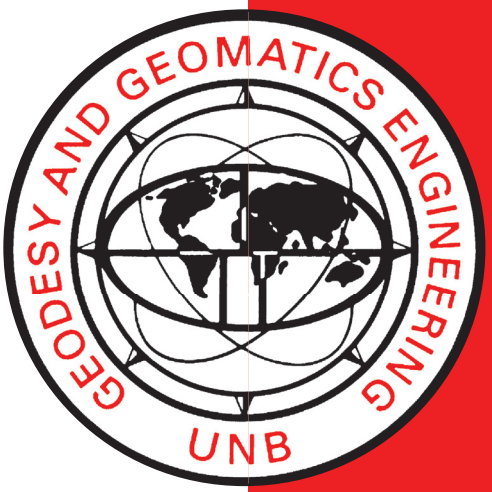
INCORPORATING THE CONCEPT OF 'COMMUNITY' INTO A SPATIALLY-WEIGHTED LOCAL REGRESSION ANALYSIS

HON SHING (RICHARD) CHAN

April 2008



TECHNICAL REPORT
NO. 256

INCORPORATING THE CONCEPT OF 'COMMUNITY' INTO A SPATIALLY-WEIGHTED LOCAL REGRESSION ANALYSIS

Hon Shing (Richard) Chan

Department of Geodesy and Geomatics Engineering
University of New Brunswick
P.O. Box 4400
Fredericton, N.B.
Canada
E3B 5A3

April 2008

PREFACE

This technical report is a reproduction of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Engineering in the Department of Geodesy and Geomatics Engineering, April 2008. The research was co-supervised by Dr. David J. Coleman and Dr. J. Douglas Willms, and support was provided by the Natural Sciences and Engineering Research Council of Canada.

As with any copyrighted material, permission to reprint or quote extensively from this report must be received from the author. The citation to this work should appear as follows:

Chan, Hon Shing (Richard) (2008). *Incorporating the Concept of 'Community' into a Spatially-weighted Local Regression Analysis.* M.Sc.E. thesis, Department of Geodesy and Geomatics Engineering Technical Report No. 256, University of New Brunswick, Fredericton, New Brunswick, Canada, 90 pp.

ABSTRACT

Linear regression has long been used to find relationships among various factors. However, when observations are spatially dependent or spatially heterogeneous the results from a linear regression model are distorted. Researchers developed Geographically Weighted Regression (GWR) to address these problems. It applies the linear regression model at a local level such that each data point has its own set of parameter estimates based on a distance-decay weighting of 'neighbouring observations'. This model, however, is susceptible to the influence of 'outliers'. A Bayesian approach of the GWR method (BGWR) was introduced to address the outlier problem by including various parameter smoothing strategies in the model. This approach provides an opportunity to incorporate the 'community' concept in social sciences to account for the community effect that cannot be addressed by the GWR or distance-based BGWR models. This thesis proposed a 'community-based' BGWR model that improves the prediction power by reducing the overall prediction errors. It also brings significant improvement in the estimation of regression parameters for certain local areas.

ACKNOWLEDGEMENT

I would like to express my gratitude to the following people who have assisted me to complete this research:

- Dr. David J. Coleman, my supervisor, for his patience, guidance and continuous support. His valuable comments have inspired me to look at the research from different perspectives.
- Dr. J. Douglas Willms, my supervisor, for his support and guidance. His expertise and encouragement opened up my eyes to the field of spatial statistics. His keen interest and insightful ideas on the topic helped shaping the direction of this research.
- Dr. Lucia Tramonte, my colleague at the Canadian Research Institute for Social Policy (CRISP) of the University of New Brunswick, for her kind assistance on all sorts of statistical problems that came up during the course of this research.
- Mahin Salmani, also my colleague at CRISP, for teaching me the complicated concepts about Bayesian Statistics.
- My mother and siblings, for their continuous encouragement and love.
- Last but not least, my dear wife and 'editor' of this thesis, Teresa M.Y. Tang, for her unconditional love and support throughout the course of the thesis, and her constructive suggestions over the process of writing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BGWR	Bayesian Geographically Weighted Regression
BGWR-Community	Bayesian Geographically Weighted Regression using a community parameter smoothing relationship
BGWR-Distance	Bayesian Geographically Weighted Regression using a distance-decay parameter smoothing relationship
DAs	Dissemination Areas
GIS	Geographical Information System
GWR	Geographically Weighted Regression
IALSS	International Adult Literacy and Skills Survey
LM	Location-related Weight Matrix
MCMC	Markov Chain Monte Carlo
OLS	Ordinary Least Squares
PCCF	Postal Code Conversion File
PM	People-related Weight Matrix
SES	Socioeconomic Status

1.0 INTROUDCTION

1.1 Background

Linear regression has long been one of the powerful tools of social scientists for finding relationships among various factors. However, when observations (or sample data) have an areal or spatial component, the observed value of the data points from nearby areas may not be independent, which violates the assumption of the linear regression model. For example, observations can be spatially dependent, which means that the observed value at one point in space depends on the values observed at other locations. Also, data can be spatially heterogeneous; that is, the relationships among variables can vary depending on the area of interest. For example, the relationship between house prices and floor area can differ among urban, suburban, and rural areas. When a dataset possesses properties like spatial dependence and spatial heterogeneity, the results from a linear regression model are distorted.

Spatially weighted local regression techniques are a relatively new approach proposed to address the effects of spatial dependence and spatial heterogeneity. One of these techniques, Geographically Weighted Regression (GWR), has attracted the attention of researchers from various fields including the social sciences (Fotheringham et al., 2001; Longley and Tobón, 2004; Malczewski and Poetz, 2005; Cahill and Mulligan, 2007), forestry (Zhang et al., 2004; Wang et al., 2005), ecology (Kupfer and Farris, 2007; Osborne et al., 2007),

and the environmental sciences (Propastin et al., 2006). GWR applies the linear regression model at the local level so that local parameters, rather than global parameters, are estimated. For each point in the dataset, it uses a subset of the data surrounding the point of interest to estimate locally linear regression parameters. Therefore, each data point has its own set of parameter estimates based on the weighted values of its 'neighbouring observations'. As a distance-decay weighting function is usually used, observations closer to the data point of interest have greater influence on the estimates.

While GWR has advantages over ordinary linear regression methods, it has its own drawbacks. One of the drawbacks is that it is more susceptible to the influences of 'outliers' than ordinary linear regression. LeSage (2004) introduced a Bayesian approach of the GWR method, coined as the Bayesian Geographically Weighted Regression (BGWR), to deal with this problem. This approach allows various kinds of parameter smoothing strategies (such as distance-decay) to be included in the model to abate the effects of outliers.

As these spatially weighted local regression techniques emphasize spatial relationships, they cannot account for complex concepts such as 'community', which comprise characteristics beyond geographical attributes. As a result, they are unable to account for the apparent local aberrant observations caused by localized effects such as a community effect. Below is an example that illustrates this situation.

In Figure 1.1, the polygons are DAs (Dissemination Areas[1]) and the red dots are their centroids. We can consider the green DAs to be areas with high socioeconomic status (SES) households while the yellow DAs are areas with low SES households. These two clusters of DAs are usually identified as two communities in social sciences research as people live in close proximity and share common characteristics, identities, or concerns tends to interact more, and hence form a community. As a result of more frequent interactions, the people from the same community are expected to share certain characteristics, more so association than the people from other communities.

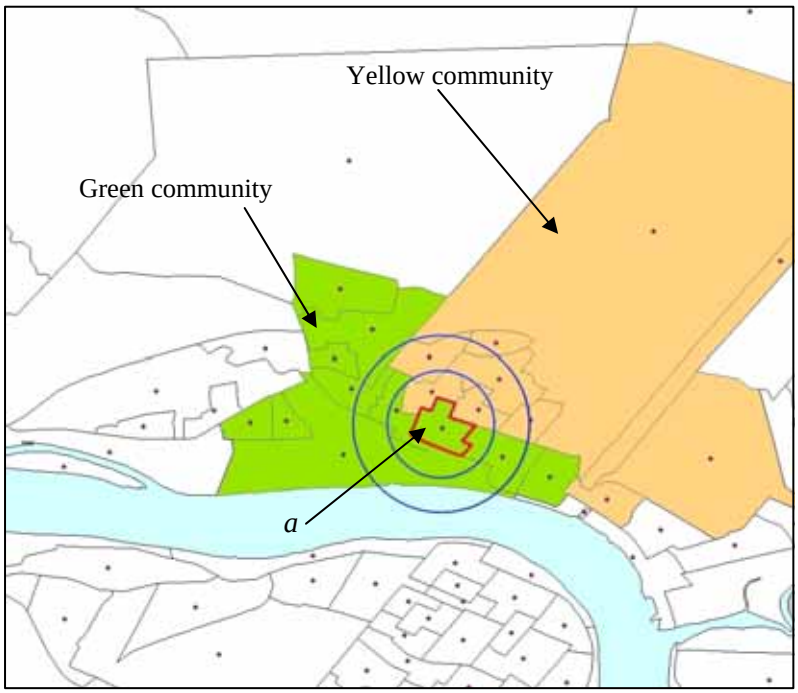


Figure 1.1: Community effect and spatially weighted local regression

[1] A Dissemination Area is a small, relatively stable geographic unit used by Statistics Canada to disseminate census data. It is composed of one or more neighbouring blocks.

Consider the centroid *a* of the highlighted DA in Figure 1.1 to be the point of interest in a local spatial regression. The two circles serve as the reference lines to consider which DA centroids should be taken into account in two scenarios. In the inner circle, there is one neighbouring observation from the green community but two observations from the yellow community. If we are considering the other scenario, represented by the outer circle, there are two neighbouring observations from the green community but four from the yellow community. As discussed earlier, both GWR and distance-based BGWR (BGWR-Distance) models assign higher weights to closer neighbours. Since the number of closer neighbours from the yellow community is greater, observations from the yellow community would have greater influence than those from the green community which is contrary to the expected results. This illustrates that these spatially weighted local regression techniques fail to account for the community effect. Therefore, an alternative model is needed in order to incorporate concepts such as community that comprises both geographical and social characteristics in the model. The BGWR approach that allows for various kinds of parameter smoothing relationship provides such an opportunity.

1.2 Research Objective

The objective of this research is to first propose a statistical model that incorporates the concept of community in a local spatial regression model, and then to assess its performance. More specifically, this study asks whether incorporating the ‘community’ concept into the Bayesian Geographically

Weighted Regression (BGWR) model will improve its performance over the purely distance-based local spatial regression models.

To achieve the objective, this research addresses the following questions:

- How can the concept of 'community' be operationalized as measurable variables that can be incorporated in the BGWR model?
- What are appropriate evaluation methods for assessing the performance of the proposed model (BGWR-Community) and the purely distance-based local spatial regression models (GWR and BGWR-Distance) under study?
- Does the incorporation of the 'community' concept into the BGWR model improve its prediction power by reducing the prediction error over the purely distance-based local spatial regression models under study?

1.3 Approaches to the Research

The author has adopted the following approaches to achieve the research objective:

- **A Literature review** pertaining to the following areas of interest has been conducted to inform the research: (i) the definitions of community and community effects as well as the techniques used in empirical studies of community effects; (ii) operationalization of the 'community' concept into measurable variables; and (iii) spatially weighted local regression methods with special emphasis on the GWR and Bayesian approach of the GWR methods.

- **Experiment design and evaluation methods** – The goal of the experiment is to assess whether incorporating the 'community' concept into the BGWR model can improve its prediction power over the purely distance-based local spatial regression models under study. The base model and the experiment dataset are selected based on previous research and availability of data. The base model is an ordinary least squares (OLS) regression model which serves as a baseline for assessment. The study area is selected according to the assumptions of the local spatial regression models. The other models under study are the GWR, BGWR-Distance, and the BGWR-Community models (i.e., the proposed model). In addition to regression statistics, three empirical evaluation methods are selected to assess the different aspects of the prediction power of the models.

- **Models Implementation** – The OLS base model is implemented with *SPSS*™ statistical package[2] while the local spatial regression models are implemented based on LeSage's (2005) Spatial Econometric Toolbox, a host of spatial econometric estimation methods implemented with *Matlab*™ [3]. To implement the 'community' concept as a parameter smoothing relationship of the BGWR model, the concept is first operationalized into measurable variables based on Galster's (2001) bundle of spatially-based attributes which capture both the physical and social characteristics of a community. Then, these measurable variables

[2] *SPSS* is a registered product name of a statistical software package produced by SPSS Inc.
[3] *Matlab* is a registered product name of a numerical computations and graphics software package produced by The MathWorks, Inc.

are used to generate two weight matrices, a geographical distance-based weight matrix and a Mahalanobis distance-based matrix which are then combined to produce a normalized weight matrix, hence the community parameter smoothing relationship.

- **Evaluate the results of the experiment** – The results of the experiment are evaluated using the above-mentioned evaluation methods and the outputs are presented as numerical indicators and charts. Further discussion of the trends and patterns of the local effects of the three local spatial regression models are illustrated with maps generated by Kriging interpolation of the prediction improvement results. These outputs are then compared and analyzed to determine whether the prediction power of the proposed model has improved over other models under study.

1.4 Scope of the Research

This research attempts to incorporate the ‘community’ concept into a local spatial regression model in order to account for the community effects that cannot be addressed by the purely distance-based local spatial regression models under study. By doing so, the prediction power of the proposed model is expected to improve over the existing models. To assess whether the proposed model improves prediction by the incorporation of the ‘community’ concept, the experiment is designed to include only the GWR and the BGWR-Distance models

for comparison. Other non-distance-based local spatial regression models such as spatial autoregressive model or spatial expansion methods are not considered.

The purpose of this research is to determine whether the proposed model can improve the prediction power by accounting for the community effects. It will not identify or assess the impact of the community effects.

Furthermore, the following assumptions are made for the experiment:

- The straight-line distances among the observations can approximate the real distances; and
- The assigned locations (i.e., the centroids of DAs or randomly assigned postal codes) of the observations are adequate replacement of the actual locations of the observations.

1.5 Significance and Contributions of the Research

The present research proposes a statistical model that incorporates the concept of ‘community’ in a local spatial regression model to account for community effect which cannot be addressed by purely distance-based local spatial regression models. By doing so, the model improves the prediction power, in comparison with the purely distance-based local spatial regression models by reducing the overall prediction errors and bringing significant improvement to certain local areas.

During the implementation process, this study demonstrates a means to operationalize a concept that captures both the location-related and other non-locational characteristics. This allows the model to address the localized effects, such as community effect, not only by geographical distance but also by other relevant attributes such as socio-economic factors.

Through the operationalization process, this research sets an example to other research areas on how to integrate concepts that are geographical in nature but have ill-defined boundaries into the local spatial regression model without pre-defining the boundaries. In other potential research areas such as forestry or ecology, these concepts may be land cover types, soil types or habitats where the boundaries are not always well-defined. The potential applications of the proposed approach are promising.

1.6 Organization of the Thesis

This thesis consists of five chapters. Chapter 1 provides an overview of the research. It describes the research problem, the objectives and approaches to the research. The significance and contributions of the research and its scope are also discussed. Chapter 2 provides the background knowledge to two major areas, (1) definition of community and operationalization of the ‘community’ concept; and (2) technical background about the GWR and BGWR models. Previous methods applied in studying community effects are also reviewed. Chapter 3 presents the design of the experiment and its considerations, including the choices

of models, selection of dataset and study area. The development of the proposed model is also described in detail. The latter part of the chapter discusses the evaluation methods of the study. Chapter 4 evaluates the results of the experiment with the methods described in Chapter 3. The evaluation results between different models are compared to determine whether incorporating the ‘community’ concept into the BGWR model brings the expected improvement to the purely distance-based local spatial regression models, followed by a detail discussion of the local improvements brought by the proposed model. Chapter 5 provides a summary of the research, discusses its limitations, and suggests future research opportunities.

2.0 COMMUNITY, COMMUNITY EFFECTS AND SPATIALLY WEIGHTED LOCAL REGRESSION

2.1 Concepts of Community in Social Sciences Research

Community has long been one of the fundamental concepts in social sciences research (Brint, 2001). However, the diversity in the definitions of 'community' is also well recognized. Hillery (1955) did a comprehensive review on the scientific literature at his time and found 94 different definitions of 'community'. Notwithstanding the diversity, he found that there was "a basic agreement that community consists of persons in social interaction within a geographic area and having one or more additional common ties" (Hillery, 1955, p. 111). From time to time, researchers revisited the definition and found that the concept of 'community' has been evolving to reflect changes in technology and social compositions (Trojanowicz and Moore, 1988; Brint, 2001; MacQueen et al., 2001; Wellman, 2001). One example is the divorce of community and geography that leads to the distinction between 'communities of place' and 'communities of interest'.

2.1.1 Definitions of Community

2.1.1.1 Communities of Interest

'Communities of interest' refers to groups of people whose members have something in common, such as political interest, hobbies, or expertise, but not necessarily conducting activities at a common place or location. These include but are not limited to unions and associations of workers, associations of

businesses, sports groups, and international professional bodies. As technologies advance, this definition also covers virtual or online communities. These groups provide their members a sense of community or identity.

2.1.1.2 Communities of Place

'Communities of place' or geographic communities are made up of the people who happen to live or take part in activities in a particular area or locality. They may or may not share a common interest but they always share certain characteristics, identities or concerns (Law, 2000). Neighbourhoods, school districts and urban regions are examples of geographic communities.

2.1.1.3 Working Definition of Community

Although there is no universally accepted definition of community, in most research, the basic elements are still location or place, social interaction, common interests and perspectives, as well as social ties (Brint, 2001; MacQueen et al., 2001). Given the focus of this research is about spatially weighted local regression, community in this research, unless stated otherwise, refers to '*communities of place in a local setting'* that resembles the meaning of neighbourhood. Small and Supple (1998, p. 3) referred neighbourhood as "a physical place defined by socially shared boundaries which includes a population of people who usually share similar life chances, socio-economic status and physical proximity". Galster (2001, p. 2112) specified neighbourhood as "a bundle of spatially-based attributes associated with cluster of residences". These

attributes include physical and social characteristics of the neighbourhood, namely structural characteristics of the buildings, infrastructural characteristics, demographic characteristics and class status characteristics of the residents, tax/public service package characteristics, environmental characteristics, proximity characteristics, political characteristics, social-interactive characteristics and sentimental characteristics.

2.1.1.4 Operationalize the 'Community' Concept in Social Sciences Research

After defining the 'community' concept, it is necessary to operationalize it as measurable attributes or variables so that they can be integrated into a statistical model. As Glaster (2001) pointed out that 'community' is spatially-based, measurement of attributes is only possible after a particular area has been specified or demarcated. Therefore, how space being delineated for measurement is part of the operationalization process.

Small and Supple (1998) found that defining meaningful spatial boundaries for study is not a straightforward task. Glaster (2001) realized that it was very difficult to define a clear boundary of a community or neighbourhood as the attributes to be measured vary over space in different patterns. Given that it has to demarcate an area to take measurement but the boundaries are most likely made arbitrarily, and may not coincide with the spatial patterns of the attributes, the best one can do is to use the smallest spatial unit of data available in order to get a 'higher resolution image' of the spatial patterns. This is in line with Dietz's

(2002) observation that space delineations in most research are constrained by the limitations of the available datasets.

With respect to choices of measurable variable to represent the 'community' concept, Small and Supple (1998) proposed a three-level framework for conceptualizing communities in terms of community effects. The first level is "the direct aggregate influences of the universe of community settings and institutions" upon the individuals in the community (Small and Supple, 1998, p. 8). These community settings and institutions include schools, health care facilities, religious institutions, and so on. The second level is the influences generated by the relationships and linkages between settings in a community. One example is cross-setting consistency which means if community settings such as schools and religious institutions share common goals and values, the influences in each setting are reinforced. Influences under the third level, such as community identity, only occur when a community reaches a critical mass. This framework considers communities as complex systems which "will not be easy to operationalize or study" (Small and Supple, 1998, p. 20).

