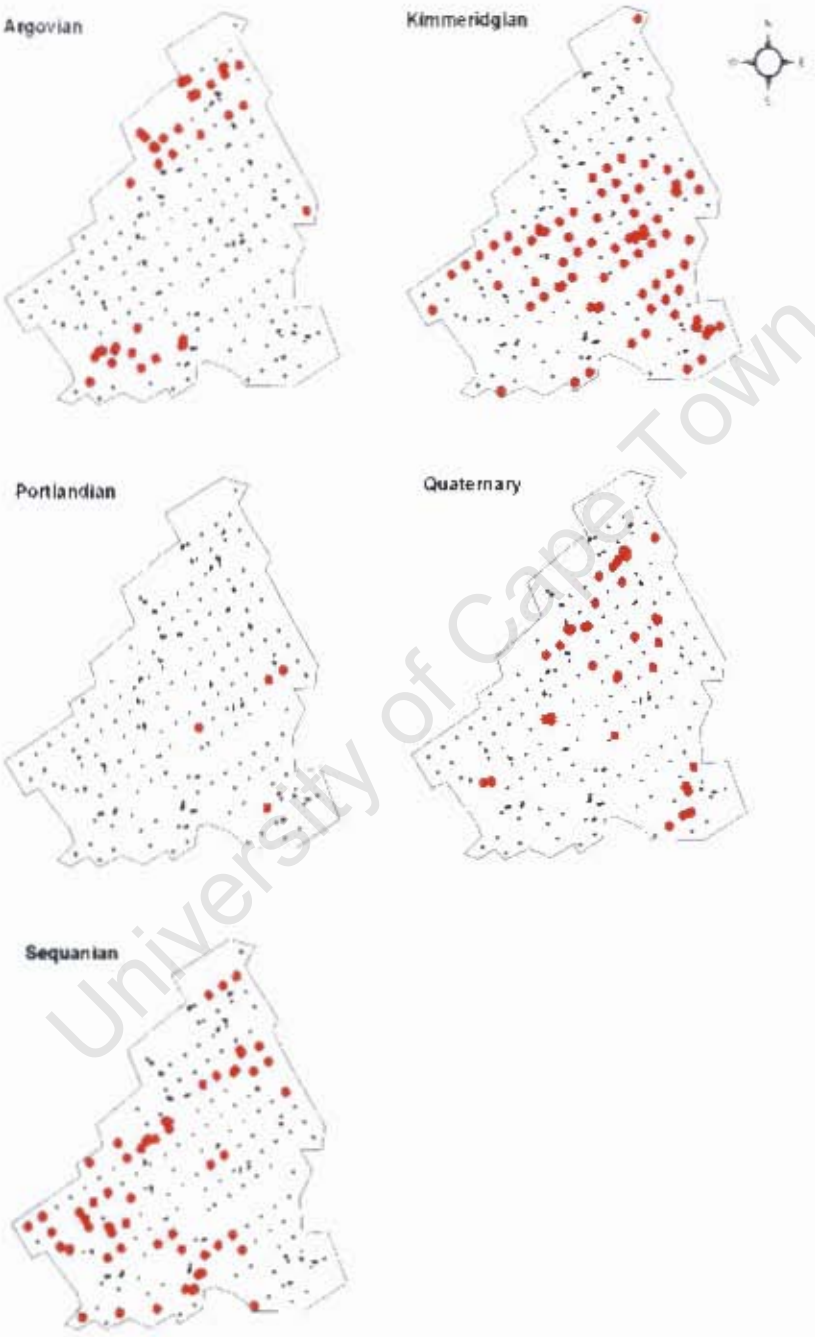
Rock type	Frequency	Frequency(%)
Argovian	47	18.15
Kimmeridgian	88	33.98
Sequanian	70	27.03
Portlandian	4	1.54
Quaternary	50	19.31

Table 5.2: Frequencies of occurrence of rock types

University of Cape Town

5.2.2 Continuous variables

Histograms of the metal concentrations expressed in parts per million for each metal are depicted in Figure 5.4. The histograms of cadmium (Cd), copper (Cu), lead (Pb) and zinc (Zn) are positively skewed indicating some particularly large concentrations of these heavy metals. The histograms of cobalt (Co), chromium (Cr) and nickel (Ni) are reasonably symmetrical. Descriptive statistics of the concentrations of each metal are given in Table 5.3. Based on the coefficient of variation, the metals with the most relative variation in concentration are copper (Cu), cadmium (Cd) and lead (Pb). The spatial distributions of the metals are mapped in Figure 5.5. The maps were created using ArcGIS Spatial Analyst software (ESRI, 2005) implementing the natural neighbour interpolation technique, which is a weighted moving average technique allowing for easier visualisation of the spatial distributions of the metals. From these maps it can be seen that the area of investigation has mostly low concentrations of cadmium (Cd), copper (Cu), lead (Pb) and zinc (Zn) with a few smaller regions for which the concentrations of these metals are high. The area is dominated by mostly mid-range to high concentration values of chromium (Cr) with a few small regions consisting of low concentrations of chromium. The distributions of cobalt (Co) and nickel (Ni) are fairly similar with both metals having large areas of mid-range values and some areas with very high and very low concentration values.



University of Cape Town

Figure 5.3: Maps showing the spatial distribution of rock type for the training data set. The bold red dots indicate the locations at which each rock type occurs.

	Cr	Cd	Co	Cu	Ni	Pb	Zn
Minimum	3.320	0.195	1.552	3.552	1.980	18.680	25.000
Median	35.000	1.100	9.960	17.040	20.600	46.000	72.080
Maximum	67.600	4.495	20.600	166.400	43.680	300.000	192.000
Mean	35.113	1.279	9.519	22.947	19.933	53.639	75.029
Std deviation	10.605	0.825	3.581	21.247	7.903	30.406	28.550
Coeff. of variation	0.302	0.645	0.376	0.926	0.396	0.567	0.381
Skewness	0.209	1.380	-0.181	3.013	-0.007	3.356	0.907

Table 5.3: Descriptive statistics of concentrations of metals in ppm

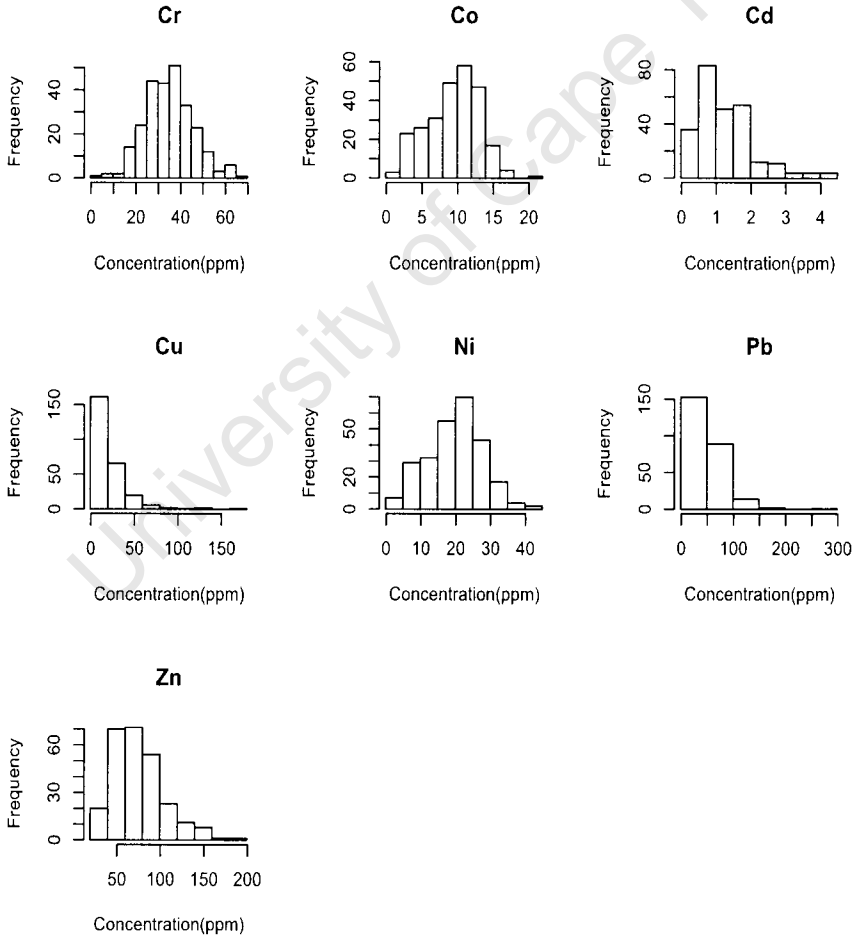
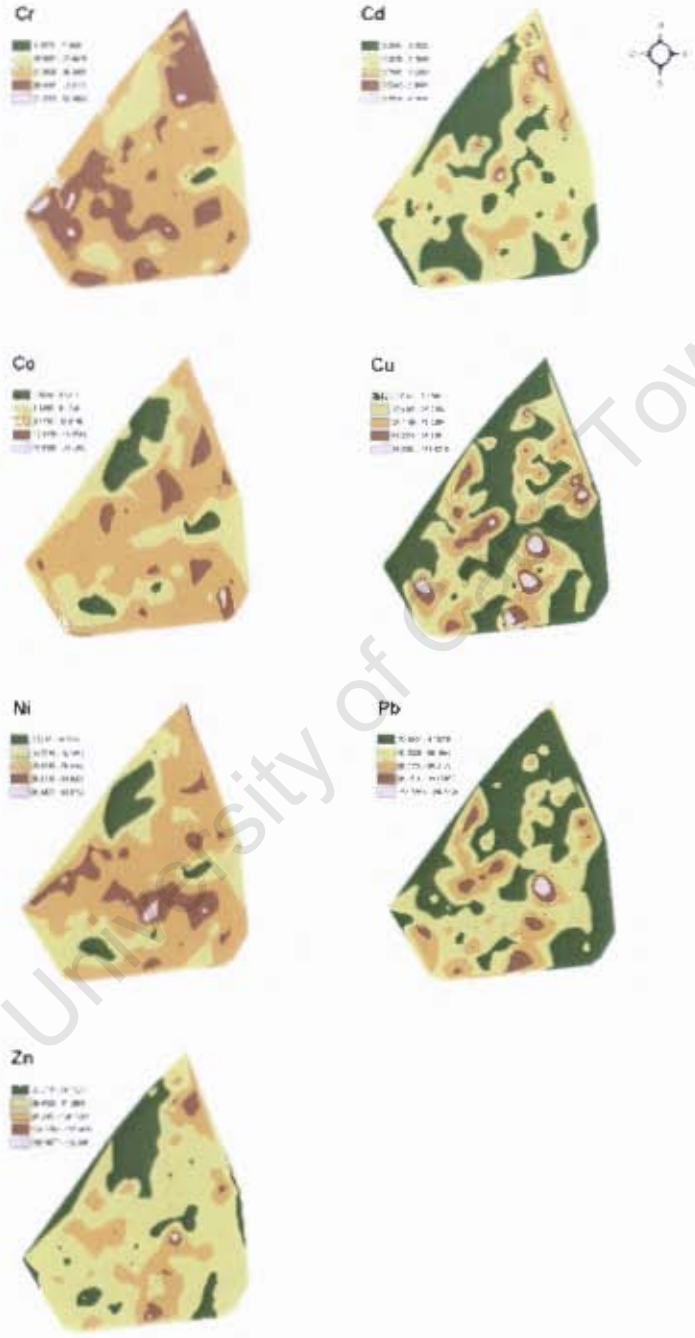


Figure 5.4: Histograms of metal concentrations

University of Cape Town



University of Cape Town

Figure 5.5: Maps showing the distributions of the concentrations of the seven metals

	Cr	Cd	Co	Cu	Ni	Pb	Zn
Cr	1	0.60	0.48	0.20	0.73	0.21	0.63
Cd	0.60	1	0.24	0.14	0.48	0.21	0.64
Co	0.48	0.24	1	0.17	0.75	0.14	0.45
Cu	0.20	0.14	0.17	1	0.22	0.79	0.61
Ni	0.73	0.48	0.75	0.22	1	0.26	0.62
Pb	0.21	0.21	0.14	0.79	0.26	1	0.62
Zn	0.63	0.64	0.45	0.61	0.62	0.62	1

Table 5.4: Correlation matrix for the concentrations of metals

The relationships between the continuous variables are quantified by their correlations as given in Table 5.4. The relationships between the variables are also shown graphically by means of a trellis plot with all the variables plotted against each other in the matrix of scatterplots given in Figure 5.6. The metals most strongly correlated with chromium are nickel, zinc and cadmium.

