Application of Geographically Weighted Regression for Mass Valuation using the Indonesian Land Agency Dataset for Bekasi, Indonesia

by

Rahmat Ganda Pandapotan Sihombing

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Science and Engineering
31 January 2019

Table of Contents

Abstract

Valid property transaction data in Indonesia is scarce because parties involved in a property transaction often report a false, lower transaction price to reduce their transaction tax liability. This has meant that in every mass valuation project administered by the National Land Agency of Indonesia (BPN RI), the sample size has never been sufficient to allow the currently employed Zonation Method to provide a complete prediction of land values across a city. A new mass valuation method that is fit for purpose when applied to a BPN RI-dataset is required. An extensive literature review was conducted to compare mass valuation techniques used worldwide, and Geographically Weighted Regression (GWR) was identified as the best potential candidate to replace the Zonation Method.

The performance of GWR modelling was tested on a typical BPN RI-dataset from the city of Bekasi in western Java. A road network dataset was required to generate data for ten of the 12 parameters listed in the current Mass Valuation Standards of BPN RI. A road network dataset had to be derived from the Land Parcel Map of Bekasi for this research because existing road networks from other sources had severe mismatches with the Land Parcel Map. Deriving a road network dataset from the Land Parcel Map was very time consuming because of the huge number of drawing errors in the Land Parcel Map that had to be corrected.

In the Bekasi case study, the GWR model had a mean absolute percentage error (MAPE) of 19.40 per cent, which was lower than the currently employed Zonation Method with a MAPE value of 10.80 per cent. Nevertheless, the GWR model solved the main problem of the Zonation Method; i.e. its inability to provide verifiable predictions for zones with fewer than three samples. Moreover, the MAPE value of 19.40 from the GWR model was well below the cut-off value of 30 per cent accuracy currently used by BPN RI.

The performance of the GWR model at non-sampled locations was estimated by out-of-sample estimation using Monte Carlo Cross Validation. The distribution of average percentage residuals from out-of-sample estimation resembles the distribution of percentage residuals from the in-sample GWR model. The correlation coefficient of the two distributions was 0.987. These two facts indicate that the GWR model does not have an issue of overfitting, and therefore it is very likely to maintain its prediction accuracy when predicting the non-sampled locations.

The main issue discovered when applying the GWR model was that a small proportion (7.51 per cent) of predictions at sampled locations had residuals greater than 50 per cent of the

actual value. In the absence of overfitting, a similar proportion of predictions at the non-sampled locations are also likely to be inaccurate. Value zones were employed to detect potentially inaccurate predictions because predicted land prices in one value zone can be expected to be similar to one another. Local Moran's I test and the coefficient of variation were employed to detect anomalous individual predictions in each value zone.

The problem of a lack of valid transaction data for mass appraisal modeling, while a big issue in Indonesia, is also a major problem in many other countries in the world. The approach taken in this study can potentially be adapted and amended in many other countries. The method is useful to provide accurate predictions at non-sampled areas. The key issue in applying the approach is the need for an accurate digital road network map.

Declaration

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

……..

Rahmat Ganda Pandapotan Sihombing

January 2019

Acknowledgements

I would like to thank my supervisor, Prof. Andrew Millington, for his guidance and intellectual contribution for my PhD. There were ups and downs during my study but he managed to guide me to finish my PhD on time.

I would also like to thank Mr. Robert Keane, the Computer Systems Officer in the School of the Environment. He wrote the scripts to run the Monte Carlo Cross Validation on the GWR model and provided technical assistance to run a number of analyses.

I thank my fellows Land Lab members for the inputs and discussions during the Land Lab meetings. New insights and corrections were often resulted from these meetings.

I also thank Dr. Cecile Cutler for the copy editing work on my PhD thesis. She put rigorous effort and gave quick responses.

I gratefully acknowledge the SPIRIT Scholarship Program as the sponsor that gave me the opportunity to conduct this PhD research.

Lastly, special thanks go to my wife, Sysca, and my two daughters, Amanda and Anggita, for their love and support during my PhD.

Rahmat Ganda Pandapotan Sihombing

Flinders University

January 2019

List of Figures

List of Tables

Abbreviations

AICc : Corrected Akaike Information Criterion

ANN : Artificial Neural Networks

BIG :*BadanInformasiGeospasial* (*Geospatial Information* Agency of Indonesia)

BPN RI : *BadanPertanahanNasional* (National Land Agency of Indonesia)

BPS : *Biro PusatStatistik* (Indonesian Bureau of Statistics)

CBR : Case-Based Reasoning

CV : Cross Validation

GA : Genetic Algorithm

GAM : Generalized Additive Model

GWR : Geographically Weighted Regression

HTM : Hierarchical Trend Modelling

IAAO : International Association of Assessing Officers

IVSC : *International Valuation Standards Council*

LOOCV : Leave-One-Out Cross Validation

LISA : Local Indicators of Spatial Association

MAPE : Mean Absolute Percentage Error

MCCV : Monte Carlo Cross Validation

MRA : Multiple Regression Analysis

NJOP : *NilaiJualObyekPajak* (Sales Value of Tax Object)

OLS : Ordinary Least Squares

PPMRA : Piecewise Parabolic Multiple Regression Analysis

RST : Rough Set Theory

SEM : Spatial Expansion Model

VIF : Variance Inflation Factor

1. INTRODUCTION

1.1. Introduction

Indonesia is the largest country in Southeast Asia. Heryani and Grant (2004) estimated that the country has around 80 million land parcels within it. The National Land Agency of Indonesia (BPN RI) started land valuation in 2006, and the mass valuation approach has been the most feasible option to have all land properly valued in as timely a manner as possible. Property sales data is the main input for mass valuation work but collecting valid property sales data in Indonesia is more difficult than in many other countries (see Tamtomo *et al.*, 2008). Parties involved in a transaction tend to report much lower transaction prices in order to lessen the transaction tax.

Because solving the issue of false declaration of transactions will require changes in policy and regulations to be adopted at different levels of administration within Indonesia, this study focuses on improving the mass valuation method. A method which is able to work well with a limited number of samples must be chosen from among the methods being used worldwide. Adjustments must also be arranged to suit the circumstances of Indonesian cities.

1.2. Land valuation in Indonesia

Land valuation was introduced in Indonesia for taxation purposes during the Dutch colonial era. Booth (1974) showed that land parcels were classified on the basis ofthe irrigation system they were part of, slope inclination, soil type, ease of cultivation, and the relative location of the village in the *kawedanan* (sub-district). Next, the average rice yield of each class of land was calculated in order to formulate the *landrente* (land tax) per hectare for each class. In 1923 and 1928, *verponding* (the first individual property taxes) were introduced; these were created primarily for urban areas (see Kelly, 2003). Booth (1974) also noted that in 1965, 20 years after Indonesian independence, the income approach was maintained for land valuation in rural areas.

Kelly (2004)observed a substantial shift in valuation methodology after the enactment of the Land and Building Tax Law in 1986:

> The mass appraisal process for land is based on a ‘similar land value zone’ approach, where land is divided up into various zones – each with an average sales price per-squared metre as determined by the tax department. All land parcels located within that zone are valued by multiplying the land area by the average per

unit price. The buildings are valued based on a cost approach using cost tables determined by the tax department. The total property value is the summation of the land and building values(Kelly, 2004, pp. 119-120).

Since then, the value of land or buildings released by the Directorate of Property Taxation has been officially called the 'sales value of tax object' *(Nilai Jual Obyek Pajak* – NJOP).

Lewis (2003) estimated that the coverage of the taxation-intended value (NJOP) had been extended to 85.6 per cent of all land parcels in urban areas and 66.0 per cent of those in rural areas. Due to its wide availability, the NJOP has been the most widely used reference for land value. Beside its main use for taxation purposes, NJOP has also been used as a reference for land values for other purposes, e.g., land acquisition planning, development planning, tariffs of services related to land ownership, and asset declaration. Yet Lewis (2003) also highlighted clearly that there is some evidence to suggest that on average, government appraisals of taxable property make up only approximately 60 per cent of real market values. Though widely available, NJOP does not represent the market values of land parcels in Indonesia.

1.3. On the need for reliable reference to land value

A market-based valuation that generates a reliable reference to land value is urgently required for development planning and taxation in Indonesia. In principle, market-based valuation is about mapping market values. This fair value will be useful to all aspects of society in support of sustainable development as a reference for the land market, land asset management, land taxes and fees, land policy making and other decisions related to land (see Tamtomo *et al.*, 2008). For example, in a study of infrastructure development in the USA, Delluchi and Murphy (2005) indicate that the availability of a reference to land value is a crucial issue when examining the feasibility of public projects. Indonesian Law No. 2 Year 2012 on Land Acquisition for Development in the Public Interest[1] obliges the land acquisition committee of an infrastructure development project to estimate land value as part of any feasibility study. The 'undervalued' NJOP has been used for land acquisition budgeting in Indonesia for decades, and its use is the main reason behind disputes about compensation. In the absence of reliable reference to land value, the National Land Agency of Indonesia (BPN RI) started land valuation based on market values in 2006 (see Tamtomo *et al.*, 2008).

[1]Undang-UndangNomor 2 TentangPengadaan Tanah Bagi Pembangunan UntukKepentinganUmum(Republic of Indonesia, 2012).

1.4. False transaction price reporting: the cause and the impact

A major problem in meeting the standards in the current mass valuation practice adopted by BPN RI is the scarcity of valid property sales data. The parties involved in a property transaction tend to give false statements about the transaction price to lessen the transaction tax. In 1997, the Indonesian government enacted a Land and Building Acquisition Tax (Law No. 21/97). This law levies a tax of five per cent applied to the 'acquisition value' of the property minus a deduction of up to 60 million IDR, and the latter threshold is determined by the regional government (see Kelly, 2003).

A value close to the NJOP, which is usually much lower than the actual transaction price, is normally declared in the sale deed.[2] In a normal circumstance, everyone is very likely to report false transaction price. Despite that, there is yet no report or study that presents the number or percentage of the false reporting on transaction prices in Indonesia.

This common practice of tax avoidance involves the property seller, the buyer, and the notary. Using NJOP as the basis of price reporting has become an 'acceptable' illegal act. In the context of mass valuation, it means that the property sales compilation at a local land office cannot be used for mass valuation because it contains false prices. Therefore field survey is required to collect actual sales data.

Advertisements are useful initial data sources in property sales surveys. They show the properties for sale, and there is usually some part of the data required for valuation. Unfortunately, well-displayed advertisements on websites generally contain only properties being sold through top-branded property brokers and therefore they cover only a small portion of properties for sale in many areas. Information from private sellers has been the main source of sales data for mass valuation practices in many cities in Indonesia. Private sellers normally place sale signs at the front of properties for sale, as shown in Figure 1.1.

In most rural areas, many people consider that it is showy to put a sale sign on a property for sale.[3] Sellers prefer to inform staff at the village office about a property for sale. That means that potential buyers often have to ask for assistance from the village office when they want to buy a property. In these cases, the asking price and the final transaction price is only known to the parties directly involved in the transaction. It is a non-transparent land market.

---

[2]Part of this information comes from the author's experience as a staff of Directorate of Land Valuation in BPN RI.

[3] Part of this information comes from the author's experience as a surveyor in a number of mass valuation projects administered by BPN RI.

The above circumstances make data collection difficult and time consuming. Because the actual sales price is considered confidential, surveyors often have to act in the guise of potential buyers or a property sales agent. Tamtomo and colleagues (2008) suggest that collecting market valuation data is more difficult in Indonesia than in many other countries.Because collecting valid data is very difficult, data scarcity has been the biggest issue for mass valuation practice in Indonesia.

Figure 1.1 Typical private seller's sale sign

Source: Mass Valuation Project of Bekasi City in 2012

Note:

- 'DIJUAL' means FOR SALE.
- 'T.P' stands for 'Tanpa Perantara' which means 'no middleman'. Private sellers prefer direct contact with potential buyers. Phone numbers are written on the sign.
- 'SHM' stands for 'Sertipikat Hak Milik' (freehold title issued by Local Land Office).
- '20x55' is the size of the land parcel in square metre.
- The price is sometimes stated but it is not common to do so.
- 081397439989 is the phone number of the private seller.

During the process of property sales data collection, surveyors have to do a door-knocking survey. Next, an interview has to be conducted in a particular manner to get sensible responses. Surveyors for a mass valuation project in a city or district are usually from the local Land Office. Surveyors from the Regional Land Office and surveyors from the headquarters of BPN RI (Land Valuation Directorate) also join the surveys in cities and districts. The official data from the Ministry of Home Affairs4 shows that there are 93 cities and 415 districts in Indonesia. For each city or district, survey for mass valuation is conducted every year because land value change rapidly in a developing country like Indonesia. Due to the limited number of surveyors at the Regional Land Offices and Land Valuation Directorate, surveyors from these offices can only take part in a limited number of surveys in a number of cities or districts each year. The author of this thesis is a surveyor from the Directorate of Land Valuation of BPN RI.

1.5. Current mass valuation technique in BPN RI: Zonation Method

The Zonation Method adopted from the Directorate General for Property Taxation of Indonesia has been employed in BPN RI. Neighbouring land parcels in an area with a dominant or homogeneous land coverage and land use are assumed to have relatively similar land values. These land parcels are used to define a closed polygon called a land value zone.

Property or land sales data are collected by means of stratified sampling. Stratified sampling is a sampling method in which a population is divided into mutually exclusive groups (called strata), and then simple random or systematic samples are selected from each of these strata (Hibberts *et al.*, 2012). In the case of Indonesia, the strata are the polygons of land value zones.

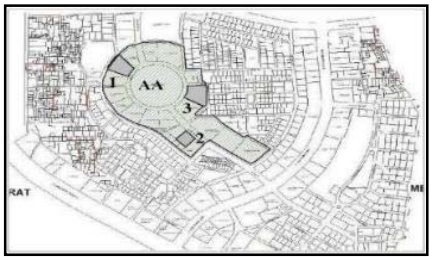
A value zone must have at least three samples (land parcels with known values) so that an average land value and its standard deviation can be calculated. The average land value will be taken as the land value for the zone if the coefficient of variation of sampled land values is lower than 30 per cent. The details about the Zonation Method are compiled in the Internal Standards for Land Valuation in BPN RI[5]. A brief description of the processes in general is shown in Figure 1.2.

---

[4]Pembentukan Daerah-Daerah Otonom Di Indonesia Sampai Dengan 2014, Kementrian Dalam Negeri Republik Indonesia (2014).

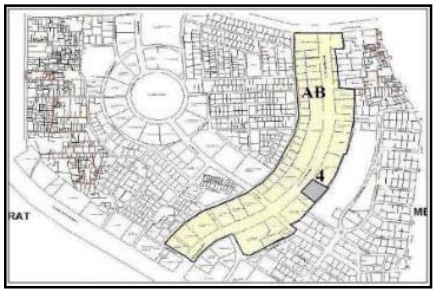
[5]Standar Operasional Prosedur Internal Survei Potensi Tanah, BPN RI(2013).

1. A zone with three or more samples



A value zone is expected to have at least three sales data. If the coefficient of variation is < 30 per cent, the value for the zone will be calculated as the average value of all the samples.

2. A zone with less than three samples



If a value zone has less than three samples, the value for the zone will not be calculated. There will be no value assigned for the zone.

Figure 1.2 Description of the Zonation Method currently employed for mass valuation in BPN RI

Source: Internal Standards for Land Valuation in BPN RI

Providing at least three samples of valid data for each zone has always been the key issue. In every mass valuation project conducted by BPN RI, sample data shortage has always been the biggest issue. The data from the 2012 Mass Valuation Project of Bekasi City has been selected in this research to give an example of the data shortage issue. The number of samples per land value zone (Table 1.1) is typical of other towns and cities surveyed by BPN RI.

Table 1.1 Breakdown of value zones based on the number of samples

Number of samples	Number of value zones	Percentage breakdown
≥ 3	50	5.1
1 or 2	390	39.8
0	540	55.1
Total	**980**	**100.0**

It should be clear from this brief introduction that the Zonation Method, the mass valuation method currently employed by BPN RI, does not work well given the circumstances in Indonesia.

1.6. Research objective and questions

As aforementioned, the major problem for accurate mass valuation practice in Indonesia is the scarcity of valid sales data. This study has been formulated to investigate effective measures in dealing with the data scarcity problem, with the main emphasis being placed on investigating geospatial modelling of existing data sets. This is because solving the issue of false declaration of transactions will require changes in policy and regulation to be adopted at different levels of administration within Indonesia.

The main objective of this study is to improve current mass valuation practice given existing data scarcity issue. This study will evaluate the most suitable geospatial modelling technique to be applied to the existing dataset (with the limited number of samples) and compare it with the current practice. A method that is able to work with a limited amount of data will be chosen among the popular mass valuation techniques being used worldwide. Next, adjustments will be incorporated into the chosen technique in order to suit the local characteristics of Indonesian cities.

Bekasi City is located in West Java Province, and it is a typical of most Indonesian cities in which road networks and arrangements of land parcels vary significantly among neighbourhoods. The 2012 Bekasi City dataset has the best distribution of sample data among all mass valuation projects administered by BPN RI, so this dataset is used for this study to give optimum exploration of the selected mass valuation technique.

The research questions related to this objective are:

1. To convert an existing BPN RI-dataset from Bekasi, Indonesia into a format that can be used in geospatial modelling of land transaction values.
2. To identify how the geospatial technique that has been used for mass valuation in other administrations performs when applied to the Bekasi dataset and compare the result with the result from the Zonation Method currently employed by BPN RI.
3. To identify what adjustments and recommendations can be made for better mass valuation practice by BPN RI in urban areas.

1.7. Thesis structure

Following this Introduction chapter, Chapter 2 reviews the existing mass valuation techniques used worldwide. The mechanisms, the advantages, and the disadvantages of each technique will be outlined.

Chapter 3 focuses on data preparation. The basic data for Bekasi that were available at the start of the research were property sales data and a land parcel map. The properties of these data and issues they pose for mass valuation will be discussed. The chapter will detail how road network dataset was derived from the land parcel map, and how a travel time map was derived from the road network dataset. This chapter focuses on Research Question 1.

Chapter 4 examines the properties of the dataset after the data preparation procedures were carried out in Chapter 3. Specifically, the focus is on correlations between variables, assessments of multicollinearity and spatial autocorrelation, and variable transformations. The goal of the research in this chapter is to optimise the performance of the prediction model using the Bekasi dataset. This chapter focuses on Research Question 1.

Chapter 5 discusses the prediction performance of the Geographically Weighted Regression (GWR) technique using the Bekasi dataset. The performance of the GWR model is compared with the Zonation Method currently employed by BPN RI. The main issue inherent in the GWR model is the extremely large prediction residuals at several locations. This chapter focuses on Research Questions2 and 3.

Chapter 6 contains explanations on how the issue of extremely large prediction residuals at several locations are tackled by controlling the input for the model. Samples are controlled using value zones. The basic idea is that land parcels in one value zone tend to have similar values. However, the analyses reveal that GWR in the unit of value zones does not solve the issue of extremely large prediction residuals at several locations. This chapter focuses on part of Research Questions 2 and 3.

Chapter 7 will explain an alternative approach to tackle the issue of extremely large prediction residuals at several locations when using the GWR model. Instead of controlling the input for the model (as conducted in Chapter 6), measures will be employed to control the prediction output of the model. The input data is taken as it is but then outliers will be detected from the prediction output. This chapter focuses on Research Question 3.

Chapter 8 is a discussion chapter in which the results from analyses in Chapters 4, 5, 6, and 7 will be discussed in relation to the research objective and research questions. These

results will also be compared with the results from other similar researches. This chapter focuses on Research Questions 2 and 3.

Chapter 9 contains conclusions from the research undertaken in the study, and makes recommendations for a better mass valuation practice in Indonesia. This chapter focuses on Research Question 3.

1.8. Researcher's positionality

The author was an analyst in the Directorate of Land Valuation of BPN RI before undertaking this research and has returned to that position after undertaking this doctoral research in Australia. The author started to work at the Directorate of Land Valuation of BPN RI in 2010, and have been involved in a number of mass valuation projects administered by BPN RI since then. The author took part in the 2012 Bekasi City Mass Valuation Project as the Survey Coordinator, and used the dataset from the 2012 Bekasi project for this study. The application of Geographically Weighted Regression (GWR) model on the 2012 Bekasi dataset is the main part of this PhD research.

2. STATE OF KNOWLEDGE

2.1. Introduction

In property valuation, the term 'mass valuation' is used inter-changeably with 'mass appraisal'; the latter term seems to be more popular amongst practitioners. Eckert (1990, cited in McCluskey *et al.*, 1997) defined mass appraisal as the systematic appraisal of groups of properties on a given date using standardised procedures and statistical testing. The goal is to value a great number of properties within a minimal time and financial budget allocations. The mass appraisal guidance from IAAO(2014)stated that mass appraisal (valuation) is required when many properties need to be valued economically and *en masse* for a purpose, such as annual property taxation.

The output of mass valuation works is widely used as the reference to property values for taxation purposes. In order to provide reliable reference for taxation, a mass valuation work must result in accurate predictions on property values. For this goal, researches to develop mass valuation techniques have been conducted worldwide. A number of mass valuation techniques will be compared in this chapter to see the advantages and disadvantages of each technique.

2.2. Mass valuation for taxation purposes

Indeed, mass valuation has been used predominantly for taxation and fee levy purposes. The idea is to tax each property based on its value. This value-based taxation system is well known as*ad valorem*, i.e., based on the estimated value of a real property which includes land and buildings either together or separately (see Portnov *et al.*, 2001). Horne and Felsenstein (2010) observed that property value is the basis for taxation in Brazil, Canada, Denmark, Great Britain, The Netherlands, The Philippines, and most states in the USA. Similar practices are found in other countries. The capital value of property is used for taxation in NSW (Australia), and land taxes are based on the assessed unimproved land value in Taipei (Taiwan) (see Chan and Chen, 2010). The valuation standards from IVSC (2011)stated that the capital value of property is generally associated with market value**,** i.e.**,** the estimated amount for which an asset or liability should be exchanged on the *valuation date* between a willing buyer and a willing seller in an arm's length transaction, after proper marketing and where the parties reach an agreement.

Although it is widely used, property value is not the only base for taxation. Countries with active and well-monitored rental markets are able to use the annual value of property to determine the tax rate. For example, property tax is calculated on the annual value of the property in Singapore; while in Hong Kong, property tax is based on the net assessable value of rental income (see Hui *et al.*, 2004).

2.3. Mass valuation techniques

Moore (2009)has traced the inception of the mass valuation concept:

> Zangerle's book on appraising, published in 1924, established a standard mass appraisal methodology that would be used for the rest of the twentieth century. This book introduced the concept of building quality classifications and included construction specifications for each classification, along with square foot unit rates for each classification.(Moore, 2009, p. 27)

Various techniques have been developed since then, each of which has tried to represent the property market value in a better way.Kauko and d'Amato (2008)classified mass appraisal techniques being used worldwide (Table 2.1).

Table 2.1 Classification of mass appraisal techniques

Approaches	Method	Examples
Orthodox approaches (based on multiple regression)	Cokriging	• Chica-Olmo (2007) • Chica-Olmo and colleagues (2013)
	Spatial Expansion Model (SEM)	• Geoghegan and colleagues (1997) • Bitter and colleagues (2007)
	Hierarchical Trend Modelling (HTM)	• Francke and Vos (2004). • Francke (2008)
	Logistic Regression	• Bolen and colleagues (1999)

In between orthodoxy and heresy (based on flexible regression)	Generalized Additive Model (GAM)	• Pace (1998) • Paceand colleagues (2002)
	Piecewise Parabolic Multiple Regression Analysis (PPMRA)	• Colwell (1998) • Colwell and Munneke (2003)
	Geographically Weighted Regression (GWR)	• Fotheringham and colleagues (1997) • Bidanset and Lombard (2014)
Heresy approaches (based on model-free estimation)	(Artificial) Neural Networks (ANN)	• McCluskey and colleagues (2012a) • Yacim and colleagues (2016)
	Genetic Algorithms (GA)	• Balmann and Happe (2000) • Jirong and colleagues (2011)
	Rule-Based Expert Systems	• Nawawi and colleagues (1997) • Kilpatrick (2011)
	Case-Based Reasoning (CBR)	• Gonzalez and Laureano-Ortiz (1992) • O'Roartyand colleagues(1997)
	Fuzzy Logic	• Bagnoli and Smith (1998) • Pagourtzi and colleagues (2003)
	Rough Set Theory (RST)	• d'Amato (2002) • d'Amato (2007)

Source: Compiled from Kauko and d'Amato (2008) and literature review

2.3.1. Multiple regression methods

Multiple regression analysis (MRA) examines the relationships between one continuous variable of interest (the dependent or criterion variable) and one or more independent (predictor) variables (Miller, 2013). Some of the geospatial methods that have employed this

approach are Cokriging, the Spatial Expansion Model, Hierarchical Trend Modelling, and Logistic Regression. These are discussed below.

Chica-Olmo (2007) suggested Cokriging could be used when house price and the predictor variables have not been sampled in the same housing area – acondition known as heterotopic data. This is a common situation in mass valuation practices worldwide, including in Granada (Spain) where the study was undertaken. First, the location price is estimated using Kriging which calculates the weighted average of neighbouring values (see Şen, 2009). Then the Cokriging method adds the effects of the predictor variables at each location. Chica-Olmo and colleagues(2013) employed Cokriging to develop a multi-equational model in which one equation explains the price of housing in terms of its explanatory variables and the other equation explains the quality of the zone by employing regional variables, e.g. air quality and environmental quality.

The Spatial Expansion Model (SEM) allows the contribution of a housing characteristic to a property's value to change over space, e.g., the value of open (green) space might be higher in urban areas, where open spaces are scarce relative to rural areas (see Geoghegan *et al.*, 1997). Bitter and colleagues (2007) let sevenhousing-attribute variables interact with nine absolute-location variables (derived from the third degree polynomial expansion of the x, y coordinates of a property). Sixty three new independent variables emanated from this analysis. They were added to the seven original variables for housing characteristic, and the nine variables for location, thereby allowing 79 variables to be included in the model. This research was undertaken in Tucson, Arizona. The prediction was most accurate in the area immediately surrounding central Tucson, where housing tends to be less dense and less heterogeneous.

Hierarchical trend modeling (HTM) is a time-series approach(see Francke and Vos, 2004). Francke (2008) stated that the Kalman filter has the ability to produce recursive predictions of the next period's observations based on information up until the present and to provide optimal revision of the trend as time proceeds. Beside this temporal aspect, Francke (2008) explained how the spatial aspect can also be analysed in the model:

> In the HTM, the spatial dependence is modeled on a cluster level basis and by specific locational characteristics. Every cluster has an individual price trend. Within clusters, the price levels may vary over different neighborhoods. The price levels are modelled as random effects within a cluster. (Francke, 2008, p. 166)

Logistic Regression was utilised by Bolen and colleagues (1999)to analyse the spatial distribution of the increase in land value by examining seven variables which were presumed

to affect land value. The aim was to determine whether the conditions that establish land values are also valid for estimating increases in land values, and to establish the probability of increases in land values related to certain characteristics.

2.3.2. Flexible regression methods

Beside the group of MRA-based approaches outlined above, Kauko and d'Amato (2008)identified a group of techniques (Table 2.1) that have an intermediate position between orthodoxy and heresy.They are based on 'flexible regression' which develops flexible functions to fit various situations. The key techniques in this group are Generalized Additive Models (GAM), Piecewise Parabolic Multiple Regression Analysis (PPMRA), and Geographically Weighted Regression (GWR).

A Generalized Additive Model (GAM) estimates the dependent variable as the sum of functions of the independent variables (see Pace, 1998). Each function of the independent variables (i.e., the regressors) is a non-parametric estimation. The functional forms of the independent variables determine the predictive accuracy of the model (Pace *et al.*, 2002). Pace (1998) applied GAM to a set of 442 houses with transactions in Memphis, Tennessee. The samples were divided into sets of modelling data and validation data, and 500 iterations were run on these data. Compared with the global model, GAM reduced the median absolute prediction error by about oneper cent in absolute terms, i.e. about USD1,000 on a house with a USD100,000 transaction price.

In Piecewise Parabolic Multiple Regression Analysis (PPMRA), land is divided into sections. The spatial location and the value of each observation are represented by the four corner vertices of each section (the barycentric coordinates). The value at each barycentric coordinate is weighted on the basis of how close it is to the observations or sales data. The barycentric coordinates are then used as the independent variables(see Colwell, 1998).Colwell and Munneke (2003) applied PPMRA to densely distributed vacant lands sales data in Chicago and suggested that this semiparametric approach was able torepresent very complex price functions. As a result, the model was capableof capturing undulations in the price surface.Kauko and d'Amato (2008)highlighted the need for large numbers of observationsin order to achieve such flexibility in the price surface output. Dense property sales data for mass appraisal work is not always available. Kauko and d'Amato (2008) also noted that the lines between the vertices in piecewise regression are straight lines, so the curve is not differentiable. This makes the model becomes not very flexible.

Brunsdon and colleagues (1996) introduced Geographically Weighted Regression (GWR). GWR allows the coefficients in the model to change at specific locations (calibration points) to comply with local variations within an area of interest. The locations of calibration points are set to give a good representation of spatial variation. Observations are weighted based on their closeness to the calibration points, and GWR provides options on the weighting function. GWR also provides options on bandwidth method to allow alteration of the number of observations involved in a calibration. Bidanset and Lombard (2014) examined how changes in weighting function and bandwidth method in order to suit the circumstances of the data, can improve the accuracy of GWR prediction.

The attractiveness of GWR is that a unique calibration exists for every calibration point in the study area, thus there is a separate regression model at each observation point. Fotheringham and colleagues (1997) noted that global models, by their very nature, are likely to be misspecifications of reality, and that GWR can help to identify the nature of the misspecification by an examination of the spatial pattern of the local parameter estimates. Furthermore, they suggested GWR as a means of incorporating 'unmeasured' effects (e.g., individuals' attitudes or tastes). Although such a factor is not included in the global model, it can be incorporated in the local regression. The effect of the ‘unmeasured’ factor will also contribute to shaping the model.

2.3.3. Model-free estimation methods

The third group of methods specified by Kauko and d’Amato (2008)is labeled ‘heretic’ (Table2.1) because the ‘model-free estimation‘ approach contrasts with the dominant framework of multiple regression analysis. Defining a formal mathematical relationship between the dependent and independent variables is not required in this approach. Among the methods included are Rule-Based Expert Systems, (Artificial) Neural Networks (ANN), Genetic Algorithms, Case-Based Reasoning, Fuzzy Logic, and Rough Set Theory.

The Self-Organizing Map (SOM) and the Multi-Layer Perceptron (MLP) are ANN-based methods that have been used for mass appraisal. White (1989) described the emphasis of the ANN being on learning procedures used to train the Artificial Neural Networks. The addition of more samples keeps the learning procedure active by forming new empirical knowledge. So, the summary of the object being learned is formed iteratively based on the sample itself. ANN provides flexibility in modelling values because it incorporates nonlinearities, and it also provides simplicity because little effort is required in pre-processing data (see McCluskey *et al.*, 2012a). White (1989) explained that empirical knowledge is encoded and converted into the weights of a suitable neural network as some function of the

sequence, thus the resulting network weights are a (vector-valued) statistic. Yacim and colleagues (2016) combined the Cuckoo Search (CS) algorithm with Levenberg-Marquardt (LM) and back propagation (BP) algorithms to reduce the issue of prediction inconsistency. This combination reduces iteration time and results in very high prediction accuracy.

Although ANN-based methods are able to give accurate predictions, Kauko and d'Amato (2008) consider this approach to be a 'black box' approach because there is no clear functional relationship between the input and output values. McCluskey and colleagues (2012a) align with this view in noting that, from an industry perspective, having a transparent and ultimately a defensible model is a prerequisite.

In Genetic Algorithms (GA), a sample is accepted to be as an individual (human being) and can be represented by a set of parameters. The potential solution of a problem is presumed to be an individual with a certain structure of parameters. These parameters are regarded in the same way as genes in a chromosome. The genes of the 'parents' (i.e., the original samples) are mixed and recombined for the production of offspring in the next generation. A 'better' chromosome (i.e., fitter for the expected solution) will create a larger number of offspring, and thus has a higher chance of surviving in the subsequent generation. The 'breeding' and 'natural selection' cycle is repeated until a desired termination criterion is reached (see Man *et al.*, 1996). Jirong and colleagues (2011) developed a hybrid of the Genetic Algorithm (GA) and support vector machine (SVM) for housing price forecasting in China, with the non-parametric kernel function employed in the SVM. The mean absolute percentage error (MAPE) is 1.94 per cent, and this result indicates that the prediction accuracy is very high.

Balmann and Happe (2000) noted that genetic algorithm-based techniques are appealing as they may not be constrained by statistical techniques related to potentially poor-fit models. However, Kauko and d'Amato (2008) underline the factthat GAs are only as good as their data inputs; and like the ANN-based techniques, issues around transparency and capability may be problems.

A Rule-Based Expert System contains information obtained from a human expert, and represents that information in the form of rules. The rules are formulated, and then used to perform operations on data to make inferences in order to reach an appropriate conclusion (see Liao, 2005). Nawawi and colleagues (1997), cited in McCluskey *et al.* (2012b),developed an expert system for mass appraisal in Malaysia, and they argued that the greatest feature of the Rule-Based Expert Systems is their ability to encapsulate rules of thumb or heuristics and generalities. Kilpatrick (2011) utilised this approach for mass

appraisal in Plaquemines Parish, Louisiana and Lomax Township, Illinois where the number of property sales data were not sufficient for regression modeling. Transactions over several years were collected and verified. Appraisal expertise was then applied to the data in order to find common themes in the valuation of properties in a particular area. Adjustment factors were developed to formulate the prediction model for property values in the area. In short, the model was developed based on the data itself along with expert judgment.