Lupton (2003) suggested four guidelines for selecting the measurable variables to represent the 'community' concept: (1) both physical and social aspects of community should be considered; (2) use appropriate boundaries for community under study; (3) reflect different relationships between individuals

and community; and (4) reflect the relationships between neighbouring communities.

Considering that different attributes vary over space in different patterns, Glaster (2001) suggested that researchers should choose only those attributes of interest for a particular type of community to avoid the discrepancies among the spatial patterns of the attributes. Dietz (2002) reviewed 39 previous studies and found that more than two thirds of the studies (24 out of 39) select socio-economic and demographic attributes as measurable variables. Only 15 studies use research specific variables.

The above discussion reveals that there is no universally accepted means to operationalize community. Based on Glaster's (2001) "bundle of spatially-based attributes" and Lupton's (2003) guidelines, the common denominator is that the attributes should include both physical and social characteristics of the community. Further discussion about the operational definition of community for this research is presented in Section 3.1.5.2(d).

2.1.2 Community Effects

In general, the term 'community effects' refers to the influences a community exerts on an individual's behaviour or socioeconomic outcomes through social interaction within that community (Dietz, 2002). Other than the effects from direct social interaction, Dietz (2002) also referred to the correlation

between individual behaviour or outcomes with the characteristics of an individual's neighbours and neighbourhood as a kind of community effect. This kind of community effect may be the result of certain social processes such as population sorting. For the purpose of this research, the term 'community effects' refers to the latter, i.e., correlation between individual behaviour or outcomes with the characteristics of an individual's neighbours and neighbourhood. In this research, the incorporation of the 'community' concept into a local spatial regression model is to account for the community effect.

2.2 Techniques Used in Empirical Studies of Community Effects

Although linear regression is a powerful tool for finding relationships among various factors, it is not effective with data showing properties of dependence such as spatial dependence. Geographical data, as Goodchild (2001) pointed out, are frequently found to be spatially dependent. Researchers studying community effects usually deal with this problem by modifying or extending the ordinary least squares (OLS) regression model. Dietz (2002) reviewed 39 previous studies and found that most researchers used OLS, two-stage OLS or multi-level regression models. Only four of them used spatial econometric or spatial auto-regressive models.

The most common way to use OLS in dealing with spatially dependent data is to introduce a dummy variable for broad classes of spatial location, such as urban, suburban and rural. Another method is to use the multi-level modeling

techniques to include a pre-defined location or community as one of the hierarchical levels. A major criticism about these approaches is that a spatial hierarchy has to be pre-defined and incorporated into the model, but the sensitivity of these models to changes in the spatial hierarchical groupings is not investigated (Brunsdon et al., 1998; Dietz, 2002).

Spatial autoregressive models, which are usually applied to areal data, simulate local spatial interaction by putting the spatially weighted dependent variable (with a spatial weight matrix) on the right side of the equation. The spatial weight matrix is typically a normalized contiguity matrix which does not consider the size or shape or absolute location of the zones. Although this type of model addresses the impact of local relationships in the data, the output is always a set of global parameter estimates (Fotheringham et al., 2002). No trend or spatial pattern of the parameters can be observed.

Gorr and Olligschlaeger (1994) used a technique called spatially adaptive filtering that based on a 'predictor-corrector' mechanism to generate the parameter 'drift' across space. Brunsdon et al. (1998) noted that a major drawback of this technique is that the validity of the assumption of variation in parameters cannot be tested statistically.

Cassetti (1972) proposed a spatial expansion method that expands the coefficients in the regression model with the explicit function of the spatial

location of the cases. It is also the expansion method that restricts how the changes of the estimated parameters can be displayed over space. For example, when x-y expansion is used, a parameter β for a variable v will be expanded from one term βv into three terms $\beta_o v$, $\beta_1 xv$, and $\beta_2 yv$. However, the trend of v can only be displayed either along the x-axis (using β_1 values) or the y-axis (using β_2 values); hence, no spatial pattern can be mapped out for exploration.

2.3 Geographically Weighted Regression

2.3.1 Basic Concept

To overcome the deficiencies of the techniques discussed above, Brunsdon et al. (1998) proposed an alternative technique, Geographically Weighted Regression (GWR). GWR is regarded as a non-parametric model so no pre-defined spatial hierarchy is necessary. It attends to spatial dependence of data with a distance-decay weight function and tackles spatial heterogeneity using a subset of the observations for each prediction point estimation. The parameters obtained are at the local level, instead of the global level, and can be mapped spatially with Geographical Information System (GIS) software for exploration of trends or patterns.

The central idea of GWR is to apply an OLS regression model locally. Unlike OLS regression that uses all observations in the parameters estimation, GWR uses a subset of the observations to estimate the parameters for each prediction point. The subset used in each estimation is defined by a moving kernel

(Figure 2.1). It also applies a pre-defined weighting function to weight the observations around the prediction point. The weighting function (Figure 2.1) is usually a distance-decay one that reflects Tobler's (1970, p. 236) First Law of Geography, *"Everything is related to everything else, but near things are more related than distant things"*.

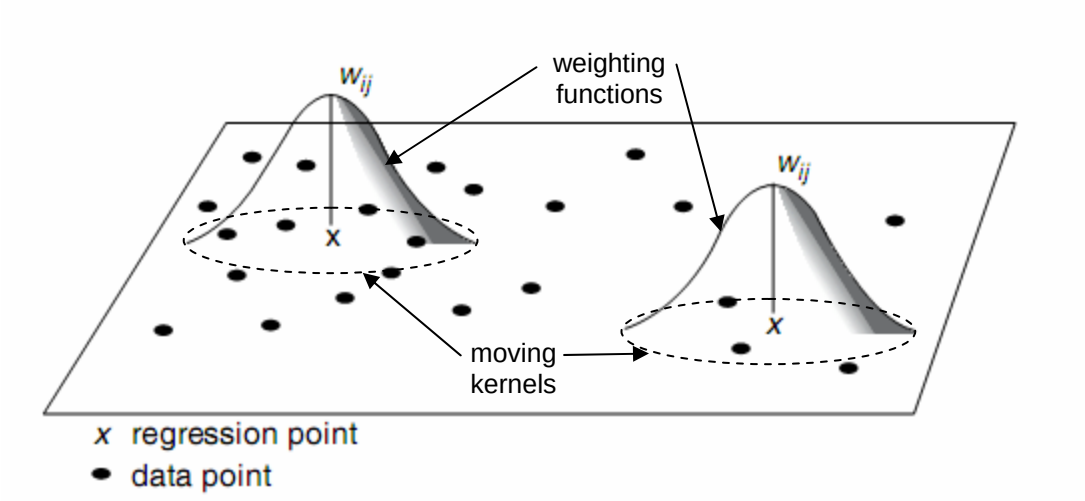


Figure 2.1: Distance-decay weighting function and moving kernel
(After Fotheringham et al. [2002, Figure 2.11])

2.3.2 The Theory

An OLS regression model can be written as:

$$y = \beta_0 + \sum_k \beta_k x_k + \varepsilon \qquad (2.1)$$

where:

y = the dependent variable
x_k = a vector of independent variables
β_0 = the intercepting constant
β_k = a vector of regression coefficients
ε = the error term whose distribution is N(0, σ^2)

In OLS, it is assumed that the parameter values (β) are constant across the study area. Any unexplained variations (including the spatial variations) are put in

the error term ε. The aim of OLS is to estimate the parameter values (β) for a regression model (which is also a statistical function) so that the function best fits a set of data (or observations) in a least squares sense. The least squares estimates of the parameter values (β) can be obtained by:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.2}$$

where:

X = a n by k matrix containing n observations of the k independent variables
X^T = transpose of X
$\hat{\beta}$ = estimated regression coefficients

Based on the OLS model represented in (2.1), the general form of a local regression can be written as (2.3).

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{2.3}$$

where:

y_i = the dependent variable of a prediction point i
x_{ik} = a vector of independent variables for prediction point i
β_0 = the intercepting constant for prediction point i
β_k = a vector of regression coefficients for prediction point i
ε_i = the error term for the estimation of prediction point i
u_i, v_i = the coordinates (or location) of prediction point i

This general form indicates that each prediction point would have its own regression coefficients being estimated. As GWR is a spatially weighted local regression method, each prediction point also has its own weight matrix. The compact form of the GWR model including the weight matrix for prediction point i are written as (2.4), with subscript i replacing (u_i, v_i) in (2.3).

$$W_i y = W_i X \beta_i + \varepsilon_i \tag{2.4}$$

where:

W_i = a n by n weight matrix for prediction point i whose off-diagonal elements are zero

y = a n by 1 vector of dependent variable observations
X = a n by k matrix containing n observations of the k independent variables
β_i = a vector of regression coefficients for prediction point i at location (u_i, v_i) containing β_0 and β_k in (2.3)
ε_i = the error term for the estimation of prediction point i

The least squares estimation scheme for (2.4) can be written as:

$$\hat{\beta}_i = (X^T W_i X)^{-1} X^T W_i y \tag{2.5}$$

where:
y = a n by 1 vector of dependent variable observations
X = a n by k matrix containing n observations of the k independent variables
X^T = transpose of X
$\hat{\beta}_i$ = a vector of estimated regression coefficients for prediction point i at location (u_i, v_i)
W_i = a n by n weight matrix for prediction point i whose off-diagonal elements are zero

By comparing (2.2) and (2.5), it is obvious that the weight matrix W_i is central to GWR. If W_i is an identity matrix, then (2.5) is equal to (2.2). As discussed in section 2.3.1, GWR put more emphasis on the observations closer to the prediction point. A distance-decay function is used to generate W_i so that higher weights are assigned to observations closer to the prediction point. W_i is also used to define the subset of observations to be used in the local regression. Therefore, obtaining a proper weight matrix is crucial to GWR modeling.

2.3.3 Spatial Weighting Function and Bandwidth

Obtaining a proper weight matrix requires an appropriate spatial weighting function and bandwidth. Spatial weighting functions can be implemented as

binary function, exponential distance-decay-based function, or kernel function (Brunsdon et al., 1998).

A binary function assigns a weight of 1 to all observations whose distances from the prediction point *i* are less than *b*; otherwise zero. In this case, the spatial weighting function defines a circular kernel of radius *b* (2.6).

$$w_{ij} = 1 \text{ if } d_{ij} < b \qquad (2.6)$$
$$w_{ij} = 0 \text{ otherwise}$$

where:
w_{ij} = weight assigned to observation *j* for the estimation of prediction point *i*
d_{ij} = distance between observation *j* and prediction point *i*
b = bandwidth

However, a binary function like (2.6) is considered as unnatural since it means that an observation which is *b* km from the prediction point is weighted 1 while the other one which is *b+0.00001* km from the prediction point is weighted 0. An exponential distance-decay function like (2.7) or a distance-decay kernel function like (2.8) is considered more appropriate as the weight changes more gradually, depending on the bandwidth being selected. The crux of these functions is the 'bandwidth' *b* which defines the behaviour of these functions.

$$w_{ij} = \exp[-½(d_{ij} / b)^2] \qquad (2.7)$$

$$w_{ij} = [1 - (d_{ij} / b)^2]^2 \text{ if } d_{ij} < b \qquad (2.8)$$
$$w_{ij} = 0 \text{ otherwise}$$

Figure 2.2 shows that a larger bandwidth (b_2) results in a flatter weight distribution. Observation point *a* is assigned a lower weight (wt_2) in the case of bandwidth b_2 than b_1. This in turn affects the goodness of fit of a GWR model. Hence,

selecting an appropriate bandwidth is actually calibrating the spatial weighting function selected for the GWR model. The most common method used for the calibration is a cross-validation approach (Fotheringham et al., 2002). Cross-validation scores for different bandwidths are computed using (2.9):

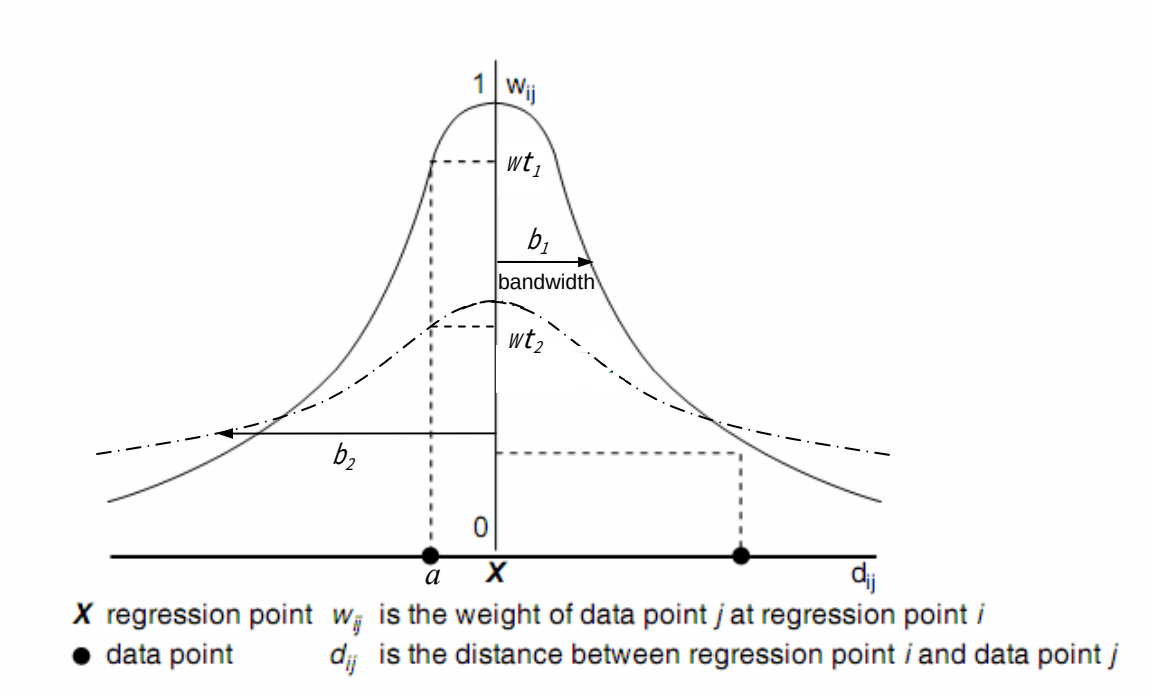


Figure 2.2: Bandwidth and spatial weighting function
(After Fotheringham et al. [2002, Figure 2.10])

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2 \tag{2.9}$$

where:

CV = cross-validation score

$\hat{y}_{\neq i}(b)$ = the predicted value of y_i from the GWR model where the i^{th} observation is omitted during the estimation process

b = bandwidth

The bandwidth that gives the least cross-validation score is considered as the most appropriate one because this indicates the model produces least prediction error at this bandwidth.

2.3.4 Outliers and Community Effects

One of the major criticisms about GWR is that the presence of any outliers would distort the results of the nearby prediction points greatly due to the nature of local regression and the distance-decay weighting function (LeSage, 2004). Fotheringham et al. (2002) suggested that outliers can be detected using the externally Studentised residual. They recommended removing the outliers from the dataset and then re-fitting the model. While such 'detect-and-remove' strategy may be useful in dealing with ordinary outliers, it cannot handle local aberrant observations or local influences caused by community effects as illustrated in the example in Section 1.1. In the next section, an alternative approach that extends the GWR model is introduced to handle local aberrant observations. This alternative approach also provides the opportunities to incorporate the 'community' concept into a spatially weighted local regression model.

2.4 Bayesian Geographically Weighted Regression Model

Apart from the outlier problem, LeSage (2004) also identified two other problems about GWR. One is about the validity of inferences for the regression parameters by traditional least squares approaches and the other is the 'weak data' problem (i.e., the effective number of observations for each estimation may be too

small). Thus, he proposed an alternative, a Bayesian approach of the GWR which he coined as Bayesian Geographically Weighted Regression (BGWR) to address the deficiencies of GWR. For brief introduction to Bayesian statistics, one may refer to Bullard (2001) and Goddard (2003) or Koch (2007) and Lynch (2007) for detailed discussions.

2.4.1 The Theory of BGWR

LeSage (2004, p. 243) extended the GWR model by expanding the parameter β_i in (2.4) with an explicit statement of what he called the "parameter smoothing relationship" such as the distance-based parameter smoothing relationship in (2.10) below.

$$\beta_i = (w_{i1} \otimes I_k \ldots w_{in} \otimes I_k) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \mu_i \tag{2.10}$$

The parameter smoothing function in (2.10) is a locally linear combination of neighbours weighted by a distance-decay function. Other parameter smoothing relationships such as contiguity, and monocentric (i.e., concentric zones to a pre-defined centre) are also possible (LeSage, 2004). The terms w_{ij} (such that j = 1 to n) in (2.10) represent the normalized distance-decay-based weights such that the sum of the row vectors (w_{i1}….w_{in}) are 1 while w_{ii} = 0.

The distribution for the error terms ε_i in (2.4) and μ_i in (2.10) are added:

$$\varepsilon_i \sim N[0, \sigma^2 V_i], \quad Vi = diag(v_1, v_2, \cdots, v_n) \tag{2.11}$$

$$\mu_i \sim N[0, \sigma^2 \delta^2 (X^T W_i^2 X)^{-1}] \quad (2.12)$$

where σ^2 is the variance of y. V_i is an unknown variance parameter introduced to accommodate spatial heterogeneity of variance. It is an n by n matrix with diagonal elements, (v_1, v_2, …, v_n), while the off-diagonal elements are zero. In order to estimate the n number of v_i terms for n observations, LeSage (2004) suggested to assign a prior distribution $\chi^2(r)$ for the n^2 terms using a hyperparameter r such that the mean of prior equals unity and the prior variance is $2/r$. This implies that when r is very large, the prior variance becomes very small and V_i become an identity matrix. Hence, the variance of ε_i become a constant variance $\sigma^2 I_n$ for all observations i (i.e., homoscedasticity). The other property of the hyperparameter r is that when it is small, say 4, it can down-weight aberrant observations or outliers (which are identified if the difference between observed values and predicted values are big) in a local regression estimation.

The term μ_i is prior uncertainty about the parameter smoothing relationship. It is assumed to follow a normal distribution with mean zero and a variance based on Zellner's g-prior, a commonly used prior in Bayesian variable selection (Berger and Pericchi, 2000; Denison et al., 2002). This prior variance is proportional to the parameter variance-covariance matrix, $\sigma^2(X^T W_i^2 X)^{-1}$. The term δ^2 is a scale factor that regulates the degree of adherence between the parameter estimates and the proposed smoothing relationship. That means when δ^2 is very small like 1 or 0.5, the smoothing relationship would impose more influence on the regression coefficient estimation. On the other hand, when δ^2 is very large

(e.g., approaching infinity), and V_i equals to identity matrix, BGWR would produce estimates very close to those by GWR.

2.4.2 Estimation with the BGWR Model

Like other Bayesian models, the estimates for the BGWR model are the multivariate posterior probability density for all of the parameters in the model. LeSage (2004) used Gibbs sampling, a technique for generating random samples from a distribution based on the Markov Chain Monte Carlo (MCMC) approach, to carry out the estimation.