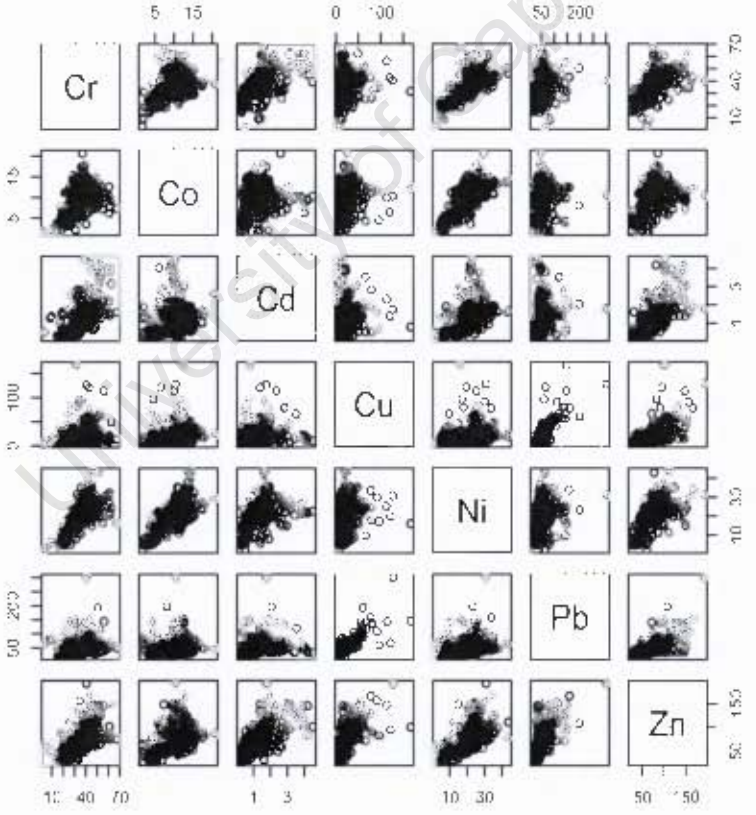


Figure 5.6: Matrix of scatterplots of concentrations of metals

University of Cape Town

Coefficient	Estimate	Std Error of estimate	p-value
$\hat{\beta}_0$	12.185	1.386	< 0.001
$\hat{\beta}_1$	7.082	1.342	< 0.001
$\hat{\beta}_2$	3.510	1.156	< 0.001
$\hat{\beta}_3$	0.766	2.604	0.769
$\hat{\beta}_4$	3.947	0.558	< 0.001
$\hat{\beta}_5$	0.708	0.058	< 0.001

Table 5.5: Global regression parameter estimates

5.3.2 Residuals

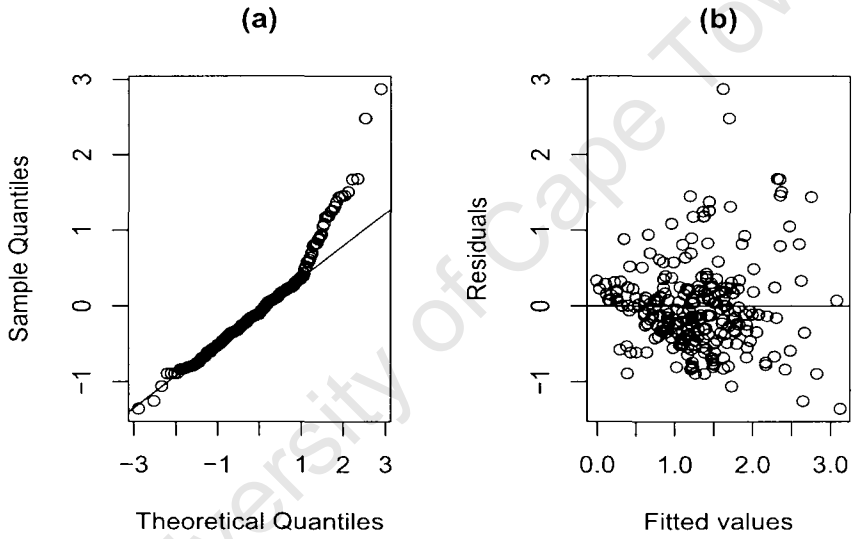


Figure 5.7: Plot of residuals against fitted values from the global model

A normal QQ plot of the residuals, defined as $e_i = y_i - \hat{y}_i$ for $i = 1, ..., 259$ where y_i are the observed values of chromium concentration and $\hat{y}_i$ the fitted values from model (5.1), is shown in Figure 5.7 (a) and a plot of the residuals against the fitted values is shown in Figure 5.7 (b). From Figure 5.7(a) there does not appear to be severe departures from normality with most points falling close to the line except for the extremes. The points appear to be fairly randomly spaced in the plot given by Figure 5.7(b). The spatial distribution of the residuals shown in Figure 5.8 however, appears to be non-random with a cluster of large negative residuals located in the east of the study region and some large positive residuals located in the south-west. A plot of the standardised residuals which is useful for outlier detection in presented is Figure 5.9. Generally standardised residuals which are greater than 2 in absolute value are considered to be potential outliers. From this plot seven observations were identified as potential outliers, namely observations

University of Cape Town

5.3 Global analysis

5.3.1 Global model

Various multiple regression models were fitted to the data with the aim of identifying the most important explanatory variables, and finding the best global model. Dummy variables were created for the categorical variables, namely land use and rock type. Rock type has 5 levels (Argovian, Kimmeridian, Sequanium, Portlandian, Quaternary) and thus 4 binary dummy variables were created with Argovian chosen as the base level. The dummy variables are assigned a value of one when an observation is a member of a specific category, and zero otherwise. The case whereby all dummy variables are assigned a value of zero implies that the base level holds. Land use has 4 levels (forest, pasture, meadow, tillage) and thus 3 binary dummy variables were created with forest chosen as the base level. Transformations of the continuous independent variables were considered but found to be unnecessary. Forward stepwise, backward stepwise as well as all subsets regression procedures were performed using the Statistica package (Statsoft, Inc., 2006) and several possible models considered. The final model was chosen in keeping with the principle of parsimony and is not necessarily the best model in terms of having the highest explanatory or predictive power. The model chosen on this basis is

$$y_i = \beta_0 + \beta_1 L_{2i} + \beta_2 L_{3i} + \beta_3 L_{4i} + \beta_4 Cd_i + \beta_5 Ni_i + e_i \qquad i = 1, ..., 259 \tag{5.1}$$

where y_i is the concentration in parts per million (ppm) of chromium, L_{2i}, L_{3i} and L_{4i} are dummy variables indicating the land type, Cd_i is the concentration in ppm of cadmium, Ni_i the concentration in ppm of nickel and e_i an independent and random error term. The model was fitted and the parameter estimates obtained, their standard errors and their p-values from the t-tests of the hypotheses,

$$H_0 : \beta_k = 0 \text{ vs. } H_1 : \beta_k \neq 0 \qquad \text{for } k = 0, ..., 5$$

are presented in the Table 5.5. The t-tests are based on the assumptions that the error terms are normally distributed and have constant variance however, t-tests are robust to deviations from those assumptions. The fit provided by the model is significant and the model has an R^2 value of 65% thus providing an adequate fit to data. The AIC for the model is 1698.34, and Residual Sums of Squares (RSS) value is 10119.45.

University of Cape Town

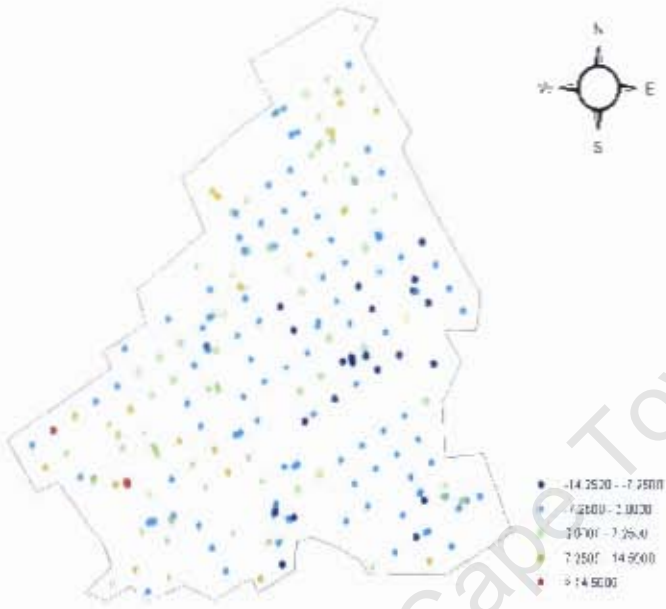


Figure 5.8: Distribution of the residuals from the global model

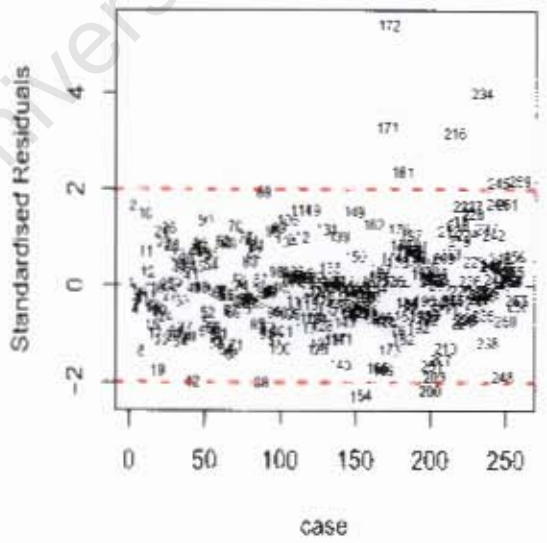


Figure 5.9: Plot of standardised residuals

University of Cape Town

171, 172, 181, 216, 154, 200 and 234. The impact of these observations on the analysis was investigated and it was decided to retain them in the remainder of the study as deleting them made little difference to the results.