McCluskey and colleagues (2012b) noted two major problems with the Rule-Based Expert Systems. First, this technique does not inherently learn but merely mirrors the actions of an expert when it should be able to deduce a solution since the problem to be addressed already contains information within the parameters of the elicited knowledge. Second, the behaviour of large rule-based systems can be difficult to predict because interactions between rules are not obvious.

Case-Based Reasoning (CBR) resembles more closely the psychological processes humans follow when trying to apply their knowledge to similar problems they handled in the past to address a current situation. In its application to mass valuation, descriptions of previously sold properties are stored in a case library. Adjustments are then applied to a property being valued based on the most similar cases (see Gonzalez and Laureano-Ortiz, 1992). O'Roarty and colleagues (1997) built a Case-Based Reasoning model for retail rent determination in Belfast, Northern Ireland, and they concluded that CBR is an effective technique to determine retail rents. CBR can be adjusted over time to incorporate new considerations, thus offering a level of flexibility. Indexing methods can also be used to score the similarities among comparable properties, so it also offers objectivity in selecting and weighting the comparable data. Though CBR provides objective and explainable processes and results, the method requires considerable data volumes (see McCluskey and Anand, 1999).

Fuzzy Logic enables gradual transition in the degree of membership of an element to a group. In the case of mass valuation, an element is sales data, whilst a group is a characteristic of the property as a variable for valuation. Logical and consistent rules are established for each variable, e.g., if the distance from principal destinations is near, then the rating number is low. Each variable and rule-based rating on the variable is converted into quantified fuzzy sets to develop the membership function (see Bagnoli and Smith, 1998). Pagourtzi and colleagues (2003) stated that one of the most important advantages of fuzzy modelling is the hierarchical ranking of the objects (e.g., buildings, lots), thus it is not an inclusion-exclusion list. Similarly, Kilpatrick (2011) considered that the concepts of Fuzzy Sets and Fuzzy Logic are the best ways to inform computers to select not just comparables which are exact matches but instead comparables which are close matches because finding

exact comparables is impossible in real estate. A sample would be the closest comparable if it had the same degree of membership with the property to value, for each valuation variable.

Rough Set Theory (RST) uses the assumption that objects can be 'seen' only through the information available about them. Objects (decision attributes) characterised by the same information (condition attributes) are indiscernible (similar) in the view of available information about them. Although those objects are not exactly the same, they appear to be the same. To deal with imprecision, RST makes use of sets – lower and upper approximations. Lower approximation consists of all objects which definitely belong to a concept, and upper approximation consists of all objects which possibly belong to the concept. Rules developed from lower approximation of the concept are certainly valid, whilst rules induced from the upper approximation are only possibly valid (see Pawlak, 1997). d'Amato (2002) 'approximated' property price (a decisional attribute) by using two conditional attributes: internal area and parking area availability. A decisional table was created to define the causal relationship between the price and the attribute through 'if….then' rules. A 'strong' relationship will derive lower approximations, whilst a 'weak' relationship will derive upper approximations. As a result, deterministic rules show how internal area and parking determine the property price. d'Amato (2007, cited in d'Amato, 2008) developed an application of RST integrated with the valued tolerance relation, and this means an integration between rough sets and fuzzy sets. This application is able to result in a crisp value on the estimated price, while the original RST application requires the estimated price to be a class instead of a crisp value.

2.4. Model choice

Kauko and d'Amato (2008) revealed intheir reviewof mass appraisal methods, that the criteria used by appraisers for selecting the mass valuation method are partly based on methodological considerations and partly on institutional considerations. Adequate and appropriate methods are required to provide accuracy in mass valuation. On the other hand, the suitability of the method for the institutional context of use is equally important. Kryvobokov (2004) concluded that the "…*usual; western valuation methods could not be applied in the Ukraine as the land market was immature",* i.e., the number of land sales was at that time insignificant and most sales involved state land being sold for the first time to private owners after the end of Soviet control. Just like in the Ukraine, a feasibility issue on using the most commonly employed mass valuation methods also prevails in Indonesia. The land market is not transparent (see Tamtomo *et al.*, 2008) and much agricultural land is

being incorporated into highly commercialised urban property markets for the first time because of the recent, rapid urbanisation and urban expansion.

Among the techniques discussed earlier in this chapter, only a few techniques have recently been employed for mass valuation research. It is very difficult to find up-to-date research on mass appraisal utilising the Hierarchical Trend Modelling (HTM), Logistic Regression, Generalized Additive Model (GAM), or Piecewise Parabolic Multiple Regression Analysis (PPMRA). Literature searches indicate that ANN-based techniques and GWR dominate most recent mass appraisal research.

ANN-based techniques offer high prediction accuracy but the processes and the results from these model-free estimation techniques are not easily explainable. This creates issues, particularly in the area of tariffs and taxes. It is a big concern in these areas to have reliable explanation on how the amounts of tariffs and taxes are determined. Moreover, model-free estimation techniques appear to require more data than regression-based techniques because inferences are made from the data themselves.

Contrary to model-free methods, all variables are well explained in orthodox techniques and can be understood by a wide cross section of people. However, these techniques only take into account the predefined information. Only the factors that are well measured are taken as variables in the model. There is the potential for missing important factors because they are not measured. This can be the main reason why Cokriging and Spatial Expansion Method (SEM) have recently been less popular than ANN or GWR. Kestens *et al.* (2006) and Bitter *et al.* (2007) compared the performance between GWR and SEM, and both studies came up with a similar conclusion saying that GWR has a better prediction accuracy.

In the area of mass appraisal, GWR is the most popular among the flexible regression techniques. The GWR approach has assumed greater prominence for price estimation because it isolates and combines spatial dependency and heterogeneity, accounting for locational or adjacency effects and market segmentation (McCluskey *et al.*, 2013).Beside the prediction performance, GWR is suitable for mass appraisal of values because the processes and results are explainable and understandable. In addition to the above reasons, the availability of GWR software within software packages such as ARCGIS adds a further reason to choose GWR. In particular, the section I manage at BPN RI has made significant investments in ARCGIS in terms of software and staff training. The decision to choose GWR for this current research project is backed up with solid reasons.

2.5. Geographically Weighted Regression (GWR)

Regression analysis is used to model the relationship between a dependent variable and its explanatory variables. Ordinary Least Squares (OLS) is a basic form of the regression model.

$$\boldsymbol{y_i = a_0 + \sum_{k=1}^{m}(a_k x_{ik} + \varepsilon_i)} \qquad \textbf{(Equation 2.1)}$$

where $y_i$is the *i*th observation of the dependent variable, x_{ik} is the *i*th observation of the *k*th independent variable, the ε_is are independent normally distributed error terms with zero means, and each a_k must be determined from a sample of n observations.

Only one regression equation is generated in OLS because the relationships are assumed to be more or less the same everywhere within the study area. Each parameter estimate describes the 'average' relationship between the dependent variable and each of the explanatory variables. In certain cases, the relationship between a dependent variable and a particular explanatory variable may differ significantly across space. Fotheringham and colleagues (2002) took an example of local variations in the relationship between a dependent variable and an explanatory variable from housing markets in England. They observed that in rural parts of England, old houses have higher prices than newer houses because they might have character and appeal, while in cities, older houses have considerably lower prices than newer houses because many of those houses were built to low standards for workers at the middle of the nineteenth century. Imposing an 'average' relationship between house price and age of the house would simply ignore the significant local variation and result in an inaccurate model.

In order to generate an accurate model, variations in the relationships between a dependent variable and its explanatory variables should be taken into account. Geographically Weighted Regression (GWR) extends the traditional regression framework of Equation (2.1) by allowing the parameter estimates to vary by location(see Brunsdon *et al.*, 1996). The regression equation then becomes:

$$\boldsymbol{y_i = a_{i0} + \sum_{k=1}^{m}(a_{ik} x_{ik} + \varepsilon_i)} \qquad \textbf{(Equation 2.2)}$$

The local regressions can take place at sampled locations or at predefined locations. Charlton and colleagues (2006)suggested that the intersections of a grid over the study area can be used in analysis with a very large volume of data. This option can reduce computing time and can be beneficial with the mapping of the results. For a local regression at location

i, a search window is used to capture the required samples. All the samples located within the search window are identified and then specified as the subset of data for location i. The search window keeps moving through the study area and stops on each sampled or predefined location. The size of the search window determines the number of samples involved in each local regression.

2.5.1. Search window

The size of the search window can be determined in advance or optimised using the data. The predetermined search window can be specified using the expected number of samples or the maximum search distance. If the predetermined search window is specified using the expected number of samples, the search window captures a constant number of samples at each local regression. If the predetermined search window is specified using the maximum search distance, the search window imposes a constant maximum search distance at each local regression.

The optimisation search window determines a specific number of samples to allow optimal fitting for the model. The bandwidth size of a kernel function acts as the maximum search window. The bandwidth size changes across space to capture the appropriate number of samples, and the changes of bandwidth size alter the shape of the kernel weighting function. So, modifying bandwidth size is a key action to reach the optimal fit of each local model.

Farber and Paez (2007) confirmed that model estimation is sensitive to bandwidth selection both in terms of goodness-of-fit and coefficient estimation. Each sample involved in a certain local estimation is weighted using the kernel function. Modifying the bandwidth will reshape the kernel function, and reshaping the kernel function will adjust the weighting scheme. As a result, the parameter estimates in the local model will change.

In conclusion, modifying the bandwidth size will alter the weighting scheme and may change the number of samples involved in a local model. A particular bandwidth, that captures a specific number of samples and shapes a specific weighting scheme, will result in an optimal fit for the local model. This bandwidth is the optimal bandwidth for a particular local model. The most commonly used bandwidth optimisation techniques for GWR are the Cross Validation (CV) and the Corrected Akaike Information Criterion (AICc) methods.

2.5.2. Cross Validation Method

The Cross Validation (CV) method selects the bandwidth with the lowest CV score as the optimal bandwidth. Páez *et al.* (2011) described the cross-validation score as the sum of squared differences between the observed value of y at i, and the value predicted by a model estimated using kernel bandwidth h after removing observation i from the sample.

CV method for GWR utilises the Leave-One-Out Cross Validation (LOOCV) technique. For a group of n observations, one data point is put aside as the test set and the rest of the data is used as training set. The process is repeated until each observation data has its turn to become a test set, so there will be n test sets and n training sets. One model is established using each training set, and the model is used to calculate the predicted value of the test set. The predicted value is compared to the observed value to measure the residual.

In GWR, the size of training set for each observation is determined by the bandwidth. A pilot bandwidth h is set for a local estimation, and a group of n samples are identified. For local estimation at location i, several steps are undertaken:

a. Observation i is taken out of the group, and the model is fitted using the rest of the samples in the group.
b. The value at location i is predicted using the fitted model.
c. The cross validation score (squared error) on bandwidth h is calculated at location i as follows:

$$\boldsymbol{CVS_i(h) = [y_i - \hat{y}_{\neq i}(h)]^2} \qquad \textbf{(Equation 2.3)}$$

d. The total CV score for bandwidth h is the sum of the local CV scores:

$$\boldsymbol{CVS(h) = \sum_{i=1}^{n} CVS_i(h)} \qquad \textbf{(Equation 2.4)}$$

The above procedures yield the total CV score for all local estimations within a specified pilot bandwidth h (CVS (h)). Multiple values of bandwidth are used to run procedures a, b, c, and d. The bandwidth that comes up with the lowest total CV score is the optimal bandwidth.

2.5.3. Corrected Akaike Information Criterion (AICc)

The Akaike Information Criterion (AIC) was designed to select the best fit model for a given set of data by letting the number of parameter estimates vary among models. Hurvich and colleagues (1998) proposed the Corrected Akaike Information Criterion (AICc), which is, in effect, AIC with a greater penalty for extra parameters. The formula is as follows:

$$\mathbf{AICc} = \boldsymbol{log}\ (\hat{\boldsymbol{\sigma}}^{\mathbf{2}}) + \mathbf{1} + \frac{\mathbf{2}(\boldsymbol{p}+\mathbf{1})}{\boldsymbol{n}-\boldsymbol{p}-\mathbf{2}} \qquad \textbf{(Equation 2.5)}$$

where: $\hat{\sigma}^2$ is the estimated variance of error

p is the number of parameters

n is the number of observations

The preferred model is the one with the minimum AIC score. AICc rewards goodness-of-fit because a smaller estimated error variance encourages a smaller AICc score. On the other hand, the AICc score will increase if the number of estimated parameters increases. This is done to avoid selecting a model that has too many parameters which leads to overfitting. The number of parameters can be calculated by making use of tr(S), which is the trace of the matrix S. The hat matrix (S) complies with Equation 2.6:

$$\hat{\boldsymbol{y}} = \boldsymbol{Sy} \qquad \textbf{(Equation 2.6)}$$

where: ŷ is the matrix of predicted (fitted) values

y is the matrix of observed values

The trace of matrix S (tr(S)) is the sum of the values in the leading diagonal of the hat matrix. In a global model, the number of parameters is indicated by tr(S), while in GWR, the effective number of parameters is estimated using the expression 2tr(S)-tr(STS) (see Charlton and Fotheringham, 2009).

For local estimation at location i, several steps are undertaken:

a. Develop a local model at each location i for a pilot bandwidth h

b. Calculate squared errors for all local models on bandwidth h

$$\boldsymbol{SE_i(h)} = [\boldsymbol{y_i} - \hat{\mathrm{y}}_{\boldsymbol{i}}\ (\boldsymbol{h})]^{\mathbf{2}} \qquad \textbf{(Equation 2.7)}$$

c. Calculate the variance of residuals for all local models on bandwidth h

$$\hat{\boldsymbol{\sigma}}^{\mathbf{2}}(\boldsymbol{h}) = \frac{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{SE_i(h)} - \overline{\boldsymbol{SE}}(\boldsymbol{h}))^{\mathbf{2}}}{\boldsymbol{n}} \qquad \textbf{(Equation 2.8)}$$

d. AICc function is applied to multiple values of bandwidth, and the bandwidth with the lowest AICc score will be the optimal bandwidth.

2.5.4. Weighting schemes

The weighting procedures in GWR comply with the first law of geography, i.e., everything is related to everything else, but near things are more related than distant things (Tobler, 1970). The weight of an observation decreases by the distance from the regression location. Páez and Wheeler (2009) observed that most applications of GWR have favoured continuous functions that produce monotonically decreasing weights, such as the negative exponential:

$$\boldsymbol{w_{ij} = \exp\left(-\frac{d_{ij}^2}{2h^2}\right)} \qquad \textbf{(Equation 2.9)}$$

or the bi-square kernel:

$$\boldsymbol{w_{ij} = \left(1 - \frac{d_{ij}^2}{h^2}\right)^2} \qquad \textbf{(Equation 2.10)}$$

where: w_{ij} is the weight assigned to observation j;

d_{ij} is the distance between observation j to regression location i; and

h is the bandwidth

A description of the weighting scheme using a negative exponential function that shapes a Gaussian kernel is shown in Figure 2.1.

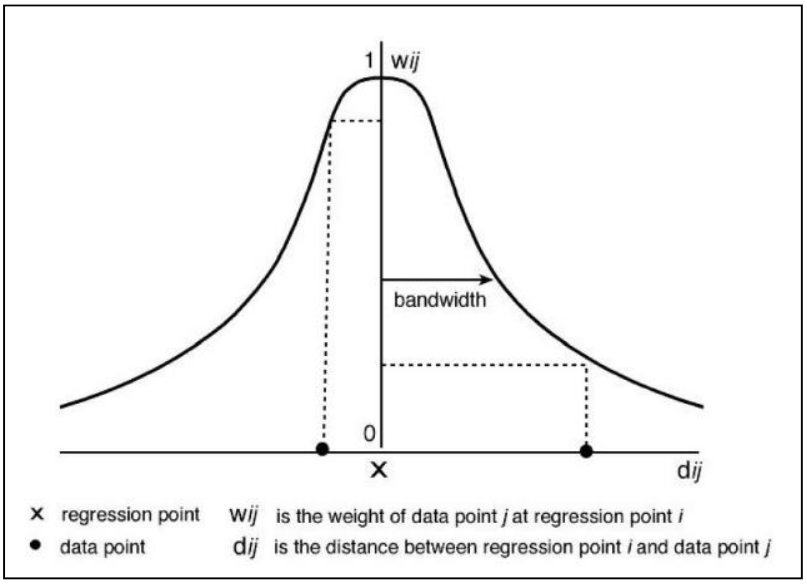


Figure 2.1 Weighting scheme at local regression

Source:Fotheringham *et al.* (2002)

Charlton and Fotheringham (2009)suggested that in terms of influencing the fit of the model, the choice of a bandwidth is more important than the type of the kernel. In ArcGIS implementation, Gaussian kernel type is used for the fixed radius kernel and bi-square kernel type is used for the adaptive kernel.

2.5.5. Weighted Local Regression

Charlton and Fotheringham (2009) specified the equation for a typical GWR version of the OLS regression model as follows:

$$\boldsymbol{y_i(u) = \beta_{0i}(u) + \beta_{1i}(u)x_{1i} + \beta_{2i}(u)x_{2i} + \cdots + \beta_{mi}(u)x_{mi}} \quad \textbf{(Equation 2.11)}$$

where: u represents the location of local regression;

i represents the i^{th} observation within a subset of n samples involved;

$\beta_0(u)$, $\beta_1(u)$, $\beta_2(u)$, ……. ,$\beta_m(u)$ are parameter estimates at location u;

x_1,x_2, x_3,…….,x_m are the independent variable; and

y is the dependent variable.

The estimator will be:

$$\boldsymbol{\hat{\beta}(u) = (X^T W(u) X)^{-1} X^T W(u) Y} \quad \textbf{(Equation 2.12)}$$

Basically, this estimator is similar to the Weighted Least Squares (WLS) global model. The weight component distinguishes GWR from WLS. The weight is calculated using a kernel function, and it applies to a specific location only (see Charlton and Fotheringham, 2009).

The X, Y, and W matrices are required for parameter estimation. For m independent variables and n observations involved in a local regression at location u, the X matrix will be:

$$\begin{bmatrix} x_{11} & x_{21} & x_{31} & \cdots\cdots & x_{m1} \\ x_{12} & x_{22} & x_{32} & \cdots\cdots & x_{m2} \\ x_{13} & x_{23} & x_{33} & \cdots\cdots & x_{m3} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{1n} & x_{2n} & x_{3n} & \cdots\cdots & x_{mn} \end{bmatrix}$$

The Y matrix contains the dependent variable, and for n observations the matrix will be:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots\ldots \\ y_n \end{bmatrix}$$

For n observations involved at location u, the W**(u)** matrix will be an n x n matrix containing the weights in the leading diagonal and 0 in the off-diagonal elements:

$$\begin{bmatrix} w_1\mathbf{(u)} & 0 & 0 & \ldots\ldots & 0 \\ 0 & w_2\mathbf{(u)} & 0 & \ldots\ldots & 0 \\ 0 & 0 & w_3\mathbf{(u)} & \ldots\ldots & 0 \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & 0 & \ldots\ldots & w_n\mathbf{(u)} \end{bmatrix}$$

In regression calculation, the weighting matrix is the key feature that differs GWR from OLS. A unique weighting matrix is formed at each local regression of GWR. For instance, distance to tollgate can be assigned as variable number 1. w1 in inner city area, is very likely to be much bigger than w1 in the countryside because distance to tollgate in inner city area is significantly more influential than in the countryside.

2.6. Summary

A mass valuation technique is employed to value a huge number of properties in a city or district using a minimum allocation of budget and time. The output of mass valuation works are predominantly used for taxes and tariffs purposes. In an *ad valorem* (value-based) taxation system, the amounts of taxes and tariffs related to a property are determined based on its value. Accurate predicted values of properties are required to ensure a proper value-based taxation system.

In order to come up with accurate predictions, various mass valuation techniques have been developed worldwide. Mass valuation techniques can be categorised into three groups, i.e. regression-based techniques, flexible regression techniques, and model-free estimation techniques. The flexible regression approach is the most favourable because it provides an examination of the relationship between price and the explanatory variables but then it develops flexible functions to fit various situations. Among the techniques in the flexible regression group, GWR has been the most popular. Another important consideration to choose GWR is that GWR software is available within the ARCGIS software package. BPN RI has made significant investments in ARCGIS in terms of software and staff training.

Unfortunately, the GWR model cannot be run on the Bekasi dataset right away because the dataset is not prepared for a regression-based analysis. The currently employed Zonation Method only uses variable price in the analysis, while GWR also takes into account the explanatory variables for land price. The data related to the explanatory variables are only recorded at the sampled locations. For the non-sampled locations, most of the data related to explanatory variables must be derived through spatial analyses.

Data preparation is undertaken to get all the data related to all of the explanatory variables become available at each parcel in the study area. Preparing the Bekasi dataset for regression-based analysis is found to be a substantial work. The basic data from Bekasi are the property sales data and the land parcel map. There are a huge number of trivial issues within these data, and they make data preparation very time consuming. The processes, the issues, and the results of data preparation are discussed in the next chapter (Chapter Three).

3. DATA PREPARATION

3.1. Introduction

The overall methodological approach undertaken in this study is described in figure 3.1.

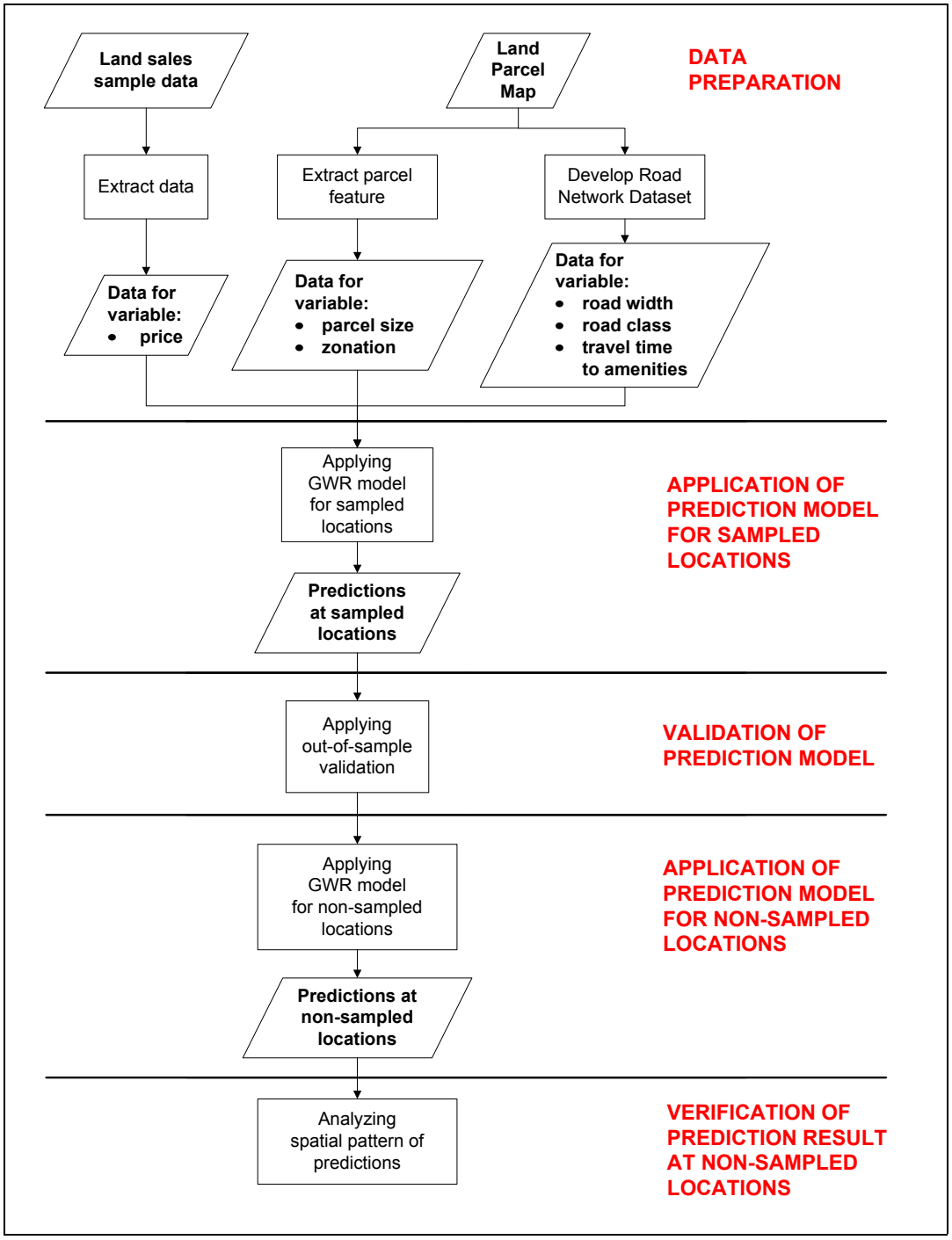


Figure 3.1. Overall methodological approach

The main datasets required for the application of GWR to mass valuation in Bekasi are property sales data, Land Parcel Map, and road network data. The property sales data contains the sales price and information related to the property. In this research, each individual sales data information will be matched with the corresponding parcel in the Land Parcel Map because the analysis will be run in the unit of land parcel. The data on variables related to information on property is obtained from the sales data compilation, while the data on variables related to accessibility (travel distance and travel time to amenities) will be determined using the road network data.

As discussed in Section 2.6, the GWR model cannot be run on the Bekasi dataset right away because the datasets are not prepared for a regression-based analysis. Adaptations are applied to the datasets. During the process of data adaptations, errors have been found in the datasets. Fixing the errors therefore becomes a dominant part of data preparation process and is outlined in this chapter.

3.2. Samples from Bekasi

As identified in Section 1.2, collecting property sales data is usually very challenging in every mass valuation project in Indonesia. Due to this reason, field surveys normally produce sparse data distribution. Distribution of samples in the 2012 Bekasi dataset is the best amongst all of the mass valuation projects administered by BPN RI.[6] This is the main reason behind the decision to use Bekasi as the study area in this research. In addition to that, the author of this research thesis was the survey coordinator in the 2012 Bekasi Mass Valuation Project.

The main information extracted from property sales data is the land value per square metre. The value of a building or a structure on a land parcel is estimated with the help from local contractors. This estimated building value is deduced from the transaction price to extract only the land value. Distribution of samples and intervals of observed land values are displayed in Figure 3.1.Samples of medium to high land values are mostly located around the city centre, while most of the samples of low valued lands are located in the south-east corner of the city. Though located in the outskirts, some samples in the west have medium to high land values. The closeness of this part of Bekasi to Jakarta (the capital city) (Figures 3.2, 3.3) is presumed to be the main reason behind this.

[6] Part of this information comes from the author's experience as a surveyor in a number of mass valuation projects administered by BPN RI.

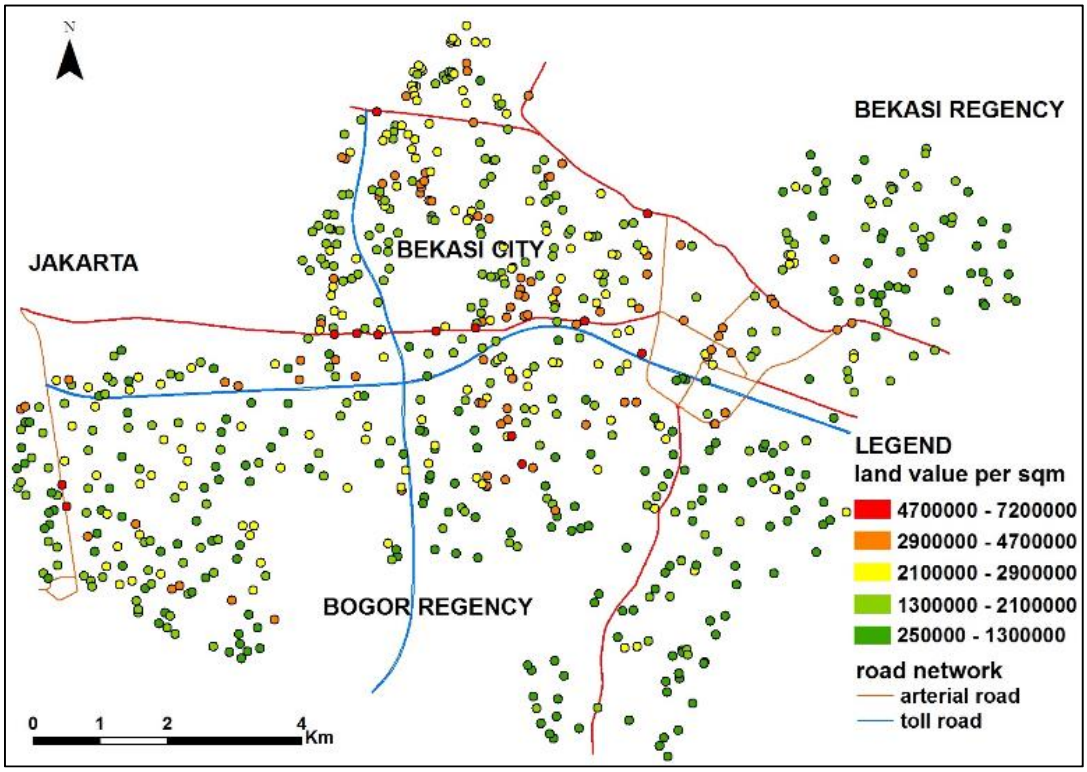


Figure 3.4 Sample distribution of 2012 Bekasi Dataset[7]

Figure 3.3 shows that Bekasi City is located in the south eastern fringes of Jakarta. Although Bekasi was once a city in its own right, Jakarta has grown towards it. A 2014 survey by BPS(2015) reported that 460,069 of Bekasi's 2.3 million residents (19.8 per cent) are commuters, of which 78.2 percent travel to Jakarta for either work or education. The survey conducted by BPS (2015) also reported that just over half of the commuters, 55.0 per cent, have to travel between one to two hours to their workplaces or education establishments, and 13.3 per cent travel for more than two hours.

East Jakarta is the part of the Jakarta metropolitan area adjacent to Bekasi (Figure 3.3). The Land Offices' Land Value Maps for 2012 show that, on average, land parcels in East Jakarta were worth almost twice as much as the land parcels in Bekasi (Table 3.1). Property prices are normally in line with land prices. Assuming this is the case in Bekasi, the property price difference between East Jakarta and Bekasi is likely to be the main reason for people who work in Jakarta to buy houses in Bekasi.

[7]The intervals of land value are rounded, and there is no sample that has a land value exactly the same with any of the break values

Figure 3.5 Location of Bekasi City

Source: BIG (2013)

Table 3.1 Land price comparison between Bekasi and East Jakarta in 2012

	Lowest land price in 2012 (IDR)	Highest land price in 2012(IDR)
Bekasi	299,000	6,338,000
East Jakarta	535,000	12,282,000
Price difference (IDR)	236,000	5,994,400
Price difference (per cent)	78.9	93.8

Note: The price is per square metre and IDR is Indonesian Rupiah

Source: Land Office of Bekasi (2012c) and Land Office of Jakarta Selatan (2012)

3.3. Land Parcel Map

An excerpt from the digital Land Parcel Map of Bekasi is displayed in Figure 3.4. The map has been maintained in the local Land Office. The main feature is the polygon of registered parcels but the map also has important real world features like roads, railways, village boundaries, etc.

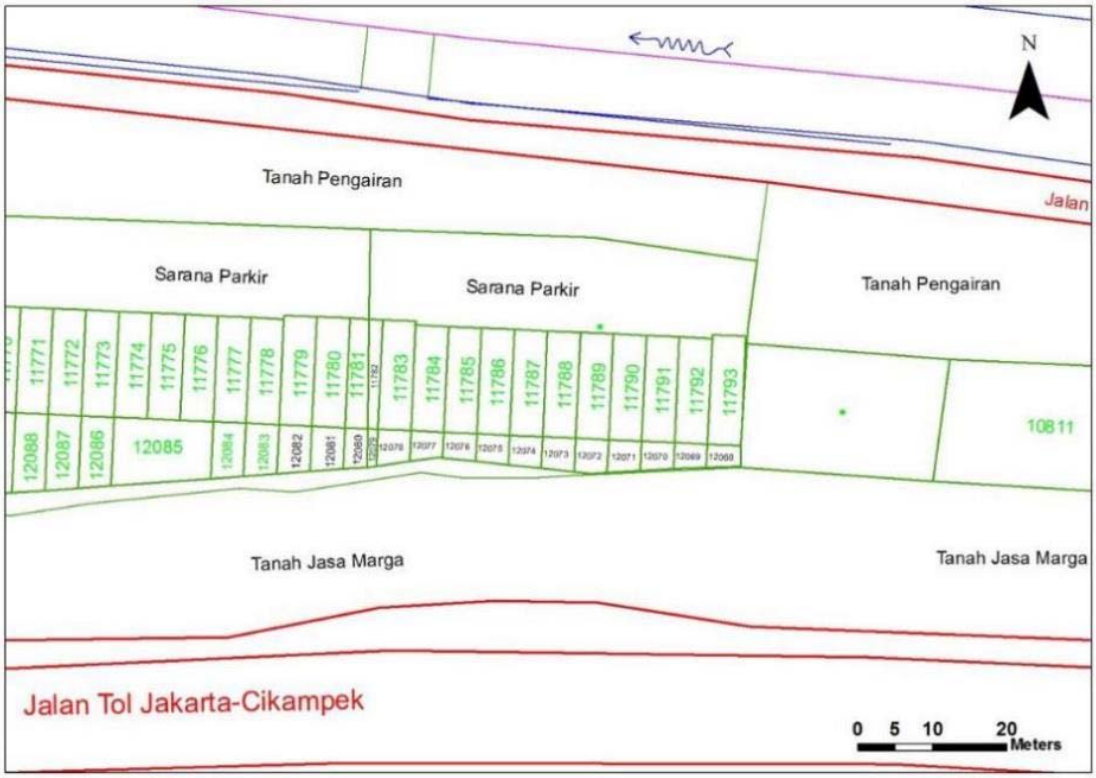


Figure 3.6 Excerpt from Land Parcel Map of Bekasi

Source: Land Office of Bekasi (2012a)

Note: 'Tanah Pengairan' is irrigation interspace
'Sarana Parkir' is parking area
'Tanah Jasa Marga' is land owned by 'Jasa Marga', a government-owned toll road operator
'Jalan Tol Jakarta-Cikampek' is a Toll Road connecting Jakarta and Cikampek
'00000'is the land parcel identification number

The map has been prepared mainly for land titling work and not for spatial analysis. The many issues encountered in using this map in this research are outlined below. For example, a land parcel is not represented as a polygon feature type but as a polyline. What makes it worse is that there is no link between the polyline feature and the attribute data of the corresponding land parcel represented by the polyline. The parcel number, which is one of the attribute data of a land parcel, has been created in an annotation layer with no link to the corresponding feature data. Other issues arise not only from the parcel layer but also from other layers. In short, the digital map is more like a precise drawing than a database. For this study, data conversions and data matching tasks were carried out to prepare the data for spatial analysis.