The procedures of the Gibbs sampling process described below are based on the parameter smoothing relationship in (2.10) (recapped below for ease of reference). A compact form of (2.10) can be written as (2.13).

$$\beta_i = (w_{i1} \otimes I_k \ldots w_{in} \otimes I_k) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \mu_i \tag{2.10}$$

$$\beta_i = J_i \gamma + \mu_i \tag{2.13}$$

where $J_i = (w_{i1} \otimes I_k \ldots w_{in} \otimes I_k)$, and $\gamma = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$

The parameters to be estimated in this process are β_i, σ, δ, V, and γ which come from (2.11), (2.12), and (2.13). Before the Gibbs sampling process starts, arbitrary values have to be assigned to the parameters β_i, σ and γ. The bandwidth for the distance-decay function in the parameter smoothing relationship is obtained from the estimates of the initial analysis of an equivalent GWR model. The prior r

(which gives V) and δ are selected based on the considerations discussed in Section 2.4.1.

The Gibbs sampling process comprises many passes, say 500 or 1,000. In each pass, a sample of each observation is drawn to compute and update certain parameters. The detailed procedures are as follows:

1. Initial values for the parameters are represented as β_i^0, σ_i^0, δ^0, V_i^0, and γ^0, where the superscript *0* indicates the pass number and subscript i indicates the observation number.
2. For each observation i = 1 to n,
    a. sample a value β_i^1 from $P(\beta_i \mid \sigma_i^0, \delta^0, V_i^0, \gamma^0)$;
    b. sample a value σ_i^1 from $P(\sigma_i \mid \beta_i^1, \delta^0, V_i^0, \gamma^0)$;
    c. sample a value V_i^1 from $P(V_i \mid \sigma_i^1, \delta^0, \beta_i^1, \gamma^0)$;
3. Update γ^0 to γ^1 with the sampled values of β_i^1 , i = 1 to n from each of the n draws in Step 2.
4. Sample a value δ^1 from $P(\delta \mid \sigma_i^1, V_i^1, \beta_i^1, \gamma^1)$.
5. Replace β_i^0, σ_i^0, δ^0, V_i^0, γ^0 in Step 1 with β_i^1, σ_i^1, δ^1, V_i^1, γ^1.
6. Steps 2 to 5 represent a single pass. Repeat for another 499 passes.

The output or estimates obtained from the Gibbs sampling process are a collection of samples of parameter values, from which the posterior probability density (or conditional posterior distribution) for the parameters can be constructed. Normally, the samples of the first 50 passes would not be used as the

initial sample values are not very stable. Therefore, for a collection of 500 samples, the process needs 550 passes.

2.4.3 Outliers and Community Effects

BGWR adopts a different strategy in handling outliers. Instead of 'detect-and-remove', BGWR mitigates the influence of the outliers by down-weighting the outliers as well as smoothing any aberrant values in the parameters of an observation with its neighbours with some pre-defined parameter smoothing relationship. The parameter smoothing relationship also provides an opportunity to incorporate the community concept into the BGWR model. A function that measures how likely observations are coming from the same community can be used in the parameter smoothing relationship to account for the community effects (to be discussed in Section 3.1.5.2(d)).

2.5 Summary

This chapter begins with a discussion on the diversity in the definitions of 'community' as well as the distinction between community of interest and community of place. Given the focus of this research, a working definition of the term 'community' is set out as communities of place in a locality setting that resembles the meaning of neighbourhood. In addition, the constraints (imposed by the available datasets) and considerations to operationalize 'community' into measurable variables are also discussed. Furthermore, the term 'community

effects' is also defined as the correlation of individual behaviour or outcomes to the characteristics of an individual's neighbours and neighbourhood.

After reviewing the techniques used in some empirical studies of community effects -- such as OLS, multi-level regression, and spatial autoregressive model -- a brief discussion about these techniques suggested that their deficiencies might be threefold. Firstly, a pre-defined spatial hierarchy is required but the effects of different definitions of the spatial hierarchy on the models are not certain. Secondly, parameters estimated are mostly at a global level. Thirdly, even though parameters are estimated at a local level, they are not easily mapped with GIS software for exploration of trends or patterns.

Geographically Weighted Regression (GWR), a non-parametric spatially weighted local regression model proposed by Brunsdon, Fotheringham and Charlton (1998), is introduced as it overcomes the above deficiencies. The central idea of this model is to apply an ordinary least squares regression locally. It uses a moving kernel to define a subset of the observations and weight them with a distance-decay function to estimate the parameters for each prediction point. Due to the nature of local regression and the properties of the distance-decay weighting function, estimates of GWR are susceptible to outliers. While the proposed 'detect-and-remove' strategy may be able to handle ordinary outliers, it cannot handle local aberrant observations or local influences caused by community effects.

LeSage (2004) proposed a Bayesian approach of the GWR, coined as BGWR (Bayesian Geographically Weighted Regression), and introduced a parameter smoothing relationship into the model to mitigate the influence of the local aberrant observations by down-weighting and smoothing them with its neighbouring values. The parameter smoothing relationship of BGWR also provides an opportunity to incorporate the concept of community into the model. A BGWR model with ‘community’ as the parameter smoothing relationship will be described in the next chapter.

3.0 EXPERIMENT DESIGN AND EVALUATION METHODS

In Chapter 1, I argued that purely distance-based local spatial regression models cannot account for the 'community' effect in social sciences research. I also argued that by incorporating the 'community' concept into the local spatial regression model, its prediction power can be improved. A scientific experiment has been designed and implemented to support this argument. This chapter discusses the details of the design of the experiment including the data requirements, procedures, and the models that have been involved in the experiment. To enable the community concept to be expressed as measurable variables, it is necessary to derive an operational definition of community. A detailed discussion about the operational definition of community is therefore included. The latter part of this chapter discusses the evaluation criteria to be used in this research.

3.1 Experiment Design

3.1.1 Background

The present experiment requires the datasets that demonstrate certain extents of spatial dependence. Therefore, it is developed on previous models that involved datasets that exhibit such properties. Combining this guideline with another key consideration, data availability, I have selected Dr. Douglas Willms' research on adult literacy as the starting point. Willms and his colleagues have used multi-level modelling techniques in several research projects to account for

the spatial dependence of adult literacy data (Willms, Chan and Tang, 2007; Willms and Tang, 2007; Willms and Murray, 2007). The base model of this study is adapted after Willms and Murray (2007).

3.1.2 Data

The dataset for this research is the International Adult Literacy and Skills Survey (IALSS). It is the same dataset that was used by Willms and Murray (2007). The IALSS studies four skills of adults from various countries, including Canada, at age 16 and older. The four skills are *prose literacy, document literacy, numeracy,* and *problem solving*. Like the research by Willms and Murray (2007), prose literacy – the knowledge and skills needed to understand and use information from text – is used in this study as the dependent variable. Proficiency of prose literacy in the IALSS is indicated on a scale ranging from 0 to 500 points. Independent variables are gender, age, age-squared, years of education, and personal income. Details about these variables are discussed in Section 3.1.5.1. In addition, data at the DA (Dissemination Area) level from the 2001 Canada Census are used as the source of spatial and community-level data (Statistics Canada, 2003a). Details about these data are discussed in Section 3.1.5.2(d).

3.1.3 Study Area

IALSS data for the Atlantic provinces, which comprise information on 4,682 adults, were initially considered for this study. However, the area was

inappropriate because one of the basic assumptions of the local spatial regression models was violated. Most local spatial regression models take straight line distances between sample points as input for computation. This, however, is not applicable to the Atlantic provinces because many parts of the provinces are separated from their neighbouring provinces by various bays and straits (see Figure 3.1). Therefore, observations regarding the fit of the models tested may vary due to the fact that straight line distances were used, and this variation would be confounded with the introduction of community characteristics with the BGWR-Community model. This problem is less prominent in the selected study area in southern Ontario (see Figure 3.2) where there are 3,709 adults in the sample after cases with missing data are removed. This limitation is discussed in the final chapter.

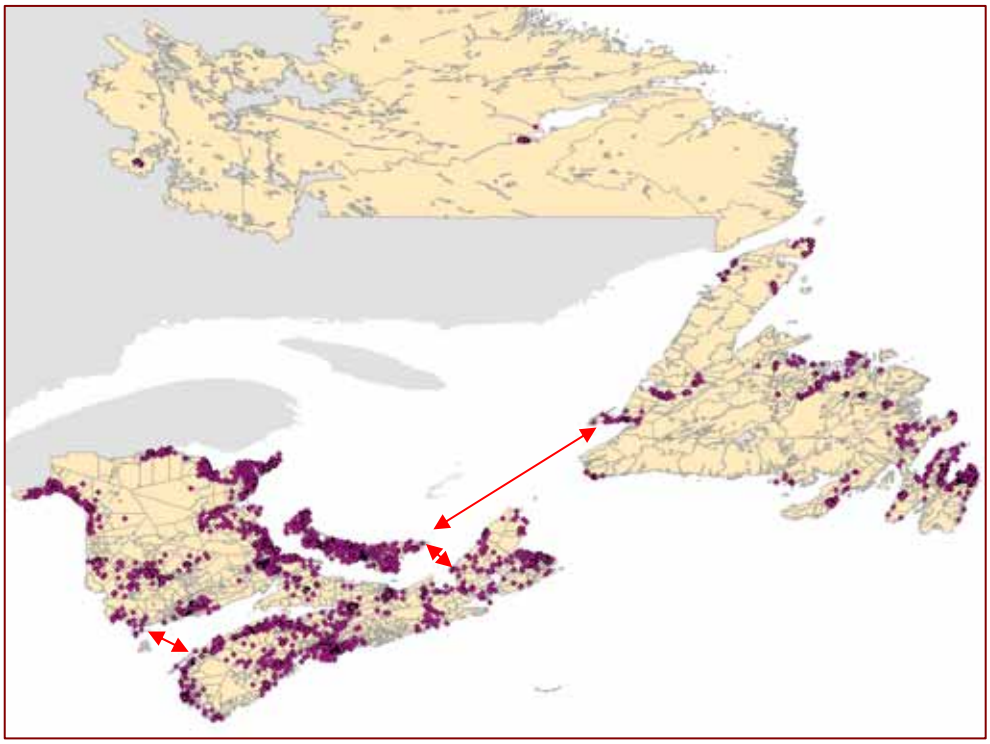


Figure 3.1: Problem of distance measurement for the four Atlantic provinces

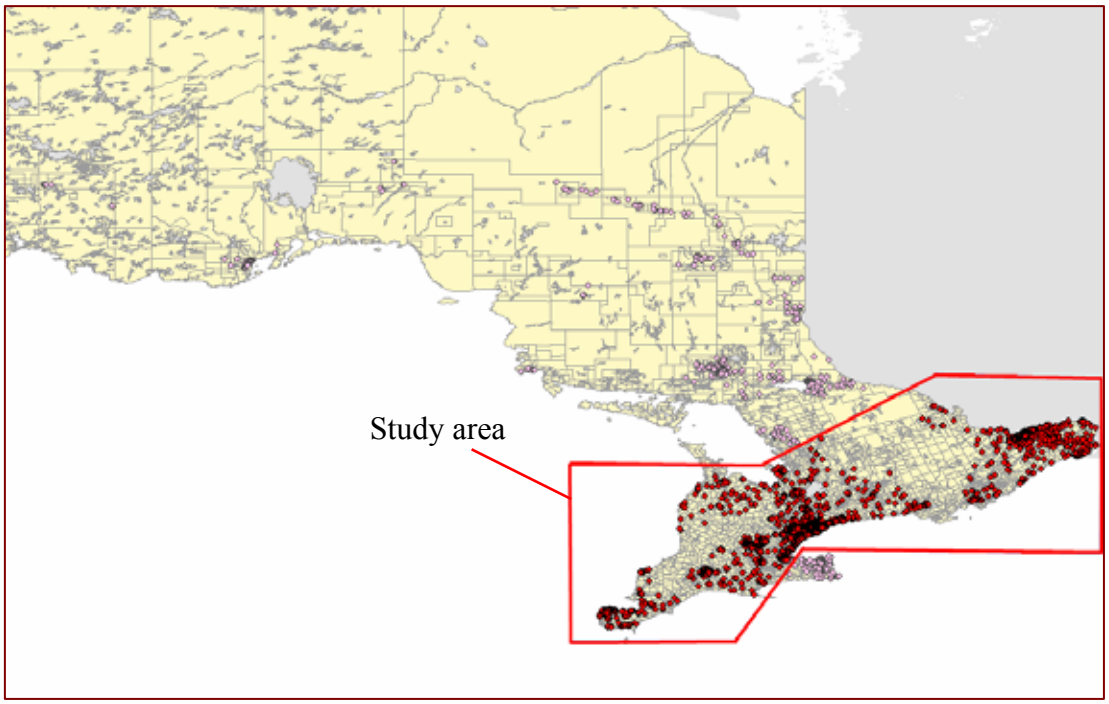


Figure 3.2: Southern Ontario as the study area

3.1.4 Preliminary Data Preparation

Firstly, the Prose Literacy scores for the adults between the age of 16 and 65 surveyed in southern Ontario were extracted. Then, the records with missing or extreme values were removed to ensure the quality of subsequent analysis. Some variables were then 'centred' such that the constant of the regression model predicts the literacy score of a typical adult of age 40, with 12 years of education, and an annual income of $30,000. Details of the data preparation process are summarized in Appendix A, while key points are highlighted where appropriate in the explanation of the base model contained in the next section.

3.1.5 Models

3.1.5.1 Base Model

The base model of this experiment is an ordinary least squares (OLS) regression model which serves as a baseline for comparison. It is represented as follows:

$$\text{Prose} = \beta_0 + \beta_1 \text{ Gender} + \beta_2 \text{ Age} + \beta_3 \text{ AgeSquared} + \beta_4 \text{ YearsEd} + \beta_5 \text{ Income}$$

where:

Prose = the prose score of the subject, ranging from 0 to 500.
Gender = re-centred value of the gender of the subject such that males were set to -0.5 and females to 0.5.
Age = re-centred value of the age of the subject. This experiment includes only adults between the age of 16 and 65. This variable was re-centered such that age 40 was set to 0, 39 to -1, 41 to 1, and so on.
AgeSquared = square of the subject's re-centred age value;
YearsEd = re-centred value of the number of years of education of the subject such that 12 years of education was set to 0, 11 to -1, 13 to 1, and so on.
Income = re-centred value of the imputed personal annual income of the subject such that annual income of $30,000 was set to 0, $31,000 to 1, $29,000 to -1, and so on.

As mentioned earlier, these data come from the IALSS. The model is implemented with the *SPSS*™ statistical package.

3.1.5.2 Local Spatial Regression Models Under Study

The major goal of this experiment is to determine whether the incorporation of community concept into the local spatial regression model can lead to improved prediction power by means of comparing the proposed model with two distance-based local spatial regression models. The local spatial regression models under study are:

(1) the Geographically Weighted Regression (GWR) using exponential distance-decay function as the spatial weighting function by Brunsdon et al. (1998);
(2) the Bayesian Geographically Weighted Regression using a distance-decay parameter smoothing relationship (BGWR-Distance) by LeSage (2004); and
(3) the Bayesian Geographically Weighted Regression using a community parameter smoothing relationship (BGWR-Community) proposed by the author.

The GWR model is selected as it is one of the most popular spatially weighted regression models that is distance-based. The BGWR-Distance model is selected as it is also a distance-based model. More importantly, it serves as a reference when comparing the results of the author's BGWR-Community model with the GWR model because it is structured in a way that offer a transition between the GWR and the BGWR-Community models. As discussed in Section 2.4, the BGWR models are a Bayesian approach of the GWR model that introduces various parameter-smoothing relationships to extend the GWR model. The major difference between the BGWR-Distance model and the GWR model is that it incorporates a normalized distance-based weight matrix in the parameter smoothing relationship. Since the 'community' concept in the BGWR-Community model is represented by a weight matrix that is comprised of a similar distance-based weight matrix and another component (to be discussed in Section

3.1.5.2(d) below), the BGWR-Distance model serves as a good reference for the GWR and BGWR-Community models.

In this experiment, the implementation of these models is based on LeSage's (2005) Spatial Econometric Toolbox which is a host of spatial econometric estimation methods implemented with *Matlab*™. In the following sections, the spatial data requirements for these models are discussed followed by detailed discussions about each model.

(a) Spatial Data Requirements and Distance Matrix

In order to generate the distance-based weight matrix mentioned above, a *n* by *n* distance matrix that captures the distance between any two sample points is required. It is used in all of the above models to determine the weight applied to each sample point during estimation. Therefore, it is pre-computed and imported into each model to avoid redundant computation. The spatial data required to generate the distance matrix includes the DA (Dissemination Area) file (which contains all the DA polygon data) of the 2001 Canada Census Spatial File (in *ArcInfo*™ [4] *.e00* format) and the Postal Code Conversion File (PCCF) from Statistics Canada. The PCCF is a text file that provides a correspondence between the postal code and Statistics Canada's standard geographical areas (such as DA) for which census data and other statistics are produced (Statistics Canada, 2003b).

[4] *ArcInfo* is a registered product name of geographical information system software produced by ESRI, Redlands, CA, USA.

To generate the distance matrix, each sample point has to be tied to a location using coordinates. In order to protect the identity of the surveyed samples, the only location information given in the IALSS dataset is the DAs where the subjects live. Hence, the coordinates of the centroid of the corresponding DA polygon of a given sample point are used. For DAs that have more than one sample point, an alternative method is applied to avoid using the same coordinates. This method randomly related each sample point to a postal code within the corresponding DA. The geographical coordinates (latitude and longitude) of the assigned postal code are then used as the location of a given sample point. Please see Appendix B for details of this procedure.

After relating each sample point to a pair of coordinates, corresponding distances between each pair of sample points are computed using a function adapted from *vdist( )*, a *Matlab* function implemented by Michael Kleder (2005). This function calculates the Great Circle distance of two points using their latitude/longitude coordinates. The distance matrix is then saved as a *.mat* file (a *Matlab* output file) as an input of the local spatial regression models.

(b) Geographically Weighted Regression Model (GWR)

As mentioned in Chapter 2, a GWR model produces locally linear regression estimates for every prediction point, using a spatial weighting function of a certain bandwidth to define the weights of the observations around the prediction point in the regression model. The spatial weighting function used in

this experiment is an exponential function as shown in (2.7) while the bandwidth is obtained through a calibration process using the cross-validation approach as described in Section 2.3.3.

$$w_{ij} = \exp[- \tfrac{1}{2}(d_{ij} / b)^2] \qquad (2.7)$$

where:

w_{ij} = weight
d_{ij} = distance between an observation and the prediction point
b = bandwidth

Figures 3.3a and 3.3b show the differences between the OLS and GWR models in terms of weights, where the 'star' is the prediction point and the size of the orange dots is proportional to the weight being assigned to all other observations. Figure 3.3a shows that the weights being assigned to all other observations are the same in the OLS model while the weights being assigned in the GWR model decrease as the distance between the prediction point and other observations increases (Figure 3.3b).