5.3.3 Global models fitted over quadrants

The area of investigation was divided into four quadrants as is shown in Figure 5.10.

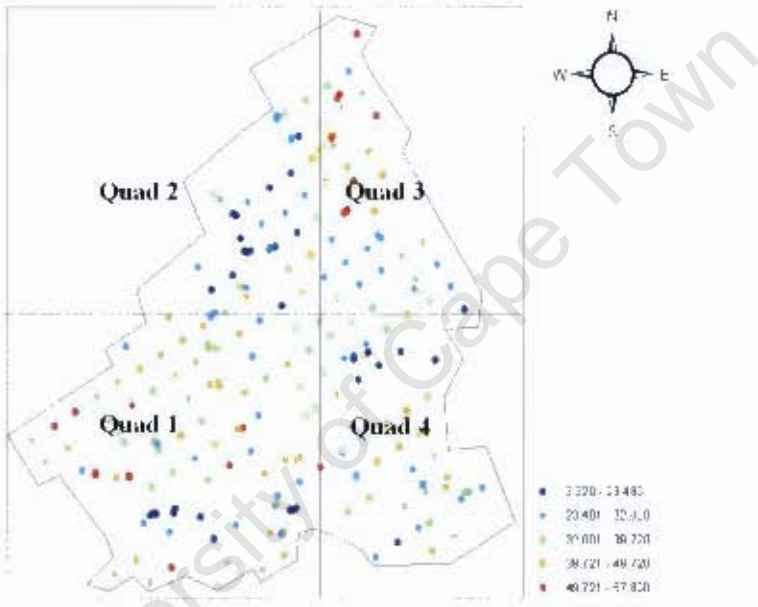


Figure 5.10: The distribution of chromium divided into 4 quadrants

Model (5.1) was fitted separately to the data for each quadrant allowing a simple way of assessing whether or not the relationship modelled between the concentration of chromium and the independent variables is likely to be stationary over space. Different models could have been fitted for each quadrant. However the 'best' model, namely that identified in Section 5.3.1, was fitted to each quadrant and used throughout the remainder of the analysis. Results of the multiple regression models fitted separately for each quadrant are presented in Table 5.6. Examination of these results reveals some differences in the parameter estimates across the different quadrants. For example, the estimate of the intercept coefficient is much lower in quadrant 4 than in the other quadrants, and the coefficient for Cd is negative in quadrants 2 and 4 and positive in quadrants 1 and 3. The models fitted to the data in quadrants 2 and 4 provided good fits with R^2 values of over 80% while the models fitted to the data in quadrants 1 and 3 provided moderate fits with R^2 values of approximately 60%. These differences suggest

University of Cape Town

		Quadrant1	Quadrant2	Quadrant3	Quadrant4
n_j		110	64	36	49
Intercept	$\hat{\beta}_0$	14.129 (1.509)	12.030 (2.626)	11.375 (5.595)	8.496 (2.105)
L_2	$\hat{\beta}_1$	-2.436 (2.223)	NA NA	-3.748 (3.032)	-1.989 (1.328)
L_3	$\hat{\beta}_2$	-0.456 (1.510)	-10.115 (2.622)	4.095 (2.868)	2.587 (2.436)
L_4	$\hat{\beta}_3$	NA NA	NA NA	0.961 (5.317)	3.236 (3.871)
Cd	$\hat{\beta}_4$	5.691 (1.040)	-9.017 (4.101)	4.145 (1.228)	-2.137 (1.049)
Ni	$\hat{\beta}_5$	0.822 (0.151)	2.385 (0.493)	0.847 (0.257)	1.273 (0.078)
R^2		0.579	0.802	0.608	0.867

Table 5.6: Results from separate regressions for each quadrant where n_j ($j = 1, \ldots, 4$) is the number of data points in each quadrant. The standard errors of the estimates are given in brackets. NA indicates that there are no data for those categories.

that the parameters are not stationary over space and that a global model is inadequate in explaining the relationship between chromium and the explanatory variables.

University of Cape Town

5.4 Application of GWR

GWR analysis was performed on the training data set using the programs written in the language R and presented in Appendix B. The weighted regression model for an observation point at s_i and centered on a regression point at s_0, as defined in equation (2.3) is given by

$$y_i = \beta_0(s_0) + \beta_1(s_0)L_{2i} + \beta_2(s_0)L_{3i} + \beta_3(s_0)L_{4i} + \beta_4(s_0)Cd_i + \beta_5(s_0)Ni_i + e_i \quad i = 1, \ldots, 259 \quad (5.2)$$

where y_i is the observed concentration in parts per million (ppm) of chromium at location s_i, L_{2i}, L_{3i} and L_{4i} are dummy variables indicating the land type, Cd_i is the concentration in ppm of cadmium, Ni_i the concentration in ppm of nickel, $\beta_i(s_0)$ for $i = 0, \ldots, 5$ are unknown parameters at location s_0 and $e_i \sim rv(0, \frac{\sigma^2}{w_{0i}})$, where w_{0i} denotes the weight given to the observed point which is a function of the distance of that point from the regression point. Model (5.2) was fitted to the data using weighted regression centered on each observation point i for $i = 1, ..., 259$. The model was fitted using a fixed Gaussian kernel with various possible values of bandwidth and the resulting CV scores calculated. The CV scores are plotted against bandwidth in Figure 5.11 to provide some guidance with bandwidth selection. From Figure 5.11 it may be seen that a minimum CV score exists for a bandwidth of approximately 0.8 km. This value for bandwidth was calculated more accurately as 0.758 km using the routine `optim` in R.

University of Cape Town

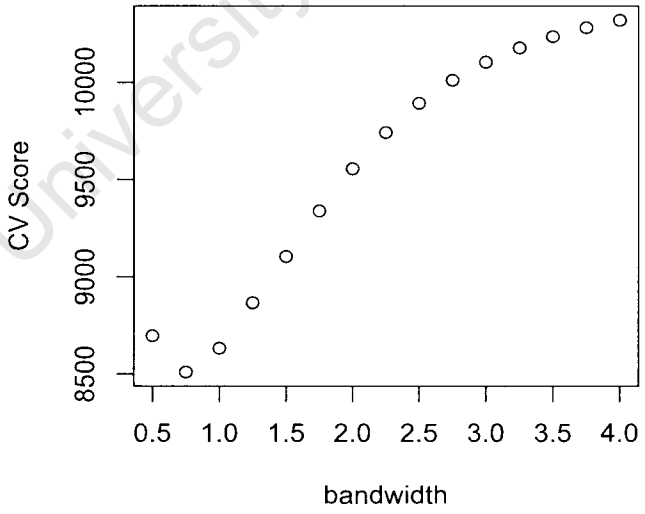


Figure 5.11: Variation in CV score with bandwidth using a Gaussian kernel. The minimum CV score exists for a bandwidth of approximately 0.8km.

Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Minimum	5.7294	-1.5585	-6.8697	-9.4150	-0.5547	0.4047
Lower quartile	8.5662	3.7891	0.6242	-2.9800	2.8221	0.7022
Median	10.4503	4.8648	2.2714	-0.7662	4.1491	0.7948
Upper quartile	15.1034	5.3904	3.3006	2.2019	5.1795	0.9574
Maximum	23.4809	7.5055	7.5708	7.6595	8.0702	1.2082

Table 5.7: Five-number summary of parameter estimates from model (5.2)

Five-number summaries of the estimated model parameters at the 259 observation point locations are given in Table 5.7. The individual parameters may be tested for significance using a pseudo t-test. The proportions of cases for which the hypothesis

$$H_0 : \beta_k = 0 \text{ vs. } H_A : \beta_k \neq 0 \text{ for } k = 0, ..., 5$$

is rejected at the 5% significance level are summarised in Table 5.8. All the parameters were found to be significantly different from zero at most of the locations, except for the coefficient associated with L_{4i} which was found to be significant at only 18% of the locations. The model has an AIC_c value of 1640.801 and a Residual Sums of Squares (RSS) value of 5522.68. The effective degrees of freedom for the model are 199.304.

	β_0	β_1	β_2	β_3	β_4	β_5
Proportion of locations where parameter is significantly different to zero	1.00	0.8	0.53	0.18	0.9	1.00

Table 5.8: Proportion of locations where the individual parameters in model (5.2) were found to be significantly different to zero.

The main output of a GWR analysis is a set of local parameter estimates that can be mapped to show how the model parameters change over space. A grid of 70×80 equally spaced points was defined over the study region. The grid is defined as the smallest rectangle enclosing the study region. Model (5.2) was fitted to the data using weighted regression centered on each grid point. Parameters were thus estimated at each of the grid points producing 5600 estimates for each parameter over space. These estimates as well as the standard errors of the estimates were mapped using ArcGIS software and are presented in Figures 5.12 to 5.14. Spatial variations are evident in all parameter estimates. The Monte Carlo method described in Section 2.6 was used to determine whether or not the parameters displayed significant non-stationarity. 1000 randomisations of the data were performed and the results used for testing the hypothesis

$H_0 : \beta_k$ is stationary across the region of interest vs.
$H_A : \beta_k$ is non-stationary across the region of interest for $k = 0, ..., 5$

University of Cape Town

are presented in Table 5.9. From these results it can be seen that parameters β_0, the intercept, β_4, the coefficient of Cd and β_5, the coefficient of Ni exhibit spatial non-stationarity, although these results are only significant at the 10% level. From Figure 5.12 (a) it can be seen that the lowest values for $\hat{\beta}_0$ are located along the eastern border of the region and highest values in the south as well is in the west. The standard errors of $\hat{\beta}_0$ are highest in the corners of the region and along the western border as can be seen in Figure 5.12 (b). From Figure 5.13 (a) it can be seen that the highest values for $\hat{\beta}_4$ are located in the middle of the region as well is in the south-west with low values located in the north-western corner of the region. The standard errors of $\hat{\beta}_4$ as shown in Figure 5.13 (b) are highest along the northern border of the region and in the south-eastern corner of the region. The highest values for $\hat{\beta}_5$ are located in the north-west of the region as can be seen in Figure 5.14 (a) and the highest values for the standard errors of $\hat{\beta}_5$ are located along the western border and in the north-western and south-eastern corners of the region. The coefficients for L_2, L_3 and L_4 are stationary and hence these parameters may be modelled as global parameters in a mixed GWR model.