In most Indonesian cities, the Land Office's Land Parcel Map is not reliable for spatial analysis because of the lack of coverage. In addition to the coverage issue, the Land Parcel Map of Bekasi also has content issues (Figure 3.5).

The Land Parcel Map only contains the land parcels registered at the local Land Office. Therefore, not all of the land parcels located in the area of interest can be involved in analysis because some of them do not exist in the map. The exact number of parcels in Bekasi cannot be determined using this map either.

Thousands of drawing errors were detected in the Land Parcel Map of Bekasi, mostly dangling and overlapping lines. These errors have not been considered as serious issues because the map was not prepared for spatial analysis. For this study, the polygons of blocks must be completed to create gaps between blocks. The gaps between 'blocks' are then used to create road segments.

Misplacement of features also occurs in the Land Parcel Map of Bekasi. The image shows some cases of misplacements in which several land parcels are placed over a toll road. The toll road is a new one and, apparently, the misplaced land parcels were expropriated for the toll road project. In this case, these parcels are going to be deleted.

Figure 3.7 Issues in Land Parcel Map

Source: Land Office of Bekasi (2012a)

3.4. The road network

In addition to providing road class and road width data for each land parcel, road network data is also used as the basis for calculating travel distance and travel time from each land parcel to each of the amenities listed in the form for data collection. For this research, road network datasets were obtained from OpenStreetMap online map and from Indonesian Geospatial Information Agency (BIG). In the latter case, the road network is one of the layers in Topographic Map. Mismatches were found when each of the road network datasets was overlain on the Land Parcel Map. Figure 3.6 shows the mismatches between road network data from OpenStreetMap online map and the road edges from the local Land Office's Land Parcel Map.

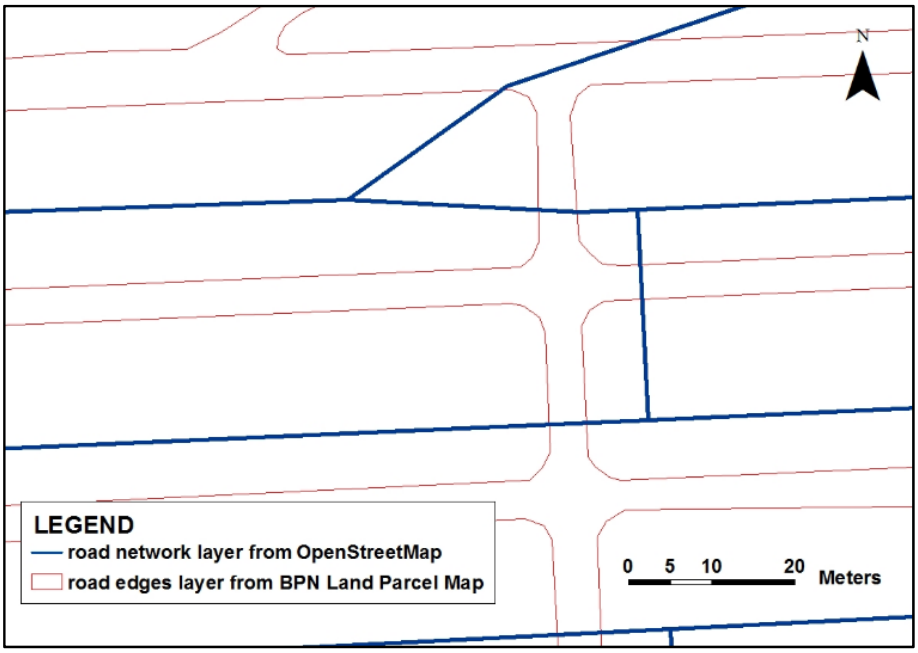


Figure 3.8 Mismatches between OpenStreetMap and Land Parcel Map

Sources: OpenStreetMap (2015) and Land Office of Bekasi (2012a)

To measure the displacements from the road network layer of OpenStreetMap, centrelines have been created from the road edges layer in the Land Parcel Map (Figure 3.7). In Figure 3.7, the shift between A and A' is significantly different from the shift between B and B', in the X and Y axes. This indicates that one measure of translation and rotation will not be able to match the road network layer of OpenStreetMap to the Land Parcel Map. A huge number of spatial adjustments will therefore be required to adjust the road network layer from OpenStreetMap to fit it to the Land Parcel Map.

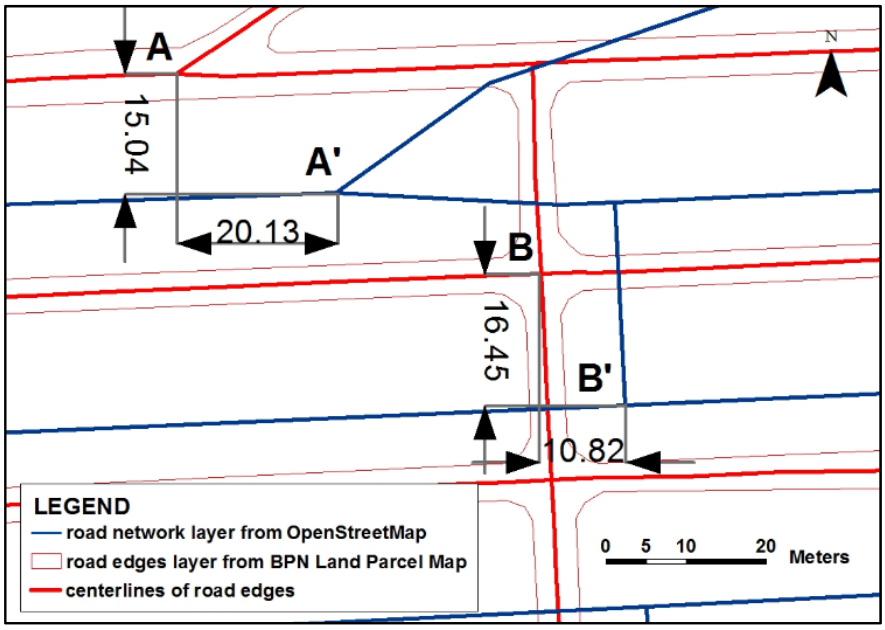


Figure 3.9 Dimension of mismatches between OpenStreetMap and Land Parcel Map

Sources: OpenStreetMap (2015) and Land Office of Bekasi (2012a)

Mismatches were also clearly identified (Figure 3.8) when the road network layer from the BIG Topographic Map was overlain on the road edges layer from the local Land Office's Land Parcel Map.

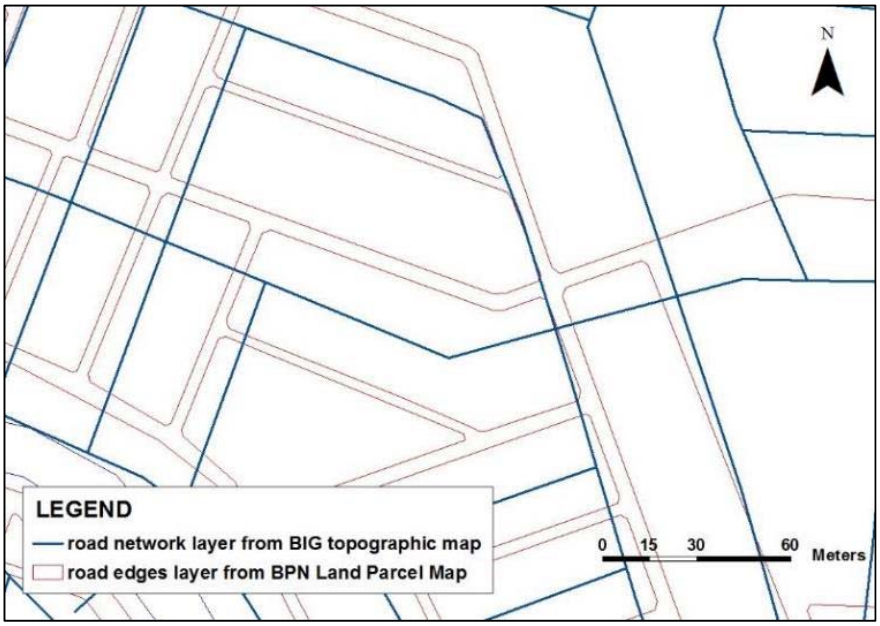


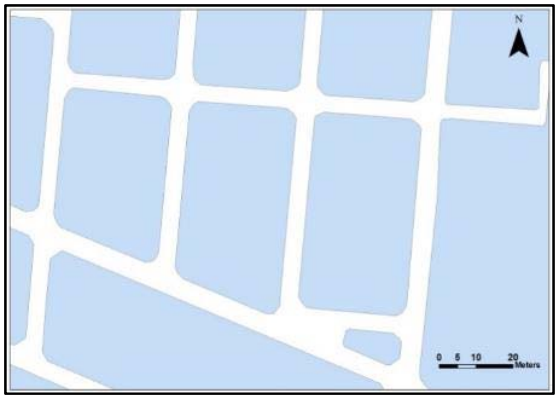
Figure 3.10 Mismatches between Topographic Map and Land Parcel Map

Sources: BIG (2013) and Land Office of Bekasi (2012a)

The mismatches differ across the entire area, and a single transformation formula cannot be used to fit the datasets to one another. Without considerable amounts of manual processing, which was not feasible in this project, nor would it be within BPN, it is obvious that the road network datasets obtained from OpenStreetMap or BIG cannot be used in this project. Therefore, the road network data was developed by making use of the Land Parcel Map.

Derivation of road network data from the cadastral map was also conducted by previous researchers. Haunert and Sester (2008) applied the method of collapsing an area into a straight skeleton to derive the centerlines of the road network from a cadastral map in Hildesheim, Germany, and most junctions (89.8 per cent) were appropriately remodelled in this work. Similar work was done by Zhang *et al.* (2010) in Barcelona, Spain, and 97 per cent of road segment reconstructions were reasonable. These previous works give more confidence to derive road network data from the Land Parcel Map in Bekasi, and the processes are shown in Figure 3.9.

The green polylines are the boundaries between land parcels, while the red polylines are the outmost borders of blocks which become the edges of road segments.



The polylines of parcels and polylines of road edges within one block are converted into polygons and dissolved into one block polygon. The results are block polygons (in blue). The idea is to capture the gaps between blocks. In the real world, these gaps are road segments (in white).

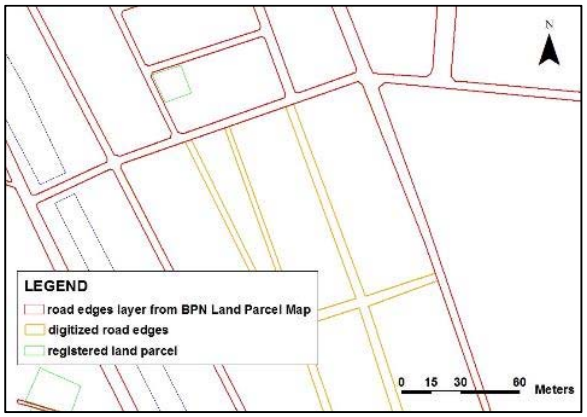
Centrelines are then derived from the gaps between the blocks. The width of each road segment is determined by measuring the distance from the centreline to the road edges. Because each centreline divides each gap into identical halves, the measured distance is half of the road segment's width.

Figure 3.11 Developing road network layer from Land Parcel Map

Sources: Land Office of Bekasi (2012a) and data analysis

Deriving the road network data using the above processes should have been completed in a relatively short time. However, the processes outlined above cannot always be accomplished without adaptations. Figure 3.10 shows one of the cases when adaptation is required.

Both land parcel layer and the road edges layer can be used to capture the gaps between blocks. In this area, road edges are the preferred choice as the number of registered land parcels is very low. Unfortunately, road edges are not delineated in several blocks. At these blocks, the absence between block features will result in no road segments being created. Taking the data as it is means missing a number of actually existing road segments.



The road network data will not represent the real world situation if a number of existing road segments are disregarded. On-screen digitation is required to delineate the road edges for these blocks. The extra polylines of road edges help to capture more gaps between blocks. With more gaps being captured, more road segments can be created.

Figure 3.12 Delineation of missing road edge features

Sources: Land Office of Bekasi (2012a)and data analysis

Beside the issue of missing road edge features, the issue of drawing errors and misplacements described in Figure 3.5 must also be addressed. Dangles, overlaps, and misplacements will cause errors in the process of generating road network dataset.

3.5. Travel distance

The distances from a property to amenities are key variables in determining land values. Euclidian distance is the simplest measure of distance because it is measured along a straight line connecting the objects. A straight segment in a rectangular x and y Cartesian coordinate system is used to represent the Euclidian distance between two points in a plane (see Li and Klette, 2011). Another common distance measure is the Manhattan distance. It is the sum of the straight line distances parallel to the x-axis and the straight line distances parallel to the y-axis (see Chan, 2005).

In this study, neither the Euclidian nor the Manhattan distance was employed to measure travel distance from a property to amenities. Euclidian distance neglects the reality that people travel through the road network. Manhattan distance cannot be used either because it is not quite appropriate at certain situations, as shown in Figure 3.11.

Between the two parcels (blue polygons),
Euclidian distance:
156.39 metres
Manhattan distance:
151.57 + 45.47 = 197.04 metres

Between the two parcels (blue polygons), the shortest route can be travelled along the road network = 1,420.33 metres

Figure 3.13 Comparing Euclidian, Manhattan, and Route distances

Sources: Land Office of Bekasi (2012a)and data analysis

Route distance is a more realistic measure as in cities, people travel through the road network which is often defined by irregularly shaped segments. Using route distance can result in considerably different distances to amenities compared with using Euclidian or Manhattan distance. The above case (Figure 3.11) shows that route distance can be over seven times larger than the Euclidian distance or Manhattan distance. This means that using the Euclidian or Manhattan distance may lead to a huge inaccuracy.

With around 200,000 road segments in the study area, multiple possible routes are available when travelling between locations. An optimum route can be computed for this research using the *route analysis* tool. It can be the shortest route, fastest route or other type of

optimum route which is set according to a specified optimisation option. A cost (impedance) must be set for a route, and this cost will be minimised in the process of selecting the optimum route. Setting distance as the cost will result in the shortest route being the optimum route, because the distance will be minimised during route selection. Travel time is also a common cost in route optimisation. Minimising travel time will give the quickest route as the optimum route.

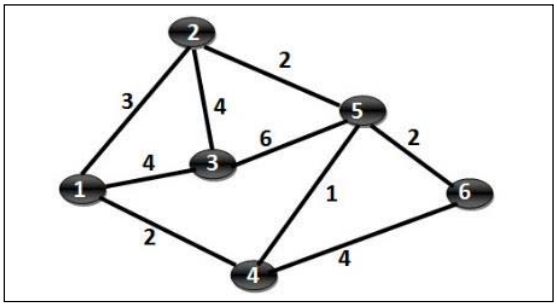
Route analysis is based on Dijkstra's algorithm. This algorithm uses the nodes in the network dataset as the unit of analysis. The basic idea is to run an iterative process to find the shortest path from one node to the rest of the nodes in the network. From a starting point, the algorithm looks for a neighbouring node with the shortest distance (v). Then, it examines each neighbour (w) of v that satisfies the following constraint:

$$\boldsymbol{d(w)} \;\leq\; \boldsymbol{d\,(v)} \;+\; \boldsymbol{c_{vw}} \qquad \textbf{(Equation 3.1)}$$

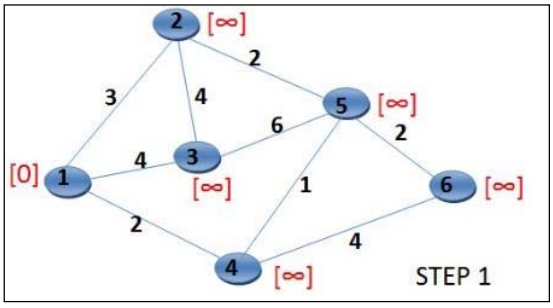
where: C_{vw} is the distance between nodes v and w.

Whenever the constraint is not satisfied, the shortest path to w is improved using the new known value d.v/Ccvw. This is repeated until each node has been marked as completed, after which the algorithm returns the vector of shortest paths (see Oliveira and Pardalos, 2011).

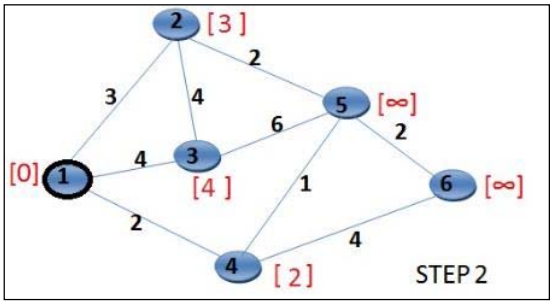
A worked example of shortest path optimisation is explained in Figure 3.12.



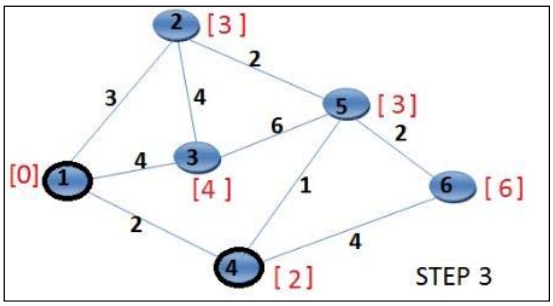
The shortest paths to all nodes in the network from node 1 are to be determined.



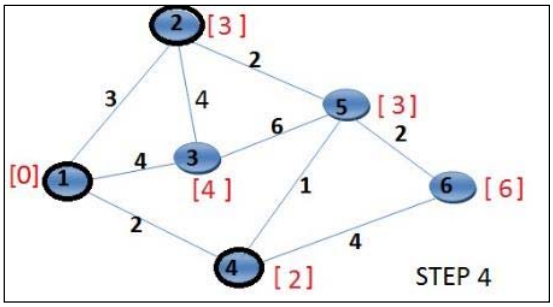
A cost of [0] is assigned to node 1, and [∞] is assigned to other nodes.



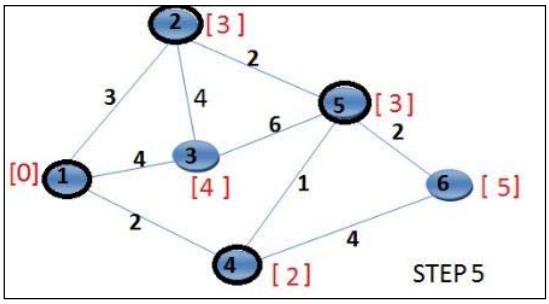
Node 1 (the starting point) is marked as visited. The costs to visit the directly connected neighbours (nodes 2, 3, and 4) are calculated. In this example, the costs are [3], [4], and [2] respectively. Node 4 has the lowest cost among the unvisited nodes, so node 4 will be visited.



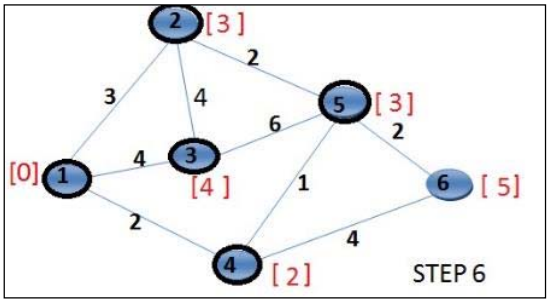
From node 1, node 4 is visited and marked. Then the total costs from node 1 via node 4 to visit the connected neighbours (nodes 5 and 6) are calculated. The total costs are [3] and [6] respectively. Nodes 2 and 5 have the lowest costs among the remaining unvisited nodes. Node 2 will be visited first as it is directly connected to node 1.



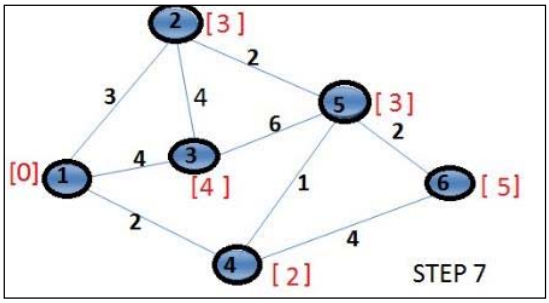
From node 1, node 2 is visited and marked. The total costs from node 1 via node 2, to the connected neighbours (nodes 3 and 5) are calculated. The total costs are [7] and [5] respectively. Previously the costs for nodes 3 and 5 were [4] and [3] respectively. The previous costs are lower, so they are retained as the minimum costs. Node 5 has the lowest cost among the unvisited nodes, so node 5 will be the next to visit.



From node 1, node 5 is visited via node 4 and marked. The total costs from node 1 to visit the unmarked neighbours (node 3 and 6) are calculated. The total costs are [9] and [5] respectively. Previously the costs at node 3 and 6 are [4] and [6] respectively. The old cost at node 3 is lower, so it will be kept as the minimum cost. The cost at node 6 is updated because the new cost is lower.



Node 3 has the lowest cost among the unvisited nodes, so node 3 will be the next to visit. Node 3 will be directly visited from node 1, and node 3 will be marked as visited as well.



Node 6 will be marked as visited because all the other nodes were already marked. The most recent cost calculation will be saved as the cost to determine the optimum route. The most optimum routes to visit each node from node 1 are:

Node 2: 1 – 2
Node 3: 1 – 3
Node 4: 1 – 4
Node 5: 1 – 4 – 5
Node 6: 1 – 4 – 5 – 6

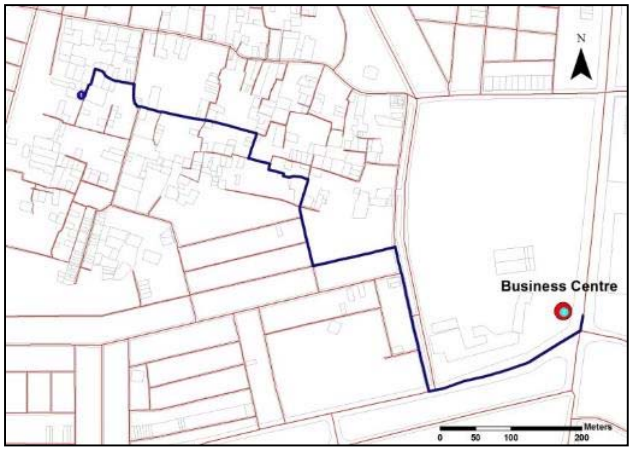
Figure 3.14 Worked example of shortest path optimisation

Source: Summarised from Prasad (2015)

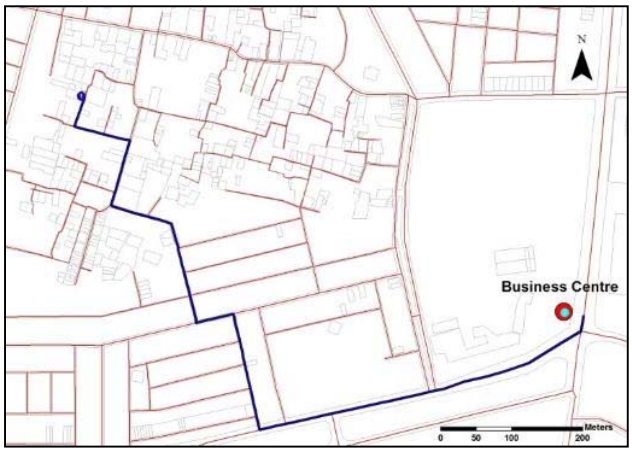
In order to run *route analysis*, the road network feature must be converted into a network dataset which contains road segments and nodes (endpoints of road segments). Turns, impedances, restrictions, stops, driving directions, and other traffic data can be accounted for in a network dataset. An issue with using distance as the measure of accessibility is that it does not account for the contribution of road classification. Driving between two locations through two routes with different road classes may result in significantly different travel times between the two routes. Therefore, travel time is a more objective measure of the accessibility of a property to amenities and services, than travel distance. In addition to that, the Internal Standards for Land Valuation in BPN RI[8] actually advises the use of travel time as a measure of accessibility.

[8]Standar Operasional Prosedur Internal Survei Potensi Tanah, BPN RI(2013).

The shortest route analysis uses distance as the travel cost, while the quickest route analysis uses travel time as the travel cost. An actual example is given to compare the optimum routes for both costs (Figure 3.13).



A network analysis was run to determine the optimum route from a land parcel to the business centre. The shortest travel distance was resolved because distance was set as the cost for the route.



For the same set of origin and destination, the optimum route differed significantly when travel time was set as the cost for the route. The optimum route was the one with minimum travel time.

Figure 3.15 Comparing routes from travel distance optimisation and travel time optimisation

Source: Data analysis

3.6. Travel time

In this study, travel time was set as the travel time using a motor vehicle. Predominantly, people travel to amenities using motor vehicles. There are amenities located within walking distance from a number of houses. There are also amenities which can be reached faster by train, from a number of houses. The most dominant travelling method, using a motor vehicle, was selected here to measure travel time to amenities.

3.6.1. Travel times in busy and quiet times

In Indonesian cities, travel time varies significantly between busy and quiet traffic conditions. In order to represent the nature of the traffic flow and congestion effects, travel time was analysed for busy and quiet times. Traffic is at its busiest when students travel to schools and workers travel to their workplaces. The common school hours in Bekasi are generally from 7 am to 1 pm, while work hours are normally from 8 am to 5 pm. Due to the movement of people associated with work and education, the busiest time in the morning is between 6.30 to 7.30 am. The traffic is less busy when people are at school and at their workplaces, this is between 9 am to 1 pm, however the traffic is sometimes busier during the lunch break. So, between 9 am and 12 pm is the time when the traffic is less busy. Most roads are relatively quiet between 11 pm to 4 am but this period is not the usual time for most people to travel. The quiet period should therefore be picked out from the common travel period which should be in line with the school or work hours.

3.6.2. Average travel speed by road class

Figure 3.14 shows the route normally taken by the residents from a sampled location number 730 to get to the nearest arterial road. They travel along a primary collector road (in yellow, orange dashed route) to get to the nearest arterial road (in red).

Figure 3.16 Example of measuring travel speed on a road class

Source: Data analysis

The amount of time to travel along the orange dashed route at busy or quiet time was obtained at the interview by the BPN RI field team (as it was for all other properties sampled), while the travel distance was acquired from the road network dataset. The speed that the resident drives along the primary collector was obtained from the distance/time ratio for both busy and quiet times.

The above procedure was repeated for 49 other locations to obtain the average travel speed along primary collector roads. Boxplot filtering was applied to determine the outliers among the samples, so that they can be excluded from the next calculations.

Average travel speeds were also calculated for the other road classes, i.e. local, secondary collector, secondary artery, primary artery, and toll road. The results are listed in Table 3.2.

Table 3.2 Average travel speed on each road class

Road class	Average travel speed – busy time (km per hour)	Average travel speed – quiet time (km per hour)
Local	7.52	13.26
Secondary collector	8.78	14.22
Primary collector	7.64	16.29
Secondary artery	7.39	23.86
Primary artery	9.92	26.75
Toll road	36.92	80.00

Source: Data Analysis

The average travel speed at quiet time increases by the road class. This is rational because a road segment of higher class is expected to allow higher travel speed than a road segment of lower class. Nevertheless, this logic does not apply for average travel speed at busy time. There is not much difference among average travel speed on local, primary collector, and secondary artery road classes. A road segment of higher class does not allow significantly higher travel speed than a road segment of lower class. The severity of traffic congestion across the city is very likely to be the main reason behind this.

A route may pass through various road classes, as exemplified in Figure 3.15.

Figure 3.17 Example of a route passing through multiple road classes

Source: Data analysis

From point 1 (purple dot), one travels to point 2 (light blue dot) through the fastest route (red line) which includes local road (black line), secondary collector road (green line), and primary collector road (yellow line). Each road class allows people to travel at a different speed. The proportion of each road class in the total travel distance differs from the proportion of each road class in the total travel time (Figure 3.16).

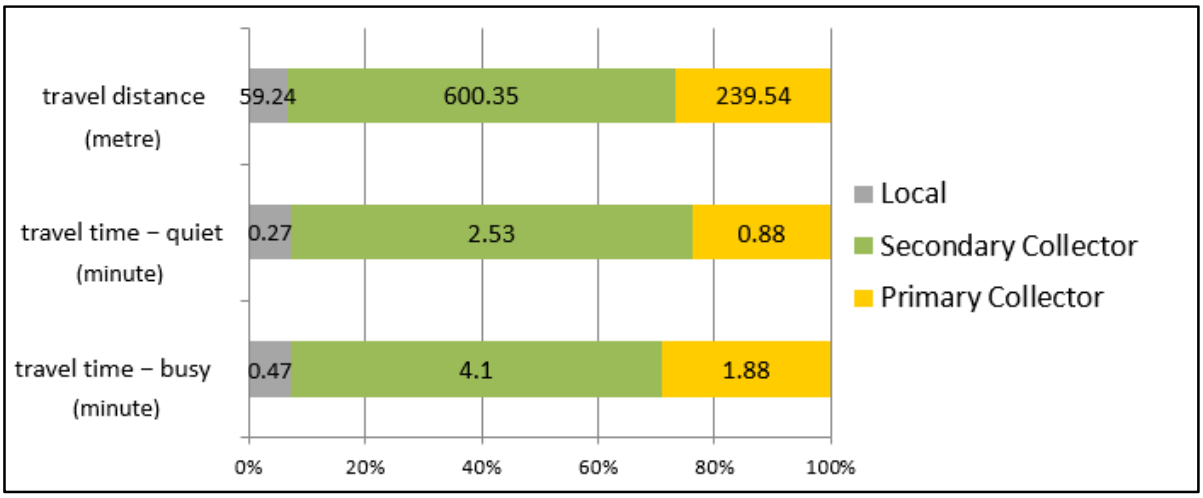


Figure 3.18 Example of proportion on travel time and travel distance of a route passing through multiple road classes

3.6.3. Assigning travel times to amenities for each land parcels

In the Land Valuation Directorate's survey form, travel times to amenities are divided into within five, 15, and 30 minutes and over 30 minutes intervals. The measures of accessibility for those intervals are defined as very good, good, moderate, and poor. Figure 3.17 shows the areas within five, 15, 30, and over 30 minutes of travel times to the business centre of Bekasi in busy times.

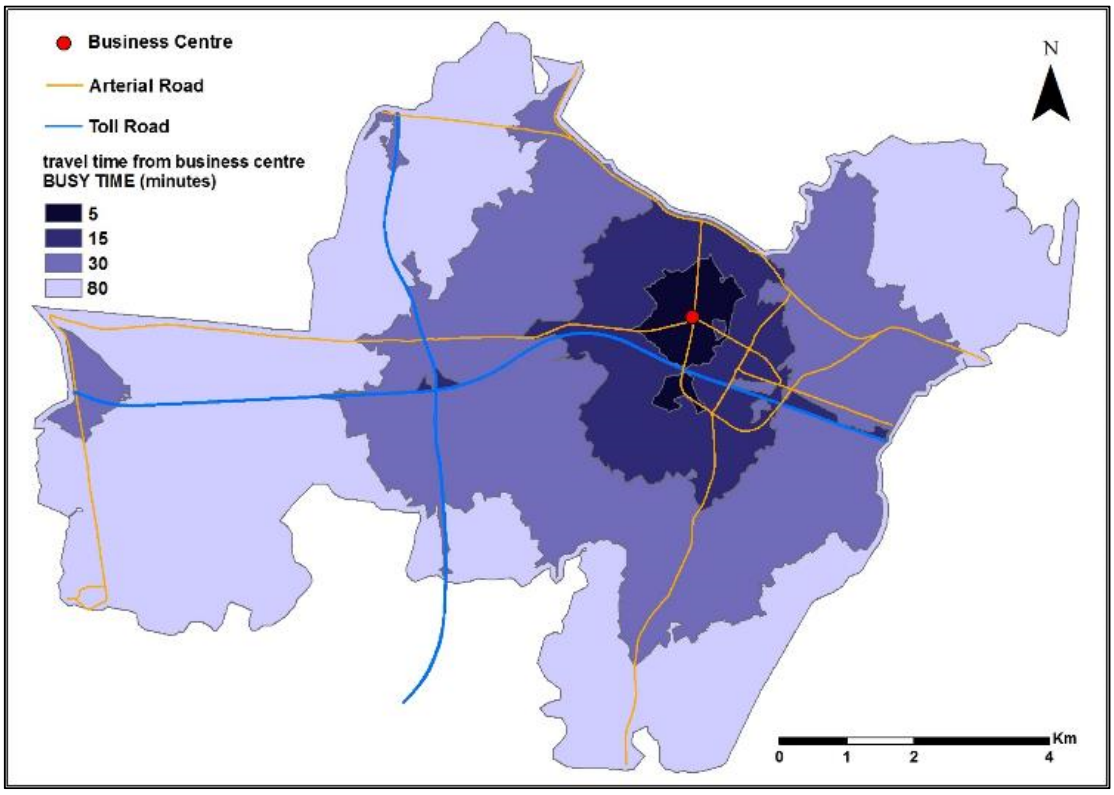


Figure 3.19 Intervals of travel time to business centrein busy time

Source: Data analysis

A property located within the '15 minutes' zone is considered to have good accessibility to the business centre. This can be contrasted to the travel time to/from the business centre in quiet time (Figure 3.18). It is clear that the zones of travel time differ significantly between busy and quiet times.