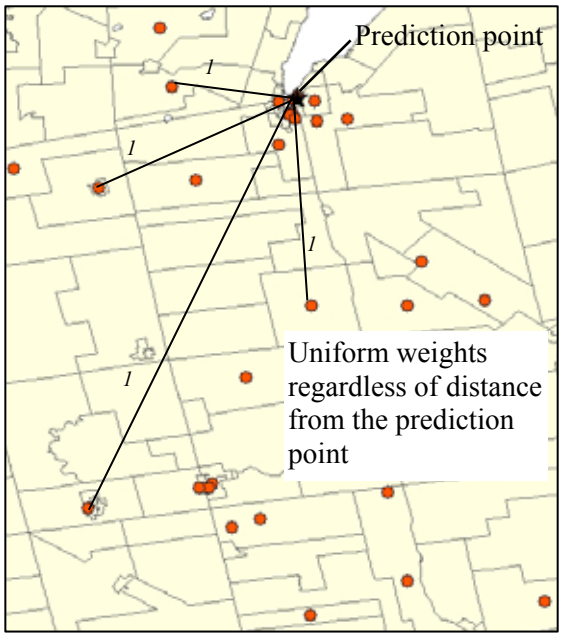


Figure 3.3a: Weights in OLS model

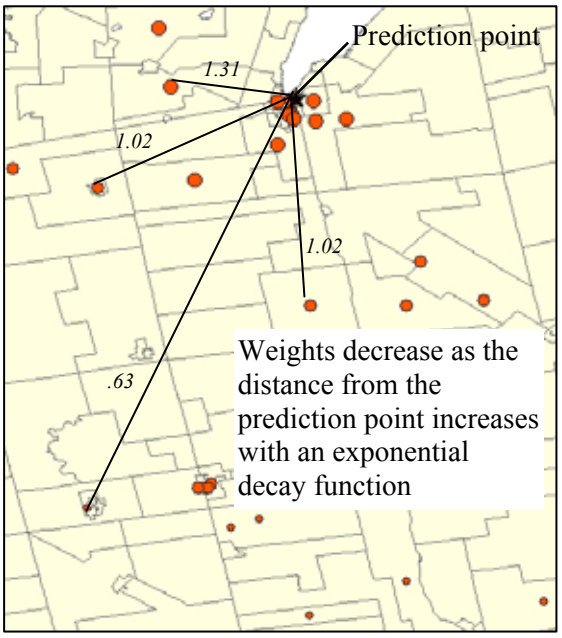


Figure 3.3b: Weights in GWR model

The implementation of the GWR model involves the following modifications of the codes from LeSage's Spatial Econometric Toolbox to enable:

- import of the pre-calculated distance matrix into the model; and
- cross-validation, i.e., during the estimation of every point *i* in the dataset, uses all observations in the dataset except point *i* itself.

(c) BGWR using a Distance-decay Parameter Smoothing Relationship (BGWR-Distance)

The BGWR-Distance model extends the GWR model with a distance-decay parameter smoothing relationship based on the assumption that parameter estimates of the local regression models of observations located close together should be more similar than those farther away. Based on this assumption, the BGWR-Distance model includes a distance-decay parameter smoothing strategy that smoothes out (or reduces) the impact of any anomaly or outlier during the estimation process.

The function for the distance-decay parameter smoothing relationship extends β_i in (2.4) and is expressed as:

$$\beta_i = (w_{i1} \otimes I_k \ldots w_{in} \otimes I_k) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \mu_i \qquad (2.10)$$

where w_{ij} (such that i = 1 to n) represents the normalized distance-decay-based weights such that the sum of the row vectors ($w_{i1}\ldots w_{in}$) are 1 while w_{ii} = 0. The function used to compute the weights w_{ij} of the parameter smoothing relationship is

the same distance-decay function in (2.7) but the weights obtained are normalized. In order to optimize the model, a similar calibration process as discussed in Section 2.3.3 is used to obtain the bandwidth for this distance-decay function.

Figures 3.4a and 3.4b illustrate the difference between the BGWR-Distance model and the GWR model in terms of parameter smoothing relationship. Figure 3.4a shows that, if there were a parameter smoothing relationship for the GWR model, all the values in the row vectors (w_{i1}….w_{in}) in (2.10) would be 0 but $w_{ii} = 1$. Hence, only the prediction point itself would be used for parameter smoothing while the other observations were discarded. For the BGWR-Distance model, its parameter smoothing relationship is a distance-decay function, the weights assigned to the parameters of the other observations decrease as the distance from the prediction point (i.e., the 'star') increases (Figure 3.4b).

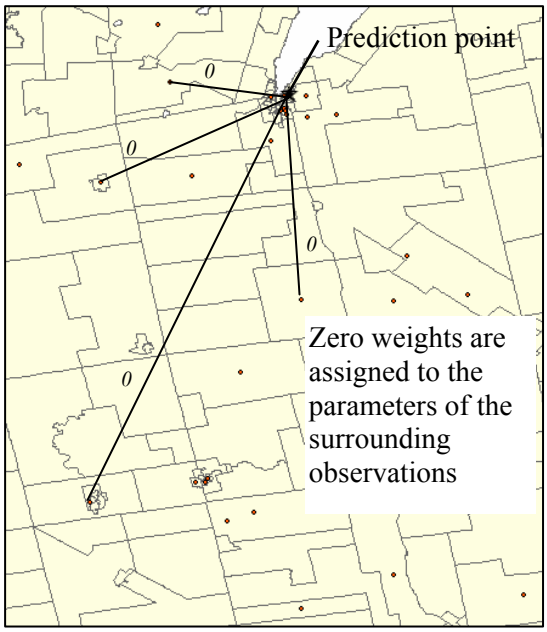


Figure 3.4a: Weights for distance parameter smoothing relationship in GWR model

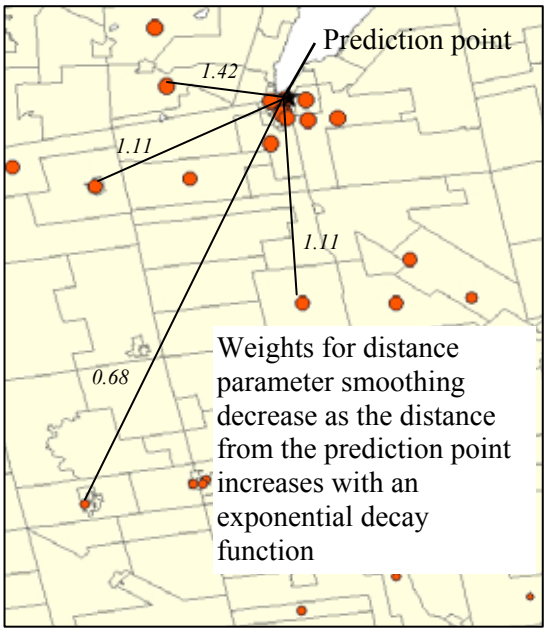


Figure 3.4b: Weights for distance parameter smoothing relationship in BGWR-Distance model

During the implementation of this model, modifications are made to LeSage's Spatial Econometric Toolbox to enable:

- import of the pre-calculated distance matrix into the model;
- cross-validation; and
- 550 passes are used in the Gibbs sampling process as discussed in Section 2.4.2 where results of the first 50 passes are to be discarded.

Additionally, the following default set up is retained:

- the hyperparameter r (as discussed in Section 2.4.1) is set to 4, meaning that the restriction of constant variance is not imposed; and
- the diffuse scale prior δ (as discussed in Section 2.4.1) is set to 1 to impose the influence of the smoothing relationship on the regression coefficient estimation.

(d) BGWR using a Community Parameter Smoothing Relationship (BGWR-Community)

For the purpose of incorporating the community concept in the BGWR model, it is necessary to develop an operational definition of 'community' so that the concept can be expressed as measurable variables.

Operational Definition of Community

The objective of incorporating the concept of community into a BGWR model is to account for the community effects that observations from the same community have higher correlation than those from other communities. This is to be encapsulated in the estimation process so that observations from the same

community of the prediction point bear a relatively higher weight than those from other communities. Hence, the operational definition of community has to be able to tell how likely it is that two observations are in fact coming from the same community. Based on the discussion in Section 2.1.1.4, two decisions have to be made during the operationalization process of the community concept: (1) how to demarcate the area for measurement, and (2) what attributes are to be included in the measurement.

However, it is difficult to demarcate the areas for measurement that fit the boundaries of communities. In fact, Dietz (2002) observed that space delineations in most research are constrained by the limitations of the available datasets. Since DA is the smallest spatial unit of Census data that is accessible through Statistics Canada, it is used as the basic spatial unit of measurement in this research.

Although the discussion in Section 2.1.1.4 reveals that there is no universally accepted methods to operationalize community, physical and social characteristics of the community are the common denominators of Glaster's (2001) "bundle of spatially-based attributes" and Lupton's (2003) guidelines. Since Glaster's (2001) bundle of spatially-based attributes captures both the physical and social characteristics of a community, it is an appropriate candidate from which the operational variables can be derived.

The attributes described by Galster (2001) can be broadly grouped into location-related attributes and people-related attributes as summarized in Table 3.1. Although these attributes provide a good framework, reliable quantitative data are not always available. For location-related attributes such as the quality of public administration is indeed quite difficult to quantify. As the objective of measuring these attributes is to obtain an indicator to tell how likely two observations are coming from the same community, this experiment uses geographical distance as a proxy for location-related attributes. This is based on two assumptions. Firstly, observations from the same community tend to be in proximity. Secondly, observations in close vicinity are likely to share similar location-related characteristics. For example, samples in vicinity are more likely to fall within the same public administrative district, hence possessing the same quality of public administration. This experiment thus uses the distance-decay value of the geographical distance between two observations as a proxy to indicate, on the location-related aspect, how likely it is for two observations to be coming from the same community. Details of how this proxy works are described in the next section.

Table 3.1: Bundle of spatially-based attributes (After Glaster [2001, p. 2112])

Location-related attributes	**People-related attributes**
• **Structural characteristics of the residential and non-residential buildings:** type, scale, material, design, state of repair, etc. • **Infrastructural characteristics:** road, sidewalk, utility services, etc. • **Environmental characteristics:** degree of land, air, water and noise pollution, topographical features, views, etc. • **Proximity characteristics:** access to major destinations of employment, entertainment, shopping, etc. • **Tax/public service package characteristics:** the quality of safety forces, public schools, parks and recreation, public administration etc., in relation to the local taxes assessed.	• **Demographic characteristics of resident population:** age distribution, family composition, racial, ethnic, and religious types, etc. • **Class status characteristics of the resident population:** income, occupation and education composition • **Political characteristics:** the degree to which local political networks are mobilised • **Social-interactive characteristics:** local friend and kin networks, degree of inter-household familiarity, resident's perceived commonality, participation in locally based voluntary associations • **Sentimental characteristics:** residents' sense of identification with place, historical significance of buildings or district. etc.

Among the people-related attributes, relevant data for the last three characteristics are not available. For demographic and class status characteristics, the following variables are given in the 2001 Canada Census data by DAs and are used in the experiment.

- **Years Education:** The average number of years of education in a DA;
- **Transience:** The percentage of people who had moved in the previous five years;
- **Income:** Average level of family income;
- **Unemployment Rate:** The percentage of people who were unemployed;

- **Percent Social Class 1 and 2:** The percentage of people who were in professional or semi-professional occupations;
- **Percent Social Classes 4, 5, 6:** The percentage of people who were in unskilled labour occupations, or unclassified occupations; and
- **Percent Recent Immigrants:** The percentage of people who had immigrated in the previous five years.

These variables are identical to those applied in the multi-level regression models in Willms' research (Willms, Chan and Tang, 2007; Willms and Tang, 2007). In this experiment, the Mahalanobis distance of these variables is used to indicate, on the people-related aspect, how likely two observations are coming from the same community. The reasons for using Mahalanobis distance are twofold. Firstly, Mahalanobis distance is being used frequently in cluster analysis problems in order to determine similarity among 'clusters' as it takes into consideration the correlations of the 'clusters' and is not dependent on the scale of measurement (Rapkin and Luke, 1993; Mimmack et al., 2001; Hagger-Johnson, 2006). Secondly, using certain type of distance as a single indicator allows the model to use a distance-decay function to create the weight matrix for the parameter smoothing relationship. In this way, no community has to be pre-defined.

In short, the geographical distance and the Mahalanobis distance of the seven selected variables thus constitute the operational definition of community (Figure 3.5). The next section will discuss the details about the implementation of the community concept.

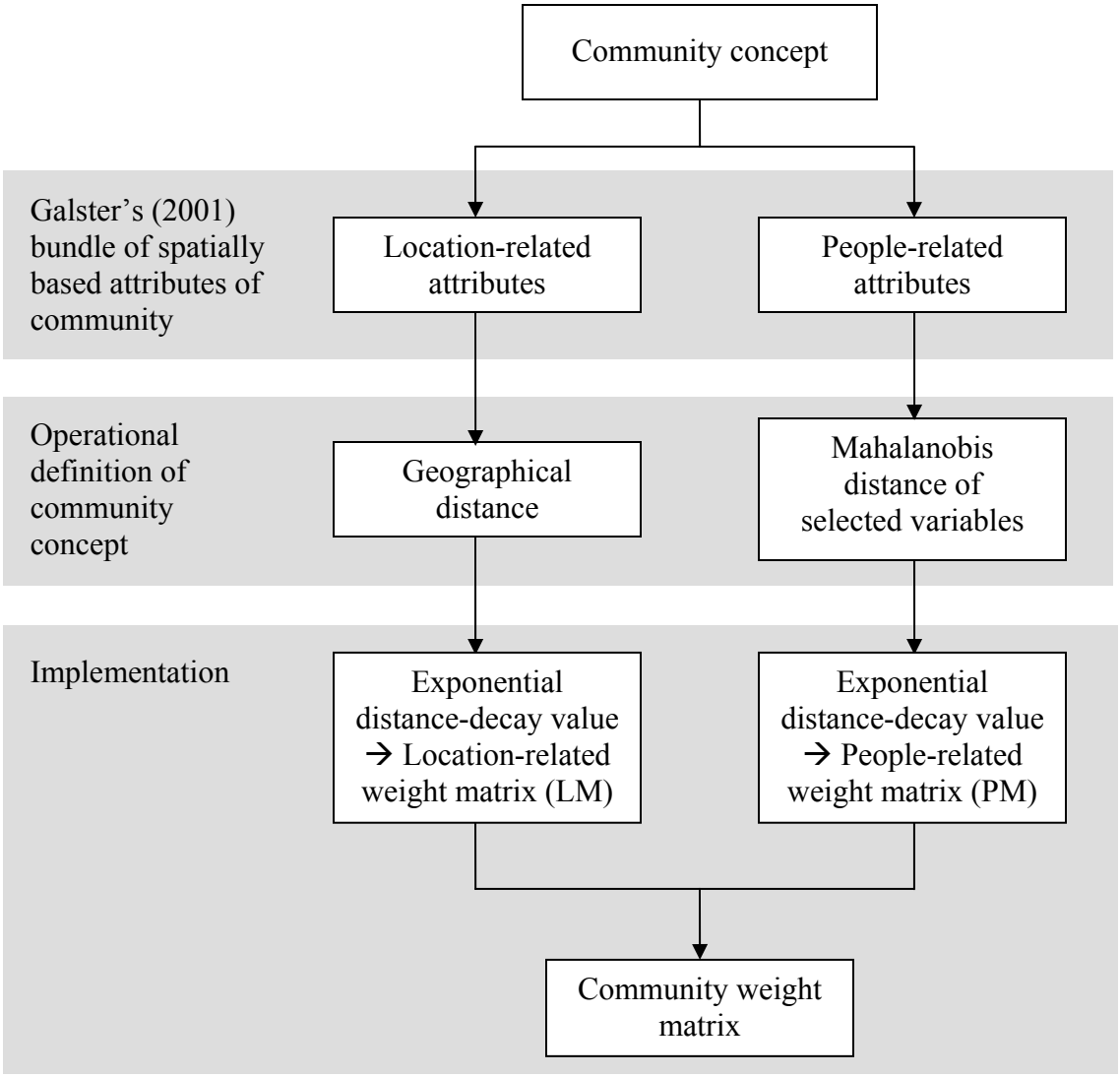


Figure 3.5: Operational definition and implementation of 'community' concept

Implementation of the Community Parameter Smoothing Relationship

Similar to the operational definition of community, the weight matrix is comprised of two parts, namely the location-related weight matrix (LM) and the people-related weight matrix (PM), as shown in Figure 3.5. The LM is an intermediate weight matrix generated by the exponential distance-decay function using the bandwidth obtained during the calibration of the GWR model. The PM and the community weight matrices are generated in three steps:

(1) A Mahalanobis distance matrix of the seven variables among all observations is generated using a *Matlab* function written by the author. Every row vector captures the Mahalanobis distances between a prediction point and the rest of the sample points.

(2) A row vector of weights for each prediction point is calculated by applying the exponential distance-decay function (as described in Section 2.3.3) to the corresponding row vector of Mahalanobis distances using several different bandwidths. The bandwidths are 0.5, 0.75 and 1.0 standard deviation of the values of the corresponding row vector of Mahalanobis distances. Hence, three PMs are generated.

(3) Multiply the LM by each of the three PMs to get three different community weight matrices which are then saved as *.mat* files for later use.

As the community parameter smoothing relationship also takes into account the people-related attributes of DAs in terms of the seven selected variables, the weights that are assigned to the other observations may not simply decrease with distance from the prediction point. As shown in Figure 3.6, some observations closer to the prediction point (the 'star') are weighted less because their people-related attributes are not similar to those of the prediction point.

Weights are determined by distance and by how similar the community is to the target community so it may not necessarily decrease as the distance increases

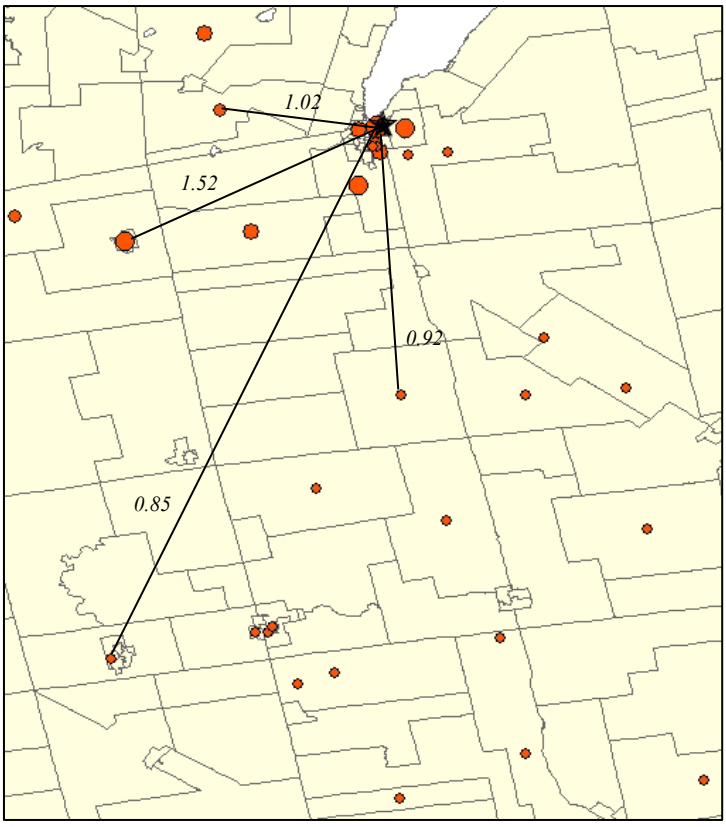


Figure 3.6: Weights for community parameter smoothing relationship in BGWR-Community model

Similar to the previous two models, the implementation of this model is also based on LeSage's Spatial Econometric Toolbox with modified set up to enable:

- import of the pre-calculated normalized community weight matrix into the model;
- cross-validation; and
- 550 passes are used in the Gibbs sampling process as discussed in Section 2.4.2 where results of the first 50 passes are to be discarded.

Additionally, the following default set up is retained:

- the hyperparameter r is set to 4; and
- the diffuse scale prior δ is set to 1.

3.2 Evaluation Methods

As indicated by Gao et al. (2006), spatial (local) regression models usually have more parameters and smaller sample sizes than ordinary linear regression models, so their degrees of freedom are reduced. Hence, even a very high R-squared value obtained by the model does not necessarily indicate that it is a good model. Therefore, instead of using a single indicator to measure and compare the prediction powers of the models, the evaluation methods used in this study compare different aspects of the prediction power of the models in order to give a more comprehensive view of the performances of the models. These include the overall performance in terms of numerical measures of the prediction errors, visual evaluation of the amount and trend of extreme predictions, and performance (in terms of prediction errors) at different error tolerance levels. Hence, in addition to the regression statistics like the R-squared values, this study adopts the following three empirical methods introduced by Gao et al. (2006) to evaluate the results obtained from the local spatial regression models under study.