Parameter	Variable	p-value
β_0	Intercept	**0.084**
β_1	L_2	0.999
β_2	L_3	0.510
β_3	L_4	0.203
β_4	Cd	**0.075**
β_5	Ni	**0.086**

Table 5.9: Monte Carlo test for non-stationarity

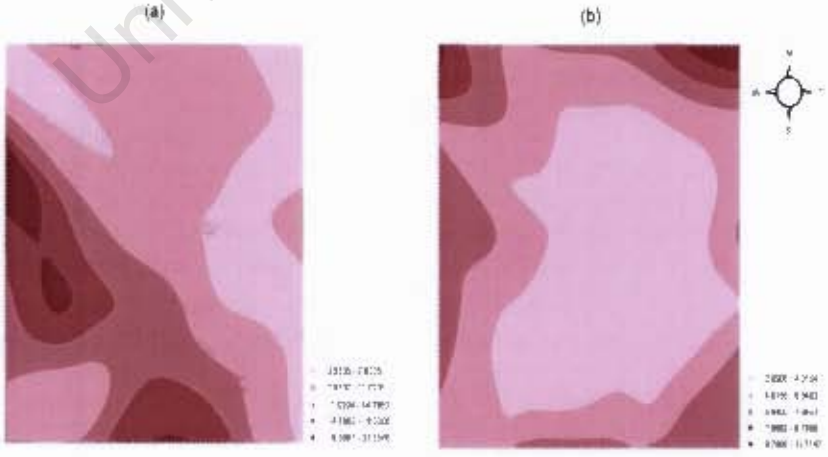


Figure 5.12: Maps of (a) $\hat{\beta}_0$ and (b) $se(\hat{\beta}_0)$ from GWR analysis

University of Cape Town

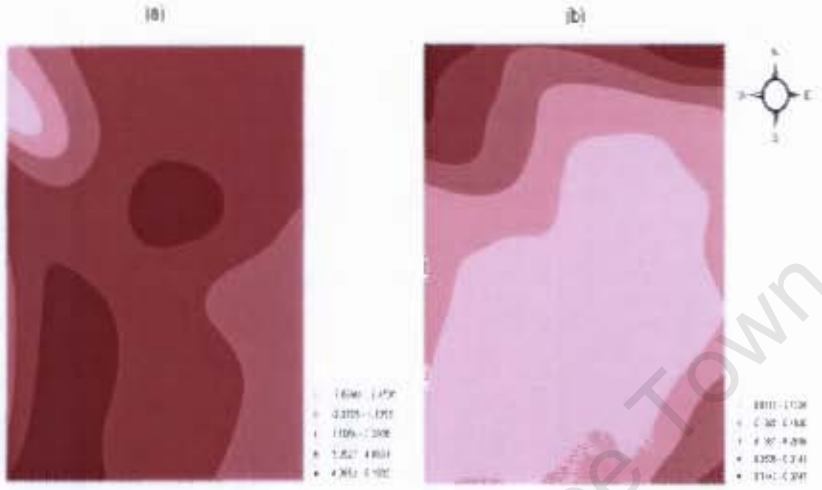


Figure 5.13: Maps of (a) $\hat{\beta}_4$ and (b) $se(\hat{\beta}_4)$ from GWR analysis

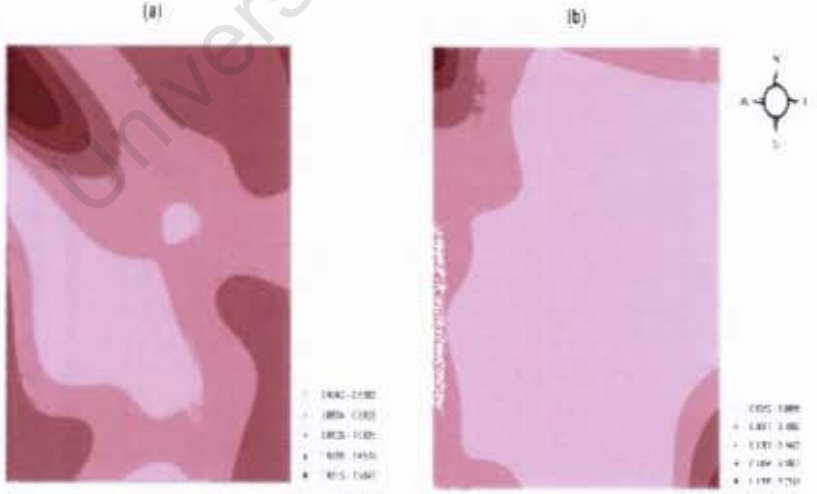


Figure 5.14: Maps of (a) $\hat{\beta}_5$ and (b) $se(\hat{\beta}_5)$ from GWR analysis

University of Cape Town

5.5 Implementation of LLGWR

LLGWR analysis was performed using programs written in the language R given in Appendix C. The weighted regression model for an observation point at location s_i and centered on a regression point at s_0, as defined in equation (3.6), is given by

$$y_i = \beta_0^*(s_0) + \beta_1^*(s_0)L_{2i} + \beta_2^*(s_0)L_{3i} + \beta_3^*(s_0)L_{4i} + \beta_4^*(s_0)Cd_i + \beta_5^*(s_0)Ni_i + e_i \qquad i = 1, \ldots, 259 \quad (5.3)$$

where

$$\beta_k^*(s_0) = \beta_k(s_0) + \beta_k^u(s_0)(u_i - u_0) + \beta_k^v(s_0)(v_i - v_0)$$

for $k = 0, 1, \ldots, 5$ and for each observation $i = 1, \ldots, 259$, are unknown parameters at location s_0, y_i is the observed concentration in parts per million (ppm) of chromium at location s_i, L_{2i}, L_{3i} and L_{4i} are dummy variables indicating the land type, Cd_i is the concentration in ppm of cadmium, Ni_i the concentration in ppm of nickel, and $e_i \sim rv(0, \frac{\sigma^2}{w_{0i}})$, where w_{0i} denotes the weight given to the observed point which is a function of the distance of that point from the regression point. Model (5.3) was fitted to the data using weighted regression centered on each observation point i for $i = 1, ..., 259$. A fixed Gaussian kernel was used to define the weights associated with each observation and bandwidth was selected based on the cross validation criterion. The model was fitted with a range of bandwidths and the resulting CV scores calculated. The CV scores are plotted against bandwidth in Figure 5.15 to provide some guidance with bandwidth selection. From Figure 5.15 it may be seen that a minimum CV score exists for a bandwidth of close to 2km. The bandwidth resulting in the minimum CV score was calculated more accurately as 1.84 km. The optimal bandwidth for LLGWR is approximately twice that calculated in the case of GWR but there is no immediate explanation for this. The model has an AIC_c value of 1603.186 and a Residual Sums of Squares (RSS) value of 6566.525. The effective degrees of freedom for the model are 221.69.

A five-number summary of the estimated model parameters at the 259 observation point locations are given in Table 5.10. The individual parameters may be tested for significance using a pseudo t-test. The proportions of cases for which the hypotheses

$$\begin{aligned} &H_0 : \beta_k = 0 \text{ vs. } H_A : \beta_k \neq 0 \\ &H_0 : \beta_k^u = 0 \text{ vs. } H_A : \beta_k \neq 0 \\ &H_0 : \beta_k^v = 0 \text{ vs. } H_A : \beta_k \neq 0 \quad \text{for k} = 0, \ldots, 5 \end{aligned}$$

University of Cape Town

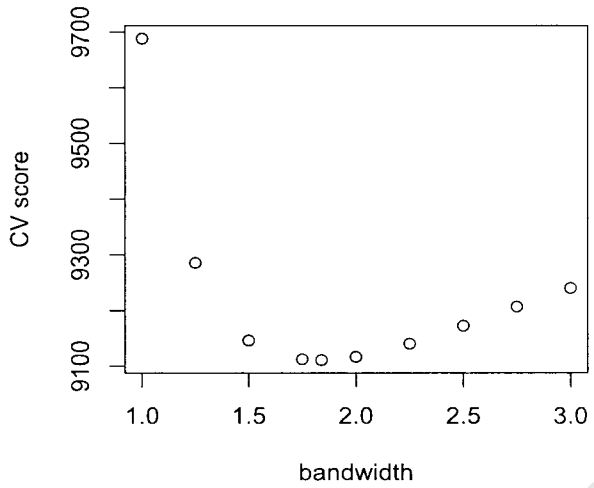


Figure 5.15: Variation in CV score with bandwidth using a Gaussian kernel. The minimum CV score exists for a bandwidth of approximately 1.8km.

are rejected at the 5% significance level are summarised in Table 5.11 along with the results of the Monte Carlo method detailed in Section 2.6 used to test the hypotheses

H_0 : β_k is stationary across the region of interest vs.
H_A : β_k is non-stationary across the region of interest

H_0 : β_k^u is stationary across the region of interest vs.
H_A : β_k^u is non-stationary across the region of interest

H_0 : β_k^v is stationary across the region of interest vs.
H_A : β_k^v is non-stationary across the region of interest for k = $0, \ldots, 5$.

From Table 5.11 it can be seen that some of the additional parameters were found to be significantly different to zero at more than half of the locations, namely β_0^u, β_4^u and β_5^u, thus suggesting that the inclusion of the linear extension may be worthwhile. Some of the parameters were found to be non-significant at most locations and hence the model may be re-fitted excluding these parameters. Results of the Monte Carlo test revealed that four of the parameters, β_0, β_4, β_5 and β_5^v displayed significant non-stationarity. Thus a mixed LLGWR model may be implemented in which the stationary parameters are modelled globally and the non-stationary parameters modelled locally. The estimates of the parameters that were found to be significantly different to zero and non-stationary, namely $\hat{\beta}_0$, $\hat{\beta}_4$ and $\hat{\beta}_5$ as well as their standard errors were mapped in order to illustrate their variation over space. A grid of 70×80 equally spaced points was defined over

University of Cape Town

	Variable	Minimum	Lower quartile	Median	Upper quartile	Maximum
$\hat{\beta}_0$	1	3.499	7.477	12.152	16.158	19.435
$\hat{\beta}_0^u$	$(u_i - u_0)$	-6.626	-5.384	-4.771	-3.392	0.092
$\hat{\beta}_0^v$	$(v_i - v_0)$	-3.970	-1.920	-1.021	-0.635	2.444
$\hat{\beta}_1$	L_2	-2.488	3.695	4.207	4.636	8.460
$\hat{\beta}_1^u$	$(u_i - u_0)L_2$	-2.973	-0.367	-0.024	0.481	0.884
$\hat{\beta}_1^v$	$(v_i - v_0)L_2$	-0.897	0.418	0.681	1.466	5.306
$\hat{\beta}_2$	L_3	-6.541	-0.594	1.620	2.300	4.296
$\hat{\beta}_2^u$	$(u_i - u_0)L_3$	-5.895	-1.494	-0.431	0.173	0.792
$\hat{\beta}_2^v$	$(v_i - v_0)L_3$	-0.638	0.569	1.408	3.155	5.099
$\hat{\beta}_3$	L_4	-9.155	-3.518	-0.378	1.682	6.362
$\hat{\beta}_3^u$	$(u_i - u_0)L_4$	-4.782	-0.501	0.003	0.395	0.645
$\hat{\beta}_3^v$	$(v_i - v_0)L_4$	-0.151	1.750	2.692	3.923	5.241
$\hat{\beta}_4$	Cd	-2.549	2.168	4.430	5.506	6.419
$\hat{\beta}_4^u$	$(u_i - u_0)Cd$	-3.380	-2.844	-2.452	-1.958	-0.364
$\hat{\beta}_4^v$	$(v_i - v_0)Cd$	0.406	0.935	1.191	1.329	1.686
$\hat{\beta}_5$	Ni	0.510	0.749	0.852	1.101	1.383
$\hat{\beta}_5^u$	$(u_i - u_0)Ni$	-0.052	0.189	0.260	0.287	0.332
$\hat{\beta}_5^v$	$(v_i - v_0)Ni$	-0.245	-0.062	-0.031	0.039	0.105