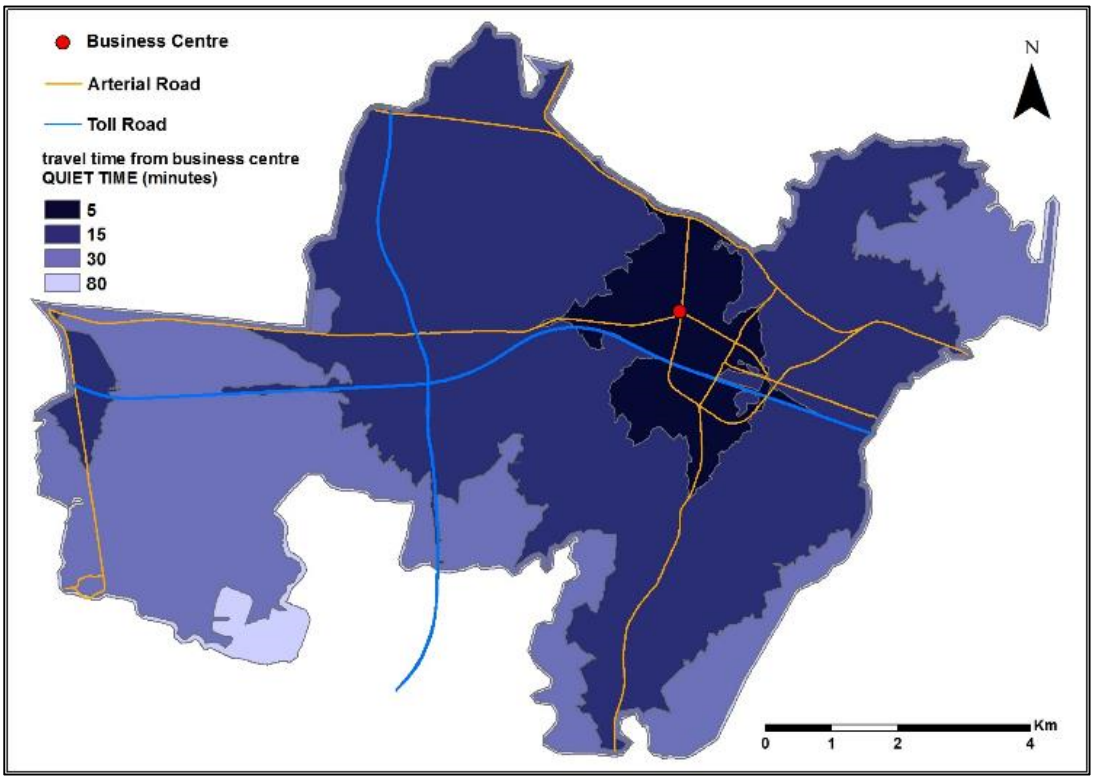


Figure 3.20 Intervals of travel time to business centrein quiet time

Source: Data analysis

Travel times were determined for all other amenities, i.e. tollgate, marketplace, hospital, school, and CBD at busy and quiet times. Maps of these travel times were created. Making use of the travel time intervals to create zones of travel time is practical, and it enables basic patterns to be visualised. However, intervals generalise travel time values because they disregard the unique travel time from each parcel to a certain amenity. For example, two parcels with six and 14 minutes travel times from the business centre, would both fall into the 'within 15 minutes' zone of travel time.

Therefore, instead of utilising zones of travel time, travel times to amenities were individually estimated for each land parcel along an optimum route. For any random land parcel, the minimum travel time to the nearest tollgate, nearest marketplace, nearest hospital, nearest school, and CBD can be retrieved at both busy and quiet times. One of the travel time maps, travel time to the business centre in the busy time, is shown in Figure 3.19.

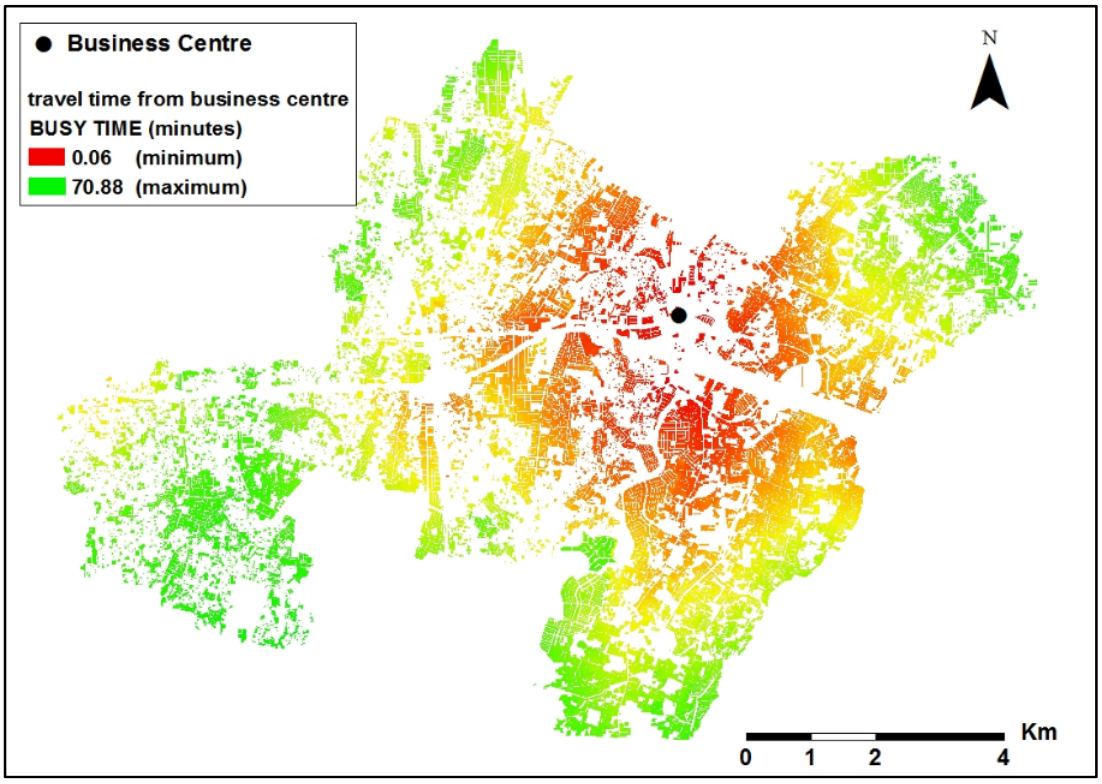


Figure 3.21 Travel times to business centrein busy time

3.7. Summary

The main datasets required for the application of GWR to mass valuation in Bekasi are property sales data, the Land Parcel Map, and the road network data. Severe mismatches were found when the Land Parcel Map was overlaid with road network datasets (from OpenStreetMap and BIG). Because of this reason, the road network data was derived from the Land Parcel Map by converting the gaps between blocks into road segments. A huge number of adaptations were applied during the process of creating road network data from the Land Parcel Map. The effort was worth doing because the road network data was utilised to derive data related to travel distances and travel times to amenities.

The travel distances and travel times from a property to amenities are key variables in determining land values. An issue with using travel distance as the measure of accessibility is that it does not account for the contribution of road classification. Two road segments of different road classes most likely allow different travel speeds when people travel through each of them. Travel time is considered to be a more objective measure of the accessibility of a property to amenities and services, than travel distance. This is in line with the guidance in the Internal Standards for Land Valuation in BPN RI.

In the Internal Standards for Land Valuation in BPN RI, travel times to amenities are divided into within five, 15, and 30 minutes and over 30 minutes intervals. The measures of accessibility for those intervals are defined as very good, good, moderate, and poor. An immediate concern is that these intervals generalise travel time values because they disregard the unique travel time from each parcel to a certain amenity. Discretion is used for this study. Instead of using intervals of travel times to amenities, the GWR model uses unique travel time from a land parcel to each of the amenities.

Before the GWR model was applied to the reconstructed dataset, the dataset had to be put through examination processes to have a clearer picture of the relationships between land price and each of the predictor variables. The most statically significant variables, multicollinearity, and spatial autocorrelation were identified during the data examination processes. These processes are discussed in the next chapter (Chapter Four).

4. DATA EXAMINATION

4.1. Introduction

In this chapter, the data related to the variables listed in the survey forms for Bekasi are examined to understand the relationship between land price and each of the explanatory variables. Only variables which are statistically significant to shape price are going to be used to form the prediction model. Multicollinearity among variables is also examined to detect any dependency among variables.

The selected explanatory variables are first used to form the OLS prediction model. The level of spatial autocorrelation is tested from the output of the OLS model. If the OLS model has a significant issue with spatial autocorrelation, the same set of explanatory variables will be used to form a GWR model.

4.2. Variables

In the prediction model, land price becomes the dependent variable and the qualities of the land parcel become the explanatory variables. There are 12 characteristics listed in the survey form that can be used as explanatory variables. Some of the original variables in the survey form are modified here in order to mimic the real world processes. Travel distances to amenities were modified into travel times to amenities. Travel distance regards similar contributions of each road segment in relation to ease of travelling to amenities, while travel time allows the unique contribution of each road class to be included. Aside from travel time to amenities, the data related to other variables remain the same as they are in the survey form. The explanatory variables to be examined here are as follows:

- Parcel size: the size of the land parcel in square metres
- Zoning: the most dominant type of land use in the zone in which a land parcel is located. The classes are:
  - class 1: agricultural area
  - class 2: irregular residential area
    The sizes and shapes of land parcels vary significantly. Most road segments are in irregular patterns. Road width varies significantly among road segments.
  - class 3: regular residential area
    Road segments are mostly in regular patterns and land parcels are mostly in regular shapes.

- class 4: residential complex

  A well planned residential area usually developed by property developer. Road segments, land parcels, facilities, and utilities are normally well arranged throughout the complex.
- class 5: commercial area

o Road class:
- class 1: lane
- class 2: local
- class 3: secondary collector
- class 4: primary collector
- class 5: secondary artery
- class 6: primary artery

o Road width, in metres

o Travel times to nearest tollgate, major road(s), business centre, marketplace, school, and health facility, in minute.

The data related to each variable from all samples were examined. The statistics of data related to variables from the parcel's features are summarised in Table 4.1, and the statistics of data related to travel times to amenities are summarised in Table 4.2.

Table 4.1 Statistics of variables from parcel's features

Features of land parcel	Mean	Median	Std. Dev.	Min	Max	Unit
Parcel size	418.10	189.50	951.40	35.25	14,787.96	Square metre
Zone	3.38	4	1.02	1	5	Class
Road width	4.81	4.28	2.17	1.50	15.85	Metre
Road Class	2.34	2	0.88	1	6	Class
Price per m^2	1,916,551	1,809,370	969,970	298,775	6,755,373	IDR

Moderate variations are found on data related to variable zoning, road width, road class, and price, while a very large variation is found on data related to variable parcel size.

Table 4.2 Statistics of travel times to amenities this format is better

Travel time to amenities	Mean	Median	Std. Dev.	Min	Max	Unit
Tollgate busy time	21.61	20.10	10.13	0.50	53.52	Minute
Tollgate quiet time	10.36	9.40	4.66	0.28	23.43	Minute
Primary Arterial Road busy time	13.35	10.85	10.21	0.11	48.49	Minute
Primary Arterial Road quiet time	6.75	5.88	4.76	0.04	24.10	Minute
Secondary Arterial Road busy time	18.89	19.26	10.44	0.02	41.31	Minute
Secondary Arterial Road quiet time	9.00	9.42	4.64	0.01	19.52	Minute
Primary Collector Road busy time	6.10	5.00	5.04	0.00	27.27	Minute
Primary Collector Road quiet time	3.21	2.79	2.52	0.00	14.82	Minute
Business Centre busy time	30.78	31.16	12.95	2.13	68.17	Minute
Business Centre quiet time	14.17	14.30	6.16	1.21	34.04	Minute
Marketplace busy time	16.47	15.65	7.80	0.53	43.60	Minute
Marketplace quiet time	8.21	7.86	3.89	0.29	21.83	Minute
Hospital busy time	14.27	12.37	8.71	0.51	38.74	Minute
Hospital quiet time	7.07	6.49	4.05	0.29	20.08	Minute
School busy time	3.32	5.50	1.38	0.22	7.49	Minute
School quiet time	1.79	2.97	0.72	0.09	2.97	Minute

Moderately low values are found on data related to travel time to nearest primary collector road and low values are found on data related to travel time to nearest school because there are so many of these two amenities. With a large number of primary collector road segments and schools distributed across Bekasi, these two amenities can be quickly reached. Nevertheless, moderate variations are found on most of the data related to all variable travel times to amenities.

4.3. Correlation

Variables were paired with one another, and scatter plots of all paired variables were produced to indicate the correlation between variables (Figure 4.1).

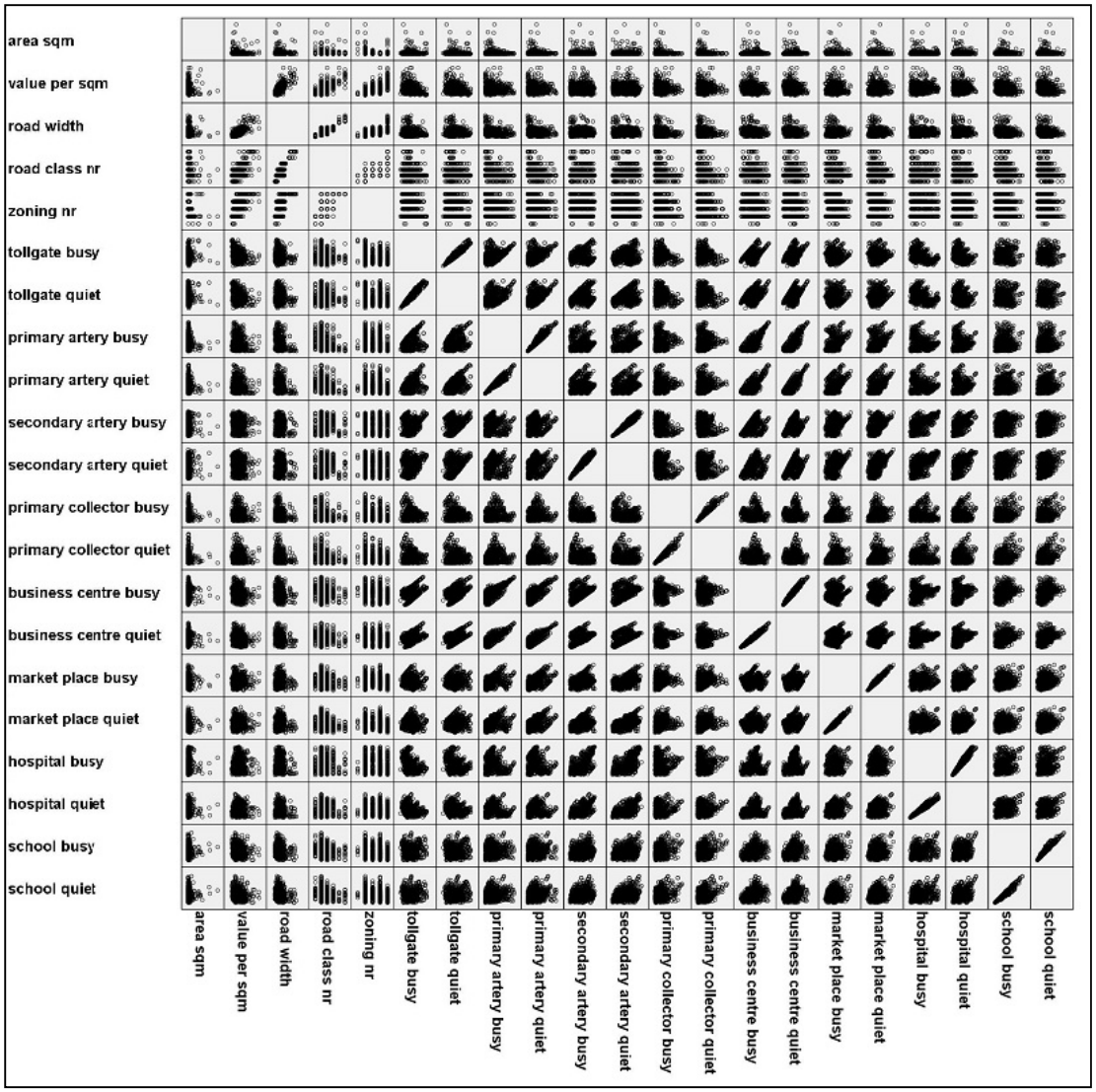


Figure 4.1 Scatter plots of all pairs of variables

The scatter plots indicate that travel times at busy times and quiet times to each amenity have positive linear relationships. If the travel time at the busy time from the business centre to parcel A is half of the travel time to parcel B, the travel time at quiet time to parcel A will be more or less half of the travel time to parcel B. In short, the busy time and quiet time schemes tend to tell the same story.

Using both, travel time at busy time and travel time at quiet time, will cause redundancy. A decision therefore had to be made as to whether to use travel time at busy time or travel time at quiet time. Travel time at busy time was considered to be more important than the travel time at quiet time because obviously the busy time is the time when most people need to travel to schools and workplaces. The travel times to amenities at quiet times were put aside, and only travel times at busy times were used for the next analyses. The relationships between land parcel price and each of the explanatory variables were extracted and are presented in Figure 4.2.

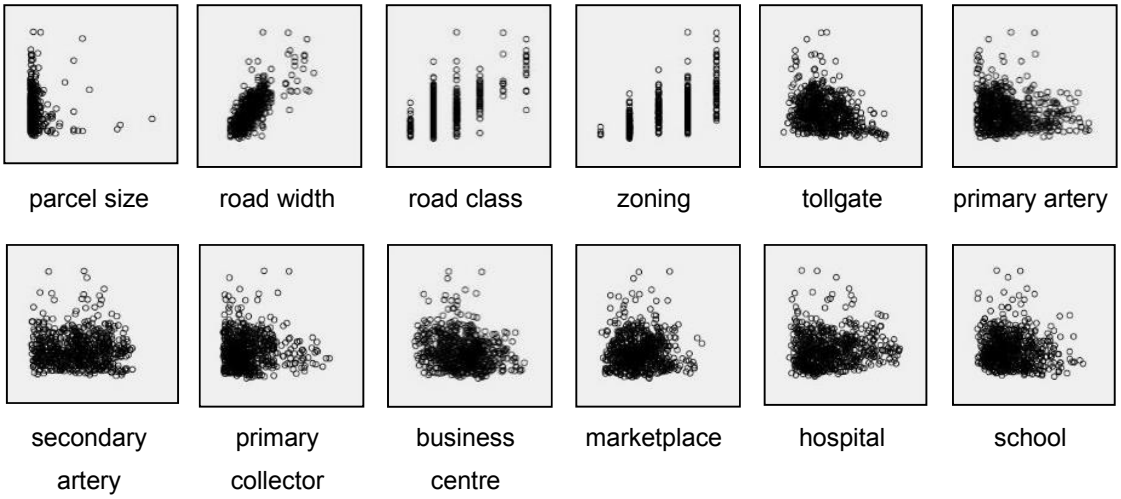


Figure 4.2 Scatter plots between price and each explanatory variable

Apparently, the parcel size does not have a significant relationship with price. Road width seems to have a linear relationship with price. Price tends to increase when the road class and zoning class increase. There is no noticeable pattern on the relationship between price and travel times to amenities. A more detailed examination of the relationships between price and each of the explanatory variables is discussed in Section 4.6.

In order to give an explicit measure on the level of association between price and each of the explanatory variables, a quantitative correlation analysis was also undertaken. A suitable correlation calculation method had to be selected based on the nature of the data. Spearman's correlation (non-parametric correlation) was chosen to suit the data in this study. The non-parametric correlation index does not try to calculate the population correlation, so it is suitable to measure the association between continuous and ordinal

types of data (see Chen and Popovich, 2002). Parcel size, road width, and travel times to amenities are continuous variables, while road class and zoning are ordinal variables. The non-parametric correlation index does not require assumptions on the data either, and bivariate normal distribution is one of the assumptions which is not required in this correlation index (see Chen and Popovich, 2002). Figure 4.3 shows that samples are likely to be normally distributed only in variable business centre and variable marketplace. Samples are likely to be not normally distributed in variable road width, variable tollgate, variable hospital, and variable school. Samples are obviously not normally distributed in variable parcel size, variable road class, variable zoning, variable primary artery, variable secondary artery, and variable primary collector.

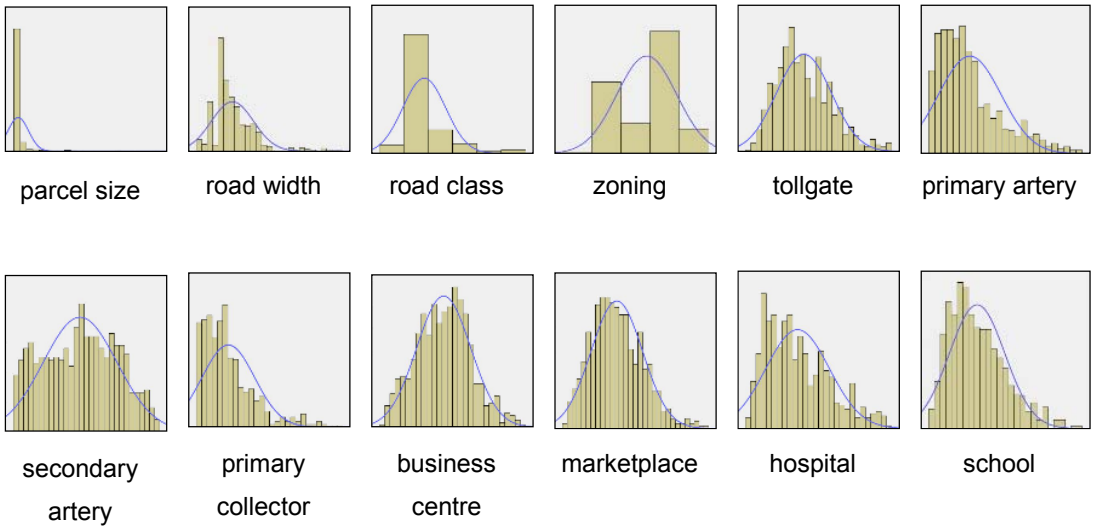


Figure 4.3 Histograms of sample distribution for each explanatory variable

In order to present a more explicit measure on normality, numerical tests of normality were undertaken. The hypothesis that the samples are normally distributed was rejected when the significance value (Sig.) was below 0.05 (Table 4.3).

Table 4.3 Summary of numerical test of normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Parcel size	.354	706	.000	.313	706	.000
Road width	.114	706	.000	.854	706	.000
Road class – number	.412	706	.000	.641	706	.000
Zoning – number	.313	706	.000	.819	706	.000
Tollgate – busy time	.071	706	.000	.969	706	.000
Primary artery – busy time	.124	706	.000	.900	706	.000
Secondary artery – busy time	.053	706	.000	.973	706	.000
Primary collector – busy time	.133	706	.000	.869	706	.000
Business centre – busy time	.031	706	.164	.993	706	.001
Marketplace – busy time	.049	706	.000	.981	706	.000
Hospital – busy time	.089	706	.000	.941	706	.000
School – busy time	.078	706	.000	.955	706	.000

a. Lilliefors Significance Correction

In the Shapiro-Wilk test, all of the variables had significance values below 0.05. In the Kolmogorov-Smirnov test, only the variable travel time to business centre at busy time, had a significance value higher than 0.05. For travel time to the business centre, the hypothesis was accepted in the Kolmogorov-Smirnov test but rejected in the Shapiro-Wilk test. These figures confirmed that data on most of variables were not normally distributed. This condition gave another reason to use the Spearman's rho correlation coefficient.

An assumption of Spearman's rho is that pairs of variables are monotonically related(see Caruso and Cliff, 1997). To judge whether this was the case in this research, lines were fitted on the scatter plots to judge whether or not the explanatory variables were monotonically related with price. The graphs are presented in Appendix A. All the fitted lines were increasing or decreasing in one way, so all the independent variables were considered to be monotonically related with price. This type of relationship meets the condition to use the Spearman's rho correlation coefficient.

The correlation coefficients are displayed in Appendix B. Among all the explanatory variables, road width and zoning type were the most significantly correlated with price. The

correlation coefficient of 0.710 indicates that land parcels located on wider roads are very likely to have higher prices than the ones whose accesses are along narrower roads. Correlation coefficient of 0.663 indicates that land parcels located in commercial areas generally have the highest average land prices, followed by those located in a residential complex. Price tends to decrease in line with the decrease on zoning class.

Road class and the travel time to nearest tollgate were shown to be moderately correlated to land value with correlation coefficients of 0.393 and -2.680 respectively. There is a moderate probability that a land parcel located on a road of higher class has a greater chance to have higher price than the ones on lower road class. The negative relationship between land value and travel time to the nearest tollgate indicates that land value decreases as travel time to the nearest tollgate increases. Other explanatory variables were only weakly correlated with land value.

4.4. Multicollinearity

There is a potential to generate collinearity among independent variables with high correlation coefficients. Collinearity exists when there are approximately linear dependence relations among independent variables, sosome variables can be nearly linear combinations of other variables (see Bingham and Fry, 2010).

Each explanatory variable has a different measure of contribution in predicting the value of a dependent variable. The contribution of each explanatory variable is represented by the variable's coefficient in the prediction model. Collinearity impairs the estimation of the coefficients of the explanatory variables. If there is redundancy between two explanatory variables, the coefficients on the explanatory variables will not represent the actual contributions to shape the dependent variable.

Before the data was run using GWR, it was first run using the Ordinary Least Squares (OLS) regression model. The idea is to find out how well the data fits a global model and to detect the existence of multicollinearity among explanatory variables and spatial autocorrelation in the residuals. The result from OLS regression includes the collinearity statistics, i.e. Variance Inflation Factor (VIF) and Tolerance (Table 4.4).

Table 4.4 Collinearity statistics of explanatory variables

Variable	Tolerance	VIF
Parcel size	.913	1.096
Road width	.246	4.070
Road class – number	.322	3.108
Zoning – number	.541	1.850
Tollgate – busy time	.362	2.764
Primary artery – busy time	.269	3.721
Secondary artery – busy time	.214	4.668
Primary collector – busy time	.627	1.596
Business centre – busy time	.272	3.681
Market place – busy time	.495	2.018
Hospital – busy time	.263	3.803
School – busy time	.771	1.297

Dependent Variable: value per sqm

The tolerance is the percentage of the variance in an explanatory variable that cannot be explained by the other explanatory variables. A lower tolerance value indicates a higher multicollinearity. For a variable with low tolerance, a cross check to the correlation coefficients to other variables is required. When a variable has a very low tolerance value and at the same time has a very high correlation coefficient with another variable, the explanatory variable should be put aside in the next analysis. However, there is no general cut-off value for tolerance and VIF. The diagnostic report from the ArcGIS software package suggests that VIF values larger than 7.5 are the ones that indicate redundancy among explanatory variables. Referring to this diagnostic report, none of the explanatory variables listed in Table 4.4 should be put aside in the next analyses.

4.5. Spatial autocorrelation

Spatial autocorrelation is the correlation among values of a single variable defined by the proximity of those values in geographic space. Positive spatial autocorrelation exists when high values tend to be located near high values, medium values near medium values, and low values near low values, while negative spatial autocorrelation shows that high values tend to be located near low values (see Griffith, 2003).

Examining the residuals of a global regression model is a common practice to observe the existence of spatial autocorrelation. Griffith (2009) suggested that spatial autocorrelation can be used as a diagnostic tool for a regression model to detect:

- 'missing variable'

  A factor that actually has a significant contribution to determine the dependent variable is not involved in the regression model as an explanatory variable. The spatial distribution of this missing variable will contribute to shape the spatial distribution of the prediction residual.

- model misspecification

  The relationship specified in the model does not represent the actual relationship between an explanatory variable and the dependent variable. Specifying the linear relationship on a nonlinear relationship is a common example of model misspecification.

- redundant information

  This situation exists when the spatial arrangement of the dependent variable allows the value at a given location to be predicted quite accurately from the values at nearby locations. Griffith (2009) explained the 'redundant information' issue using the example from house pricing. Building an expensive house near an inexpensive house in a neighbourhood tends to reduce the value of the expensive house while increasing the value of the inexpensive house. Beside the building qualities, all other information are quite the same for both houses. This information duplication emerges from locational closeness. In turn, this situation allows inference of nearby values once the value for a given location is known.

- failure to capture spatial processes mechanism

  Fotheringham (2009) explained how spatial autocorrelation is caused by applying a global model to a spatially varying process. If spatial autocorrelation is caused by spatial non-stationary, calibration of local regression will remove the spatial autocorrelation problem.

- areal unit problem

  The standard eight-by-eight checkerboard is a common example to explain the areal unit problem. A completely negative spatial autocorrelation is shown by the distribution of the red and the black squares on a checkerboard. If the squares are aggregated into bigger squares made of four original squares, there will be bigger squares with dark red colour. The spatial autocorrelation will be positive.

Besides mapping the residuals, a spatial autocorrelation test can also be run on the residuals to examine the level of spatial autocorrelation. Moran's I test was run on the residuals from the OLS model, and the result is shown in Table 4.5.

Table 4.5 OLS model's Moran's I test report

Item	Value
Moran's Index	0.251577
Expected Index	-0.001418
Variance	0.000562
Z-score	10.675667
P-value	0.000000

Moran's Index values fall between -1 and +1. Moran's Index of +0.25 indicates that there is a moderate level of positive spatial autocorrelation in the model. Residuals of high values are moderately clustered, and so are residuals of low values.

The null hypothesis in the test is that the distribution of residual results from random processes. A very low p-value is associated with a very high z-score (positive or negative), and these are found in the tails of a normal distribution. The combination of the extremely low p-value and the high z-score rejects the null hypothesis. There is a very low likelihood that any clustering or dispersion pattern could be the result of random process. The spatial autocorrelation issue does exist in the model.

Each of the factors causing spatial autocorrelation may contribute differently to the moderate positive spatial autocorrelation. Out of the five factors causing spatial autocorrelation, tackling 'spatial process mechanism' can be the most feasible action. To some extent, the spatial autocorrelation issue in the Bekasi dataset can be contributed by the spatially varying processes. This kind of process can be modelled using GWR. Tackling this issue is expected to reduce the level of spatial autocorrelation by a significant extent.

4.6. Variable transformation

As noted in Section 4.5, model misspecification has the potential to cause spatial autocorrelation. Griffith (2009) observed that specifying linear relationship on nonlinear relationship is a common example of model misspecification. OLS and GWR assume linear relationship between the dependent variable and each of the independent variables. In

reality, this relationship is not always linear. Transforming an independent variable is sometimes required to improve the linearity of the correlation between the dependent variable and the independent variable.

The SPSS software package provides 11 options of transformation models. For each relationship between an independent variable and the dependent variable, the R-square value is used to indicate the linearity of relationship between land value and an explanatory variable. The highest R-square value indicates the most linear relationship between land value and explanatory variables. The names and formulas of the transformation models are shown in Table 4.6.

Table 4.6 Variable transformation models

Transformation model	Equation
Linear	$Y = b0 + (b1 * t)$
Logarithmic	$Y = b0 + (b1 * \ln(t))$
Inverse	$Y = b0 + (b1/\ t)$
Quadratic	$Y = b0 + (b1 * t) + (b2 * t^2)$
Cubic	$Y = b0 + (b1 * t) + (b2 * t^2) + \ (b3 * t^3)$
Compound	$Y = b0 * (b1^t)$
Power	$Y = b0 * (t^{b1})$
S-curve	$Y = e^{(b0+\left(\frac{b1}{t}\right))}$
Growth	$Y = e^{(b0+(b1 * t))}$
Exponential	$Y = bo \ * \ e^{(b1*t)}$
Logistic	$Y = 1/(1/u \ + (b0 * (b1^t),$ where u is the upper boundary value

Source: IBM Corp. (2015)

The quadratic model adds $b2$ in the transformation formula, while the Cubic model adds $b2$ and $b3$. Adding these extra parameter estimates decreases parsimony in the regression equation. Next, using Power, Compound, S, Growth, Exponential, and Logistic models could probably complicate the relationship between land value and each explanatory variable. The

formulas for these transformation models even include exponential functions involving the parameter estimates. The excessive bending may fit the data well but it may impose an equation on the natural variation in the data. Out of the 11 transformation models (Table 4.6), only linear, logarithmic, and inverse transformation models were employed for the test. The summary of curve estimations for all explanatory variables in relation to price is displayed in Table 4.7.