3.2.1 Numerical Cross-Validation Criteria

The criteria to be included in this method are:

- the mean of squares of prediction errors, $1/n\sum(y_i - \hat{y}_{\neq i})^2$;
- the mean of absolute deviations, $1/n\sum|y_i - \hat{y}_{\neq i}|$; and
- average error rate, $\frac{1}{n}\sum\frac{|y_i - \hat{y}_{\neq i}|}{y_i}$.

where y_i and $y_{\neq i}$ denote the observed and the predicted values at point i. It is called 'cross-validation' criteria as the '$\neq i$' symbol indicates that during the prediction of point i (that is the point of interest), the model uses all sample points except point i itself. As these criteria measure different aspects of the prediction errors, relatively smaller values indicate better performance of the model. Table 3.2 shows a sample output from the numerical cross-validation criteria.

Table 3.2: Sample output from numerical cross-validation criteria
(After Gao et al. [2006, Table 1])

Comparison of models with numerical criteria

Model	Numerical criteria			
	$\frac{1}{n}\sum(y_i - \hat{y}_{\neq i})^2$	$\frac{1}{n}\sum\lvert y_i - \hat{y}_{\neq i}\rvert$	$\frac{1}{n}\sum\frac{\lvert y_i - \hat{y}_{\neq i}\rvert}{y_i}$ (%)	Correl. coef.
Basic model	0.0103	0.0805	10.40	0.836
Spatial dependency model	0.0101	0.0806	10.40	0.835
GWR model	0.0095	0.0768	9.87	0.851
Mixed model	0.0093	0.0754	9.84	0.853

3.2.2 Scatter Plots for Observed and Predicted Values

Figure 3.7 depicts a sample scatter plot between the observed and predicted values. From the distribution of the points in the scatter plot and the 45-degree diagonal (which indicates the perfect prediction), the prediction power of the models can be observed by comparing how spread out or how close these points are to the diagonal. A 97.5% density ellipse (which encloses 97.5% of the points) for each model can also be added to facilitate the comparison. When comparing the shapes of the ellipses, a narrower ellipse generally indicates a better model. The scatter plots also provide a good picture of the extent and pattern of the poorly predicted samples by showing how far from the diagonal and where they are.

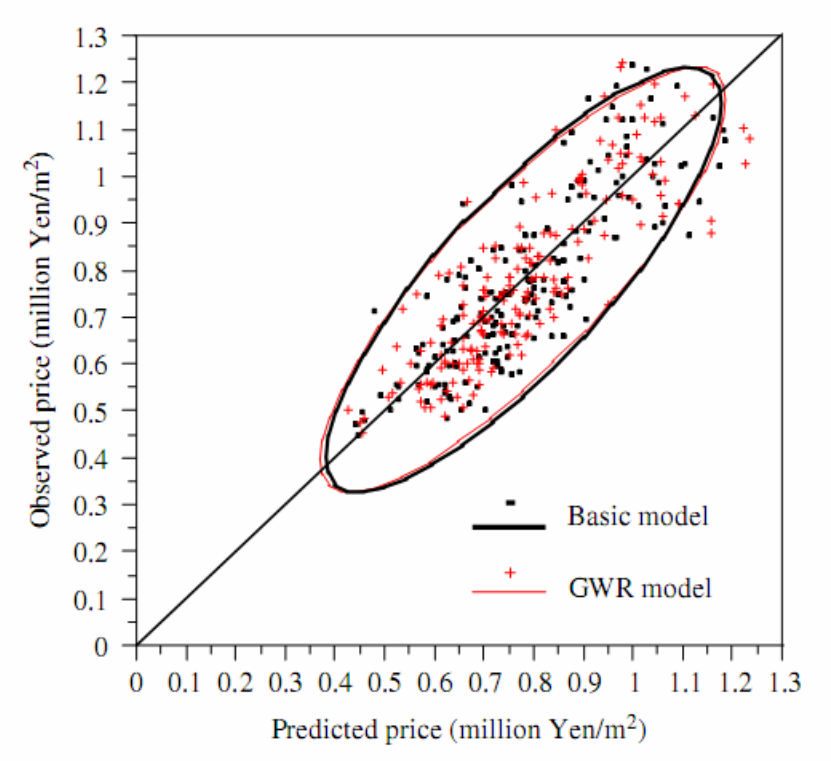


Figure 3.7: Scatter plot for observed and predicted values
(After Gao et al. [2006, Figure 2])

The drawback of this method is that visual comparison becomes overwhelming when more than two sets of results are showing in the same chart. Hence, comparisons across several models have to be done in pairs.

3.2.3 Prediction Rate Curve

A prediction rate curve is a graph showing the rate of samples against the prediction errors (Figure 3.8). In the vertical axis, the values are the accumulated rate of samples (i.e., 1/n, 2/n, …, (n-1)/n, 1; where n = total number of samples). In the horizontal axis, the values are the sorted prediction errors ($|y_i - \hat{y}_{\neq i}|$) in ascending order. In other words, data points are sorted according to how 'well-

predicted' they are. The area formed by the prediction rate curve and the vertical axis indicates the aggregation of the prediction errors. The smaller the area, the better the prediction power of a given model. As depicted in Figure 3.8a, one can say that Model 1 is a better model than Model 2 as the area formed by the prediction rate curve of Model 1 is smaller than that of Model 2. In Figure 3.8b, the areas formed by the two curves with the vertical axis are more or less the same indicating that the overall prediction powers of the two models are close. However, when a tolerance level e_1 is set, 70% of the predicted values (or a prediction rate of 70%) from Model 1 are below the tolerance level while only 60% from Model 2 are. For a higher tolerance level (e_2), Model 2 has better performance.

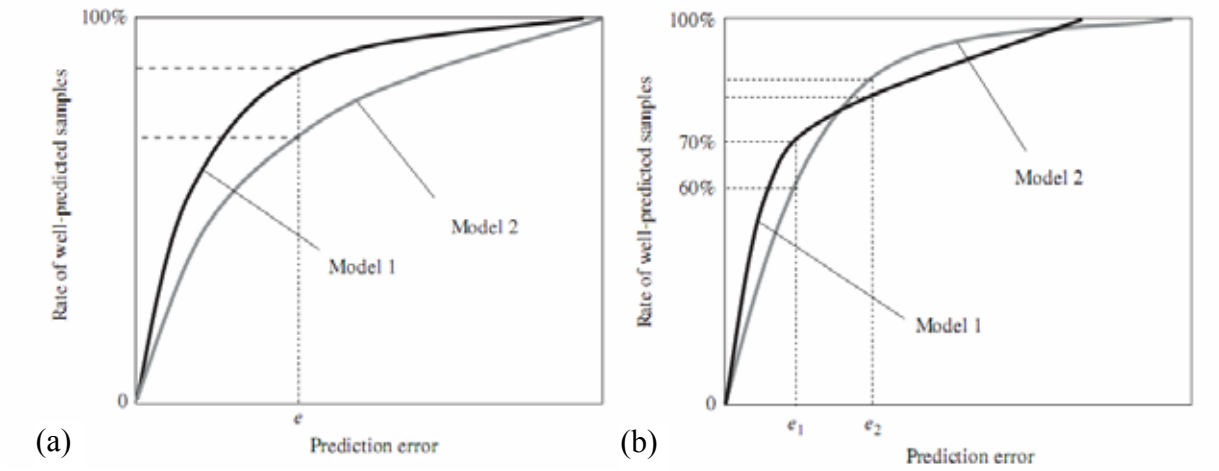


Figure 3.8: Sample prediction rate curves
(After Gao et al. [2006, Figure 3])

3.2.4 Summing Up

While the numerical cross-validation criteria give an indication on the overall performance of the models, the prediction rate curves give more details about the performance of the models at various tolerance levels. The scatter plots

are good for visual examination of the overall performance of the models as well as identification of extreme predictions.

3.3 Summary

This chapter begins with a detailed discussion on the experiment design, starting with the considerations for the selection of dataset (International Adult Literacy and Skill Survey), study area (southern Ontario) and base model. The base model of this experiment is an ordinary least square (OLS) regression model which serves as a baseline for comparison with the three local spatial regression models under study: (1) GWR, (2) BGWR-Distance, and (3) BGWR-Community.

Before getting into the details of the three models, the source and preparation of the required spatial data and the distance matrix are described. The three local spatial regression models are implemented by modifying a set of base programs from LeSage's Spatial Econometric Toolbox, a host of spatial econometric estimation methods implemented with *Matlab*. By describing the characteristics of the spatial weighting function of GWR and the parameter smoothing relationship of the three BGWR models, the differences between the three models are highlighted. Modifications to LeSage's base program and the model settings are also described. As the BGWR-Community model is a new model introduced by the author, the process of developing 'community' from a concept into an operational definition is explained in details, including the additional data requirements.

The latter part of this chapter described the evaluation criteria to be used in this study, namely (1) regression statistics; (2) numerical cross-validation criteria; (3) scatter plots of the observed and predicted values; and (4) prediction rate curve.

4.0 SUMMARY AND ANALYSIS OF RESULTS

In this chapter the results of the experiment are presented in two parts. The first part presents the results in the form of regression statistics to give a general idea about the relative performance of the models. Then, the results and findings of each evaluation method are presented, followed by the concluding remarks.

4.1 Regression Statistics

Table 4.1 below summarizes the beta values and estimates of R-squared of the models under study. R-squared is an indicator of how well a model fits the data. It is the proportion of the variance in the data that can be explained by a regression (Warner, 2008). The higher the R-squared value, the better the model. Table 4.1 shows that all of the selected local spatial regression models have higher R-squared values over the OLS base model. The BGWR-Distance model shows the least improvement (increased by 6.90%) while the BGWR-Community model shows the greatest improvement (increased by 10.63%). It is also noted that the R-squared value of the BGWR-Distance model is smaller than that of the GWR model, suggesting that using the BGWR approach does not guarantee a better R-squared result.

Table 4.1: Beta values and R-squared of the models under study

Parameters	Beta*			
	OLS	**GWR**	**BGWR-Distance**	**BGWR - Community**
Constant	257.31	255.42	255.62	257.19
GENDER	12.52	11.42	9.66	10.62
AGE	-0.84	-0.88	-0.88	-0.89
AGE_SQ	0.00	0.00	0.00	0.00
YRS_ED	6.84	7.05	7.32	7.22
IMP_INC	0.46	0.47	0.47	0.46
R-squared	0.4139	0.4453	0.4425	**0.4579**
Improvement of R-squared over base model (OLS)	**-**	**7.59%**	**6.90%**	**10.63%**

Note: *The beta values of the local spatial regression models (i.e., GWR, BGWR-Distance and BGWR-Community models) are the means of the beta values obtained from the local regression equations.

4.2 Numerical Cross-Validation Criteria

Numerical cross-validation criteria give an indication of the overall performance of different models in terms of their prediction errors. The better model is the one with the lowest scores in all three criteria. Table 4.2 shows that the BGWR-Community model scores the lowest for all three criteria. While the scores of the BGWR-Distance and GWR models are smaller than those of the OLS model, the differences between the two local spatial regression models are very small.

Table 4.2: Comparison of models with numerical cross-validation criteria

Model	Numerical cross-validation criteria		
	$1/n\sum(y_i-\hat{y}_{\neq i})^2$	$1/n\sum\lvert y_i-\hat{y}_{\neq i}\rvert$	$\frac{1}{n}\sum\frac{\lvert y_i-\hat{y}_{\neq i}\rvert}{y_i}$
OLS	2001.88	35.62	15.74
GWR	1894.56	34.52	13.54
BGWR-Distance	1904.00	34.39	13.70
BGWR-Community	**1851.34**	**33.53**	**13.33**

To determine whether the differences in prediction errors (or residuals) among these models are statistically significant, F-tests for each pair of models were carried out and the results are summarized in Table 4.3.

Table 4.3: Results of F-test of prediction errors among different models

	***p*-value**			
	OLS	**GWR**	**BGWR-Distance**	**BGWR-Community**
OLS	-	0.0523	0.1254	**0.0008**
GWR	-	-	0.6901	0.1499
BGWR-Distance	-	-	-	0.0678

The alpha value for the F-test was set at 0.05. Therefore, if the *p*-value of the F-test is less than 0.05, the difference in the prediction errors of the two models under testing is considered to be statistically significant. The results indicate that the difference between the base model (OLS) and the BGWR-Community model is statistically significant while other differences are not. This suggests that the improvement of the BGWR-Community model in the reduction of prediction errors is not random. It is also found that although the figures of the three numerical criteria of the BGWR-Distance and GWR models are very close, the *p*-value of the GWR model (i.e., 0.0523) is much smaller than the BGWR-Distance model and close to the pre-defined alpha value of 0.05.

From the above discussion, the following conclusions are drawn:

- the relatively better results of the GWR model over the BGWR-Distance model confirms the suggestion that using the BGWR approach does not necessarily improve the performance; and

- the improvement made by the BGWR-Community model over the GWR and BGWR-Distance models suggests that incorporating the community concept into the BGWR model can improve the overall performance of the model by reducing the prediction errors.

4.3 Scatter Plots

For ease of comparison, only two models are compared at a time. The order of comparisons is as follows:

(1) The three local spatial regression models compare with OLS one by one;

(2) The two BGWR models compare with GWR model one by one; and

(3) The two BGWR models compare with each other.

In each scatter plot, the X-axis represents the predicted values (*yhat*) while the Y-axis represents the observed value (*y*). A 45-degree diagonal is included to give reference to perfect prediction. Each scatter plot contains two point sets. One in blue at the back acts as the baseline and one in red in the front for comparison. There are also two density ellipses, one in blue and one in red, covering 97.5% of the points of the corresponding point set.

4.3.1 OLS Model as the Baseline

Figures 4.1 to 4.3 show the OLS model in blue and the others for comparison in red. It is noticed that the long axis of all the ellipses are forming angles with the diagonal, while that of the blue ellipse (for the OLS model) forms

a larger angle than the other models. This indicates that the OLS model under-predicts when y-values are larger than the mean (roughly 250) but over-predicts when y-values are smaller than the mean. Smaller angles between the long axes of the red ellipses (Figures 4.1 to Figure 4.3 refer) and the pink diagonals indicate that the local spatial regression models have less degree of over- and under-prediction problems. It is evident that the red points at the lower score regions are closer to the diagonal than the corresponding blue points (Figures 4.1 to Figure 4.3 refer). In general, the red points are tighter and located closer to the diagonal than the blue points meaning that the local spatial regression models fit better than the OLS model.

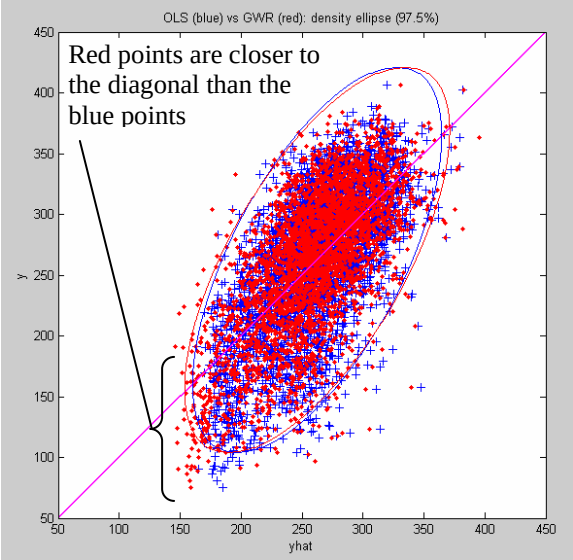


Figure 4.1: Scatter plot of OLS vs. GWR

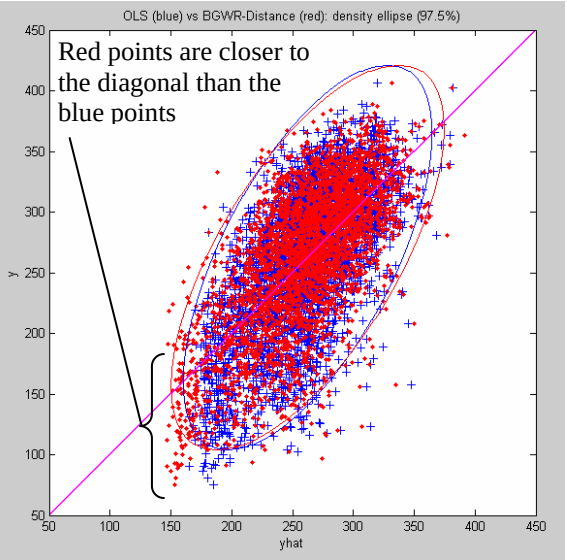


Figure 4.2: Scatter plot of OLS vs. BGWR-Distance

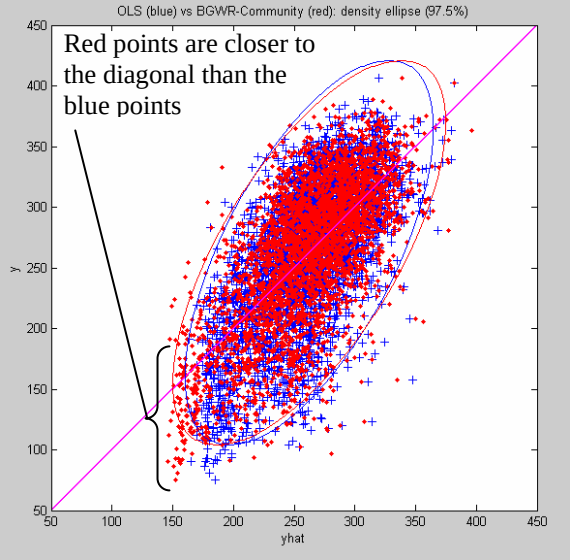


Figure 4.3: Scatter plot of OLS vs. BGWR-Community

4.3.2 GWR Model as the Baseline

When the GWR model is acting as the baseline to compare with the two BGWR models (see Figures 4.4 and 4.5), it is found that the long axis of the GWR ellipse (the blue one) forms a slightly larger angle with the diagonal than the two BGWR models although the differences between the GWR and the two BGWR models are much less than that between the OLS and the others. Visually, the distribution of the red points and the blue points does not show much difference. The number of extreme prediction points is also very close.

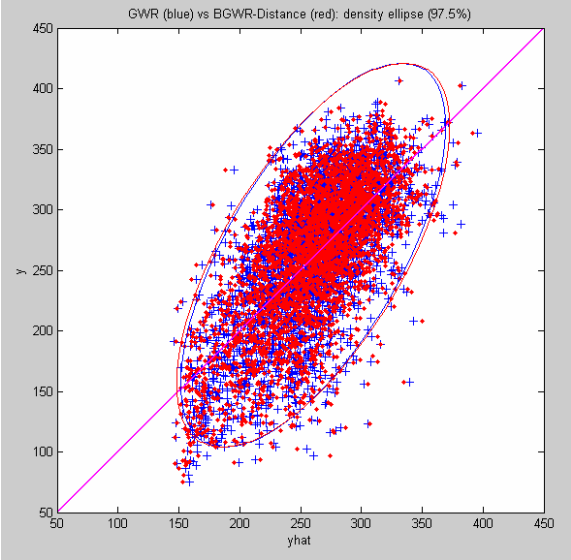


Figure 4.4: Scatter plot of GWR vs. BGWR-Distance

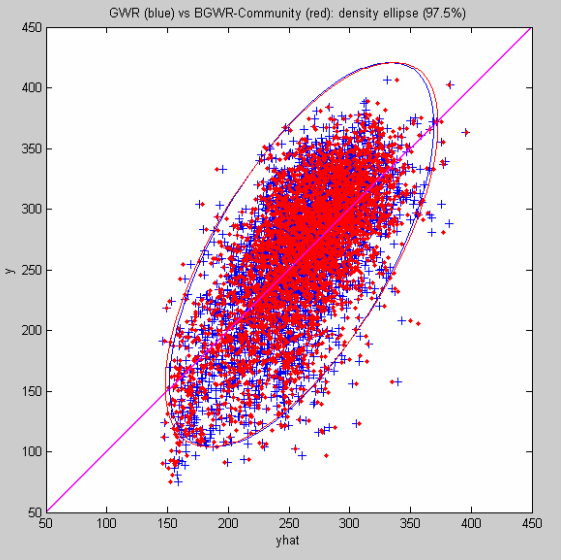


Figure 4.5: Scatter plot of GWR vs. BGWR-Community

4.3.3 Comparison of the Two BGWR Models

When comparing the density ellipses of the two BGWR models in Figure 4.6, it is found that they are almost identical, although the density ellipse for the BGWR-Community model is slightly slimmer than the BGWR-Distance model. The distribution of the red and blue points is very similar and the number of extreme prediction points is also very close.