Table 5.10: Descriptive statistics of parameter estimates from model (5.3)

University of Cape Town

	Variable	Proportion of locations where parameter is significantly different to zero	p-values from Monte Carlo test
β_0	1	0.96	0.028
β_0^u	$(u_i - u_0)$	0.66	0.329
β_0^v	$(v_i - v_0)$	0.03	0.441
β_1	L_2	0.67	0.897
β_1^u	$(u_i - u_0)L_2$	0.00	0.748
β_1^v	$(v_i - v_0)L_2$	0.00	0.529
β_2	L_3	0.14	0.504
β_2^u	$(u_i - u_0)L_3$	0.07	0.319
β_2^v	$(v_i - v_0)L_3$	0.18	0.134
β_3	L_4	0.00	0.433
β_3^u	$(u_i - u_0)L_4$	0.00	0.706
β_3^v	$(u_i - u_0)L_4$	0.00	0.449
β_4	Cd	0.81	0.004
β_4^u	$(u_i - u_0)Cd$	0.83	0.473
β_4^v	$(v_i - v_0)Cd$	0.45	0.814
β_5	Ni	1.00	0.002
β_5^u	$(u_i - u_0)Ni$	0.79	0.103
β_5^v	$(v_i - v_0)Ni$	0.07	0.038

Table 5.11: Results of significance test of individual parameters in model (5.3) and Monte carlo test for non-stationarity

University of Cape Town

the study region and model (5.3) was fitted to the data using weighted regression centered on each grid point as for GWR. Maps of the parameter estimates were produced using the ArcGIS software and are presented in Figures 5.16, 5.17 and 5.18. The spatial patterns of the parameter estimates are evident from these figures. From Figure 5.16 (a) it can be seen that the lowest values of $\hat{\beta}_0$ are located in the eastern half of the region and the highest values located in the north-west. From Figure 5.17 (a) it can be seen that the western half of the field is dominated by high values of $\hat{\beta}_4$ with low values located in the north-eastern corner and along the eastern border. Much of the region is dominated by high values of $\hat{\beta}_5$ as is shown in Figure 5.18 (a) with some low values located in the north-western corner of the region. The standard errors of the estimates as shown in Figures 5.16 (b), 5.17 (b) and 5.18 (b) exhibit similar spatial patterns with low values concentrated in the center of the region with values increasing towards the borders of the region.

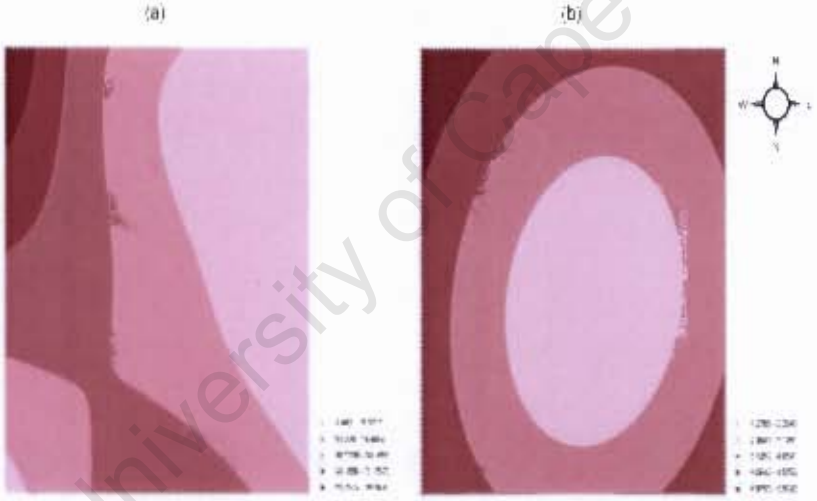


Figure 5.16: Maps of (a) $\hat{\beta}_0$ and (b) $se(\hat{\beta}_0)$ from LLGWR analysis

University of Cape Town

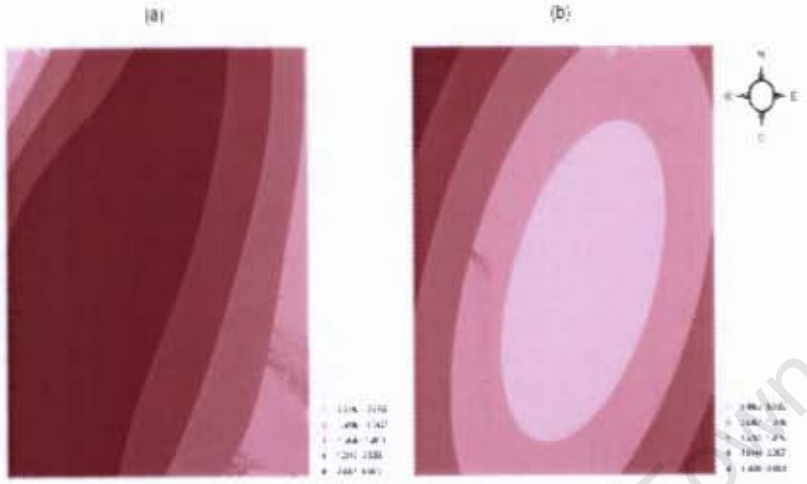


Figure 5.17: Maps of (a) $\hat{\beta}_4$ and (b) $se(\hat{\beta}_4)$ from LLGWR analysis

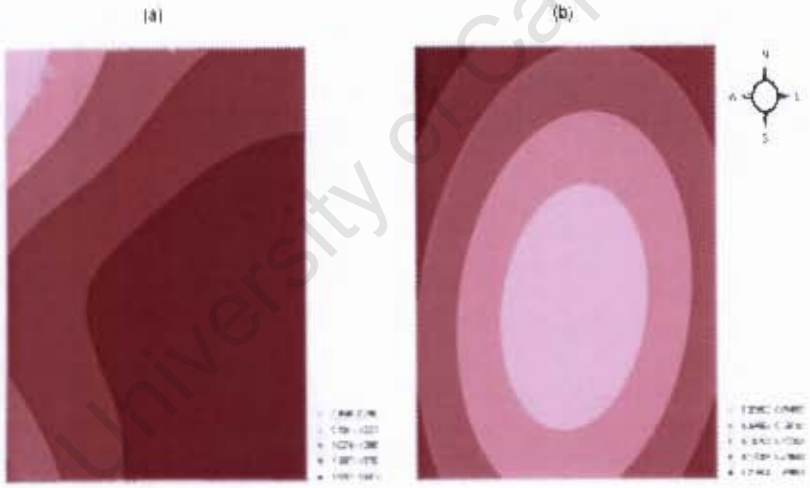


Figure 5.18: Maps of (a) $\hat{\beta}_5$ and (b) $se(\hat{\beta}_5)$ from LLGWR analysis

5.6 Kriging

Kriging was performed using ArcGIS Geostatistical Analyst in order to predict the concentrations of chromium. The use of kriging on the current data set is only for comparative purposes and mathematical details of the kriging model are not presented. Several types of theoretical semivariograms were fitted to the data and the exponential model for the spatial covariance structure was chosen. Ordinary kriging and universal kriging with linear, quadratic and cubic trends

University of Cape Town

were performed. Cokriging incorporating cadmium and nickel in the prediction of chromium was also performed. The results of the various models are presented in Table 5.12. From these results it can be seen that the ordinary cokriging model was the best in terms of having the smallest Root Mean Square Prediction Error (RMSPE). Predicted values of chromium obtained from this model as well as the standard errors of predictions are mapped in Figure 5.19. From Figure 5.19 (a) it can be seen that high values of chromium concentration are predicted in the north-eastern and south-western regions.

	RMSPE
Ordinary kriging	7.5650
Universal kriging with:	
linear trend	20.6400
quadratic trend	9.7010
cubic trend	109.6000
Ordinary cokriging	**5.4760**
Universal cokriging with:	
linear trend	6.0680
quadratic trend	6.8970
cubic trend	201.2000

Table 5.12: RMSPE for kriging models fitted to the data

University of Cape Town

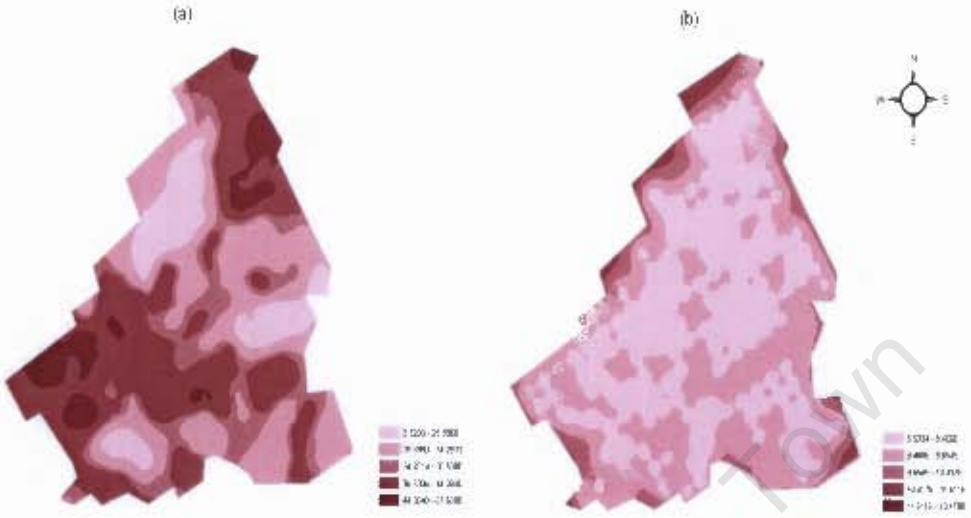


Figure 5.19: Maps of (a) Predicted values of chromium concentration and (b) Standard errors of predictions from ordinary cokriging model

5.7 Comparative results for training data set

The results on goodness of fit obtained from the analyses based on the four models, namely global regression, GWR, LLGWR and ordinary cokriging, are summarised and presented in Table 5.13. The global model has the largest AIC and RSS values and hence is inferior to the local models in modelling the relationship between chromium and the explanatory variables. The GWR model has the lowest RSS values followed by LLGWR and then the kriging model. Boxplots of the residuals from the four models are presented in Figure 5.20. From these plots it may be seen that that the GWR and LLGWR models produced less extreme residuals compared to the global and kriging models. Overall the GWR model seems to be the best model as it has the highest R^2 value and lowest AIC_c and RSS values. Based on these results it does not appear as if the introduction of the linear extension provides any improvement to the GWR model.