Table 4.7 Summary of curve estimations

Explanatory variables	R-square value		
	Linear	Logarithmic	Inverse
Parcel size	0.002	0.005	**0.007**
Road width	**0.542**	0.507	0.367
Road class – number	**0.297**	0.257	0.186
Zoning – number	**0.403**	0.379	0.336
Tollgate	**0.084**	0.053	0.002
Primary artery	0.033	**0.043**	0.024
Secondary artery	0.000	**0.001**	0.000
Primary collector	0.000	0.000	**0.001**
Business centre	**0.037**	0.035	0.022
Marketplace	0.000 (sig =0.991)	**0.000 (sig=0.589)**	0.000 (sig=0.777)
Hospital	**0.008**	0.002	0.000
School	0.010	0.010	**0.012**

The new Bekasi dataset therefore contains original survey data and transformed data. For explanatory variables whose relationships with land value were kept linear, the data were kept as they were. For explanatory variables whose relationships with land value were transformed using logarithmic or inverse model, the data were transformed using the corresponding transformation model.

4.7. Prediction model

The Bekasi dataset was first examined using the OLS model, and the diagnostic report is given in Table 4.8.

Table 4.8 OLS diagnostic report

Number of Observations: 706	Akaike's Information Criterion (AICc) [d]: 20775.722305
Multiple R-squared: 0.639791	Adjusted R-squared: 0.633553
Joint F-Statistic [e]: 102.573501	Prob (>F), (12,693) degrees of freedom: 0.000000*
Joint Wald Statistic [e]: 1020.551054	Prob (>chi-squared), (12) degrees of freedom: 0.000000*
Koenker (BP) Statistic [f]: 55.490287	Prob (>chi-squared), (12) degrees of freedom: 0.000000*
Jarque-Bera Statistic [g]: 805.957589	Prob (>chi-squared), (2) degrees of freedom: 0.000000*

The points to be discussed further are:

- The Koenker (BP) Statistic is statistically significant ($p<0.01$), and it indicates that the relationships between price and the explanatory variables are not consistent. The prediction accuracy tends to vary significantly among different locations.
- The Jarque-Bera Statistic is statistically significant ($p<0.01$), and it explains that the residuals are not normally distributed.
- The Joint Wald Statistic is statistically significant ($p<0.01$). It is a sign that the explanatory variables in the model are effective and that the model is statistically significant.

Although the explanatory variables are effective overall in the OLS model, not all of the 12 variables are statistically significant (Table 4.9).

Table 4.4 Summary of OLS variables

Explanatory variables	Robust probability	Coefficient
1/ Parcel size	0.736390	-2294449.50
Road width	*0.000000	268153.43
Road class	0.118443	-71859.84
Zoning	*0.000000	273909.23
Tollgate	*0.000000	-20539.84
Ln Primary artery	0.695113	12179.47
Ln Secondary artery	0.886986	4299.91
1/ Primary collector	0.156733	-767.98
Business centre	0.331173	2306.96
Ln Marketplace	0.114985	68145.95
Hospital	0.775157	-1166.18
1/ School	0.598277	-36391.26

* explanatory variable is statistically significant

Because the Koenker (BP) Statistic is statistically significant, the robust probability value was used to assess the statistical significance of each explanatory variable. When the robust probability is very small, the chance of the coefficient being zero is also small.

The above result indicates that only three out of 12 explanatory variables are effective for prediction. In order to come up with solid inference on variable selection, *backward elimination* technique, which is one of the approaches used in *stepwise regression*, was undertaken. In backward elimination, the first model uses all of the available variables. The variable with the smallest *F*-statistics will be removed from the model if the *F*-statistics is less than the *F*-out threshold. The procedure was continued until the smallest *F*-statistics was bigger than *F*-out or all of the variables were eliminated (see Bingham and Fry, 2010).

Eight explanatory variables were eliminated and four variables were kept in the final model. The R-squared values of the original model with 12 variables (step 1) and the model with four chosen variables (step 9) are 0.640 and 0.637 respectively. The standard errors of the estimate model in step 1 and step 9 are 587,586.03 and 586,582.68 respectively. Eliminating eight variables did not increase the overall performance of the model, as a whole, significantly. However, the effect of multicollinearity was significantly reduced as indicated by the lower VIF values in step 9 (Table 4.10).

Table 4.5 Collinearity statistic from backward elimination regression

Variables	VIF	
	Step 1	Step 9
1/ Parcel size	1.165	
Road width	4.141	1.629
Road class – number	3.095	
Zoning – number	1.929	1.607
Tollgate	2.359	1.027
Ln Primary artery	1.947	
Ln Secondary artery	1.909	
1/ Primary collector	1.019	
Business centre	2.016	
Ln Marketplace	1.334	1.015
Hospital	2.450	
1/ School	1.050	

The OLS results suggest that there are only three statistically significant explanatory variables, i.e. road width, zoning, and travel time to nearest tollgate. The backward elimination regression result suggests that beside the abovementioned three significant variables, travel time to nearest marketplace is also an important explanatory variable.

In order to decide whether to use three or four explanatory variables, the level of spatial autocorrelation was then examined. First, an OLS model was built using three explanatory variables and then another OLS model was built using four explanatory variables. The Moran's I test was applied on the residuals from both models, and the results are given in Table 4.11.

Table 4.6 OLS models' Moran's I test summary

	OLS model with three explanatory variables	OLS model with four explanatory variables
Moran's Index	0.297553	0.286344
Expected Index	-0.001418	-0.001418
Variance	0.000561	0.000561
Z-score	12.621235	12.145635
P-value	0.000000	0.000000

The combinations of very low p-values and moderately high z-scores (Table 4.11) indicate that the clustering patterns on residuals are less likely to be resulted by random process. Both models have an issue of spatial autocorrelation. Because the OLS model cannot be used for prediction, the same dataset was run using the GWR model.

The robust probability values from the OLS model and the backward elimination stepwise regression come up with different suggestions whether to use three or four explanatory variables. Next, the results from Moran's I tests indicate that OLS models using three and four explanatory variables have a quite similar level of spatial autocorrelation. Those results still raise questions of whether to use the set of three or the set of four explanatory variables in the GWR model. Instead of keeping on analysing this issue, both sets of explanatory variables were then run in the GWR model.

4.8. Summary

OLS and GWR assume a linear relationship between land price and each of the predictor variables. The curve estimations on the data reveal that land price is more likely to have linear relationships with only six out of the 12 listed variables, i.e. road width, road class, zoning, travel time to nearest tollgate, travel time to business centre, and travel times to nearest hospital. Data related to other explanatory variables must be transformed to improve the linearity to price.

Being put through the backward elimination regression, none of the explanatory variables indicates a significant issue of multicollinearity as represented by the low VIF values. However, this analysis suggested that only the variables road width, zoning, tollgate, and marketplace are significant for the model. This conclusion is slightly different from the conclusion of the OLS model which suggested only the variables road width, zoning, and tollgate are significant for the model.

In order to decide whether to use three or four explanatory variables, an OLS model was formed using three explanatory variables and another OLS model was formed using four explanatory variables. Moran's I test was run on the output of each model, and the results indicate that both models have a significant issue of spatial autocorrelation at a quite similar level. With no significant difference on the level of spatial autocorrelation, the decision whether to use three or four explanatory variables cannot be made yet. Both sets of explanatory variables are therefore used in the next analysis.

When an OLS model has a significant issue of spatial autocorrelation, the set of variables forming the OLS model are potential candidates to form a GWR model. As discussed in Section 4.5, the failure to capture the spatial process mechanism is one of the factors that causes spatial autocorrelation. GWR, with the arrangements of the weighting scheme and local regressions, has the capability to capture local variations of the data. Therefore, GWR model is expected to come up with a very low spatial autocorrelation level when using the set of explanatory variables previously used in the OLS model. The performance of the GWR model using the abovementioned sets of explanatory variables will be the main part of the discussion in the next chapter (Chapter Five).

5. GEOGRAPHICALLY WEIGHTED REGRESSION (GWR) MODELLING WITH INDIVIDUAL LOCATIONS

5.1. Introduction

Geographically Weighted Regression (GWR) has never been applied for mass valuation practice in The National Land Agency of Indonesia (BPN RI). As explained in Section 1.5, the mass valuation method currently employed by BPN RI is the Zonation Method.The Zonation Method has always had issues with data shortages because it requires a considerably largevolume of data. Despite the fact that the large data requirement is rarely met, it yields relatively accurate predictions. GWR is expected to overcome the data shortage issue and at the same time to deliver accurate predictions.

Mass appraisal on the Bekasi dataset using the Zonation Method was undertaken within the Land Valuation Directorate of BPN RI in 2012. The result is presented at the beginning of this chapter so that it can be used as a benchmark to assess the performance of the GWR model developed in this study. Comparison between the results from the Zonation Method and the GWR model is expected to give an objective assessment on the advantages and disadvantages of each method because both methods are applied to the same dataset.

5.2. Results from applying the Zonation method

A small portion of the study areawasselected to describe the calculations run using the ZonationMethod. Figure 5. 1 shows that zone 460 has only one sample (sample number 660), while zone 686 has four samples (sample numbers 115, 116, 117, and 118).

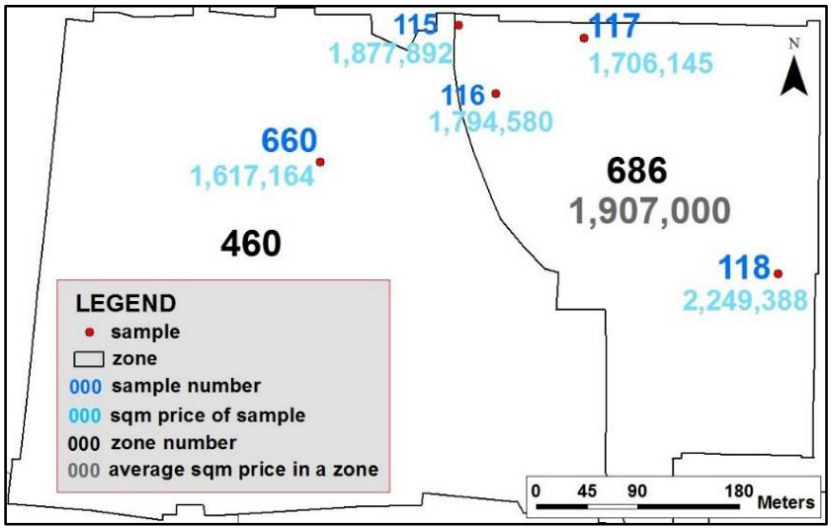


Figure 5.1 Value zone and sample

The value of a zone is the average value of samples located within it.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textbf{(Equation 5.1)}$$

Where: n is the number of samples in a zone

$x_i$is the observed value of sample i

The coefficient of variation is calculated for each zone for quality assessment.

$$c_v = \frac{\sqrt[2]{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\overline{x}} \quad \textbf{(Equation 5.2)}$$

A valid zone must contain at least three samples. Therefore zone 460 is regarded as a non-valid zone, whilst zone 686 is valid. Average value and coefficient of variation of samples are only calculated for a valid zone. If the coefficient of variation is less than 30 per cent, the average value is taken as the zone value. The prediction residual is calculated by comparing the observed value and the predicted value. In this case, the observed value is the sample value and the predicted value is the zone value.

$$\varepsilon_i = \overline{x} - x_i \quad \textbf{(Equation 5.3)}$$

As a measure of accuracy, the percentage residual can be a more objective measure than the actual prediction residual. The percentage residual gives a relative measure of a prediction residual compared with the corresponding observation value. From the zones and samples in Figure 5.1, prediction residuals and percentage residuals were calculated as shown in Table 5.1.

Table 5.1 Calculation of prediction residual for Zonation Method applied to zones 460 and 686

Zone number	Zone value $\hat{x}$	Sample number	Sample value x_i	Prediction residual $x_i - \hat{x}$	Percentage residual $\frac{x_i - \hat{x}}{x_i}$x 100 per cent
460	N/A	660	1.167.000	N/A	N/A
686	1.907.000	115	1.878.000	-29.000	-1.5
		116	1.795.000	-112.000	-6.2
		117	1.706.000	-201.000	-11.8
		118	2.249.000	342.000	15.2

In the Bekasi dataset, only 209 out of 706 samples are located within valid zones (i.e., a zone containing at least three samples). Prediction residuals can be calculated for these 209 samples, which account for 29.6 per cent of all samples. The magnitudes of the percentage residuals for these 209 sampled locations are shown in Figure 5.2.

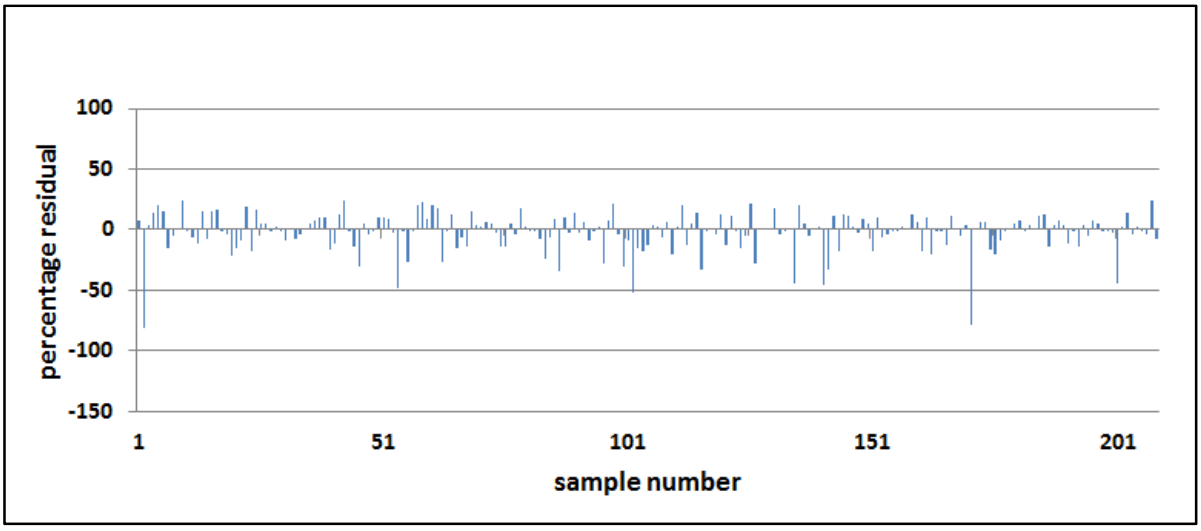


Figure 5.2 Distribution of percentage residuals for the 209 samples in 50 valid zones using the Zonation Method

The mean absolute percentage error (MAPE) is 10.8 per cent. There are only three sampled locations which come up with percentage residuals larger than 50 per cent, i.e. location numbers 2, 102, and 171. A large negative percentage residual indicates that an observation is significantly smaller than the average value of all observations within the corresponding zone. Locations 2, 102, and 171 are most likely to be undervalued.

5.3. Results from applying the GWR model to individual locations

The performance of the GWR model using individual samples from the Bekasi dataset was assessed using in-sample estimation and out-of-sample estimation. In the first assessment, GWR analysis was run using the whole dataset. All samples were used for developing the model, and then all samples were again used as the validation locations. Therefore, prediction residuals were computed using the internal data (in-sample estimation).In the second assessment, the 706 sample dataset were split into two subsets, i.e.,training and validation subsets. The training dataset was used to develop the model, and the second as

an independent validation dataset which was to obtain a more objective assessment on the prediction accuracy through out-of-sample estimation.

5.3.1. In-sample estimation of the GWR model with individual locations

GWR was run using all of the 706 samples. The adaptive kernel was chosen because the samples are not regularly distributed. Both bandwidth optimisation methods, CV and AICc, were tested to find out whether or not there wasa significant difference in prediction accuracy in relation to bandwidth method. GWR was run using three explanatory variables and four explanatory variables, and the diagnostic reportsarepresented in Table 5.2.

Table 5.2 Diagnostic report of GWR model using individual samples

a. GWR with three explanatory variables and AICc bandwidth method

Neighbours	67
ResidualSquares	111,955,400,000,000
EffectiveNumber	125.20
Sigma	439,044.02
AICc	20,435.63
R-squared	0.83
R- squared Adjusted	0.80

b. GWR with three explanatory variables and CV bandwidth method

Neighbours	73
ResidualSquares	115,607,500,000,000
EffectiveNumber	115.62
Sigma	442,512.74
AICc	20,438.52
R- squared	0.83
R- square Adjusted	0.80

c. GWR with four explanatory variables and AICc bandwidth method

Neighbours	210
ResidualSquares	157,883,300,000,000
EffectiveNumber	48.96
Sigma	490,198.14
AICc	20,535.36
R- squared	0.76
R- squared Adjusted	0.74

d.GWR with four explanatory variables and CV bandwidth method

Neighbours	209
ResidualSquares	157,659,100,000,000
EffectiveNumber	49.08
Sigma	490,267.01
AICc	20,506.61
R- squared	0.76
R- squared Adjusted	0.75

The diagnostic reports (Table 5.2) revealed that there is not much difference between the results from AICc and CV bandwidth methods in the Bekasi dataset. Since this dataset represents the common circumstances of mass valuation datasets from Indonesia, it appears that choosing the bandwidth method should not be a crucial issue in mass valuation work elsewhere in Indonesia. Only the results from analysis using the AICc bandwidth method are discussed from this point onward.

Compared with the GWR model using four explanatory variables, the GWR model using three explanatory variables has a higher R-squared value and lower total squared residuals. In general, this model performs better than the one using four explanatory variables. Though with this indication, each model will be tested against Moran's I test to examine the level of spatial autocorrelation. The results from this test are shown in Table 5.3.

Table 5.3 Summary of Moran's I test for GWR model using individual samples

a. GWR with three explanatory variables and AICc bandwidth method

Item	Value
Moran's Index	0.030555
Expected Index	-0.001418
Variance	0.000562
Z-score	1.349163
P-value	0.177285

b. GWR with four explanatory variables and AICc bandwidth method

Item	Value
Moran's Index	0.135886
Expected Index	-0.001418
Variance	0.000560
Z-score	5.800581
P-value	0.000000

The Moran's Index ranges from -1 to 1. In the GWR model using three explanatory variables, there is a very low index for clustering; 0.030555. The combination of a low p-value (0.177285) and a low z-score (1.349163) suggests that the low clustering pattern could be random. In short, spatial autocorrelation is not an issue in this model. In the GWR model using four explanatory variables, the Moran's Index is also low (0.135886) but the combination of very low p-value (0.00000) and moderately high z-score (5.800581) suggests that the moderately low clustering pattern is most likely not a random process. Although moderately low, spatial autocorrelation is an issue in the GWR model using four explanatory variables. Moran's I results provide a statistical basis to choose the GWR model using three

explanatory variables over the four-variable model. Only the result from the GWR model using three explanatory variables are discussed from this point onward.

As discussed in Section 4.6, the dataset is first run using the OLS model. When using the GWR model, the issue of 'model misspecification' is tackled. OLS imposes a linear relationship between the dependent variable and each explanatory variable, while GWR models the non-linear function between the dependent variable and each explanatory variable by weighted regression which gives a unique weight to each sample based on the proximity to the local regression point (see Brunsdon *et al.*, 1996). GWR also tackles the issue of 'failure to capture spatially varying processes' from the OLS model by the arrangement of weighting scheme and local regression (see Páez and Wheeler, 2009). The Moran's Index is 0.297553 in the OLS model and is 0.030555 in GWR model, so the GWR model reduced the Moran's Index by 89.73 per cent. This very significant change indicates that 'model misspecification' and 'failure to capture spatially varying processes' are key factors causing spatial autocorrelation in the OLS model. Two out of the five factors behind spatial autocorrelation issue which were listed in Section 4.4 have been addressed. The Moran's Iindex of 0.030555in the GWR the model is very likely to be a combination of the three other factors, i.e. redundant information, areal unit problem, and missing variables.

Recalling the example from house prices in Section 4.5, the nature of land value distribution tends to exhibit redundant information. Houses located in a certain neighbourhood share common neighbourhood characteristics, and sometimes have common house characteristics. Moreover, they normally have quite similar land values. In neighbourhoods like this, redundant information is inevitable. Like the factor of redundant information, the factor of areal unit problem is not tackled in this study. Prediction is undertaken in the unit of individual land parcels, and the land parcels have various shapes and sizes. In addition to that, sampled land parcels are not located in a regular pattern. The last factor behind the spatial autocorrelation issue which is not yet addressed is the factor of 'missing variables'.

The explanatory variables involved in the GWR model are extracted from the Bekasi dataset, and they are based on the Internal Standards for Land Valuation within BPN RI. It is actually possible to introduce new explanatory variables for the prediction model but it is out of the focus area of this study. Moreover, Moran's Index of 0.030555 indicates a very weak clustering pattern among prediction residuals. Disregarding the factor of redundant information, areal unit problem, and 'missing variables' does not cause significant spatial autocorrelation issue.

The map of standardised residuals confirms the results from local Moran's I test, as shown in Figure 5.3. There is no significant pattern of clustering or dispersion among the standardised residual.

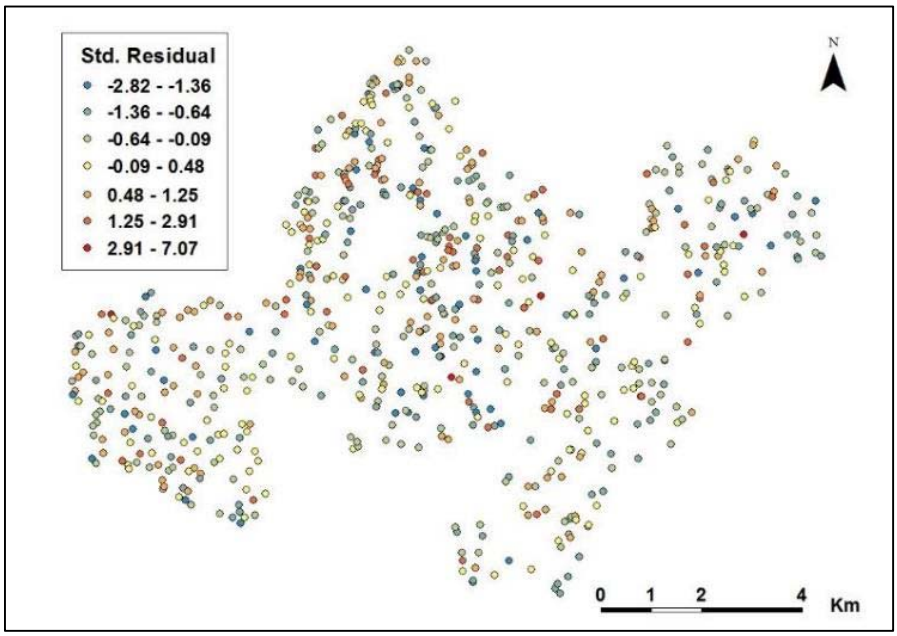


Figure 5.3Standardised residual of GWR model using individual samples

There is no clear pattern of clustering or dispersion of the standardised residual in Figure 5.3. The distribution of under-predicted and over-predicted locations appears to be quite random. Therefore, no further examination is required on the distribution of residuals. Having examined the issue of spatial autocorrelation, the following discussion concerns local R-squared values.

The diagnostic report of the GWR model (Table 5.2) gives a global R-squared value for the model as a whole, while the goodness-of-fit for each local model is indicated by a local R-squared value. Local R-squared values range from 0.19 to 0.92, with an average value of 0.67: 13.9 per cent of all local regressions have local R-squared values lower than 0.5. Attention was paid to locations with very low local R-squared values when examining prediction accuracy and precision. The highest local R-squared values are found in the south-west and the east central parts of Bekasi (Figure 5.4). In these areas, high variances in land value can be predicted using the three explanatory variables, indicating land values are strongly associated with road width, zoning, and travel time to the nearest tollgate.

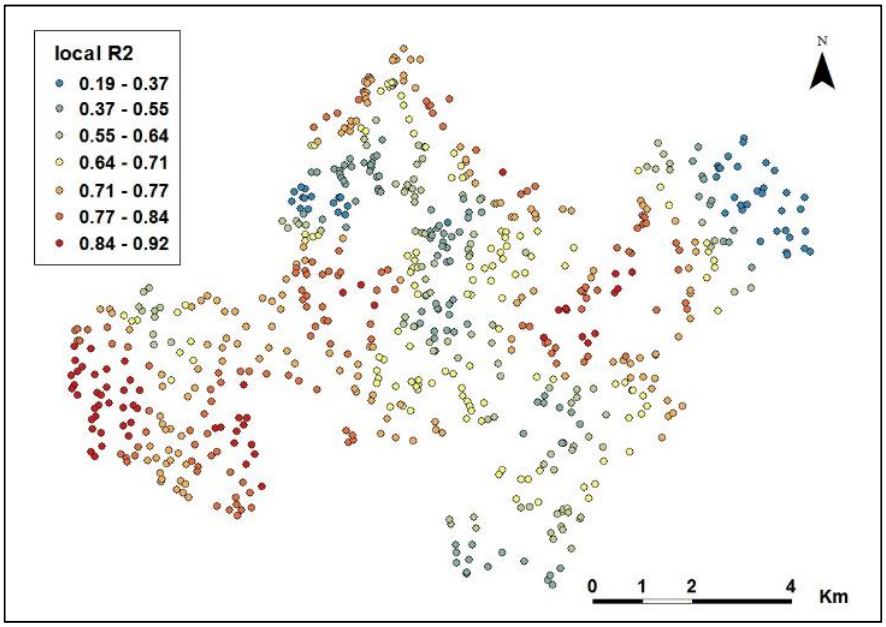


Figure 5.4 Local R-squared value of GWR model using individual samples

Locations with the lowest local R-squared values are found in the mid-west and north-east. Here, road width, zoning, and travel time to the nearest tollgate (as a set of explanatory variables) are not quite as effective for land value prediction as in the south-west and east central parts of the city. Other variables influence land values in these locations. Having considered the issues of spatial autocorrelation and the spatial distribution of R-squared values, the next discussion focusses on the spatial variation of parameter estimates.

Parameter estimatesare actually the main output of a GWR model. The distribution of coefficients for all explanatory variables is plotted in Figures 5.5, 5.9, and 5.12. In the case of variable road width, the lowest coefficients were for locations in the central east and in the north-east. Low coefficients for variable road width were also found in most of the east of the city. These areas are the farthest from Jakarta, which is located immediately to the west, and where many residents commute to by car. Bekasi has grown as an expansion of development from Jakarta. Because northern and eastern Bekasi are the farthest parts of the city from Jakarta, they are the latest to be exposed to the gradual expansion of the capital city. These areas are covered by relatively new residential complexes built from the 1990s until recently, on areas that were farmland and small villages. The sampled locations reveal that the coefficient of variation of road width in residential complexes is 25.38 per cent. Variation of road width in the residential complexes is lower than in other zoning types, so road width has least significance in this type of zoning. This can be one of the most

reasonable explanations on the low parameter estimates in relation to road width in these areas.

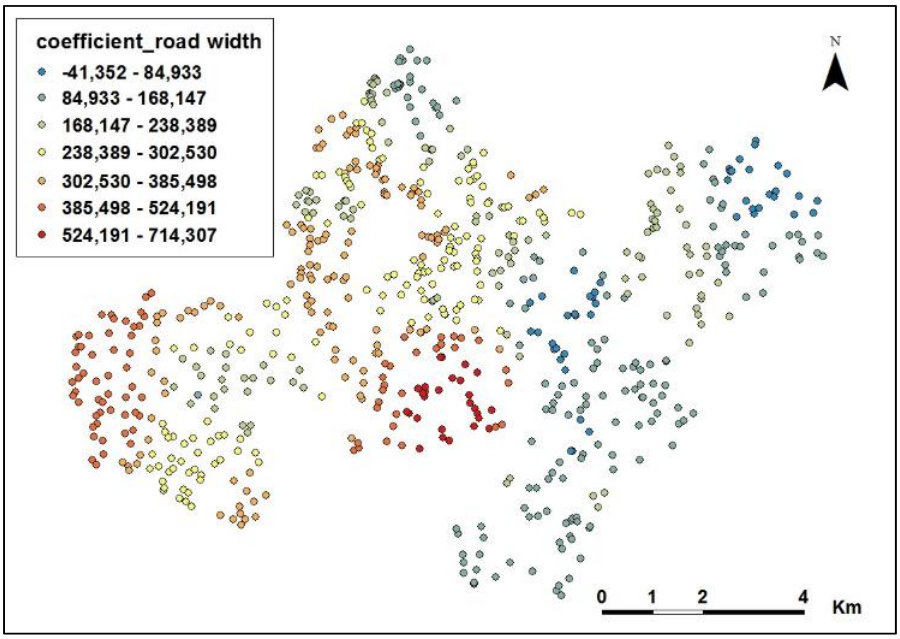


Figure 5.5 Coefficient estimate for variable road width in GWR model using individual samples

The largest positive coefficients for variable road width are found in the central south of Bekasi City. Interestingly, some of the locations with the largest positive coefficients are in residential complexes. This is in contrast to the earlier argument saying that road width has its least significance in residential complexes. This arises because many of these residential complexes are surrounded by irregular residential areas. Therefore, the argument derived for the new residential complexes in the north and east does not apply in the central south because road widths vary more significantly within irregular residential areas as indicated by the coefficient of variation of 33.29 per cent. Consequently, in these irregular residential areas, road width has a higher chance to influence land values. The irregular residential areas have a dominant contribution in determining local relationships between road width and land value in these areas.

GWR computes a coefficient's standard error at every location for each explanatory variable. The standard error of a coefficient can be used as a measure of precision of parameter estimate. A small standard error indicates high precision of parameter estimation in a local regression, while a large standard error indicates low precision. Thirty four locations in the central south and the middle west have the largest coefficient's standard errors in relation to

road width as a predictor variable (Figure 5.6). Although with large coefficient standard errors at a number of locations, predictions at all locations are considered reliable. A measure and a cut-off value to judge whether or not prediction at one location is reliable, is provided within the GWR package in ArcGIS. A location with a condition number larger than 30 is considered to have unreliable prediction. None of these locations had condition numbers greater than 30.

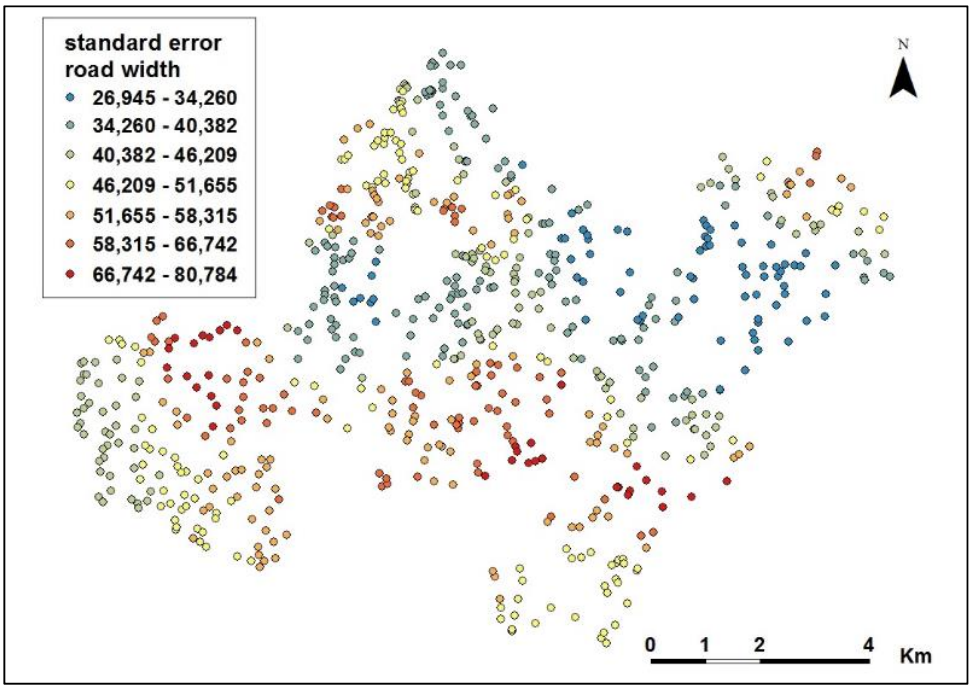


Figure 5.6Coefficient's standard error for variable road width in GWR model using individual samples

In GWR model, a local linear regression is run at each location. An output of a linear regression normally contains a t-statistic and a p-value to diagnose the significance level of each explanatory variable. However, the GWR package in ArcGIS does not report these diagnostic measures. Charlton and Fotheringham (2009) suggested that the p-value as a measure of significance of parameter estimate in a global model is not appropriate to be used in GWR, and they considered thatthe Benjamini-Hochberg False Discovery Rate (FDR) is a more appropriate approach, but it has not been incorporated in the GWR model.

But as there is no measure of significance for parameter estimates available in the GWR model, assessment of the significance of parameter estimates was not undertaken in this work. Nevertheless, assessment of the precision levels of parameter estimates was performed by computing local t-statistic values. The t-statistic simply compares the actual

value of a coefficient to its standard error. Both inputs, coefficient and the coefficient's standard error are provided in the output table of the GWR model.

The local t-statistic value helps to indicate the level of reliability of a parameter estimate. On the one hand a small coefficient's standard error for a large coefficient results in a large t value, which indicates high confidence in the parameter estimate. On the other hand, a large coefficient's standard error for a small coefficient results in a low t, indicating low confidence on the estimate of the parameter.

For variable road width, 653 of 706 (92.49 per cent) locations have t-statistic values larger than two which means that most locations have actual coefficient values over two times the corresponding coefficient's standard errors (Figure 5.7). Therefore, most locations have reliable coefficients for road width. This inference from the GWR model is in line with the results from the initial data examination using the OLS model and backward elimination stepwise regression which gave a high t-value for road width.