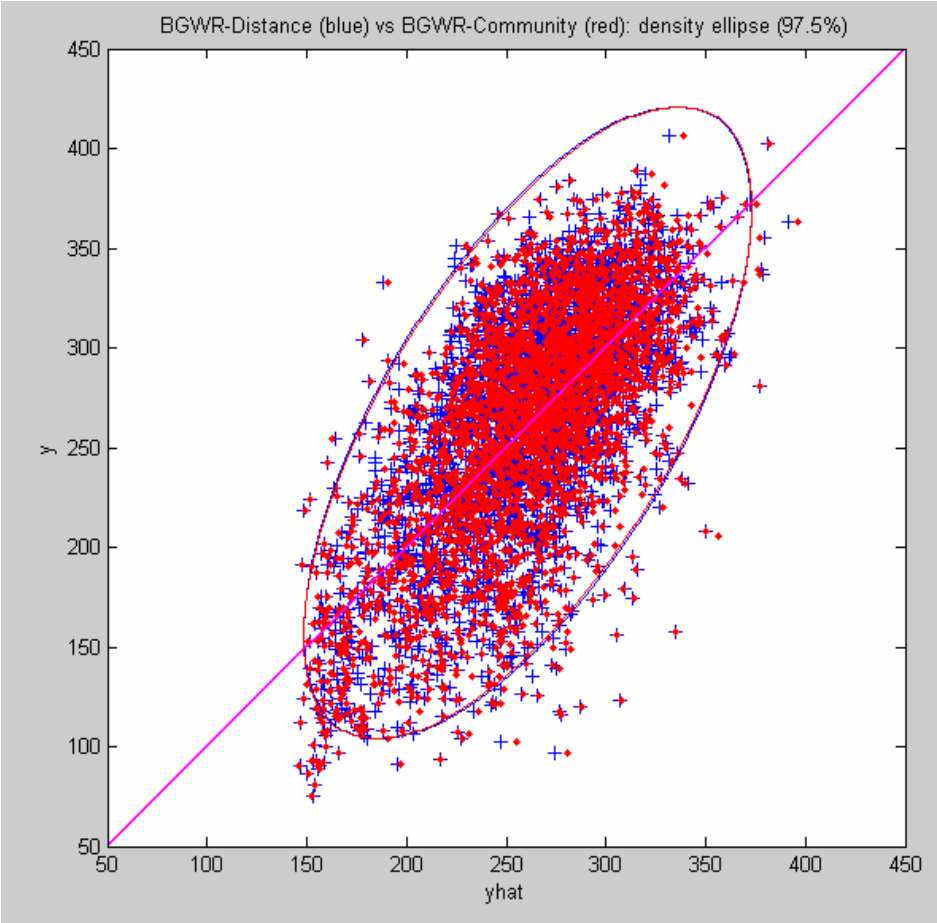


Figure 4.6: Scatter plot of BGWR-Distance vs. BGWR-Community

4.3.4 Summing up

Major improvements of the local spatial regression models over the OLS model are evident from the scatter plots with less over-prediction and under-prediction problems. The two BGWR models also have minor improvement on this aspect over the GWR model. The two BGWR models have little differences in terms of the shape of the ellipse and the distribution of the points. Visually, there is no apparent improvement of the BGWR-Community model over the GWR and the BGWR-Distance models in this evaluation.

4.4 Prediction Rate Curve

The prediction rate curves of the models under study are shown in Figure 4.7. Although the curves are quite close to each other, it is obvious that, in the range of prediction errors 5 to 50, the curve of the BGWR-Community (red line) models is at the top-left side of other curves. This means that the area formed by the curve of this model and the vertical axis is smaller than other models; hence, the BGWR-Community model has better prediction power in this range.

In order to get a closer look at these curves, the prediction rate curve of the OLS model is used as the baseline. The difference between the prediction rate curve of a local spatial regression model and that of the OLS model is regarded as the prediction rate improvement. The prediction rate improvement of the three local spatial regression models are plotted against the prediction error and shown in Figure 4.8. It is found that there is a large gap between the BGWR-Community model and the GWR model between the prediction errors of 5 and 50 (which corresponds to 12% to 76% prediction rates in Figure 4.7). Then, the gap starts to narrow. Around prediction error of 80 (or 94% prediction rate in Figure 4.7), the two curves cross. That means the BGWR-Community model has better prediction power than the GWR model up to the 94% prediction rate (or prediction error of 80).

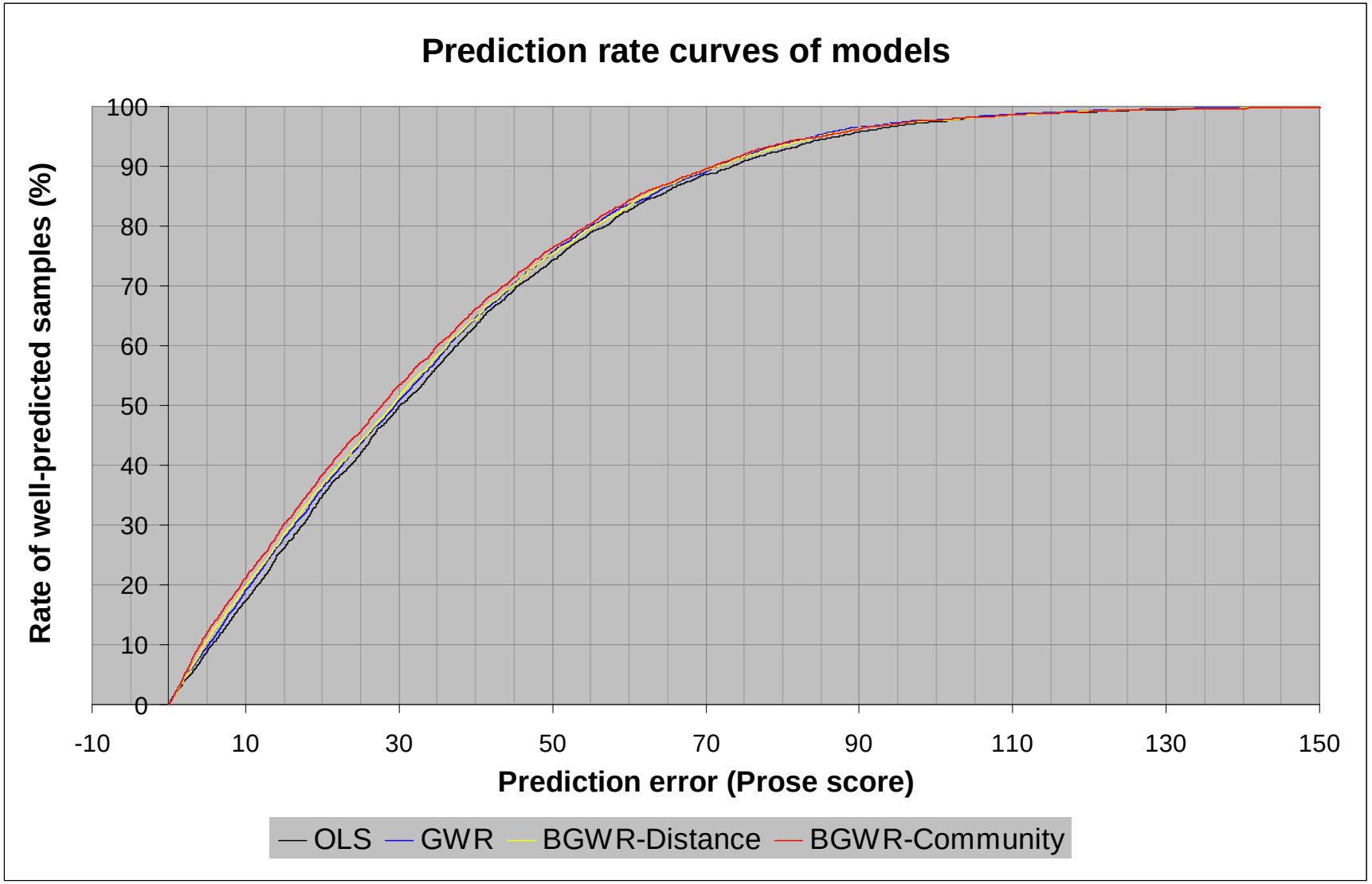


Figure 4.7: Comparing the prediction curves of the models

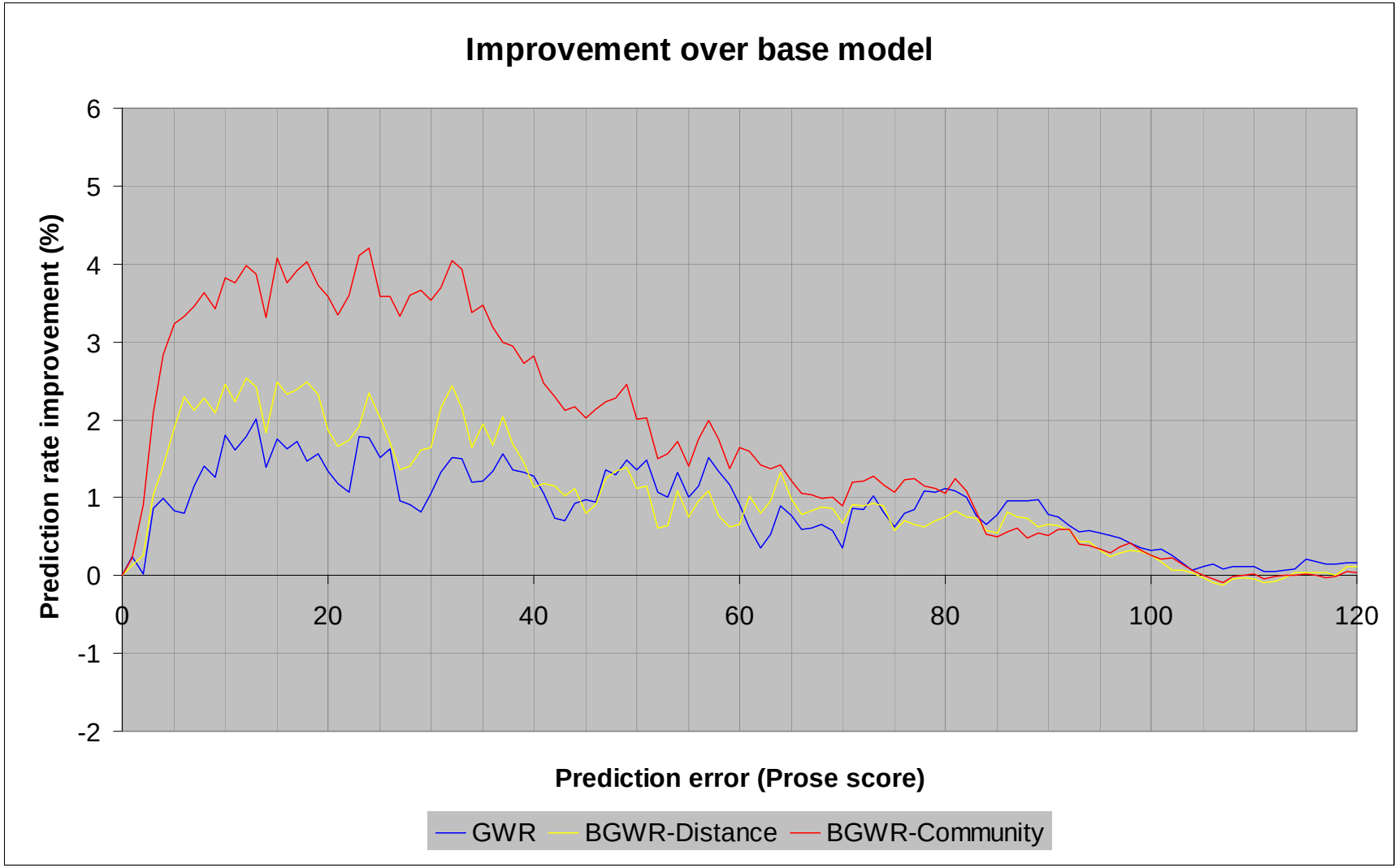


Figure 4.8: Prediction improvements of the local spatial regression models over the base model (OLS model)

A very similar pattern is observed between the BGWR-Community and the BGWR-Distance models although the gap between them is relatively smaller (Figure 4.8). When comparing the BGWR-Distance model with the GWR model, the former has greater improvement over the latter up to the prediction error of 40 (or 65% prediction rate) and then the two lines are very close and cross each other from time to time. In short, the BGWR-Distance model has better prediction power than the GWR model only up to the 65% prediction rate (or prediction error of 40).

From the above discussion, it is noted that the BGWR-Community model maintains a higher prediction rate up to a very high tolerance level while the BGWR-Distance model performs a little bit better than the GWR model at low tolerance level.

4.5 Discussion

From the results of the numeric cross-validation criteria and prediction rate curves and in comparison with the two other models, the BGWR-Community model has a better overall prediction power and maintains a higher prediction rate up to a very high tolerance level. The results in the scatter plots show that the BGWR-Community model does not generate more extreme prediction points than the other models. These results lead to the conclusion that by incorporating the ‘community’ concept into the BGWR model, the prediction power of the model improves over the purely distance-based local spatial regression models.

Nevertheless, from the results in Table 4.1, it is also noted that the improvement (in terms of R-squared values) made by the BGWR-Community model over the GWR model (which is about 3%) is not as large as that made by the GWR model over the OLS model (which is 7.6%). This is in line with the expectation that the impact of community effect is localized and lesser than the impact of Tobler's First Law of Geography (Section 2.3.1 refers), or the First Law of Geography, to certain extent, has accounted for the community effect.

In order to visualize the local improvement brought by the BGWR-Community model, a series of maps have been prepared using a spatial interpolation method called Ordinary Kriging (Fotheringham et al., 2000) to generate the gradient surfaces of the prediction improvement of the local spatial regression models over the OLS model as shown in Figures 4.9 to 4.11.

The prediction improvement of a prediction point *i* is the difference between the prediction errors of *i* obtained by the OLS model and that of the local spatial regression models divided by the observed value of *i* and multiplied by 100 to express as a percentage (4.1). A positive value indicates that the local spatial regression model has a smaller prediction error than the OLS model and vice versa.

$$\frac{\text{prediction error of OLS} - \text{prediction error of local spatial regression model}}{\text{observed value}} \times 100 \qquad (4.1)$$

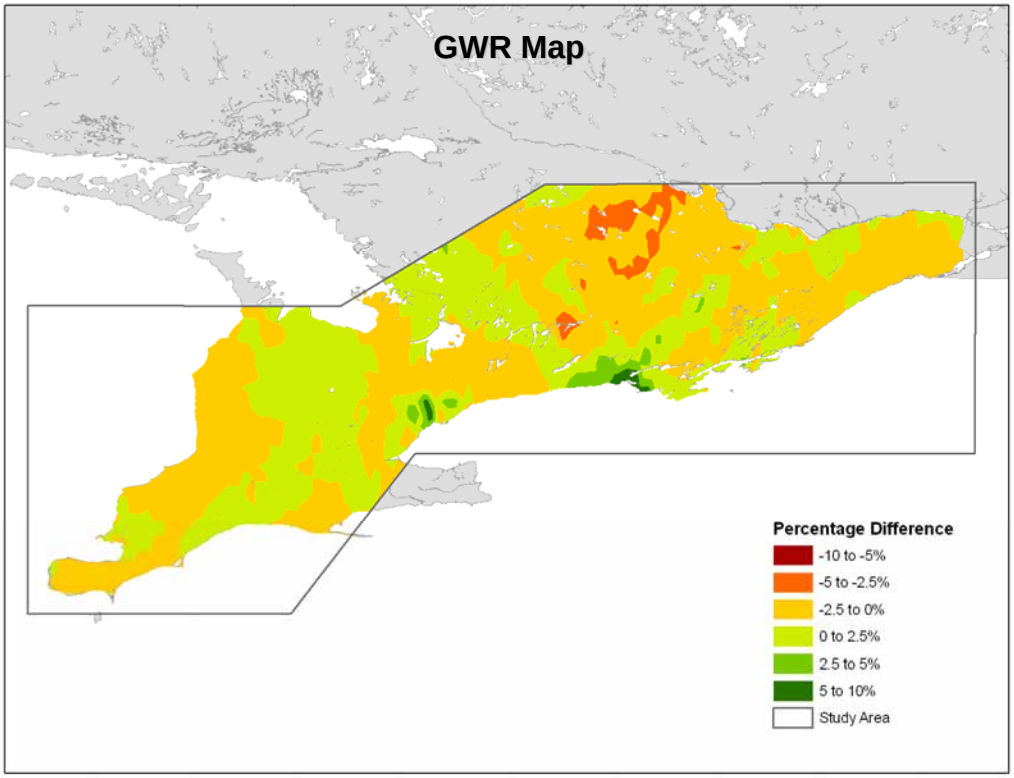


Figure 4.9: Prediction improvement of GWR model over OLS by percentage

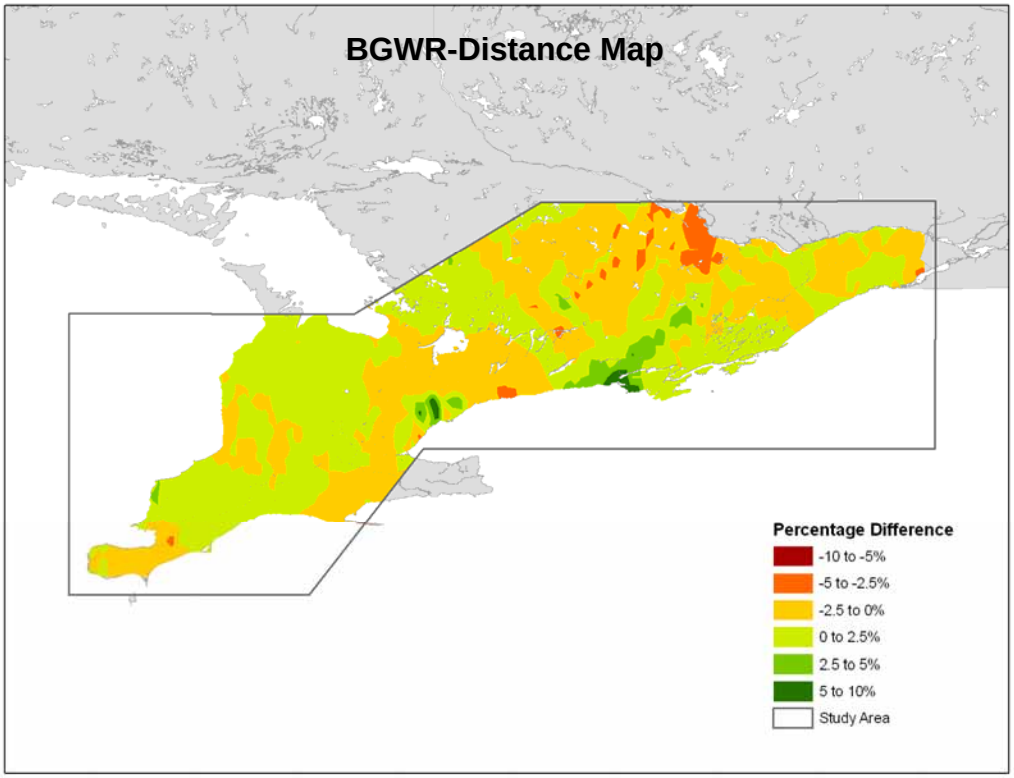


Figure 4.10: Prediction improvement of BGWR-Distance model over OLS by percentage