University of Cape Town

Method	AIC_C	R^2	RSS
Global	1698.34	65%	10119.45
GWR	1640.80	80.10%	5522.68
LLGWR	1645.29	77.36%	6566.52
Ordinary cokriging	-	-	7766.92

Table 5.13: Comparative results of the four models for the training data set. AIC_c and R^2 are not relevant for Ordinary kriging

University of Cape Town

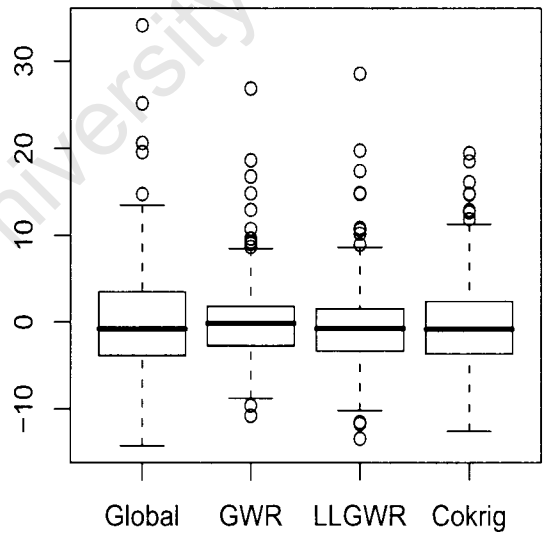


Figure 5.20: Boxplots of residuals from the training data set

5.8 Results of the validation data set

In order to compare the models in terms of prediction capability, 100 observations were excluded from the model calibration and these formed the validation data set. The global model identified as having the best fit to the training data, given by equation (5.1), was used to predict the concentration of chromium at the 100 locations of the validation set using the known values of the explanatory variables at those locations. In the cases of GWR and LLGWR, the models given by equation (5.2) and equation (5.3) respectively, were fitted to the data using weighted regression centered on each validation point and using the values for the bandwidth which were found to provide an optimal fit to the training data. Parameters were thus estimated at each validation point and since values of the explanatory variables are known at those locations, the concentration of chromium at these locations may be predicted. Plots of predicted chromium values against observed chromium values for each model are presented in Figure 5.21. From these plots it can be seen that the GWR and LLGWR models provide better fits compared to the global and kriging models. The residuals from the four models were calculated and boxplots of the residuals are presented in Figure 5.22. From these plots it is evident that the residuals from the GWR and LLGWR models are less variable relative to the global and kriging models. The RSS values defined as

$$RSS = \sum_{k=1}^{100} (y_i - \hat{y}_i)^2$$

where y_i is the observed value of chromium and $\hat{y}_i$ is the predicted value of chromium from the validation data set, were calculated for each model and presented in Table 5.14. These results indicate that the GWR model is the best performing model in terms of predicting the concentration of chromium and that the linear extension does not seem to provide any improvement to the GWR model. The kriging model, with the highest RSS value was the the poorest model in terms of predicting the chromium concentration.

	Global	GWR	LLGWR	Ordinary cokriging
RSS	7806.91	4735.53	5212.64	9456.24

Table 5.14: Residual Sums of Squares (RSS) for the four models for the validation data set

University of Cape Town

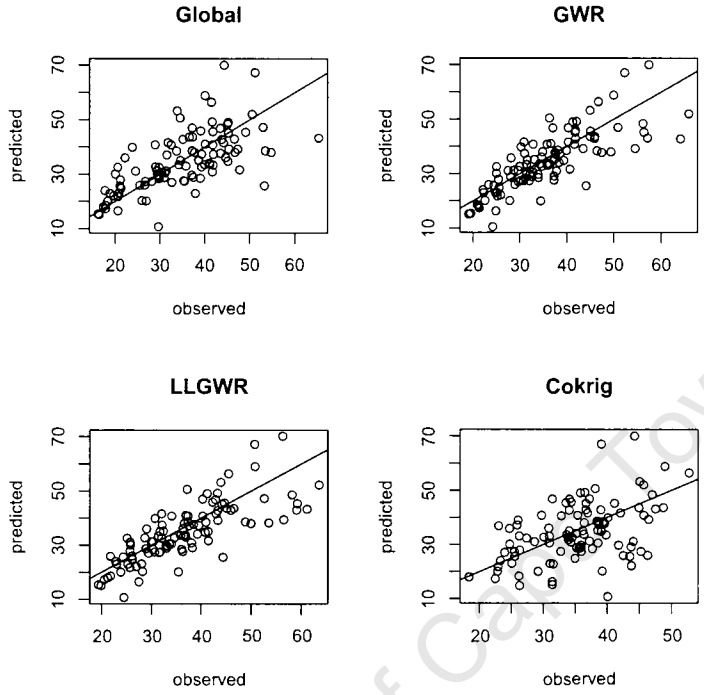


Figure 5.21: Plots of predicted against observed values of chromium from OLS, GWR, LLGWR and Ordinary cokriging models for the validation data set

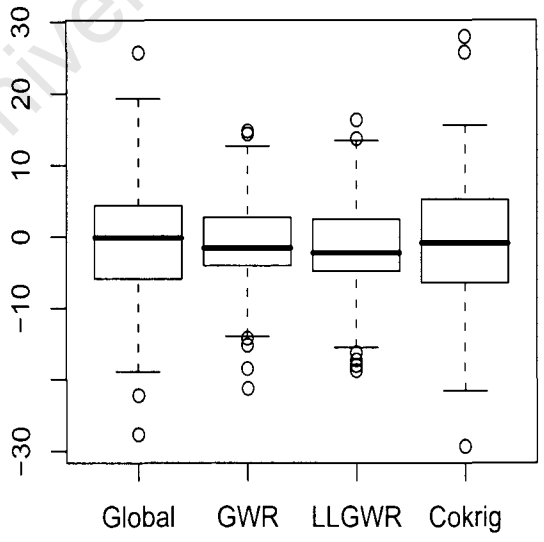


Figure 5.22: Boxplots of residuals from the validation data set

University of Cape Town

Chapter 6

Conclusions

University of Cape Town

Main aims achieved

In this study, the background to spatial regression modelling and its development has been explored. The more recently developed technique of Geographically Weighted Regression (GWR), which has been widely used in the analysis of spatial data across various disciplines, has been expressed in statistical terms. The GWR model has been noted as being a varying-coefficient model in which the coefficients vary as functions of location.

A new idea of extending the GWR model has been proposed, the aim of the extension being to increase parameter flexibility and to capture more of the variation in spatial data set. The extended model, termed Local Linear GWR (LLGWR) has been shown to be easily formulated and has the advantage of being fitted in the same way as the GWR model.

The feasibility and effectiveness of LLGWR has been examined for two data sets, a small data set taken from soil science and a large data set taken from geology. It has been shown that the LLGWR technique is straightforward to implement. Programs have been written in the language R to perform GWR and LLGWR, and are supplied. The only drawback of the LLGWR model is the large number of parameters. It may be possible however, to reduce this number of parameters by fitting a mixed LLGWR model whereby stationary parameters are modelled globally, and only the non-stationary parameters are modelled locally.

The results produced from the two sets of analyses were conflicting, with the small data set showing some evidence of an improvement in using LLGWR model over the GWR model, but the large data set showing no improvement. The lack of improvement in the case of the large data set may well be due to the fact that the data set exhibited little non-stationarity. Alternatively, it could be

that the nature of the variability of the regression coefficients in the two data sets favour different models. It is thus arguable as to whether the extension to the GWR model is worthwhile and further investigation is required.

Recommendations and future work

Local Linear Geographically Weighted Regression (LLGWR) is easy to implement and it may add value in the analysis of certain data sets and can be easily included in the GWR repertoire. Further investigation involving the analysis of more data sets is required, particularly data sets which show strong non-stationarity. An investigation into Mixed LLGWR models whereby stationary parameters are modelled globally and non-stationary parameters modelled locally is also required. Furthermore, the LLGWR model may be extended to Poisson regression models for count data as Nakaya, Fotheringham, Brunsdon and Charlton (2005) have down for GWR, but this is beyond the scope of this dissertation.

University of Cape Town

Bibliography

Aitken, M. (1996). A General Maximum Likelihood Analysis of overdispersion in Generalised Linear Models. *Statistics in Computing*, 6, 251-262.

Anselin, L. (1993). Discrete Space Autoregressive Models. In: Goodchild, M., Parks, B. and Steyaert, L., Editors, Environmental Modelling with GIS. Oxford University Press. 454-469.

Bivand, R. and Yu, D. (2007). The spgwr Package. [Online], Available: http://cran.r-project.org

Brunsdon, C., Fotheringham, A. and Charlton, M. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Non-stationarity. *Geographical Analysis*, 28, 281-289.

Brunsdon, C., Fotheringham, A. and Charlton, M. (1998). Geographically Weighted Regression: modelling spatial non-stationarity. *The Statistician*, 47, 431-443.

Brunsdon, C., Aitkin, M., Fotheringham, S. and Charlton, M. (1999). A comparison of Random Coefficient Modelling and Geographically Weighted Regression for Spatially Non-stationary Regression Problems. *Geographical and Environmental Modelling*, 3, 47-62.

Brunsdon, C., Fotheringham, A.S., and Charlton, M.E. (1999). Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science*, 39(3), 497-524.

Brunsdon, C. (2006). Personal communication.

Byrd, H., Lu, P., Nocedal, J., and Zhu, C.Y. (1995). A Limited Memory Algorithm for Bound Constrained Optimisation. *SIAM Journal on Scientiic Computing*, 16(5), 1190-1208.

University of Cape Town

Casetti, E. (1972). Generating Models by the Expansion Method: Applications to Geographic Research. *Geographical Analysis*,4, 81-91.

Clarke, G.P.Y. and Dane, J.H. (1991). A simplified Theory of Point Kriging and its extension to Cokriging and sampling optimization. Bulletin 609, Alabama Agricultural Experiment Station, Auburn University, Alabama.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

Cressie, N.A. (1993). Statistics for Spatial Data. Revised Edition. Wiley.

Dent, M.C., Lynch, S.D. and Schulze, R.E. (1988). Mapping Mean Annual and other Rainfall Statistics over Southern Africa. University of Natal, Dept. of Agricultural Engineering, ACRU Report, 27, Water and Research Commision, Pretoria, SA. Report No. 109/1/89.

Eldridge, J.D., and Jones, J.P. III (1991). Warped Space: A Geography of Distance Decay. *Professional Geographer*, 43, 500-511.

ESRI (2005). ArcGIS Software: The complete enterprise system, Version 9.1., URL: www.esri.com/software/arcgis/index.html

Foody, G.M. (2003). Geographical Weighting as a further refinement to Regression Modelling: An example focused on the NDVI-rainfall relationship. *Remote Sensing of Environment*, 88, 283-293.

Foster, S.A. and Gorr, W.L. (1986). An Adaption Filter for Estimating Spatially-varying Parameters: Application to Modelling Police hours spent in response to calls for service. *Management Science*, 32, 878-889.

Fotheringham, A.S., Brunsdon, C., Charlton, M. (2000). Quantitative Geography: Perspectives on spatial data analysis. Sage Publications, London.

Fotheringham, A.S., Brunsdon, C., Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester: Wiley.