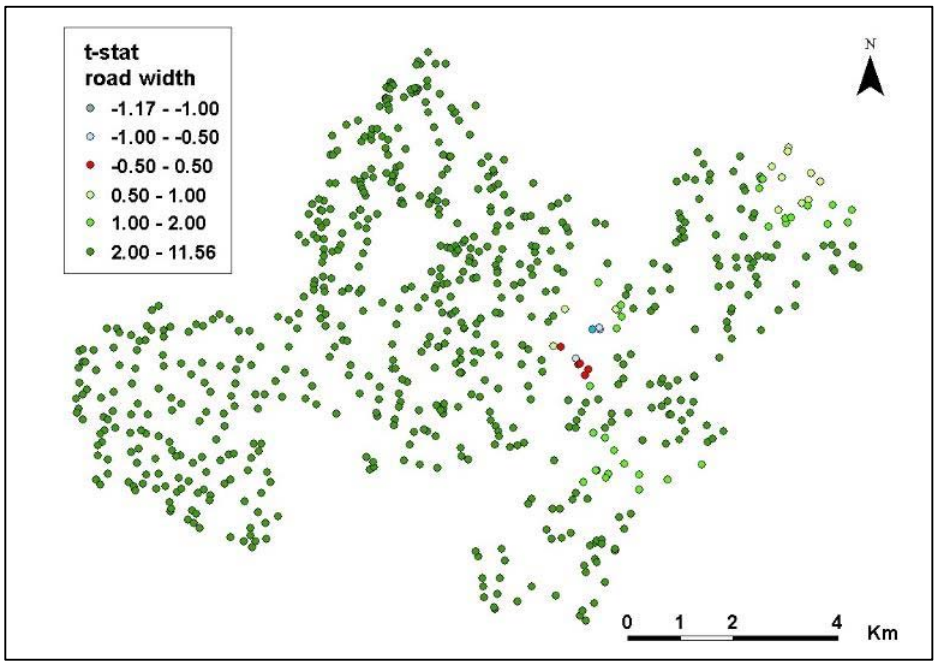


Figure 5.7 T-statistic value for variable road width in GWR model using individual samples

Locations with the smallest t values (positive or negative) are in the central west, as indicated by the red dots in Figure 5.7. In these locations, coefficients for variable road width are the least reliable because the coefficient's standard errors are more than twice the corresponding actual coefficient values. Figure 5.8 shows examples of these locations.

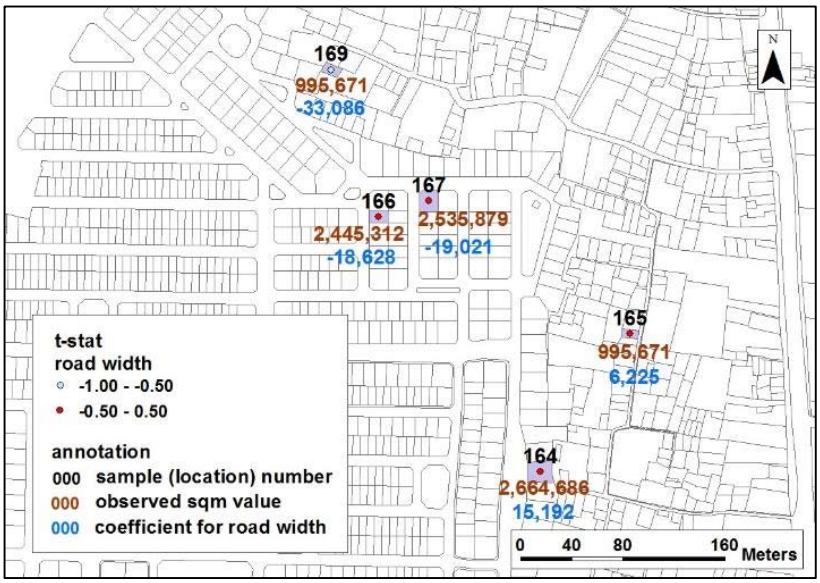


Figure 5.8 Locations with very low t values for variable road width in the GWR model using individual samples

In this study, the term sqm value is used to represent the land price per square metre. Locations 164, 166, and 167 have quite similar observed sqm values (2,664,686, 2,445,312, and 2,535,879 respectively). They are located in a residential complex, so they also have the same value for variable zoning. As they are located close to one another, they also have similar ravel times to nearest tollgate – another explanatory variable. The travel times are 19.31, 18.93, and 19.37 minutes respectively. For variable road width, they have a moderately low variation. The road widths are 6.73, 7.21, and 7.32 metres respectively. It is surprising then that location 164 has a road width coefficient of 15.192 which is remarkably different from locations 166 and 167 which have road width coefficients of -18.628 and -19.021 respectively. This is because GWR allows variation in the dataset to emerge with very different coefficients for neighbouring locations whose values related to explanatory and dependent variables are similar.

As explained earlier in this section, 67 of 706 total samples are involved in each local regression. The 67 samples used in the local regression at location 164 have resulted in a different inference from the 67 samples involved at either locations 166 or 167, although most of the samples involved in local regression at location 164 are the same as those involved in local regressions at locations 166 or 167. The weighting scheme at each local regression assigns different weights to each mutually used samples, and the variation of weights must be the main reason behind these large differences of local regression results. Locations 165 and 169 (Figure 5.8) are another pair of samples that show this kind of difference.

With similar observed sqm values and explanatory variables, locations 164, 166, and 167 would be expected to have similar predicted sqm values as well. However, the coefficients related to road width for those locations do not contribute in the same way to the prediction of land values. The wider road width at location 164 leads to a higher land value because the coefficient on variable road width is positive. In contrast, the wider road widths lead to lower land values at locations 166 and 167 because the coefficients on variable road width are negative. The situation at locations 166 and 167 is unusual. This will be examined further during the evaluation of prediction accuracy.

Local variations are also found in the parameter estimates for variable zoning (Figure 5.9). Zoning makes its highest contribution in determining land value in the city centre and its immediate vicinity. If two land parcels have similar road widths and travel time to the nearest tollgate, any difference in prices in the area will be mainly related to the zoning class. The lowest parameter estimates are found in the central-west, the part of Bekasi that is closest to Jakarta. The variation of zoning class does not cause significant variation of predicted land prices in this area.

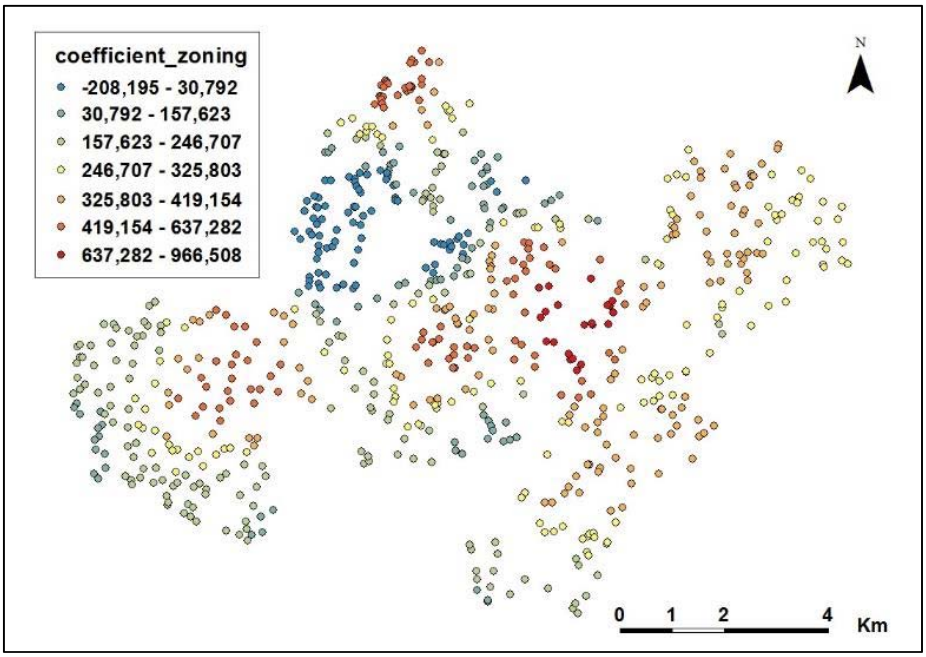


Figure 5.9 Coefficient estimate for variable zoning in the GWR model using individual samples

It was noted earlier that condition numbers are lower than 30 for all locations, and they do not exhibit local collinearity. However, significant differences in standard errors can be seen

in Figure 5.10. Locations in the central-north and the north have the highest standard errors in relation to zoning class, indicating zoning is least reliable at those locations.

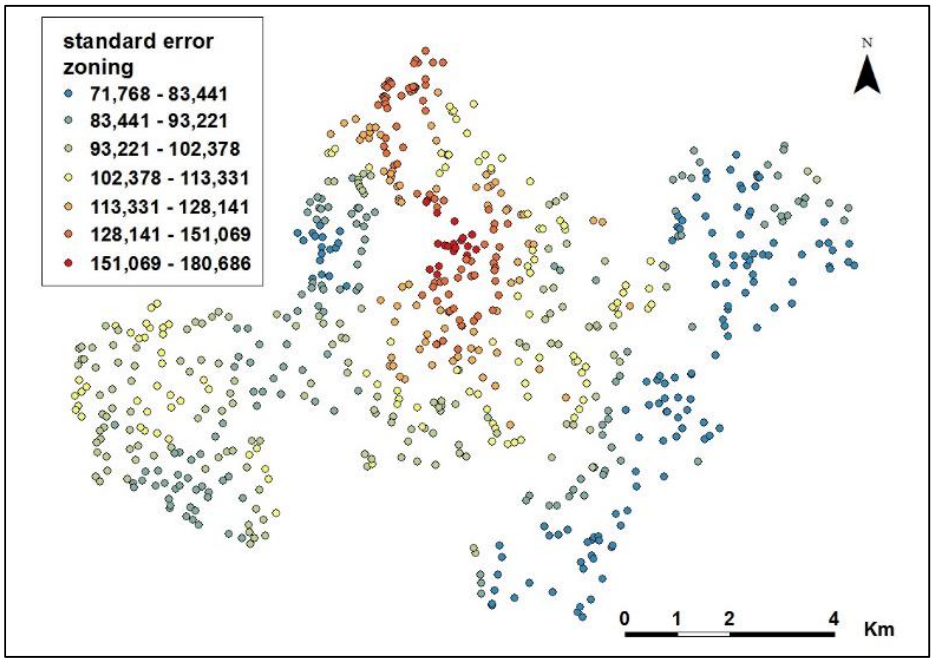


Figure 5.10 Coefficient's standard error for variable zoning in the GWR model using individual samples

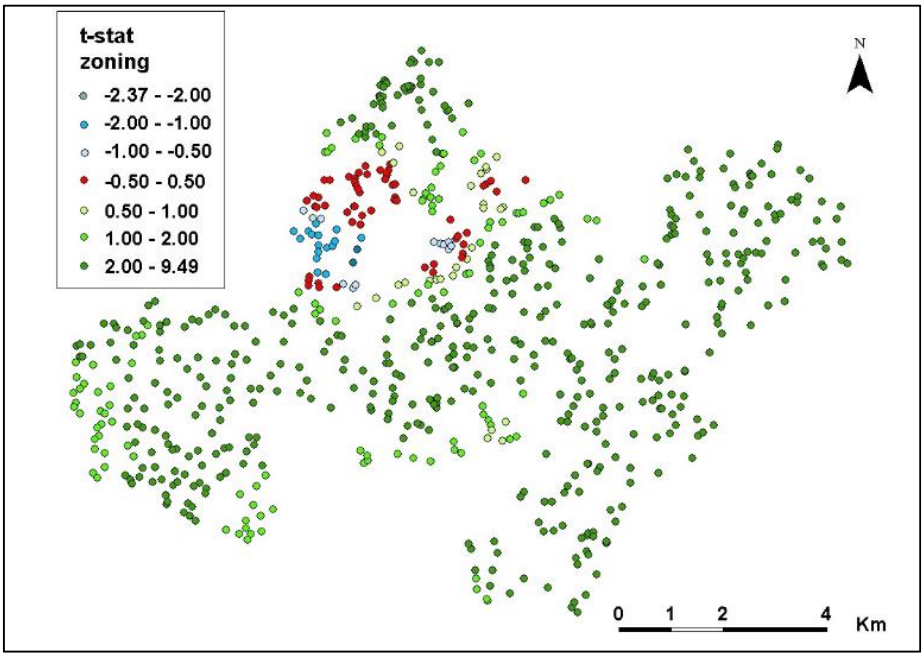


Figure 5.11 T-statistic value for variable zoning in the GWR model using individual samples

The t-statistic value shows different locations of low reliability of parameter estimation (Figure 5.11) compared with the map of standard error (Figure 5.10). The lesson learned from analysing road width is that t-statistic values are more meaningful than the standard errors. Locations with moderately high and high t-values form a dominant portion of the area studied (Figure 5.11). This confirms the result of the initial data examination using the OLS model and the backward elimination stepwise regression which indicated zoning as the second-most influential variable after road width.

The travel distance to the nearest tollgate ('tollgate' from this point onwards) displays interesting statistical behaviour. Out of 706 locations, 490 locations come up with negative coefficients which supports the idea that a shorter travel time to the nearest tollgate increases land value. However, 290 locations have positive coefficients (Figure 5.12). The importance of travel time to the nearest tollgate does vary by location, and the variation in variable tollgate is greater than either road width or zoning.

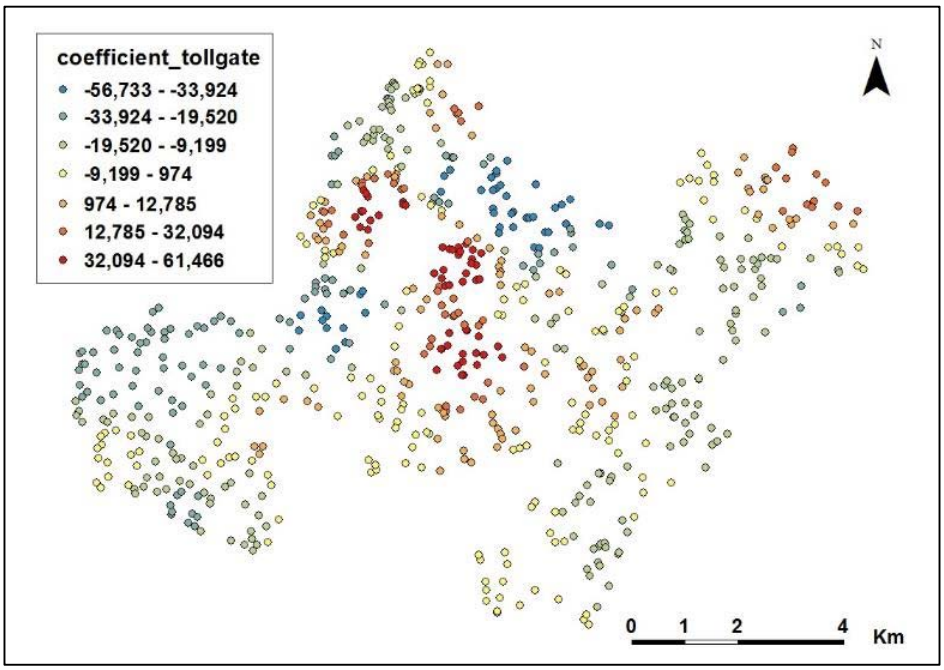


Figure 5.12Coefficient estimate for variable tollgate in GWR model using individual samples

Locations in the central-west have the highest positive coefficients, and this indicates that a shorter travel time to the nearest tollgate decreases land value. This inference is counterintuitive to the relationship between travel time and land value derived in the global

OLS model and the backward elimination stepwise regression, i.e., that shorter travel time to the nearest tollgate increases land value. Therefore the locations with high positive coefficients were checked, see the examples in Figure 5.13.

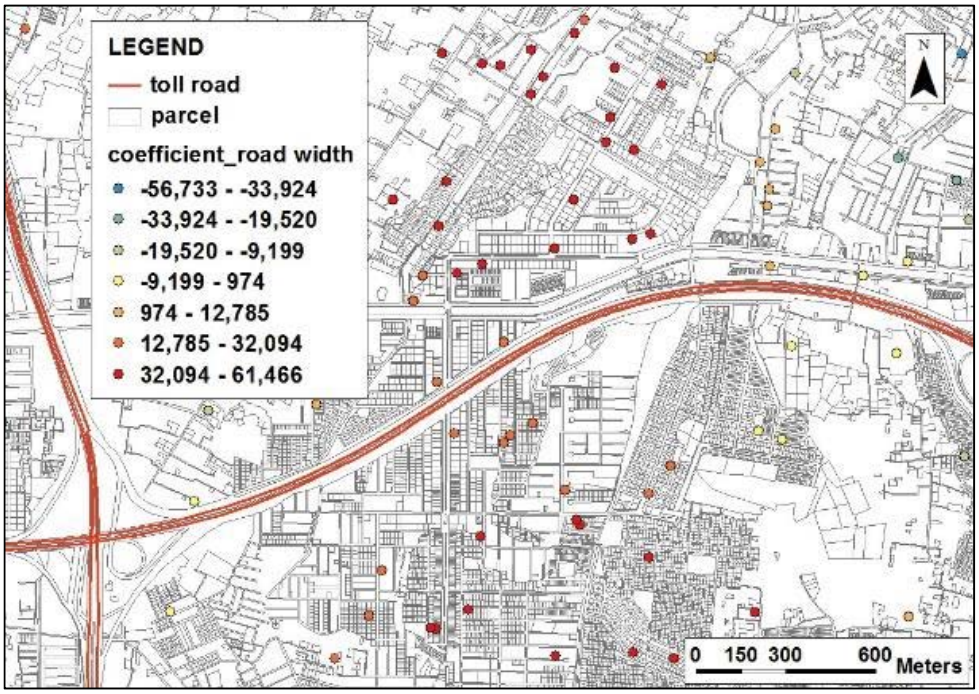


Figure 5.13 Locations with the highest positive coefficients for variable tollgate in GWR model using individual samples

Figure 5.13 shows that most of the locations with large positive coefficients are located within large residential complexes with very easy access to a toll road. It is clear that shorter travel time to the nearest tollgate is not an important matter in these areas, although proximity to the centre of the residential complex seems to increase land value at these locations. This could be an 'unmeasured' effect which is not taken into account in the model but is captured by GWR.

Locations with the largest standard errors forthe tollgate are in the central-west and the middle-west (Figure 5.14) and compared to the range of coefficient values, the range of standard errors is moderately high.A more objective comparison by location is given in the map of t values (Figure 5.15). The number of locations with small t values for the tollgate is quite high, and the locations are well distributed across the study area. The variable tollgate is quite different from variables road width and zoning, in which only few locations have low t values and these locations are somewhat clustered.

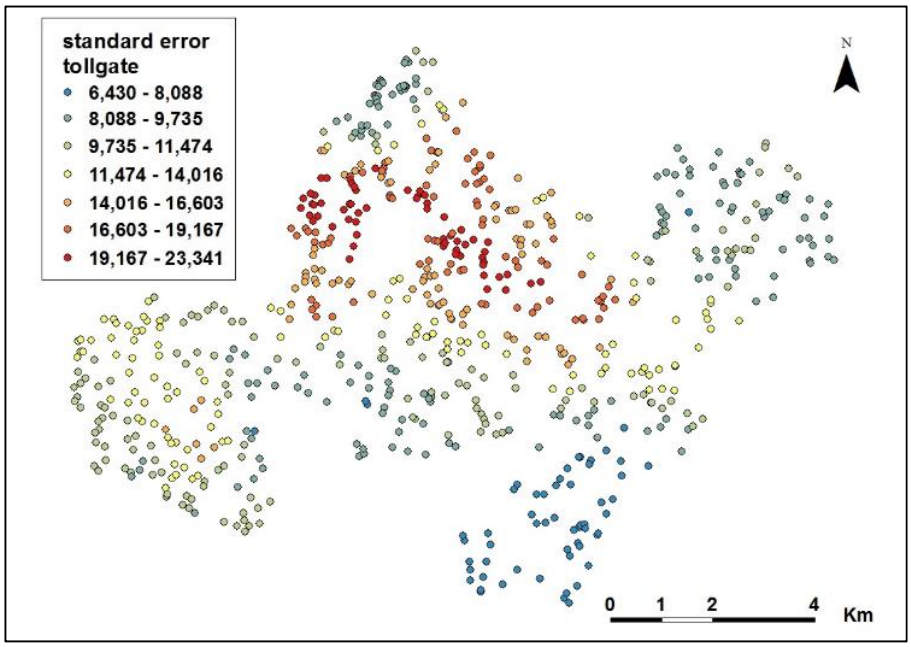


Figure 5.14 Coefficient standard error for variable tollgate in GWR model using individual samples

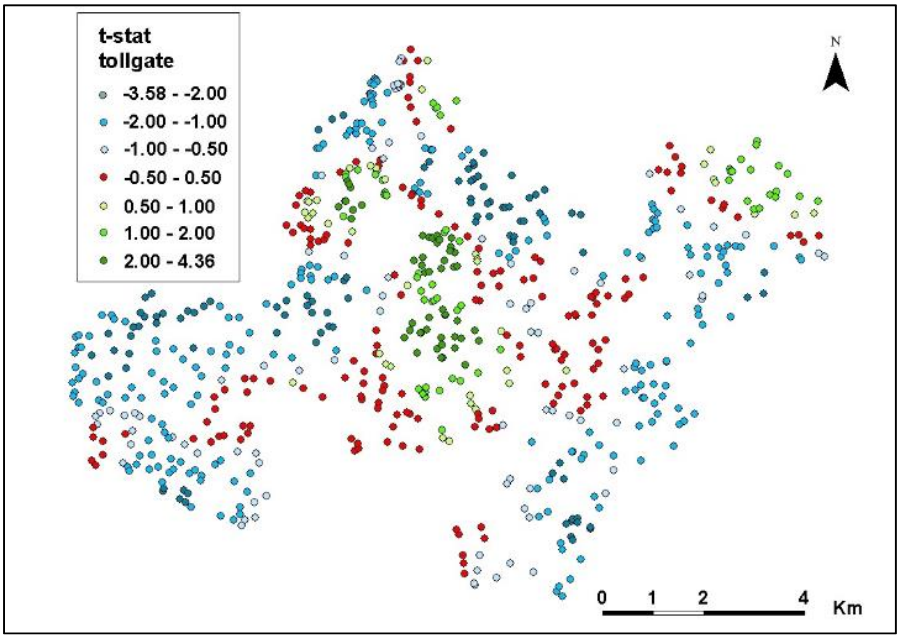


Figure 5.15 T-statistic value for variable tollgate in GWR model using individual samples

There is no clear pattern in the distribution of t values for variable tollgate (Figure 5.15). The level of reliability of the coefficient appears to vary randomly across Bekasi. This suggests

that assessments of the coefficient's reliability to decide which locations to discard from the model in order to improve prediction accuracy, need to be made at individual locations.

At the end of this section, prediction accuracy will be checked for locations with very low t values for one explanatory variable, two explanatory variables, or all three of the explanatory variables, with the goal of finding out whether or not the low precision in parameter estimation also results in low prediction accuracy. In order to makethese assessments by location, the magnitude of prediction residual at each location was calculated. The prediction residuals atall of thesampled locations aredisplayed in Figure 5.16.

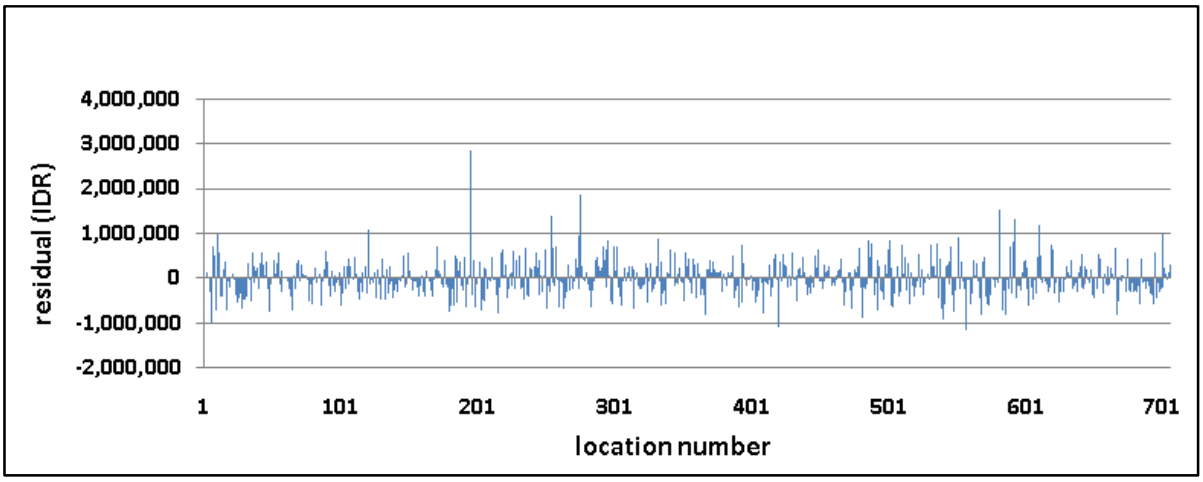


Figure 5.16 Plot of prediction residual in GWR model using individual samples

The five locations with the largest prediction residuals respectively are locations 196, 254, 276, 582, and 593, all prediction residuals at these five locations are positive. Comparison of the observed values and residuals for these locations shows that location 196 has the largest residual (2,847,766), which is 42.16 per cent of the observed value. Location 254 has the fourth largest residual (1,381,410), but it is 46.98 per cent of the observed value.

The observed land values are the benchmarks against which the residuals need to be measured. That means each residual needs to be measured against a different benchmark, and these benchmarks will vary considerably. Because of this, it is evident that the residual value is not the most appropriate measure to assess prediction accuracy in this study. Therefore, an approach was adopted which compares each residual with its own observed value as this will be more objective. The values in Figure 5.17 revealed that distribution of percentage residuals is quite different from the distribution of residuals in Figure 5.16. Some locations have extremely largepercentage residuals. Locations 126, 133, 180, 263, 306, and

694 have percentage residuals of 152.49, 108.70, 122.00, 119.70, 106.99, and 130.81 respectively. In order to investigate this issue, distribution of standardised residuals was analysed (Figure 5.18).

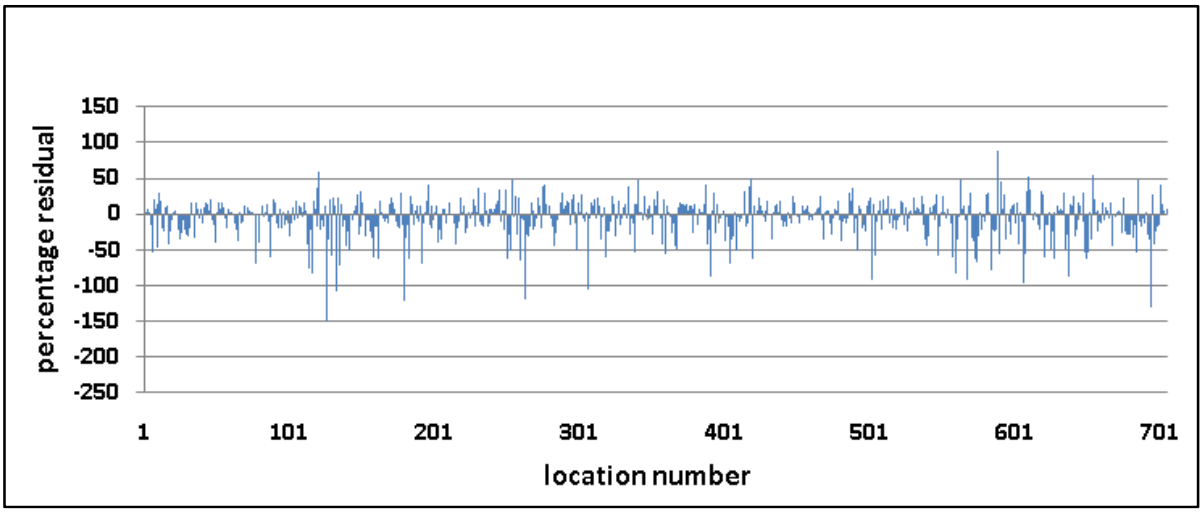


Figure 5.17Plot of percentage residual in GWR model using individual samples

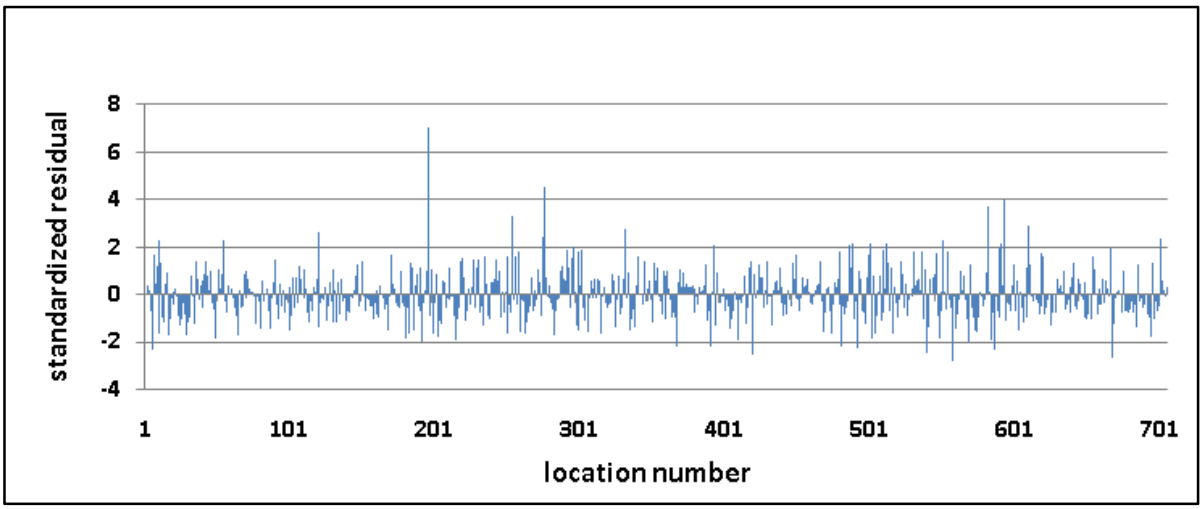


Figure 5.18 Plot of standardised residual in GWR model using individual samples

Standardised residual is a comparison between the prediction residual and the standard error of the local model at the same location. The standardised residual graph (Figure 5.18) looks quite similar to the residual graph (Figure 5.16) because the variation of standard errors among local models is relatively low. This can be explained by the coefficient of variation of standard errors among local models which is only 7.90 per cent. With nearly

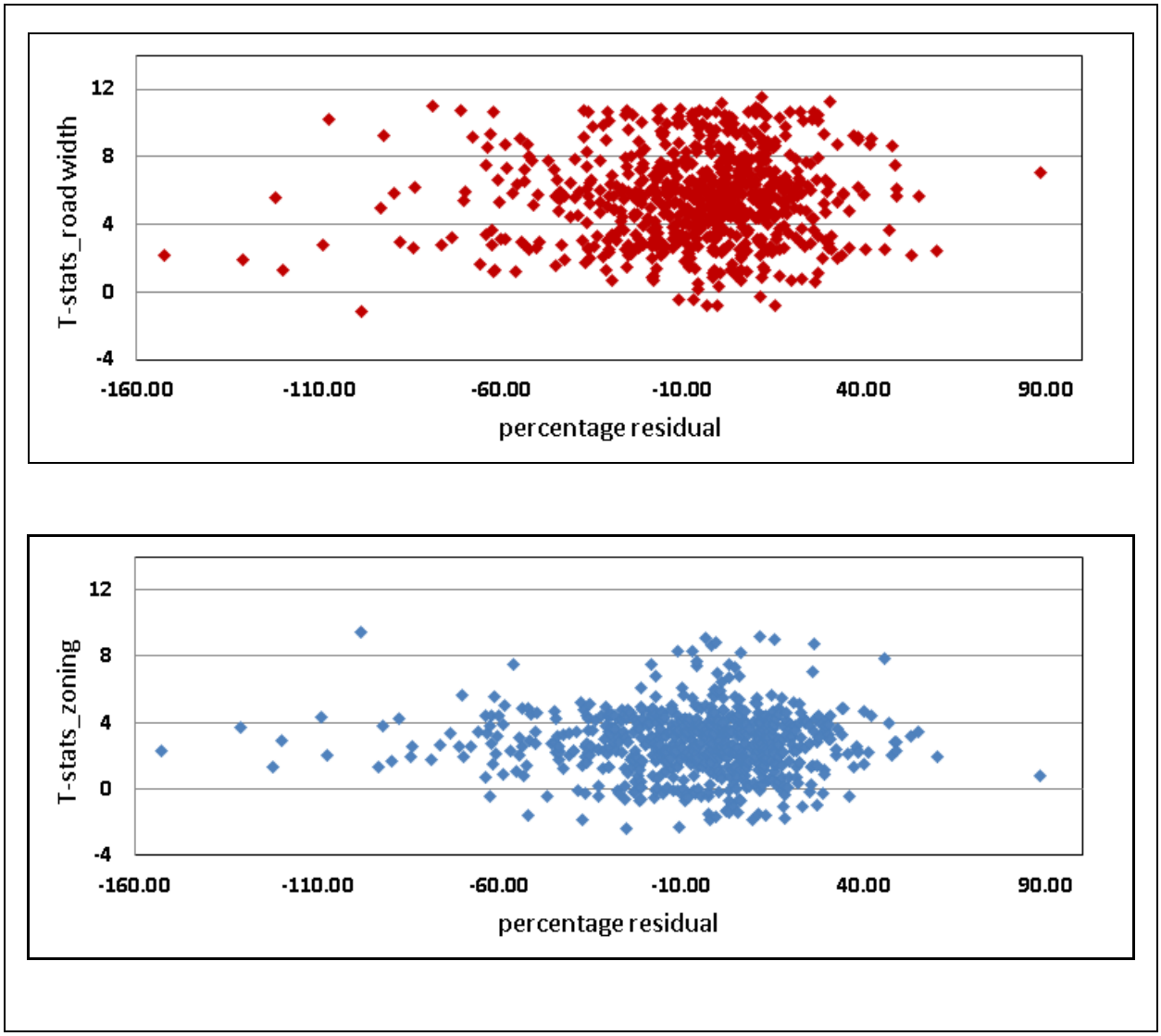
homogenous standard errors at all local models, the distribution of residuals becomes a dominating factor in shaping the distribution of standardised residuals.