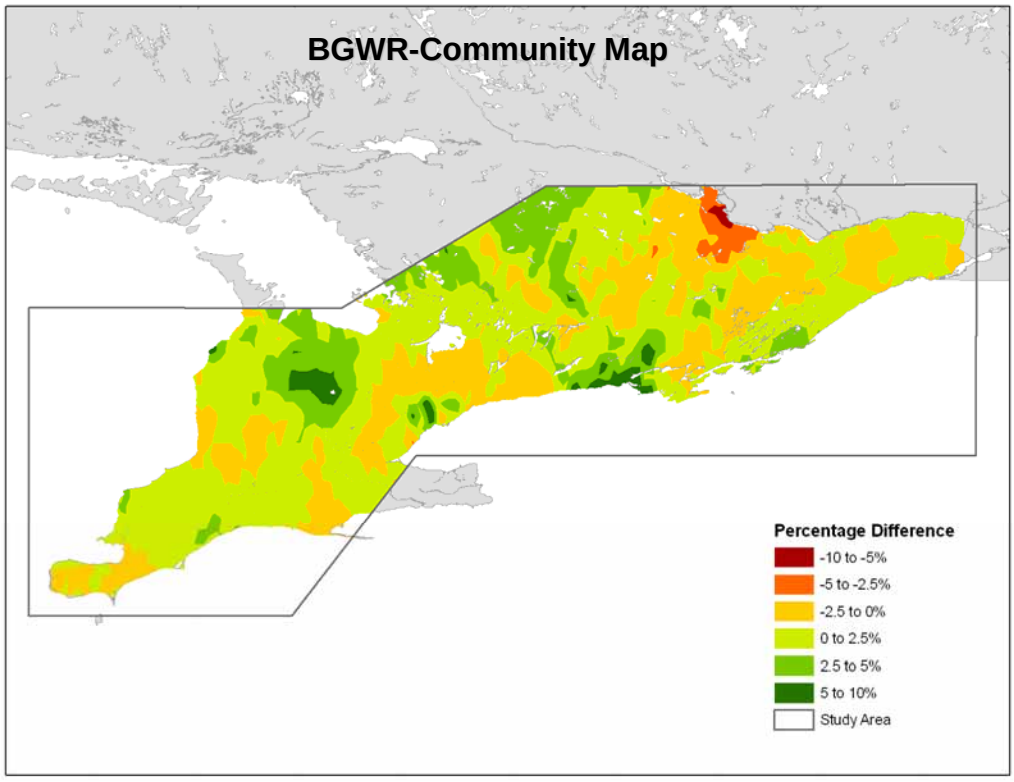


Figure 4.11: Prediction improvement of BGWR-Community model over OLS by percentage

Ordinary Kriging is an interpolation method that predicts the value at a certain location using the observed values around it. It weights the surrounding observed values by considering their distances from the prediction point as well as the spatial correlation of the observations. The maps in Figures 4.9 to 4.11 were created with *ArcGIS*™ Ordinary Kriging function that uses the prediction improvements of the local spatial regression models as input. On these maps, improvements in estimation appear as areas shaded in green, with darker green areas indicating greater improvement. To facilitate discussion, the GWR, BGWR-Distance, and BGWR-Community Maps are shown side by side in Figure 4.12 with areas highlighted by boxes that are numbered. Figure 4.13 shows the distribution of the observations with the BGWR-Community Map as the backdrop.

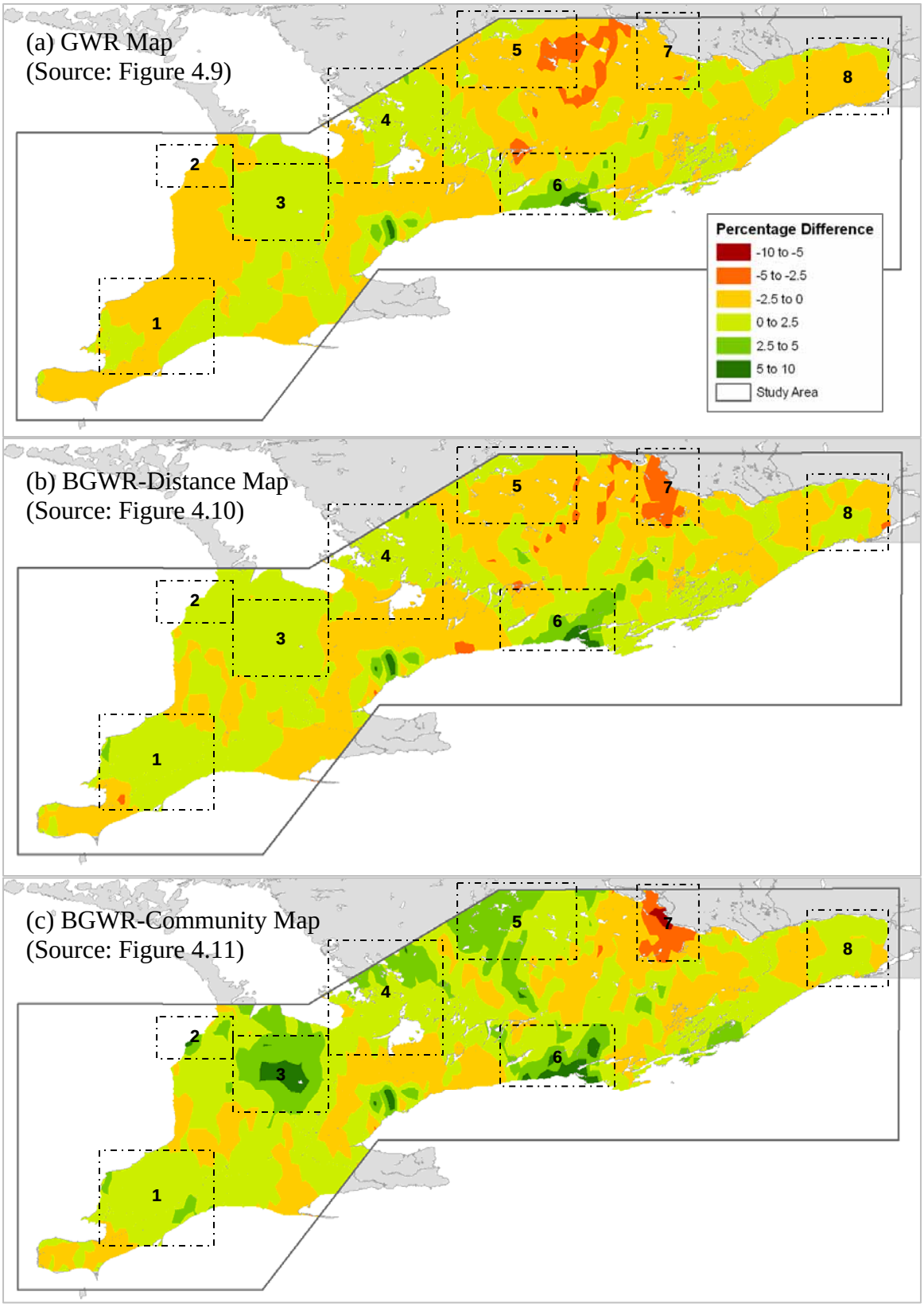


Figure 4.12: Comparison of the local prediction improvement of the three local spatial regression models

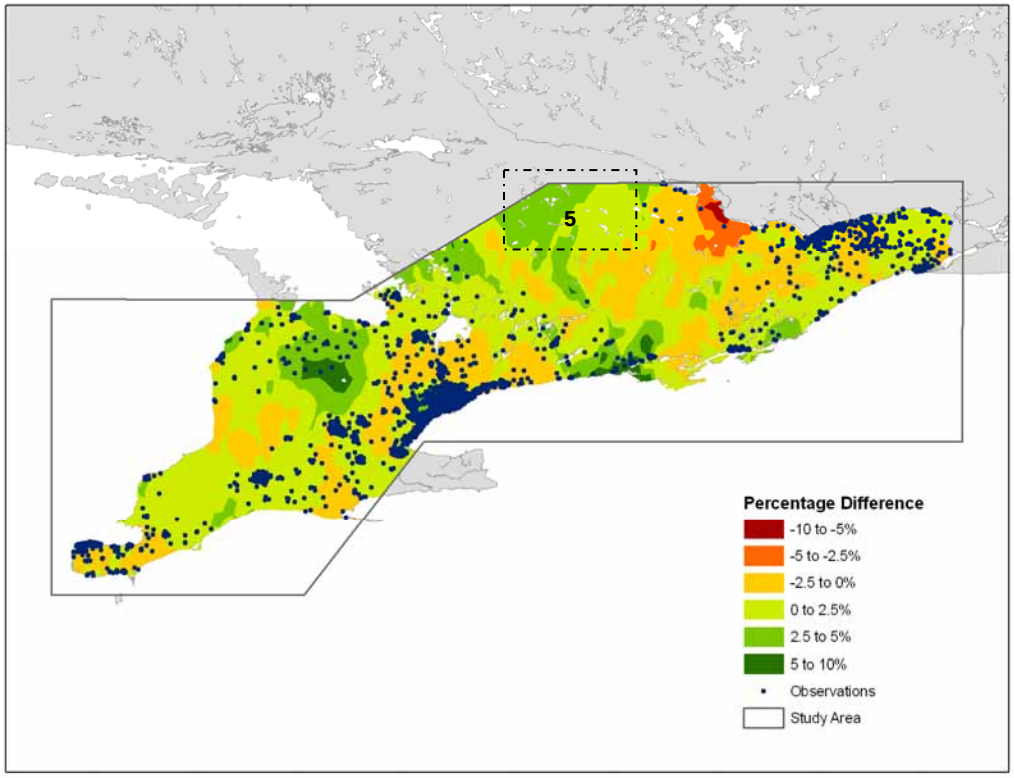


Figure 4.13: Distribution of observations with BGWR-Community Map as the backdrop

By cross-referencing Figures 4.12 and 4.13, the following results are observed:

- In the GWR Map, the proportion between the areas of positive values (green areas) and negative values (yellow, orange and red areas) are generally the same (Figure 4.12(a)). In the BGWR-Distance Map, there are relatively more green areas such as Areas 1, 2, 4 and 8 (Figure 4.12(b)). In the BGWR-Community Map, most parts of the study area are green (Figure 4.12(c)). It also has larger green areas and higher positive prediction improvement than the BGWR-Distance Map in Areas 1, 2, 4 and 8 (Figure 4.12(b)).

- When comparing Figure 4.12(a) the GWR Map with Figure 4.12(c) the BGWR-Community Map, major local improvements are found in Areas 3 and 6 where the BGWR-Community model improves the prediction performance from the range of 0 to 2.5% less prediction error than the OLS model to the range of 5 to 10% less.

- A close inspection of the corresponding location of Area 5 in Figure 4.13 found that there is no observation in that area. The interpolated improvements at that area of the BGWR-Community map are mainly due to the prediction improvements of the observations to the left and right of that area.

- Although the BGWR-Community model improves the local performance in most parts of the study area, it increases the prediction error in Area 7 from the range of 0 to 2.5% more prediction error than the OLS in the GWR Map (Figure 4.12(a)) to the range of 2.5 to 10% in the BGWR-Community Map (Figure 4.12(c)). Figure 4.14 shows a close-up of this location with the available observations around that area. It is found that only two observations are in the orange region and no observation is in the red region. The poor performance of the BGWR-Community model in this area may be due to insufficient neighbours, suggesting a potential limitation of this model.

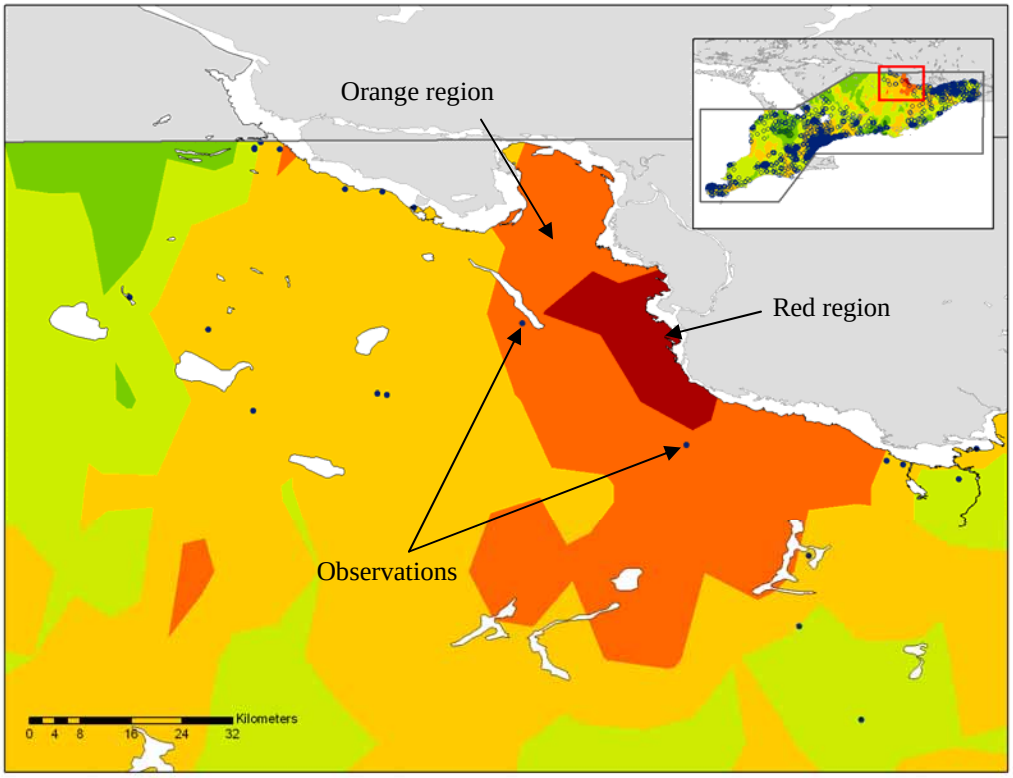


Figure 4.14: A close-up of Area 7 in the BGWR-Community Map

The above observations reveal that even though the contribution of the BGWR-Community model to an indicator like R-squared are not great, it improves the prediction performance in most parts of the study area. More importantly, it has substantial impacts on certain local areas. Like any other models, the BGWR-Community model has its limitations. As it is a Bayesian model, it took relatively long time (around 2 hours on a computer with *Intel*™ *Xeon*™ CPU[5]) to run. Besides, it needs sufficient data at the local level to deliver good prediction results; hence, this method is not suitable for dataset with sparse distribution of observations such as the northern Ontario (as shown in Figure 3.2), Nunavut and Northwest Territories.

[5] The configuration of the computer is Microsoft Window XP Professional running on Dell Precision PWS690 Intel® Xeon® CPU 5160 at 3.00GHz, 8Gb Ram.

In the next chapter, a summary of the works completed for the research will be presented, followed by the future research opportunities.

5.0 CONCLUSIONS

This chapter summarizes the work and findings presented in the previous chapters, followed by a discussion of the opportunities for future research and the concluding remarks.

5.1 Summary of Work Completed

The objective of this research was to first propose a statistical model that incorporated the concept of community in a local spatial regression model, and then to assess its performance. More specifically, the study asked whether incorporating the 'community' concept into the Bayesian Geographically Weighted Regression (BGWR) model would improve its performance over the purely distance-based local spatial regression models. The concept of community was introduced by including characteristics of local neighbourhoods using data from the Census. The following tasks have been carried out to accomplish this objective.

5.1.1 Acquiring Background Knowledge About the Research

An extensive literature review was carried out on the following areas of interest to acquire background knowledge about the research:

- the definitions of community and community effects as well as the techniques used in empirical studies of community effects;
- operationalization of the 'community' concept into measurable variables;

- the Geographically Weighted Regression (GWR) model; and
- the Bayesian approach of the GWR method (BGWR).

An important finding in the literature is that the GWR model is susceptible to the influence of 'outliers'. LeSage (2004) proposed the BGWR approach that allowed for various kinds of parameter smoothing relationships to tame the outlier problem. This also provided an opportunity to incorporate the 'community' concept into the BGWR model to account for the community effect that was not addressed by the purely distance-based local regression models. Therefore, this thesis proposed a 'community-based' Bayesian geographically-weighted regression model.

5.1.2 Designing the Experiment and Finding Appropriate Evaluation Criteria

A scientific experiment was designed to compare the prediction power of the proposed model with those of the distance-based local spatial regression models under study.

The base model of the experiment was an ordinary least squares (OLS) regression model which served as a baseline for comparison. The dependent variable of the model was the prose literacy scores of adults aged 16 to 65, while the independent variables were gender, age, age-squared, years of education, and personal income. The data for these variables came from the International Adult Literacy and Skills Survey (IALSS). The selected study area was southern Ontario

(Figure 3.2 refers). Furthermore, the data for the measurable variables of the 'community' concept and those for computing the distance matrix used in the local spatial regression models came from the Profile of Dissemination Areas, 2001 Canada Census. The data at this level included the following variables measured at the level of the Dissemination Area: (a) the average number of years of education in a DA, (b) the percentage of people who had moved in the previous five years, (c) average level of family income, (d) the percentage of people who were unemployed, (e) the percentage of people who were in professional or semi-professional occupations, (f) the percentage of people who were in unskilled labour occupations, or unclassified occupations, and (g) the percentage of people who had immigrated in the previous five years.

The two distance-based local spatial regression models under study were the GWR and the BGWR model using a distance-decay parameter smoothing relationship (BGWR-Distance). The GWR model was selected as it was one of the most popular spatially weighted regression models that is distance-based. The BGWR-Distance model was selected as this model is structured in a way that offers a transition between the GWR model and the proposed model. The proposed model, called the BGWR-Community model, was a BGWR model using a community parameter smoothing relationship.

Four specific evaluation methods were selected in order to give a broad view of the performance of the models. The methods included regression statistics,

numerical measures of the prediction errors to evaluate the overall performance, scatterplots for visual evaluation of the amount and trend of extreme predictions, and prediction rate curves to evaluate the performance (in terms of prediction errors) at different error tolerance levels.

5.1.3 Implementing the Models

The OLS base model was implemented with *SPSS*™ , while the local spatial regression models were implemented based on LeSage's (2005) "Spatial Econometric Toolbox", a host of spatial econometric estimation methods implemented with *Matlab*™.

To implement the 'community' concept as parameter smoothing relationship of the BGWR model, the concept was first operationalized into measurable variables based on Galster's (2001) bundle of spatially-based attributes which captured both the physical and social characteristics of a community. Then, these measurable variables were used to generate two weight matrices. One was a geographical distance-based weight matrix that represented the physical characteristics of the community. The other was a Mahalanobis distance-based matrix that represented the community characteristics. Finally, the community parameter smoothing relationship was implemented as a normalized weight matrix that combined the above two matrices.