Goovaerts, P. (1997). Geostatistics for Natural Resources Evaluation. Sage Publications, London

Gorr, W.L. and Olligschlaeger, A.M. (1994). Weighted adaptive filter-

University of Cape Town

ing: Monte Carlo Studies and Application to Illicit Drug Marke Modelling. *Geographical Analysis*, 26(1), 67-87.

Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. Chapman and Hall, London.

Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society*, Series B, 55, 757-796.

Hastie, T.J., Tibshirani, R.J. and Friedman, J. (2001). The Elements of Statistical Learning. Springer, New York.

Hurvich, C.M. and Tsai, C.L. (1989). Regression and Time Series Model Selection in small samples. *Biometrika*, 76, 297-304.

Kam, S., Hossain, M., Bose, M.L. and Villano, L.S. (2005). Spatial Patterns of Rural Poverty and their relationship with welfare-influencing factors in Bangladesh. *Food Policy*, 30(5-6), 551-567.

Lobo, J.M and Martin-Piera, F. (2002). Searching for a Predictive Model for Species Richness of Iberian Dung Beetle Based on Spatial and Environmental Variables. *Conservation Biology*, 16, 158-173.

Loader, C. (1999). Local Regression and Likelihood. Springer.

McMiller, D.P. (1996). One Hundred and fifty years of land values in Chicago: A nonparametric approach. *Journal of Urban Economics*, 40(1), 100-124.

Nakaya, T., Fotheringham, A.S., Brunsdon, C. and Charlton, M. (2005). Geographically Weighted Poisson Regression for Disease Association Mapping. *Statistics in Medicine*, 24, 2695-2717.

Nelson, A. (2000). Spatial structire and multivariate analysis in Hillside Agro-ecosystems. Internal document, PE-4/GIS, CIAT, Cali, Colombia.

R Development Core Team (2006). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

Rawlings, J.O., Pantula, S.G. and Dickey, D.A. (1998). Applied Regression Analysis: A research tool. 2nd Edition. Springer.

University of Cape Town

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

StatSoft, Inc. (2006). Statistica Version 7., URL: http://www.statsoft.com.

Swamy, P.A.V.B (1971). Statistical Inference in Random Coefficient Regression Models. Springer-Verlag, Berlin.

Tobler, W. (1970). A Computer Movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.

Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman and Hall, London.

Zhao, F., Chow, L., Li, M. and Liu, X. (2005). A Transit Ridership Model based on Geographically Weighted Regression. Lehman Center for Transportation Research, Florida International University, Florida (MA). [Online], Available: http://lctr.eng.fiu.edu/re-project-link/finalDO97591BW.pdf

University of Cape Town

Appendix A

Soil Science Data

University of Cape Town

u	v	Water	Clay
0.0000	0.0000	0.2205	15.40
6.2500	0.0000	0.2352	19.20
12.5000	0.0000	0.2430	17.60
18.7500	0.0000	0.2251	17.70
25.0000	0.0000	0.2436	17.40
37.5000	0.0000	0.2410	17.70
10.9375	3.1250	0.2038	13.40
17.1875	3.1250	0.2328	18.40
23.4375	3.1250	0.2290	17.30
29.6875	3.1250	0.2400	17.10
6.2500	6.2500	0.2301	18.10
34.3750	6.2500	0.2587	19.40
17.1875	9.3750	0.2462	21.50
12.5000	12.5000	0.2499	20.70
25.0000	12.5000	0.2621	25.00
7.8125	15.6250	0.2432	19.50
3.1250	18.7500	0.2433	21.50
4.6875	21.8750	0.2655	19.40
0.0000	25.0000	0.2558	19.20
12.5000	25.0000	0.2704	22.00
18.7500	25.0000	0.2740	24.00
31.2500	25.0000	0.2610	23.00
37.5000	25.0000	0.2560	22.50
4.6875	28.1250	0.2363	15.30
10.9375	28.1250	0.2767	21.40
3.1250	31.2500	0.2807	26.60
9.3750	37.5000	0.3017	29.90
12.5000	37.5000	0.2956	28.00

u	v	Water	Clay
25.0000	37.5000	0.2786	24.80
3.1250	43.7500	0.2969	28.90
7.8125	46.8750	0.2946	24.70
14.0625	46.8750	0.2876	27.00
20.3125	46.8750	0.2909	27.60
0.0000	50.0000	0.2661	24.20
12.5000	50.0000	0.3063	32.00
25.0000	50.0000	0.2685	26.00
31.2500	50.0000	0.2740	20.70
37.5000	50.0000	0.2389	24.40
7.8125	53.1250	0.3025	25.50
14.0625	53.1250	0.3136	26.80
3.1250	56.2500	0.2991	26.00
18.7500	56.2500	0.2691	23.60
21.8750	56.2500	0.3077	32.90
12.5000	62.5000	0.2912	22.80
15.6250	62.5000	0.2936	24.30
10.9375	65.6250	0.2672	22.00
3.1250	68.7500	0.2922	28.20
18.7500	68.7500	0.3023	25.60
21.8750	68.7500	0.2893	22.70
34.3750	68.7500	0.2341	16.00
1.5625	71.8750	0.2868	20.70
7.8125	71.8750	0.3033	23.50
14.0625	71.8750	0.2999	26.50
29.6875	71.8750	0.2856	29.50
0.0000	75.0000	0.2957	26.90
12.5000	75.0000	0.3053	25.70
25.0000	75.0000	0.2800	20.10
37.5000	75.0000	0.2554	19.80

Table A.1: Data comprising water content(cm^3/cm^3), clay content (%) and geographical coordinates

University of Cape Town

Appendix B

GWR code

```
########################## BETA FUNCTION ##########################
#Beta function calculates betas, y-hats, RSS, R^2, sig^2, effective #
#degrees of freedom, AIC and AIC_c for the GWR model                #
###################################################################

Beta=function(y,x,loc,b)
# y = vector of observations on dependent variable
# x= matrix of explanatory variables
# b = bandwidth
#loc = (nx2) matrix of geographical coordinates
{
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
int=rep(1,n)
x=cbind(int,x)
y=as.matrix(y)
betas=matrix(0,nrow=n,ncol=1+p)
yhats=matrix(0,nrow=n,ncol=1)
W=matrix(0,n,n)  #weight matrix
rss=0
sume=0
d=dist(loc)
d=as.matrix(d)
ybar=mean(y)
S=matrix(0,n,n)

for (j in 1:n)
{
W=diag((exp(-1*((d[j,]/b)^2))),n,n)  #weight matrix with   #Guassian kernel
     S[j,]=x[j,]%*%(solve((t(x))%*%W%*%x))%*%(t(x))%*%W
bhat=(solve((t(x))%*%W%*%x))%*%(t(x))%*%W%*%y #bhatj
yhat=x[j,]%*%bhat
e=y[j]-yhat
e2=e^2
```

University of Cape Town

```
sume= e + sume
rss=e^2+rss
yhats[j,1]=yhat

for (k in 1:p) #set up matrix of betas
{
betas[j,k]=bhat[k,1]
betas[j,k+1]=bhat[k+1,1]
}

}
v1=sum(diag(S))
S2=t(S)%*%S
v2=sum(diag(S2))
effdf= n-2*v1+v2
sig2=rss/effdf
sig2ml=rss/n
AICc=2*n*log(sqrt(sig2ml))+n*log(2*pi)+ n*((n+v1)/(n-2-v1))
AIC=2*n*log(sqrt(sig2ml))+n*log(2*pi)+n+v1

sst=0
for (j in 1:n)
{
       W=diag((exp(-1*((d[j,]/b)^2))),n,n)

sst=(W[j,j]*(y[j]-ybar)^2)+sst
}

r2=(sst-rss)/sst


list(betas=betas,yhats=yhats, rss=rss,r2=r2, sig2=sig2,effdf=effdf,sig2ml=sig2ml,AICc=AICc,
AIC=AIC)
}

a=Beta(y=ydata,x=xdata,loc=loc,b=17.282)
mybetas=a$betas

############################   STD ERRORS  ############################
# Calculates the std errors of estimates from GWR model             #
#####################################################################


StdErr=function(y,x,loc,b)
{
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
int=rep(1,n)
x=cbind(int,x)
y=as.matrix(y)
```

University of Cape Town

```
W=matrix(0,n,n)
d=dist(loc)
d=as.matrix(d)
S=matrix(0,n,n)
rss=0
var=matrix(0,nrow=n,ncol=1+p)
stderr=matrix(0,nrow=n,ncol=1+p)

for (j in 1:n)
{
       W=diag((exp(-1*((d[j,]/b)^2))),n,n)
S[j,]=x[j,]%*%(solve((t(x))%*%W%*%x))%*%(t(x))%*%W
bhat=(solve((t(x))%*%W%*%x))%*%(t(x))%*%W%*%y #bhatj
yhat=x[j,]%*%bhat
e=y[j]-yhat
e2=e^2
rss=e^2+rss

}

v1=sum(diag(S))
S2=t(S)%*%S
v2=sum(diag(S2))
effdf= n-2*v1+v2
sig2=rss/effdf

for (j in 1:n)
{
W=diag((exp(-1*((d[j,]/b)^2))),58,58)
A=(solve((t(x))%*%W%*%x))%*%(t(x))%*%W
varB=A%*%t(A)*sig2[1]

for (k in 1:p)
{
var[j,k]=varB[k,k]
stderr[j,k]=(var[j,k])^0.5
var[j,k+1]=varB[k+1,k+1]
stderr[j,k+1]=(var[j,k+1])^0.5
}
}


list(stderr=stderr)
}
s=StdErr(y=ydata,x=xdata,loc=loc,b=17.282)
mystderr=s$stderr



############################    CV SCORE   ############################
# Calculates the CV score of GWR model                                #
```

University of Cape Town

```
#########################################################################

CV=function(y,x,loc,b)
{
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
int=rep(1,n)
x=cbind(int,x)
y=as.matrix(y)
W=matrix(0,n,n)
d=dist(loc)
d=as.matrix(d)

cv=0
for (j in 1:n)
{
     W=diag((exp(-1*((d[j,]/b)^2))),n,n)
     W[j,j]=0
      bhat=(solve((t(x))%*%W%*%x))%*%(t(x))%*%W%*%y
     cv=(y[j]-x[j,]%*%bhat)^2 + cv
}

list(cv=cv)
}

CV(y=ydata, x=xdata, loc=loc,b=.758)



############################    BETA GRID   ############################
# Calculates betas at grid locations for GWR model                     #
# Makes use of a program written by Brunsdon to create the grid        #
#########################################################################

BetaGrid =function(y,x,loc,b,gr.x,gr.y)
#gr.x and gr.y specifies the size of the grid
{
source("c:/Karen/Masters/GWR/Rprogs/gwr4_1.txt")
gr=nice.grid(loc,c(gr.x,gr.y))
grN=gr.x*gr.y
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
int=rep(1,n)
x=cbind(int,x)
y=as.matrix(y)
betas=matrix(0,nrow=grN,ncol=1+p)
W=matrix(0,n,n)
d=matrix(0,grN,n)
```