The output table of the GWR model contains local R-squared values and standard errors which can be used to assess the goodness-of-fit of the local models. The local R-squared value provides a measure on how well the local model fits the data involved in a local regression, while the standard error provides a measure on precision of prediction at a location when predicted using the local model. There is no performance indicator that can be used to explain the extremely low prediction accuracy at several locations. Therefore, parameter estimates at each prediction location were examined to find out whether or not low prediction accuracy is related to low precision of parameter estimation in the local model. This examination was made with reference to t values (Table 5.4) as they indicate the reliability of the coefficients at each location.

Table 5.4 T-statistic values at locations with the largest percentage residuals in GWR model using individual samples

Location number	Percentage residual (%)	Variable	T value
126	-152.49	road width	2.16
		zoning	2.26
		tollgate	-0.30
694	-130.81	road width	1.94
		zoning	3.71
		tollgate	0.34
180	-122.00	road width	5.55
		zoning	1.30
		tollgate	-0.52
263	-119.70	road width	1.32
		zoning	2.91
		tollgate	1.39
133	-108.70	road width	2.81
		zoning	4.32
		tollgate	-0.85
306	-106.99	road width	10.22
		zoning	2.03
		tollgate	-1.92

A t value close to zero means the coefficient's standard error compared is remarkably larger than the actual coefficient value, so reliability is low. At the six locations with the largest percentage residuals, variable tollgate has small t values, but road width and zoning have moderate to large t values. Therefore, in these six cases, large residuals are not always a product of low reliability of coefficients. In other words, low prediction accuracy is not always related to low precision of parameter estimates.The next question is to what extent does this inference apply to all the sampled locations in Bekasi? To answer this, the scatter plots between percentage residuals and the t-statistic values at all sampled locations are examined for each explanatory variable (Figure 5.19).



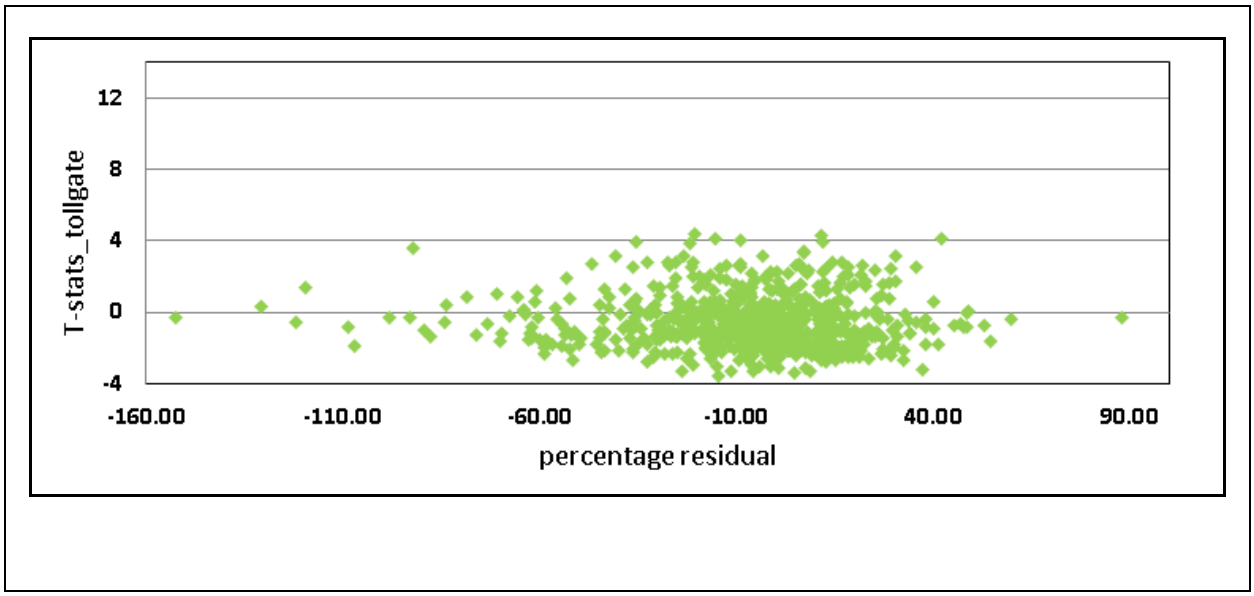


Figure 5.19 Scatterplots of T-statistic value and percentage residual in GWR model using individual samples

The graphs in Figure 5.19 show that there are no clear relationships between t values and prediction residuals for any of the three explanatory variables. This supports the earlier inference that low prediction accuracy isnot always related to low precision of the parameter estimates. A much broader inference from Figure 5.19 is that prediction accuracy is not related to the precision of parameter estimates at all.

In order to investigate this new inference, associations between percentage residuals and t values were examined for each explanatory variable. The correlation coefficients between percentage residuals and the t values for road width, zoning, and tollgate (0.067, 0.018 and -0.047 respectively) were very low. These results lend weight to the inference that prediction accuracy is not related to the precision of the parameter estimates for any of the explanatory variables.

As there were no measures from the GWR output table that could be used to explain the extremely large prediction residuals, locations with these types of residuals were examined individually. Location 133 exemplifies this. It is located between 132 and 137 along a local street (Figure 5.20). The road width varies along this street and at the locations 132, 133 and 137 the widths are 2.29 m, 4.50 m, and 3.67 m respectively.

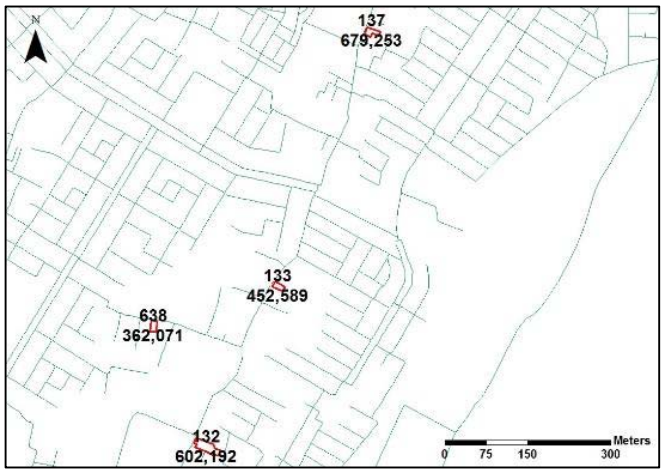


Figure 5.20 Example of a location (location 133) with a high prediction residual examined individually

In Indonesian cities generally, land along wider roads has higher value than land along narrower roads; and most of the coefficients for road width in the GWR local regressions confirm that road width usually has a positive relationship with land value as displayed in Figure 5.5 earlier in this section. Using this logic, location 133 should have the highest land value of the three samples and 132 the lowest. However, the observed data shows that 133 has the lowest land value, in fact it is 28 and 40 per cent lower than that at 132 and 137 respectively (Table 5.5).

Table 5.5 Predicted and observed land values for locations 132, 133 and 137

Sample number	Observed value (in IDR)	GWR Model	
		Prediction residual (%)	Predicted value (in IDR)
133	452,589	-108.70	944,569
132	602,192	11.57	532,510
137	679,253	-18.37	804,064

Although the observed data at location 133 goes against the predominant relationship, the GWR model maintains this relationship as most of the samples show a strong positive relationship between road width and land value. Therefore, location 133 with its large prediction residual is an outlier whose land value is influenced by unknown factors.

Like location 133, a number of other locations have large residuals. Six out of 706 locations come up with percentage residuals larger than 100 per cent, and 47 locations come up with percentage residuals between 50 to 100 per cent. Despite the issue of very large percentage residuals at a number of locations, the model performs moderately well as a whole with a mean absolute percentage error (MAPE) of 19.40 per cent which is classed as moderately good. MAPE is a measure of average prediction accuracy.

$$\boldsymbol{MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{O_i - P_i}{O_i}\right|} \qquad \textbf{(Equation 5.4)}$$

Where: n is the number of observations

O_i is the observed value at location i

P_i is the predicted value at location i

A cut-off value is required to assess each prediction location, and the cut-off value of 30 per cent was adopted from the Zonation Method which is currently employed in BPN RI. The GWR model, as a whole, is quite reliable for mass valuation when the cut-off value of 30 per cent is used. It is true that the MAPE value of 19.40 per cent is much lower than 30 per cent but there are a number of locations with percentage residual significantly larger than 30 per cent. At individual locations, 137 out of 706 locations (19.41 per cent) do not comply with this cut-off value.

The above results indicate that the in-sample GWR model, as a whole, has moderately good prediction accuracy and prediction precision. Hastie et al. (2009) described an in-sample model as a training set model which may overfit the training data in order to minimise the training error, and the optimally minimised training error is not a good estimate of the expected actual prediction error. In the case of the in-sample GWR model explained above, the model was built using all the sampled locations. During the process of minimising the error, the model was optimally fitted to all these sampled locations. Being only optimally fitted to a set of sampled locations, as indicated by the minimum training error, the model's prediction ability maybe overrated in the way Hastie *et al.* (2009) explained. Therefore, a GWR model with an out-of-sample validation was also undertaken on the data.

5.3.2. Out-of-sample estimation of GWR model with individual locations

There are a number of out-of-sample validation techniques. K-Fold Cross Validation, Monte-Carlo Cross Validation, and Bootstrap are well-known, and all of these techniques share a common basic procedure, i.e., the dataset is resampled to create a training subset and validation subset (see Anguita *et al.*, 2012). The training subset is used to build the model, while the other subset is used to validate the predictions.

In a K-Fold Cross Validation (K-Fold CV), all samples are divided into K equally sized subsets. First, subset 1 is used as the validation subset and all other subsets are used as a training subset. Next, subset 2 is used as the validation subset and all other subsets are used as the training subset. This process is continued until subset K is used as the validation, and this gives K iterations of the validation process(see Hastie *et al.*, 2009).

Monte-Carlo Cross Validation (MCCV) is also known as the Repeated Random Sub-sampling Validation or the Repeated Hold-out Method. The concept was introduced by Burman (1989), and was originally called Repeated Learning-Testing Method. In this approach, the original dataset is split into training and validation subsets multiple times. Each time, samples for the training subset and samples for the validation subset are randomly selected. The number of samples in the training subset is the same for each iteration, as is the size of the validation subset. The number of possible iterations is calculated as follows.

$$\textbf{Number of iterations} = \frac{\boldsymbol{n}!}{\boldsymbol{t}!\boldsymbol{x}\ \boldsymbol{v}!} \qquad \textbf{(Equation 5.5)}$$

Where: $n!$ is the number of total samples-factorial

$t!$is the number of training samples-factorial

$v!$is the number of validation samples-factorial

As a result of random selection, a sample can be used in a number of training or validation subsets. Conversely, it is possible that a sample never joins even one training subset or validation subset if the number of iterations undertaken is less than the maximum possible iterations.

Efron (1979) introduced the Bootstrap method, in which a sample is randomly picked up from the original dataset to form a bootstrap dataset but then that sample is returned to the original dataset allowing it to be selected again. Bootstrap method performs a resampling with the replacement that allows a sample to be picked multiple times in a bootstrap dataset

and there is no limit on the number of possible bootstrap datasets. When the performance of a certain model is to be examined, the model is fitted to the bootstrap datasets and these bootstrap datasets are used as the training datasets. The bootstrap models are then utilised to predict the original dataset, which acts as the validation dataset.

The bootstrap datasets and the validation dataset have samples in common, and the overlap results in good predictions but may overestimate the prediction ability of the model. The .632+Bootstrap method is designed to overcome this overfitting problem (see Hastie *et al.*, 2009). The idea is similar to the Leave-One-Out Cross Validation (LOOCV). For each location, only predictions from bootstrap datasets not containing the sample at that particular location are taken into account.

Expected to solve the overfitting issue while maintaining low bias, the .632+Bootstrap method has a larger bias problem than the Cross Validation (CV) methods (Kim, 2009), which have lower bias but higher variance than .632+Bootstrap. Consequently, there is a trade-off between bias and variance in selecting the method to use. Putting aside the trade-off between bias and variance, CV methods are preferable to Bootstrap method for this work because CV methods completely separate the samples in the training and validation subsets. By doing so, CV methods are expected to give a more objective assessment of the model's performance. The next step is to choose between K-Fold CV and MCCV.

Molinaro *et al.* (2005) found that MCCV has a slightly lower bias than K-Fold CV. MCCVs have heavy computation loads, though this is not as big an issue as it used to be, given increases in computing power. There are at least two more reasons to choose MCCV over K-Fold CV. First, a much larger number of iterations can be run with MCCV. Second, samples are randomly split into training and validation subsets in MCCV, whereas in K-Fold CV the samples are split into K groups manually. Added to which, with a large number of iterations in MCCV a large number of variations can be captured. This is an advantage over K-Fold CV in which each sample stays in the same group in all iterations, so there is less freedom to combine samples to form the training and validation subsets. When all these factors were taken into account MCCV was chosen for this research.

There is no exact prescribed ratio between the numbers of samples in training or validation subsets in MCCV. It was decided to use 80 per cent of the samples (565) to build the model, and 20 per cent (141) for validation in each iteration. The number of iterations that can be run is determined by the number of combinations that can be made from the samples. When selecting 565 training samples out of 706 in total, the number of combinations is:

$$\frac{706!}{565! \, x \, 141!}$$

The number of possible combinations cannot be calculated for this work because it is extremely large, and it is far greater than normal software packages can calculate. Therefore, an arbitrary decision was made to run 1,000 iterations of the GWR model which is built using the 565 training samples and 141 locations of validation data (Figure 5.21). All 1,000 iterations were successful. It is possible to obtain multiple predictions to compare against one observation value at a location given the number of iterations (Figure 5.22).

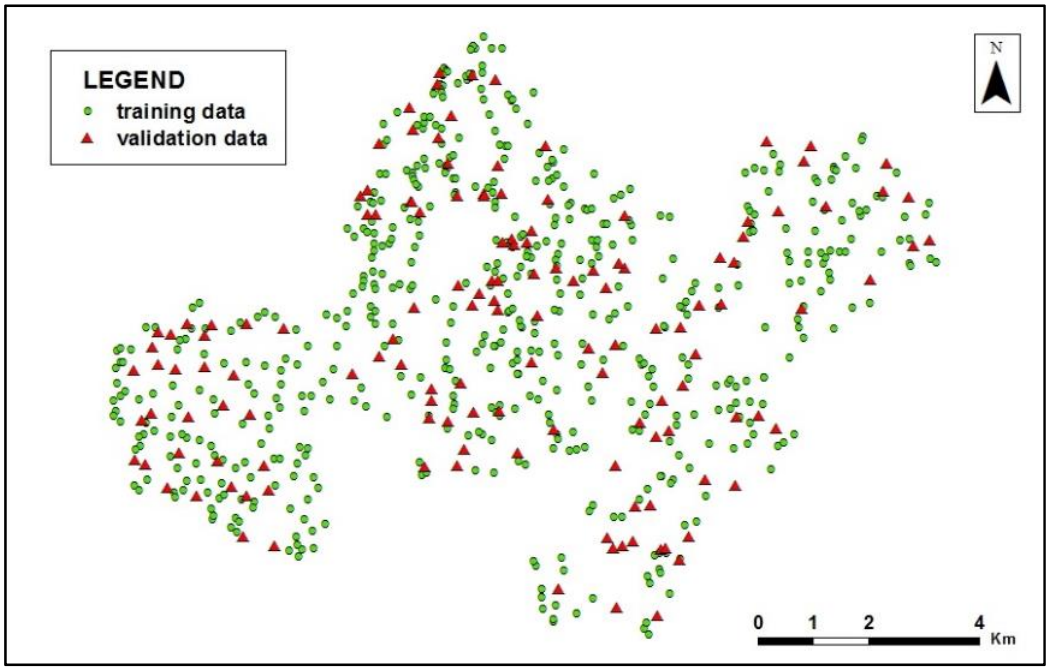


Figure 5.19 Distribution of training and validation samples for the first iteration of MCCV on GWR prediction using individual samples

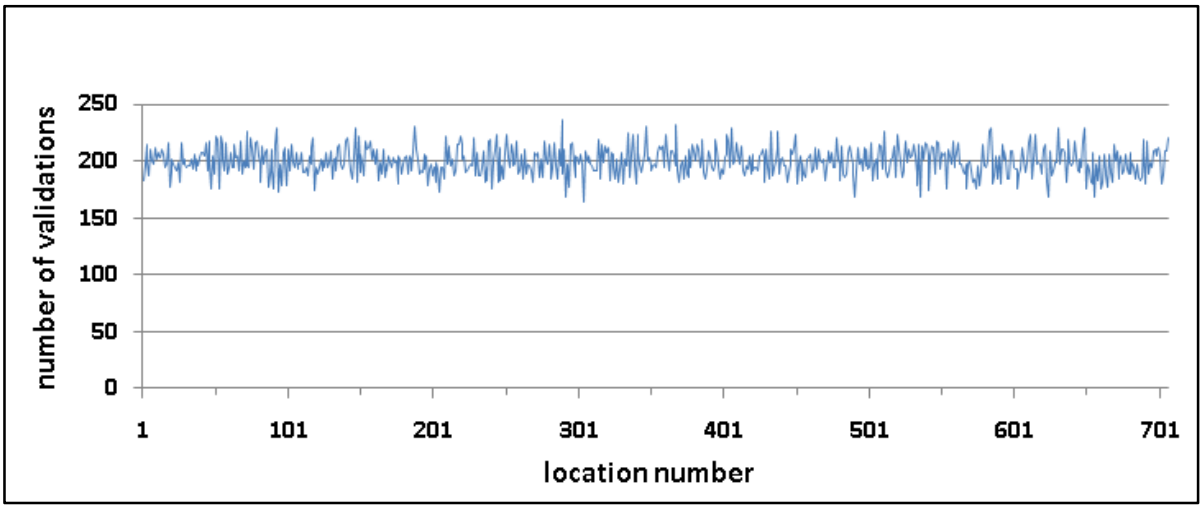


Figure 5.20 The number of predictions that were validated for each location in MCCV on GWR prediction using individual samples

The lowest frequency of validation at a location is 164 times and the highest is 237. On average, a location is predicted 200 times using 200 different GWR models with randomly selected inputs. Although the sample selection for training and validation subsets is designed to be random, samples (sampled locations) have relatively equal chances to be validated. Multiple predictions at each location were compared to the observed value at the location, so there are multiple prediction residuals at each location. From these multiple prediction residuals, an average prediction residual was calculated to provide the average prediction accuracy at each location. An immediate concern is that positive and negative residuals at one location counteract each other. In order to have a clearer picture of the distribution of residuals, the range and the average of percentage residuals at each location are displayed at the same time (Figure 5.23).

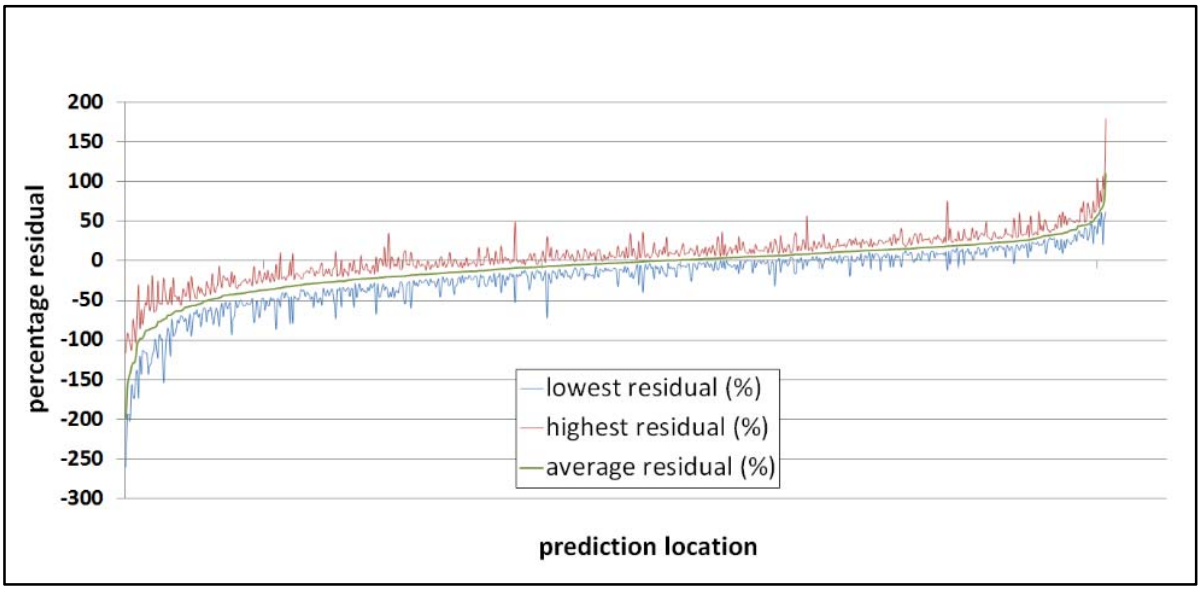


Figure 5.21 Range and average of percentage residual in MCCV on GWR prediction using individual samples, plotted in increasing value of the average residual

For each location, the average percentage residual represents the prediction accuracy and the range of percentage residuals represents the prediction precision. It is hoped a location will have an average percentage residual close to zero and a small range, and these can only be achieved when the around 200 predictions are close to the observed value at the corresponding location. But a significant number of locations do not meet the ideal expectation, as revealed by Figure 5.23. There is no clear pattern in the relationship between the average percentage residual and the range of percentage residual. Among locations with the largest ranges of percentage residuals, those with the largest negative or positive average percentage residuals and the one with average percentage residual closest to zero, were selected to be discussed. Among locations with the smallest ranges of percentage residuals, the ones with largest negative or positive average percentage residuals and the one with average percentage residual closest to zero were also selected to be discussed (Figure 5.24).

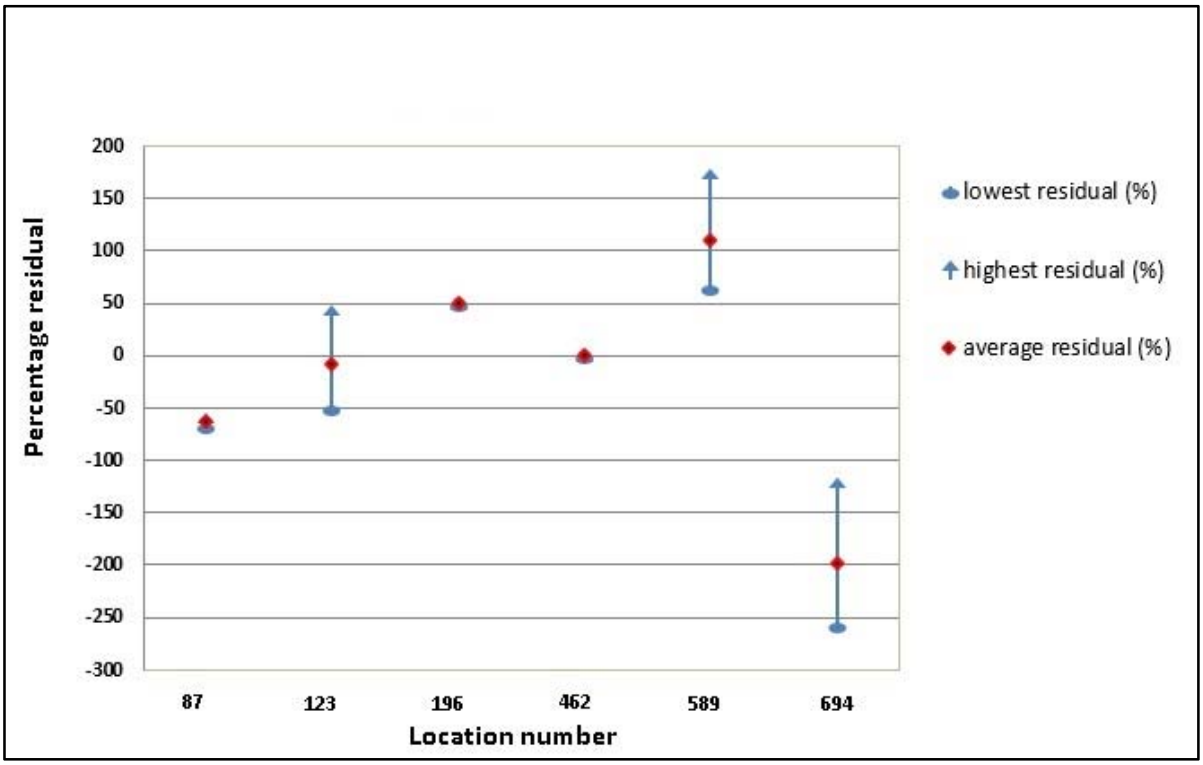


Figure 5.24 Ranges and average percentage residual at selected locations in MCCV on GWR prediction using individual samples

Locations 87,196, and 462 have small ranges of percentage residuals. Out of these three locations, only 462 has a high prediction accuracy. It was validated 193 times, and all residuals were close to zero. The prediction model works well for this location. In contrast, the predicted values were always larger than the observed value at location 87; whereas at 196 they were always under predicted. Both locations have high prediction precision, like 462, but the prediction accuracies are quite low. The average percentage residuals are -62.55 and 49.30 for locations 87 and 196 respectively.

Locations 196 have around 200 predictions which are similar to one another, so this location has high prediction precision. However, all of the predictions are quite different from the observed value. This raises questions about the quality of the observed data, and Location 196 is discussed further (Figure 5.25).

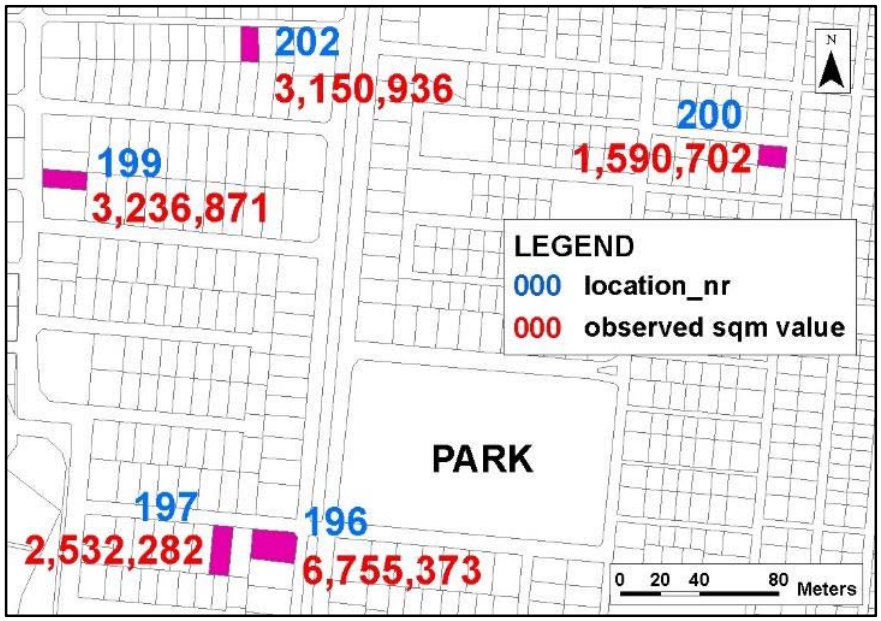


Figure 5.22 An example of a location (location 196) with a large average and a small range of percentage residuals

The locations in Figure 5.25 are within a residential complex, so the value for variable zoning is the same for all five locations. In addition, travel times to the nearest tollgate are quite similar. The map shows that the five samples are located on road segments of various widths and, in fact, this is the only variablewith any meaningful variation.

Location 196 is located on a wider road segment compared with other locations, and the observed land value is also higher. However, even though land value and road width are positively correlated in the model, the observed value for location 196 is far greater than the model predicts. The highest predicted value from 204 validations was 3,625,000 IDR; which is around half of the observed value of 6,755,000 IDR. There are several possible reasons for this large difference. First, the recorded observed value can actually be higher than the market value. In which case, 196 can be considered as a positive outlier. Another possibility is that there are ‘unmeasured variables’ that are contributing significantly to price variations in this location. In fact 196 is located on a boulevard in the residential complex, on the corner of a block and across the main park. These factors may have a significant contribution in shaping land price in this residential complex but they were not incorporated in the prediction model. Because of these possibilities, a further examination including a field check may be required to sort out cases like this in mass valuation practice by BPN in the future.

A quite different situation is found at locations 123, 589, and 694 where prediction residuals are 101.81 per cent of the observed value (location 123), 117.41 per cent (location 589) and 144.25 per cent (location 694). In fact location 694 has the largest range of percentage residuals of all locations. Besides the issue of a large range of prediction residuals, locations 589 and 694 also have large average prediction residuals; 589 is under predicted by 109.84 per cent value while 694 is over predicted by 198.59 per cent. With extremely low prediction accuracies and extremely large ranges of prediction residual, it can be inferred that the prediction model does not work well for these locations. However, even though 123 has a large range of residuals, the average residual at location 123 is only -8.90 per cent of the observed value. A closer look at location 123 is therefore warranted.

Figure 5.26 shows that many predictions are quite accurate. Recalling the cut-off value to assess accuracy in the in-sample GWR estimation, 186 out of 202 (92.1 per cent) predictions have prediction residuals between -30 to +30 per cent of the observed value, so these predictions are considered valid. Concluding that the prediction model performs poorly at location 123 by the large range of percentage residuals alone, will disregard the above fact. Though with a large range of percentage residuals, a very dominant portion of all predictions are considered valid.

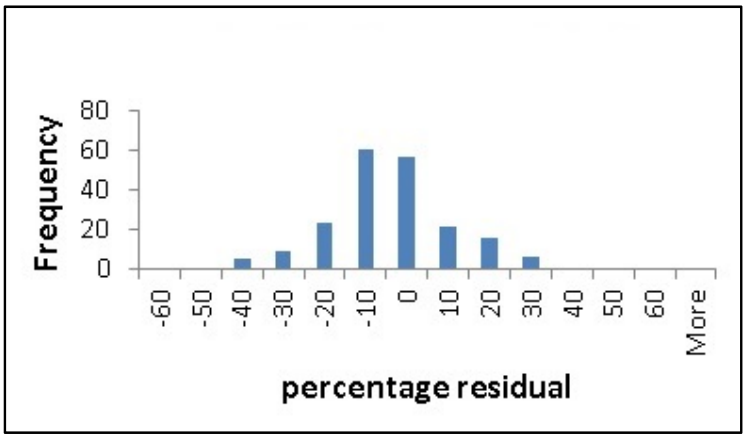


Figure 5.23 Distribution of percentage residuals at an example of location (location 123) with a small average percentage residual and a large range of percentage residuals

Careful examination is required when using the range of percentage residuals to assess the variability of residuals at any location. The standard deviation of the percentage residual is a better option to use in these examinations. The standard deviation of the residual for location 123 is 15.18 per cent of the observed value, which is far below the cut-off value of 30 per

cent adopted from BPN standards. The magnitudes of the standard deviations of percentage residuals at all locations are illustrated in Figure 5.27.

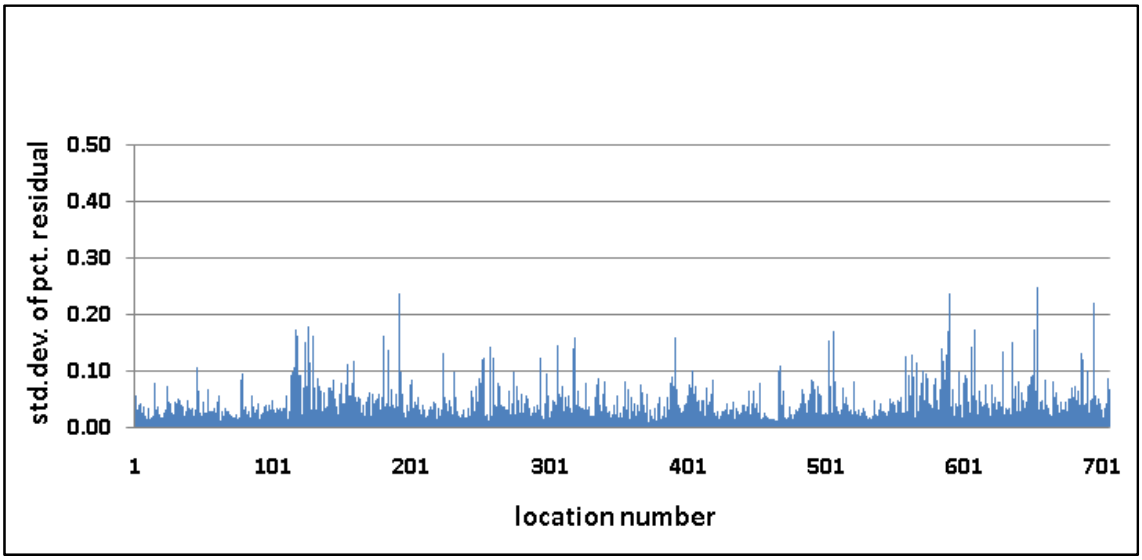


Figure 5.24 Plot of standard deviation of percentage residuals in MCCV on GWR prediction using individual samples

The above graph clearly shows that all 706 locations have standard deviations less than the 30 per cent cut-off value. Hence all locations are considered to have reliable prediction precision, and assessment of prediction performance can be based on prediction accuracy. The prediction model was considered to work well at locations with average percentage residuals between -30 to +30 per cent of the observed value. If the distribution of absolute average percentage residuals is displayed (Figure 5.28), instead of the distribution of average percentage residuals, locations with extremely large average residuals are more readily apparent.