5.1.4 Analyzing Experiment Results

The evaluation methods found that – compared with other models – the BGWR-Community model had a better overall prediction power and maintained a higher prediction rate up to a very high tolerance level. In addition, it did not produce more extreme prediction points than the other models. Therefore, it was concluded that by incorporating the 'community' concept into the BGWR model the prediction power of the model improved over the purely distance-based local spatial regression models.

In order to visualize the local improvements brought by the BGWR-Community model, the gradient surfaces of the prediction improvement of the local spatial regression models over the OLS model were created with a spatial interpolation method called Ordinary Kriging and presented as maps for visual comparison. It was observed that the BGWR-Community model could improve the prediction performance in wide range of areas and brought significant improvement at certain local areas.

5.2 Opportunities for Future Research

The present research demonstrated a means to incorporate concepts that are geographical in nature, even if the boundaries are ill-defined, into the parameter smoothing strategy of a local spatial regression model without pre-defining the boundaries. This research demonstrated its applicability in accounting for the community effect on the adult literacy scores, but the approach

could potentially be applied to other branches of social sciences, as well as other research areas including forestry, environmental science and ecology where concepts like land cover types, habitats, soil types are geographical in nature but have ill-defined or ‘fuzzy’ boundaries.

As discussed in Section 3.1.3, using straight line distance as distance measurement method may render the local spatial regression models not applicable in certain geographical areas. Another direction for future research that is worth examining is to compare the impact of using straight line distance with other distance measurement methods like the distance based on road networks, either in terms of the geographical distance or a cost function such as travel time.

5.3 Concluding Remarks

This research demonstrates that the incorporation of the ‘community’ concept into the local spatial regression model can improve the prediction power over the purely distance-based models by reducing the overall prediction errors. Furthermore, it shows that even though the contribution of the proposed model to an indicator like R-squared is not great, it can still bring significant prediction improvement to certain local areas. The research also sets an example to other research areas on how to integrate concepts that are geographical in nature but with ill-defined boundaries into the local spatial regression model to improve the prediction performance of the model.

Nevertheless, like any other models, the proposed model has its limitations. It was found that like other spatial models, the BGWR-Community model needs sufficient data at the local level to deliver good prediction results. Therefore, it is not suitable for datasets with a sparse distribution. As the proposed model is a Bayesian statistical model, it is computation intensive and takes a relatively long time to run. Exploring ways in which to optimize its performance under specific circumstances may also be a useful topic for future research.

6.0 REFERENCES

Berger, J. O., and L. R. Pericchi (2000). "Objective Bayesian methods for model selection: introduction and comparison." Working Paper, Duke University. [On-line] March 2008. http://ftp.stat.duke.edu/WorkingPapers/00-09.ps

Brint, S. (2001). "Gemeinschaft revisited: a critique and reconstruction of the community concept." *Sociological Theory*, Vol. 19, No. 1, pp. 1-23.

Brunsdon, C., S. Fotheringham, and M. Charlton (1998). "Geographically weighted regression – modelling spatial non-stationarity." *Journal of the Royal Statistical Society*, Series D-The Statistician, Vol. 47, Issue 3, pp. 431-443.

Bullard, F. (2001). "A brief introduction to Bayesian Statistics." Talks and Papers, Department of Mathematics and Computer Science, The North Carolina School of Science and Mathematics. [On-line] August 2007. http://courses.ncssm.edu/math/TALKS/PDFS/BullardNCTM2001.pdf

Cahill, M., and G. Mulligan (2007). "Using geographically weighted regression to explore local crime patterns." *Social Science Computer Review*, Vol. 25, No. 2, pp. 174-193.

Cassetti, E. (1972). "Generating models by the expansion method: applications to geographic research." *Geographical Analysis*, Vol. 4, pp. 81-91.

Denison, D. G. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley & Sons, Chichester, England.

Dietz, R. D. (2002). "The estimation of neighborhood effects in the social sciences: an interdisciplinary approach." *Social Science Research*, Vol. 31, Issue 4, pp. 539-575.

Fotheringham, A. S., C. Brunsdon, and M. Charlton (2000). *Quantitative Geography: Perspective on Spatial Data Analysis*, SAGE Publications, London.

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

Fotheringham, S., M. E. Charlton, and C. Brunsdon (2001). "Spatial variations in school performance: a local analysis using geographically weighted regression." *Geographical and Environmental Modelling*, Vol. 5, Issue 1, pp. 43-66.

Galster G. (2001). "On the nature of neighbourhood." *Urban Studies*, Vol. 38, Issue 12, pp. 2111-2124.

Gao, X., Y. Asami, and C.-J. F. Chung (2006). "An empirical evaluation of spatial regression models." *Computers & Geosciences*, Vol. 32, Issue 8, pp. 1040-1051.

Goddard, M. (2003). "Introduction to Bayesian Statistics." Course material for Armidale Animal Breeding Summer Course 2003, Part 2, Models and Methods for Genetic Analysis, University of New England, Armidale, Australia, 17-27 February 2003. [On-line] March 2008. http://www-personal.une.edu.au/~jvanderw/Introduction_to_Bayesian_Statistics1.pdf

Goodchild, M. F. (2001). "A geographer looks at spatial information theory." Paper presented at the Conference on Spatial Information Theory (COSIT), Morro Bay, CA U.S.A., 19-23 September.

Gorr, W. L., and A. M. Olligschlaeger (1994). "Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modelling." *Geographical Analysis*, Vol. 26, pp. 67-87.

Hagger-Johnson, G. (2006). "Seven useful features of the Mahalanobis distance statistic for psychologists." *PsyPAG Quarterly*, Issue 58, pp. 28-33.

Hillery, G. A. (1955). "Definitions of community: areas of agreement." *Rural Sociology*, Vol. 20, pp. 111-123.

Kleder, M. (2005). *Vectorized geodetic distance and azimuth on the WGS84 earth ellipsoid*. [On-line] 16 October, 2007. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=8607&objectType=File

Koch, K.-R. (2007). *Introduction to Bayesian Statistics.* 2nd ed., Springer-Verlag, Berlin.

Kupfer, J. A., and C. A. Farris (2007). "Incorporating spatial non-stationarity of regression coefficients into predictive vegetation models." *Landscape Ecology*, Vol. 22, No. 6, pp. 837-852.

Law, S. Y. J. (2000). *Spatial Analysis of Multivariate Demographic Data for Identifying Communities in GIS*. Ph.D. thesis, Department of Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, N.B., Canada, 259 pp.

LeSage, J. P. (2004). "A family of geographically weighted regression models." In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, Eds. L. Anselin, R. J. G. M. Florax, and S. J. Rey, Springer-Verlag, Berlin, pp. 241-264.

LeSage, J. P. (2005). *Spatial Econometric Toolbox*. [On-line] 10 March, 2008. http://www.spatial-econometrics.com

Longley, P. A., and C. Tobón (2004). "Spatial dependence and heterogeneity in patterns of hardship: an intra-urban analysis." *Annals of the Association of American Geographers*, Vol. 94, Issue 3, pp. 503-519.

Lupton, R. (2003). "'Neighbourhood effects': can we measure them and does it matter?" CASE Paper 73, Centre for Analysis of Social Exclusion, London School of Economics. [On-line] 8 March, 2008. http://sticerd.lse.ac.uk/dps/case/cp/CASEpaper73.pdf

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York.

MacQueen, K. M., E. McLellan, D. S. Metzger, S. Kegeles, R. P. Strauss, R. Scotti, L. Blanchard, and R. T. Trotter II (2001). "What is community? An evidence-based definition for participatory public health." *American Journal of Public Health*, Vol. 91, Issue 12, pp. 1929-1938.

Malczewski, J., and A. Poetz (2005). "Residential burglaries and neighborhood socioeconomic context in London, Ontario: global and local regression analysis." *Professional Geographer*, Vol. 57, No. 4, pp. 516-529.

Mimmack, G. M., S. J. Mason, and J. S. Galpin (2001). "Choice of distance matrices in cluster analysis: defining regions." *Journal of Climate*, Vol. 14, Issue 12, pp. 2790-2797.

Osborne, P. E., G. M. Foody, and S. Suárez-Seoane (2007). "Non-stationarity and local approaches to modelling the distributions of wildlife." *Diversity and Distributions*, Vol. 13, Issue 3, pp. 313-323.

Propastin, P., N. Muratova, and M. Kappas (2006). "Reducing uncertainty in analysis of relationship between vegetation patterns and precipitation." *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science*, Eds. M. Caetano, and M. Painho. Lisbon, 3-6 July, pp. 459-468.

Rapkin, B. D., and D. A. Luke (1993). "Cluster analysis in community research: Epistemology and practice." *American Journal of Community Psychology*, Vol. 21, No. 2, pp. 247-277.

Small, S., and A. Supple (1998). "Communities as systems: is a community more than the sum of its parts?" Paper presented at the national forum on community effects on children, adolescents and families, Penn State University, State College, PA, U.S.A., September.

Statistics Canada (2003a). "Profile of Dissemination Areas in Canada, 2001 Census." Statistics Canada, Catalogue No. 95F0495XCB01002.

Statistics Canada (2003b). "Postal code conversion file: January 2003 postal codes reference guide." Statistics Canada, Catalogue No. 92F-0153-GIE.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region." *Economic Geography*, Vol. 46, Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods. pp. 234-240.

Trojanowicz, R., and M. Moore (1988). *The Meaning of Community in Community Policing*. Michigan State University: National Neighborhood Foot Patrol Center. [On-line] 10 September 2007. http://cpt4cops.org/pdf/Meaning%20of%20Community%20in%20CP.PDF

Wang, Q., J. Ni, and J. Tenhunen (2005). "Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems." *Global Ecology and Biogeography*, Vol. 14, Issue 4, pp. 379-393.

Warner, R. M. (2008). *Applied Statistics*: *from Bivariate through Mutlivariate Techniques*. SAGE Publications, Los Angeles.

Wellman, B., (2001). "The persistence and transformation of community: from neighbourhood groups to social networks." Report to the Law Commission of Canada. [On-line] September 2007. http://www.chass.utoronto.ca/~wellman/publications/lawcomm/lawcomm7.PDF

Willms, J. D. and T. S. Murray (2007). "Gaining and losing literacy skills over the lifecourse." Statistics Canada, Catalogue No. 89-552-XIE, No. 16.

Willms, J. D. and T. Tang (2007). "Mapping of literacy as a determinant of health: final report." Unpublished report of the Canadian Research Institute for Social Policy, University of New Brunswick, and the National Collaborating Centre for Determinants of Health, Canada.

Willms, J. D., R. Chan and T. Tang (2007). "Geographical distribution of adult literacy skill in Canada based on local area estimates." Human Resources and Social Development Canada, Catalogue No. HS28-118/2007-MRC.

Zhang, L., H. Bi, P. Cheng, and C. J. Davis (2004). "Modeling spatial variation in tree diameter–height relationships." *Forest Ecology and Management*, Vol. 189, Issues 1-3, pp. 317-329.

APPENDIX A DATA PREPARATION

The data preparation process involved the IALSS, the Dissemination Area (DA) profile data (Statistics Canada, 2003a), the DA polygon files, and the postal code conversion file (PCCF) (Statistics Canada, 2003b) from Statistics Canada. Figure A.1 below illustrates the details of the preparation process.

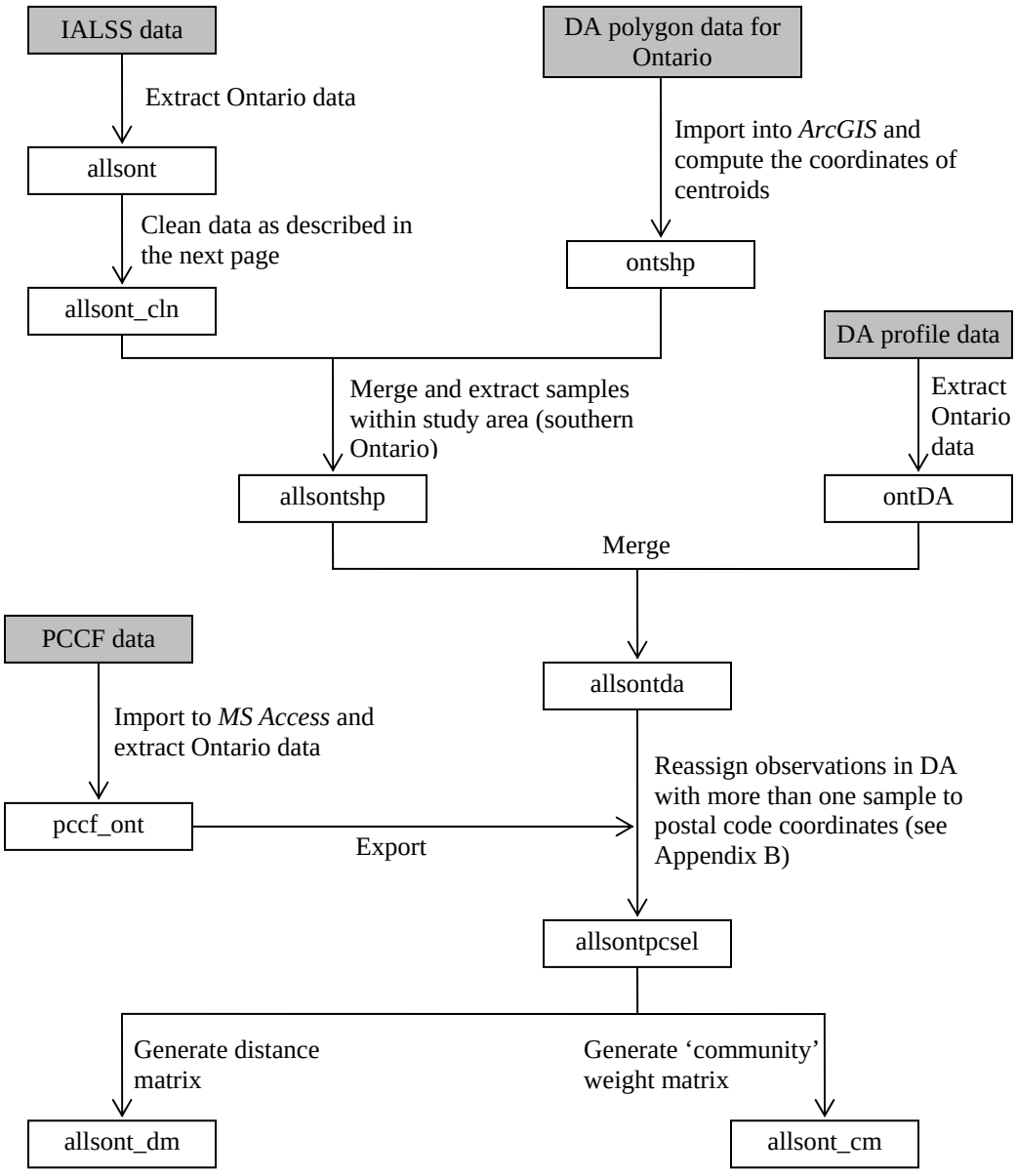


Figure A.1: Data preparation flow chart

After **allsont.csv** was imported into *MS Excel*, the following steps were carried out to clean, re-centre and/or scale the data:

a. Gender:
    i. Re-centre gender (*gender*) to give *rcgender* such that male = -0.5, female = 0.5

b. Age:
    i. Select data of age 16-65 (*age_resp*).
    ii. Re-centre age to give *rcage* such that *rcage* = *age_resp* – 40

c. Age square:
    i. Create new variable *rcagesq* = *rcage***rcage*

d. Years of Education:
    i. Recode year of education (*a3*) such that
        - If (*a3*<= 6) *a3* = 6
        - If(*a3*>=21) *a3* = 21
    ii. Re-centre year of education (*a3*) such that *rcyrsed* = *a3* – 12

e. Personal income:
    i. Re-scale imputed personal income (*K6i*) such that *pincome* = (*K6i*/100000)
    ii. Recode *pincome* such that:
        - If (*pincome* > 150) *pincome* = 150
    iii. Re-centre *pincome* such that *rcincome* = *pincome* – 30

APPENDIX B RANDOM ASSIGNMENT OF POSTAL CODES TO SAMPLES

For Census DAs with more than one observation, the method illustrated in Figure B.1 below was used to randomly assign postal codes to the observations.

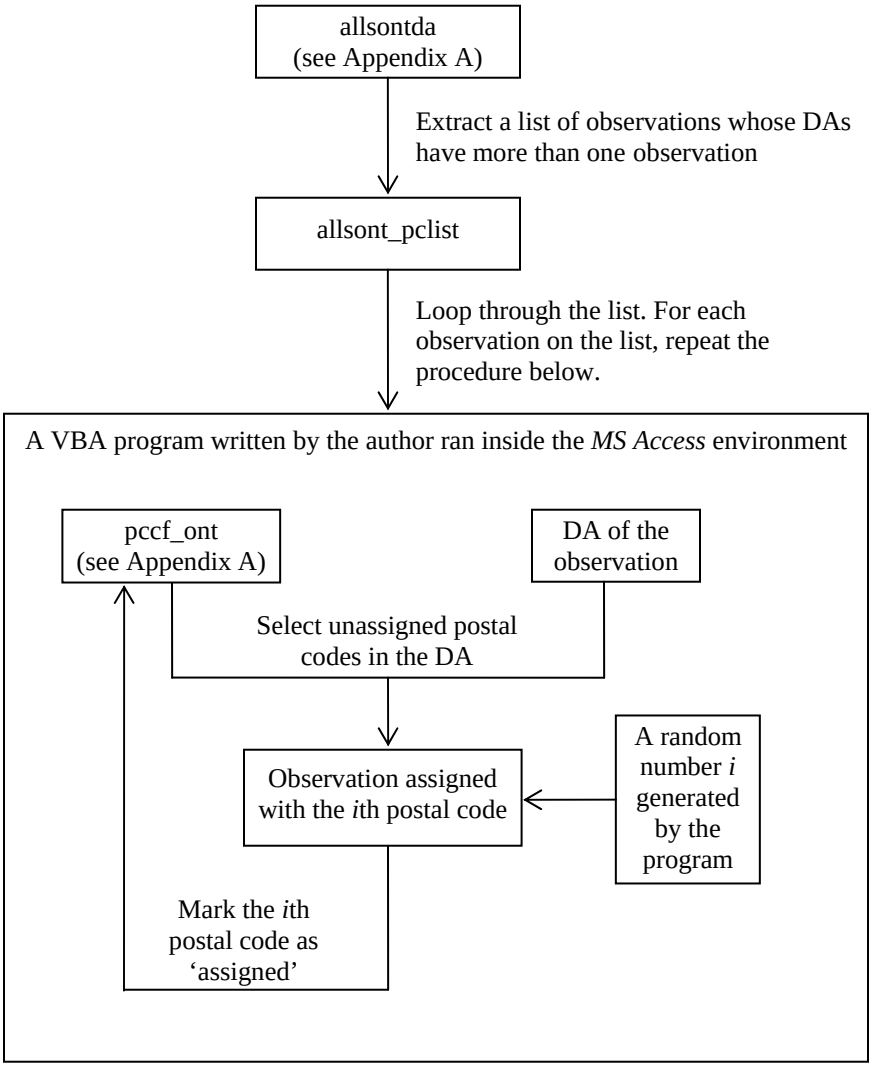


Figure B.1: Postal code assignment process

CURRICULUM VITAE

Candidate's full name: Hon Shing (Richard) Chan

Universities attended: The Hong Kong Polytechnic University, Postgraduate Diploma in Geo-information Systems, 2000
University of Hong Kong, M.Sc.(Urban Planning), 1996
University of Hong Kong, B.Sc.(Quantity Surveying), 1987

Conference Presentations:

Presenter, "*Close Range GIS - A New Way of Visualizing Geographical Information*", 2003 Student Technical Conference, Geodesy and Geomatics Engineering, University of New Brunswick, on 21 and 24 March 2003.