University of Cape Town

```
#calc distance matrix
for (i in 1:grN)
{
   for (j in 1:n)
         {
d[i,j]=((gr[i,1]-loc[j,1])^2+(gr[i,2]-loc[j,2])^2)^0.5
   }
}


for (j in 1:grN)
{
W=diag((exp(-1*((d[j,]/b)^2))),n,n)

bhat=(solve((t(x))%*%W%*%x))%*%(t(x))%*%W%*%y #bhatj
#betas[j,1]=bhat[1,1]
#betas[j,2]=bhat[2,1]


for (k in 1:p)
{
betas[j,k]=bhat[k,1]
betas[j,k+1]=bhat[k+1,1]
}



}



list(betas=betas)
}

a=Beta(y=ydata,x=xdata,loc=loc,b=17.282,gr.x=50,gr.y=100)
mybetas=a$betas

############################ MONTE CARLO TEST ##########################
# Monte carlo test for stationarity                                      #
##########################################################################

#First randomise locations

Randomloc=function(loc)
{
n=nrow(loc)
newloc=matrix(0,n,2)
ind=1:n
s1=sample(ind)
s1=as.matrix(s1)
```

University of Cape Town

```
for (i in 1:n)
{
newloc[i,1]=loc[(s1[i]),1]
newloc[i,2]=loc[(s1[i]),2]
}
list(newloc=newloc)
}
testing=Randomloc(loc=loc)
newloc=testing$newloc
#Function to calculate the variance of the betas. This function is
#called in the actual MC test function

VarP=function(betamatrix)
{
n=nrow(mybetas)
p=ncol(mybetas)
Vb=matrix(0,1,p)

for (j in 1:p)
{
sumB=sum(mybetas[,j])
sumSq=0

for (i in 1:n)
{
sumSq=(mybetas[i,j]-(sumB/n))^2+sumSq
}
sumSq
Vb[1,j]=sumSq/n
}
list(Vb=Vb)
}
VarP(betamatrix=mybetas)


###MC test function. This function calls the Beta and Variance functions ###

MC=function(y,x,loc,b,nruns)
{
x=as.matrix(x)
p=ncol(x)+1
allvar=matrix(0,nruns,p)

for (l in 1:nruns)
{

a=Randomloc(loc=loc)
newloc=a$newloc
d=Beta(y=y,x=x,loc=newloc,b=b)
mybetas=as.matrix(d$betas)
```

University of Cape Town

```
e=VarP(betamatrix=mybetas)
varbeta=as.matrix(e$Vb)
allvar[l,]=varbeta
}

list(allvar=allvar)
}

res=MC(y=ydata,x=xdata,loc=loc,b=2.11,nruns=1000)
res
```

University of Cape Town

Appendix C

LLGWR code

University of Cape Town

```
############################ BETA FUNCTION ############################
#Beta function calculates betas, y-hats, RSS, R^2, sig^2, effective #
#degrees of freedom, AIC and AIC_c for the LLGWR model              #
#######################################################################

llBeta=function(y,x,loc,b)

{
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
q=(p+1)*3
int=rep(1,n)
delta_u=matrix(0,n,1)
delta_v=matrix(0,n,1)
y=as.matrix(y)
betas=matrix(0,nrow=n,ncol=q)
yhats=matrix(0,nrow=n,ncol=1)
W=matrix(0,n,n)
rss=0
sume=0
d=dist(loc)
d=as.matrix(d)
ybar=mean(y)
S=matrix(0,n,n)

for (j in 1:n)    #calculate delta_u =(u_i-u_0) and
#delta_v =(v_i-v_0)
#the additional parameters for linear extension
{
W=diag((exp(-1*((d[j,]/b)^2))),n,n)

for (k in 1:n)
{
delta_u[k]=loc[k,1]-loc[j,1]
```

```
delta_v[k]=loc[k,2]-loc[j,2]
}


t=p*2
s=p-1
prod=matrix(0,nrow=n,ncol=t)

for (i in 0:s)
{
for(l in 1:n)
{
prod[l,(2*i+1)]=x[l,i+1]*delta_u[l]
prod[l,(2*i+2)]=x[l,i+1]*delta_v[l]
}
}


newX=cbind(int,x,delta_u,delta_v,prod)
newX=as.matrix(newX)



bhat=(solve((t(newX))%*%W%*%newX))%*%(t(newX))%*%W%*%y #bhatj
yhat=newX[j,]%*%bhat
e=y[j]-yhat
e2=e^2
sume= e + sume
rss=e^2+rss

yhats[j,1]=yhat

for (k in 1:q)
{
betas[j,k]=bhat[k,1]

}



S[j,]=newX[j,]%*%(solve((t(newX))%*%W%*%newX))%*%(t(newX))%*%W


}

v1=sum(diag(S))
S2=t(S)%*%S
v2=sum(diag(S2))
effdf= n-2*v1+v2
sig2=rss/effdf
sig2ml=rss/n
```

University of Cape Town

```
AICc=2*n*log(sqrt(sig2ml))+n*log(2*pi)+ n*((n+v1)/(n-2-v1))
AIC=2*n*log(sqrt(sig2ml))+n*log(2*pi)+n+v1

sst=0

for (j in 1:n)
{
       W=diag((exp(-1*((d[j,]/b)^2))),n,n)

sst=(W[j,j]*(y[j]-ybar)^2)+sst
}

r2=(sst-rss)/sst



list(betas=betas,yhats=yhats, rss=rss,r2=r2,v1=v1,v2=v2,effdf=effdf,sig2=sig2,AICc=AICc,AIC=AIC)



}

a=llBeta(y=ydata,x=xdata,loc=loc,b=1.84)

############################   STD ERRORS  ############################
# Calculates the std errors of estimates from LLGWR model           #
#####################################################################
StdErr=function(y,x,loc,b)
{
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
q=(p+1)*3
delta_u=matrix(0,n,1)
delta_v=matrix(0,n,1)
int=rep(1,n)
y=as.matrix(y)
var=matrix(0,nrow=n,ncol=q)
stderr=matrix(0,nrow=n,ncol=q)
W=matrix(0,n,n)
d=dist(loc)
d=as.matrix(d)

for (j in 1:n)
{
W=diag((exp(-1*((d[j,]/b)^2))),n,n)

for (k in 1:n)
{
delta_u[k]=loc[k,1]-loc[j,1]
delta_v[k]=loc[k,2]-loc[j,2]
```

University of Cape Town

```
}


t=p*2
s=p-1
prod=matrix(0,nrow=n,ncol=t)

for (i in 0:s)
{
for(l in 1:n)
{
prod[l,(2*i+1)]=x[l,i+1]*delta_u[l]
prod[l,(2*i+2)]=x[l,i+1]*delta_v[l]
}
}


newX=cbind(int,x,delta_u,delta_v,prod)
newX=as.matrix(newX)

#print(newX)



bhat=(solve((t(newX))%*%W%*%newX))%*%(t(newX))%*%W%*%y #bhatj
yhati=newX%*%bhat
ei=y-yhati
rssi=sum(ei^2)
sig2hati=(t(y-yhati))%*%W%*%(y-yhati)/(n-(p+1))
varB=(solve(t(newX)%*%W%*%newX))*sig2hati[1]


for (k in 1:q)
{
var[j,k]=varB[k,k]
stderr[j,k]=(var[j,k])^0.5

}



}

list(stderr=stderr)
}
s=StdErr(y=ydata,x=xdata,loc=loc,b=30.69)
mystderr=s$stderr

############################    CV SCORE   ############################
# Calculates the CV score of LLGWR model                              #
```

University of Cape Town

```
#########################################################################

llCV=function(y,x,loc,b)

{
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
q=(p+1)*3
int=rep(1,n)
delta_u=matrix(0,n,1)
delta_v=matrix(0,n,1)
y=as.matrix(y)
#betas=matrix(0,nrow=n,ncol=q)
#yhats=matrix(0,nrow=n,ncol=1)
W=matrix(0,n,n)
rss=0
sume=0
d=dist(loc)
d=as.matrix(d)
ybar=mean(y)
cv=0

for (j in 1:n)    #calculate delta_u and delta_v
{
W=diag((exp(-1*((d[j,]/b)^2))),n,n)
W[j,j]=0
for (k in 1:n)
{
delta_u[k]=loc[k,1]-loc[j,1]
delta_v[k]=loc[k,2]-loc[j,2]
}


t=p*2
s=p-1
prod=matrix(0,nrow=n,ncol=t)

for (i in 0:s)
{
for(l in 1:n)
{
prod[l,(2*i+1)]=x[l,i+1]*delta_u[l]
prod[l,(2*i+2)]=x[l,i+1]*delta_v[l]
}
}


newX=cbind(int,x,delta_u,delta_v,prod)
newX=as.matrix(newX)
```

University of Cape Town

```
bhat=(solve((t(newX))%*%W%*%newX))%*%(t(newX))%*%W%*%y #bhatj
cv=(y[j]-newX[j,]%*%bhat)^2 + cv


}

list(cv=cv)


}
llCV(y=ydata, x=xdata, loc=loc,b=30.69)

############################    BETA GRID   ############################
# Calculates betas at grid locations for LLGWR model                   #
# Makes use of a program written by Brunsdon to create the grid        #
########################################################################

llBeta=function(y,x,loc,b,gr.x,gr.y)

{
source("c:/Karen/Masters/GWR/Rprogs/gwr4_1.txt")
gr=nice.grid(loc,c(gr.x,gr.y))
grN=gr.x*gr.y
x=as.matrix(x)
n=nrow(x)
p=ncol(x)
q=(p+1)*3
int=rep(1,n)
delta_u=matrix(0,n,1)
delta_v=matrix(0,n,1)
y=as.matrix(y)
betas=matrix(0,nrow=grN,ncol=q)
W=matrix(0,n,n)
rss=0
sume=0
d=matrix(0,grN,n)
ybar=mean(y)


#calc distance matrix
for (i in 1:grN)
{
   for (j in 1:n)
         {
d[i,j]=((gr[i,1]-loc[j,1])^2+(gr[i,2]-loc[j,2])^2)^0.5
   }
}


for (j in 1:grN)
```

University of Cape Town

```
{
W=diag((exp(-1*((d[j,]/b)^2))),n,n)

for (k in 1:n)
{
delta_u[k]=loc[k,1]-gr[j,1]
delta_v[k]=loc[k,2]-gr[j,2]
}


t=p*2
s=p-1
prod=matrix(0,nrow=n,ncol=t)

for (i in 0:s)
{
for(l in 1:n)
{
prod[l,(2*i+1)]=x[l,i+1]*delta_u[l]
prod[l,(2*i+2)]=x[l,i+1]*delta_v[l]
}
}


newX=cbind(int,x,delta_u,delta_v,prod)
newX=as.matrix(newX)

#print(newX)

bhat=(solve((t(newX))%*%W%*%newX))%*%(t(newX))%*%W%*%y #bhatj

for (k in 1:q)
{
betas[j,k]=bhat[k,1]
  }


}


list(betas=betas)


}

a=llBeta(y=ydata,x=xdata,loc=loc,b=30.69,gr.x=50,gr.y=100)
```

University of Cape Town