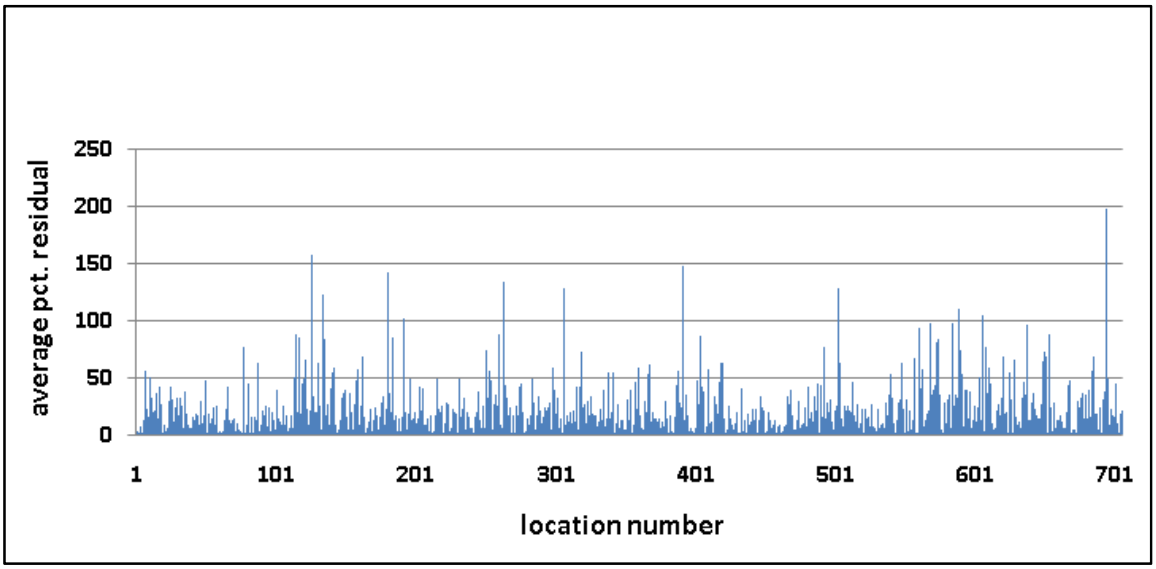


Figure 5.25 Plot of absolute average percentage residuals in MCCV on GWR prediction using individual samples

Given the distribution in Figure 5.28 the prediction model is considered to work well at 530 locations (75.07 per cent out of all locations), i.e., locations with absolute average percentage residuals smaller than 30 per cent of the corresponding observed values. With nearly a quarter of all locations having accuracy issues, it can be said that the prediction model does not work really well on the Bekasi dataset.

The main lesson learned from the above analysis of location 196, is that a location with a large average percentage residual (low accuracy) and small standard deviation of percentage residual (high precision) must also be carefully examined. Out of 175 locations with average percentage residuals larger than 30 per cent, 74 have standard deviations of residuals smaller than five per cent of the observed values and 70 locations have standard deviations between five to ten per cent of the observed values. Hence a high proportion of locations with low prediction accuracy require further examination to find out whether there are significant effects of ‘unmeasured variables’ in the predictions or whether the observations are inaccurate.

5.4. Summary

5.4.1. Prediction accuracy

The mean absolute percentage error (MAPE) value is used to indicate the prediction accuracy of a method as a whole. With in-sample validation of the GWR model,the MAPE was 19.40 per cent. This is moderately larger than the MAPE from the Zonation Method currently employed by BPN, which when applied to this dataset is 10.8 per cent. Therefore, although the GWR model solves the main problem of the Zonation Method, i.e. non-verifiable prediction for zones with fewer than three samples, the Zonation Method outperforms the GWR model in terms of accuracy for the Bekasi dataset.

The cut-off value of 30 per cent, which is currently used for mass valuation in BPN, was adopted to assess prediction performance at individual locations from the GWR model. Just over 80 per cent (569 of 706) of locations were less than the cut-off value in terms of the absolute percentage residuals from the in-sample GWR model. A slightly smaller number of locations, 530, had absolute average percentage residuals less than 30 per cent when using GWR with out-of-sample estimation. When 1,000 iterations of GWR model were run using MCCV, the average results were quite similar with the results from the original GWR model (Figure 5.29).

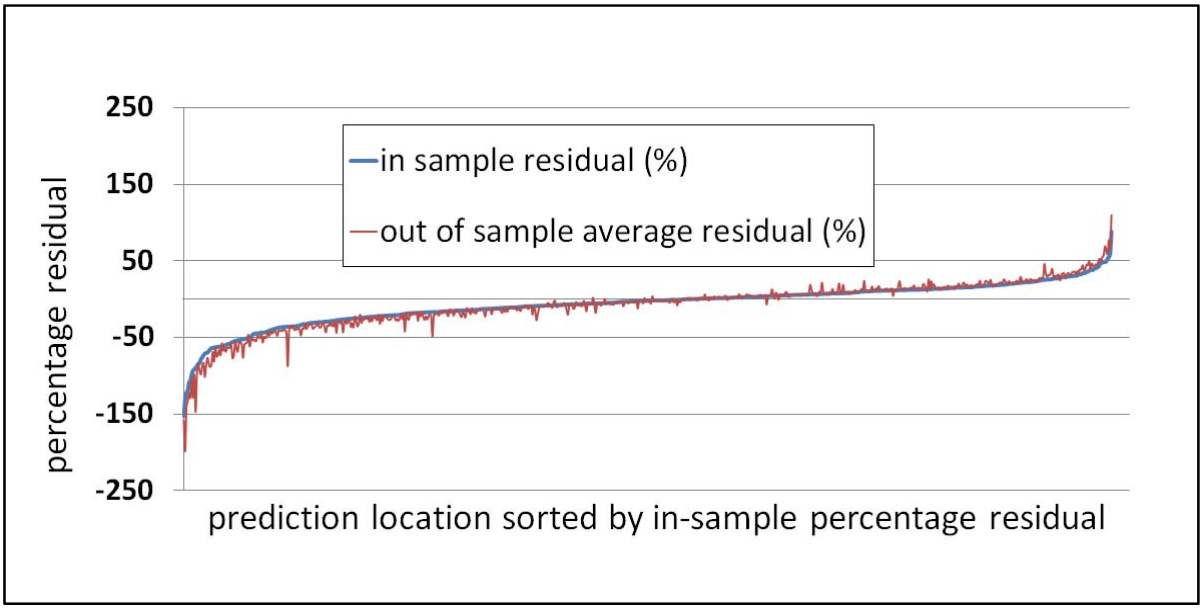


Figure 5.26Average percentage residuals from out-of-sample estimation and percentage residual from in-sample estimation, plotted in increasing value of the in-sample residual

The correlation coefficient between out-of-sample average percentage residuals and in-sample percentage residuals was 0.987. This high correlation coefficient indicates that the GWR prediction model does not have a significant issue of overfitting; an issue that arises when a model overfits the sampled locations to minimise the training error. The objective of conducting an out-of-sample estimation was to explore the prediction performance of the prediction model when applied to non-sampled locations. With no significant overfitting in the GWR prediction model, the prediction accuracy at the non-sampled locations was expected to be similar with the prediction accuracy at the sampled locations.

5.4.2. Prediction precision

Prediction precision was only examined for the out-of-sample estimation as there are multiple predictions for each location. The range of percentage residuals gives a clear picture of the extent of percentage residuals at each location, and in turn indicates the level of prediction's uncertainty at a specific location. However, it tends to give a pessimistic assessment on precision because it focuses on the highest and lowest values. In fact, there are normally only a small number of predictions that are close to the ones with the highest or lowest percentage residuals. With a huge number of iterations at each location, percentage residuals tend to be normally distributed. The number of predictions with percentage residuals close to the average percentage residual are normally much higher than the ones where percentage residuals deviate far from the average percentage residual. Standard deviation is considered to be a more appropriate measure of prediction precision at each location in this study.

All locations had standard deviations of percentage residuals lower than the cut-off value of 30 per cent. Of the 706 locations, 482 (68.27 per cent) had standard deviations of percentage residuals lower than five per cent of the corresponding observed values. Approximately a quarter of locations, 178, had standard deviations of percentage residuals between five to ten per cent of the corresponding observed values. These proportions indicated that the GWR prediction model is quite likely to be able to maintain consistent predictions at non-sampled locations.

5.4.3. Extremely large residual at several locations

In the in-sample validation analysis using the GWR prediction model, several locations were detected with extremely large percentage residuals. These locations contribute significantly

to reducing the overall accuracy of the model as whole. Five per cent of samples with the largest absolute percentage residuals contribute 20.69 per cent of the total absolute percentage residual. This is the main issue with the GWR prediction model.

An extremely large percentage residual at one location is very likely to be caused by anomalous observation at that location. The sales price of a sampled land parcel can be significantly lower or higher than the market price for various reasons. When compared to the predicted price, this anomalous observed price will result in a large percentage residual. Deleting the sample at a location with large percentage residual is an easy way to solve the problem. However, this should be done very carefully with solid reasoning.

A lesson learned from the Zonation Method is that neighbouring land parcels in one value zone tend to have similar prices, so the observed prices at the sampled locations in one zone are expected to be similar. An anomalous observation can cause a high variation of price among observations in a zone. The coefficient of variation on price among samples in a zone can be a sensible measure to use to detect anomalous observations. A zone with large variation on observed price is considered to be less reliable than a zone with small variation on observed price.

An immediate concern is that the coefficient of variation is not a proper measure in a zone containing only one or two samples. In a zone containing only one or two samples, it is difficult to detect the anomalous sample. For this reason, observations in a zone containing at least three samples with low variation on price are considered to be more reliable than observations in a zone containing only one or two samples. The number of samplesin a zone can be used as a measure of reliability among zones. A zone containing a large number of samples is considered to be more reliable than a zone containing a small number of samples.

The value zones can be employed to control the observation data effectively, so the GWR model will be run using value zones. In order to minimise the effect of undetected anomalous observations in the GWR model, a weighting scheme will be applied on zones based on the number of samples. If anomalous observations from zones containing one or two samples are not detected and therefore are involved in the model, their contributions to shape the model will be less than the normal observations. The GWR model using value zones is expected to overcome the issue of extremely large percentage residuals at several locations. The processes and results will be thoroughly examined in the next chapter (Chapter Six).

6. GEOGRAPHICALLY WEIGHTED REGRESSION (GWR) MODELLING WITH VALUE ZONES

6.1. Introduction

As discussed in Section 5.4, the main issue with the Geographically Weighted Regression (GWR) model using individual locations is the extremely large residuals at a number of locations. This issue is going to be addressed by controlling the input data using value zones. The feature to be used for GWR analysis is the valid value zone. In the Standard Operating Procedure within the Land Valuation Directorate of the National Land Agency of Indonesia (BPN RI), avalid land value zone must contain at least threesales data records. Hundreds of valid value zones are required because GWR works well for a dataset of several hundred features. The distribution of value zones for the Bekasi dataset is shown in Figure 6.1 and the composition of zones based on the number of samples is shown in Table 6.1.

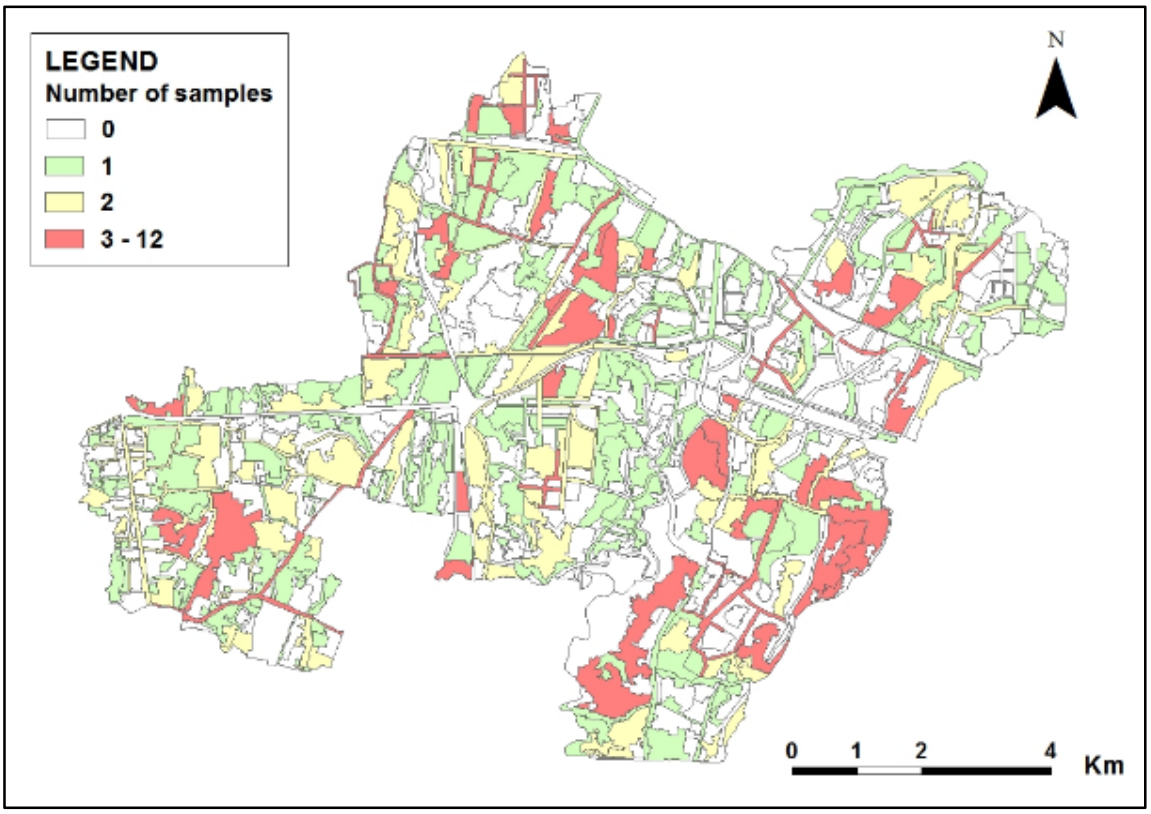


Figure 6.1 Distribution of value zones in Bekasi with the number of samples in each zone

Source:Land Office of Bekasi (2012b)

Table 6.1 Composition of value zones in Bekasi based on the number of samples

Number of samples	Number of value zones	Percentage breakdown
≥ 3	50	5.1
2	110	11.2
1	280	28.6
0	540	55.1
Total	**980**	**100.0**

Source: Land Office of Bekasi (2012b)

Unfortunately, the number of valid value zones in the Bekasi dataset is too small to run with GWR. Referring to the operational procedure of BPN, only 50 zones can be analysed further. An exception is made in this PhD study, and use is made of zones containing two samples where the difference between the sqm values of the two samples in such zones should be lower than 30 per cent of the average value of both samples. This increases the number of zones that can be run with the GWR model to 160. Another exception is made also to make use of the value zones containing only one sample. The objective is simply to enlarge the number of input featuresas the GWR model works best with hundreds of features, and this exception increases the valid value zone to 440.

Weighting among zones

A zone containing at least three samples has a higher reliability level than a zone containing only one sample. The number of samples in a zone determines the level of importance of each zone: those with more samples contribute more significantly to shaping the model compared with zones with fewer samples. For this reason, the number of samples is set as the weight for each zone.

Explanatory variables

The set of explanatory variables for the GWR model using value zones is different from that in Chapter 5. Road width, road class, and parcel size are not used in this GWR model because a value zone normally has a number of land parcels with various road widths, road classes, and parcel sizes. Travel times to major roads are also set aside because some zones even have segments of major roads located within them. The explanatory variables used are zoning type and travel times to amenities. All the variables related to travel times to

amenities listed in the data collection survey form are put into the statistical test to find out which are statistically significant for the model and do not cause multicollinearity.

Location

A value zone consists of a number of individual land parcels, but one location is required to represent the cluster of parcels. The centroid of each zone is used for simplicity and consistency, however, some centroids fall outside the zones. For these cases, the ArcGIS facility to determine the most representative point within the polygon is used. These central locations are used as starting points to calculate travel times to amenities.

6.2. Data examination

The input data was examined to identify statistically significant explanatory variables, and to address the potential issues of multicollinearity and spatial autocorrelation. Before undertaking the processes of data examination, the linearity between price and each explanatory variable was assessed.

6.2.1. Variable transformation

The dataset was first examined using a global Ordinary Least Square(OLS) model. The OLS model assumes linear relationships between the dependent variable and each of the independent variables. Transforming the independent variables is sometimes required to improve the linearity of the relationship. Each explanatory variable was paired with the dependent variable and tested against a number of transformation models. For each relationship between an independent and dependent variable, the highest R-squared value was used to indicate the most linear relationship between land value and each explanatory variable.

As discussed in Chapter 4, Power, Compound, S, Growth, Exponential, and Logistic transformation models could probably complicate the relationship between land value and the explanatory variables. Only linear, logarithmic, and inverse transformation models were used for the linearity estimation test. The results of the transformation models are given in Table 6.2, with the models selected in bold.

Table 6.2R-squared values for each transformation model

Explanatory variables	R-squared values for each transformation model		
	Linear	Logarithmic	Inverse
Zoning number	**0.649**	0.531	0.322
Tollgate	**0.075**	0.063	0.029
Business centre	0.037	0.050	**0.061**
Marketplace	0.001	**0.003**	0.002
Hospital	**0.018**	0.016	0.012
School	0.000	0.001	**0.002**

Source: Data analysis

6.2.2. Prediction Model

An OLS model was first run on the data, and the diagnostic report is presented in Table 6.3.

Table 6.3 OLS diagnostic summary

Number of Observations: 440	Akaike's Information Criterion (AICc) [d]: 12836.743004
Multiple R-squared: 0.683568	Adjusted R-squared: 0.679183
Joint F-Statistic [e]: 155.896750	Prob (>F), (12,693) degrees of freedom: 0.000000*
Joint Wald Statistic [e]: 599.741703	Prob (>chi-squared), (12) degrees of freedom: 0.000000*
Koenker (BP) Statistic [f]: 54.017445	Prob (>chi-squared), (12) degrees of freedom: 0.000000*
Jarque-Bera Statistic [g]: 81.032578	Prob (>chi-squared), (2) degrees of freedom: 0.000000*

Although all explanatory variables were effective in the OLS model, not all are statistically significant. Table 6.4 shows that zoning type, travel time to nearest tollgate, inverse travel time to business centre, and travel time to nearest hospital are significant. From this point onward, the short forms used for these variables are zoning, tollgate, business centre (CBD), and hospital.

Table 6.4 Summary of OLS model variables

Explanatory variable	Robust probability	Coefficient
Zoning nr	*0.000000	221311.68
Tollgate	*0.008275	-7302.74
1/ Business centre	*0.032026	927362.69
Ln marketplace	0.387001	-32295.03
Hospital	*0.004328	9626.49
1/ School	0.384332	-19385.38

* explanatory variable is statistically significant at 99 per cent confidence level

6.2.3. Multicollinearity

It is very important to ensure that there is no significant dependency among explanatory variables. Collinearity exists when there are nearly linear dependence relations among independent variables, sosome variables can be nearly linear combinations of other variables (see Bingham and Fry, 2010). In order to come up with solid inferences about the explanatory variables to be selected for the prediction model, a backward elimination regression was also run on the data (Table 6.5).

Table 6.5 Summary of collinearity statistics from backward elimination regression

Variable	Variance Inflation Factor (VIF)		
	Step 1	Step 2	Step 3
Zoning number	1.051	1.051	1.050
Tollgate	1.358	1.357	1.357
1/ Business centre	1.127	1.127	1.119
Ln marketplace	1.058	1.057	
Hospital	1.372	1.368	1.317
1/ School	1.008		

The results from the backward elimination regression (Table 6.5) confirm the results from the OLS model. The school and marketplace variables were not statistically significant. All of the variables have low Variance Inflation Factor (VIF) values in the original model, so the original model had a relatively low level of multicollinearity among variables. Therefore, although not

statistically significant, variables school and marketplace would not harm the original model in terms of multicollinearity. Eliminating these variables does not change the VIF values of the remaining variables significantly. The decision whether or not to eliminate variables school and marketplace will be made after analysing the results from spatial autocorrelation test.

6.2.4. Spatial autocorrelation

Griffith (2009) suggested that the existence of spatial autocorrelation can be used as an indication of the existence of other issues in the model; issues of missing variables, model misspecification, redundant information, failure to capture spatial processes, and areal unit problem. The level of spatial autocorrelation in the model must be assessed to give an indication of the intensity of the above issues. Moran's I test was applied to the residuals from the OLS model to examine the level of spatial autocorrelation in the prediction model, and the results are given in Table 6.6.

The combinations of very low p-values and moderately high z-scores indicate that the clustering patterns on residuals are less likely to result from random processes. Although the issue of spatial autocorrelation from the OLS model is at a very low level, it should not be ignored. A lesson learned from the GWR modelling with individual locations outlined in Chapter 5, was the inability of the OLS model to capture the spatially varying mechanism which is a significant factor causing spatial autocorrelation. A GWR model with a spatial weighting scheme and arrangement of local regressions is a potentially effective method to capture a spatially varying mechanism, and in turn will be able to reduce or even remove the effect of spatial autocorrelation.

Table 6.6 Moran's I test reports on the OLS models

	OLS with six explanatory variables	OLS with four explanatory variables
Moran's Index:	0.085586	0.087188
Expected Index:	-0.002278	-0.002278
Variance:	0.000683	0.000683
Z-score:	3.361349	3.422737
P-value:	0.000776	0.000620

The

examination of multicollinearity in Section 6.2.3 reveals that both OLS models, using six and four explanatory variables, have similar levels of multicollinearity. The level of spatial autocorrelation is also quite similar for both sets of explanatory variables. The OLS model and backward elimination regression indicate that only four explanatory variables are statistically significant, so a prediction model using four explanatory variables was run in the GWR model.

6.3. In-sample estimation of GWR model with value zones

The adaptive kernel is utilised in the GWR model because the zones are not located in a regular pattern. The GWR analysis in Chapter 5 revealed that there was not much difference between the results from the Corrected Akaike Information Criterion (AICc) and Cross Validation (CV) bandwidth methods. Hence only one method, the AICc bandwidth method, is discussed in this chapter. The diagnostic report from the GWR model is provided in Table 6.7.

Table 6.7 Diagnostic report from GWR model using value zones

Neighbours	170
ResidualSquares	146,561,400,000,000
EffectiveNumber	34.82
Sigma	601,431.62
AICc	12,982.54
R-squared	0.76
R-squared Adjusted	0.74

One hundred and seventy out of 440 (38.64per cent) of the total features are involved in each local regression. As discussed in Chapter 2, the optimum number of neighbours yields the lowest sum of residual squares. This number gives the highest prediction accuracy of the model as a whole. However, having more than one third of the total number of features involved in each local regression may reduce the ability of local models to capture local variations.

The results from the GWR model were also tested for spatial autocorrelation (Table 6.8).

Table 6.8Moran's I test report for GWR model using value zones

Item	Value
Moran's Index	0.031151
Expected Index	-0.002278
Variance	0.000684
Z-score	1.278355
P-value	0.201124

The very low positive Moran's I index signifies that there is a very low level of clustering of residuals. The combination of the low p-value and the low z-score indicates that the very low clustering pattern is not very different from that of random processes. GWR removes the issue of spatial autocorrelation which exists in the OLS model. GWR models the non-linear function between dependent variable and each explanatory variable by weighted regression which gives a unique weight to each sample based on the proximity to the local regression point (see Brunsdon *et al.*, 1996). By doing so, GWR tackles the issue of 'model misspecification' in the OLS model which imposes a linear relationship between dependent variable and each explanatory variable.

A map of standardised residual was created (Figure 6.2) which confirms the result from the Moran's I test. There is no significant pattern of clustering or dispersion among standardised residuals.

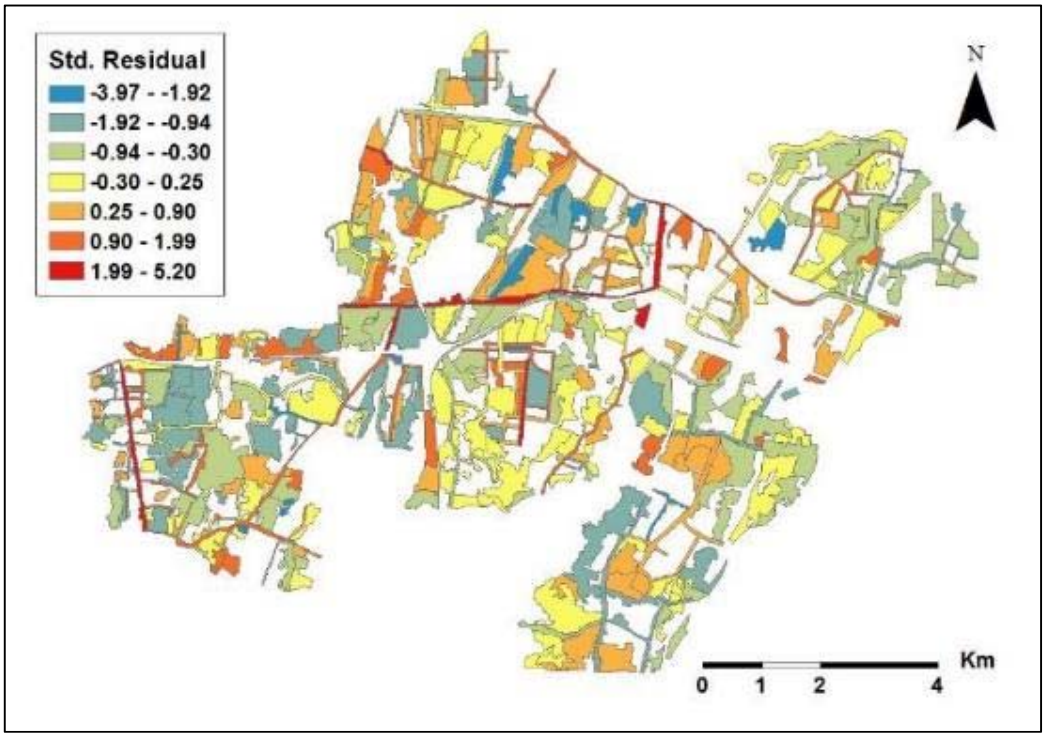


Figure 6.2 Spatial distribution of standardised residuals from GWR model using value zones

The distribution of colours representing intervals of standardised residual over the study area in Figure 6.2 looks random. Under-predicted zones or over-predicted zones do not show a clear pattern of clustering or dispersion. The results from global Moran's I test, and the distribution of standardised residuals indicate that spatial autocorrelation is not an issue in the model. The next discussion will be on the spatial distribution of local R-squared values.

The spatial distribution of localR-squared values was examined (Figure 6.3). The highest local R-squared values are found in mid-south and the north-east of Bekasi, while the lowest are found in the mid-north. Despite the distinct spatial distribution, the range of variation is moderately low. 86.6 per cent of local R-squared values are between 0.6 and 0.8. Overall, this set of explanatory variables is quite effective for prediction.

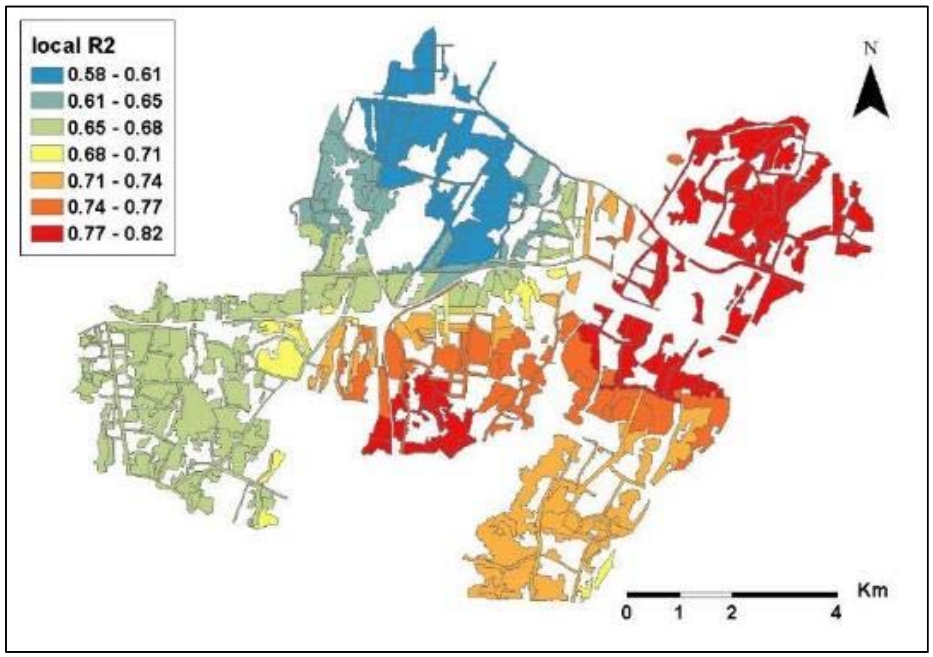


Figure 6.3 Spatial distribution of localR-squared values from GWR model using value zones

Parameter estimates are the main output of a GWR model. In this section maps of each explanatory variable are discussed. The map of local coefficients for variable zoning is presented in Figure 6.4.

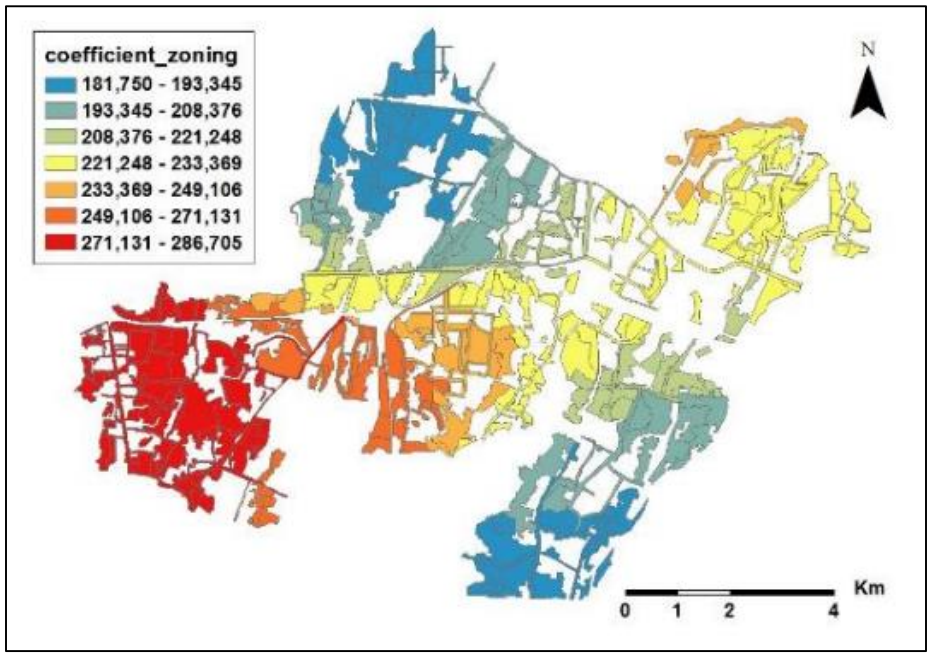


Figure 6.4 Spatial distribution of coefficient estimates for variable zoning in GWR model using value zones

All zones have positive coefficients for variable zoning, meaning that zoning has a positive correlation with land value over the whole study area. The largest positive correlations are found in the south-west, so variable zoning pays its greatest contribution to shape land price in this area. The lowest correlations are found in the mid-north and south-east. The mid-north is predominantly residential complexes, while the south-east corner is mostly industrial precincts. With a dominant zoning type at each area, variable zoning pays its least contribution to shape land price in each of these areas.

A lesson learnt from GWR in Chapter Five was that the GWR package in ArcGIS does not report the p-value to indicate the significance of each explanatory variable at each location. Charlton and Fotheringham (2009) suggested that using the p-value as a measure of significance of the parameter estimate is not appropriate in GWR. Rather the Benjamini-Hochberg False Discovery Rate (FDR) is considered a more appropriate approach; but it has not yet been incorporated within the GWR model. With no measure of the significance of parameter estimates available in the GWR model, assessment of a parameter estimate's significance level cannot be undertaken.

Instead, assessments are made on the precision levels of parameter estimates by computing local t values. The t-statistic value simply compares the actual value of a coefficient to its standard error. Both inputs, coefficient and coefficient's standard error, are provided in the output table of a GWR model. The local t value helps to indicate the level of reliability of a parameter estimate. A small standard error for a large coefficient, results in a large t value, which gives high confidence in the parameter estimation. In all of the zones, the standard errors are relatively small compared with the actual coefficients of zoning. As the result, t values are large (Figure 6.5); and the confidence in using zoning as an explanatory variable is high for all zones.

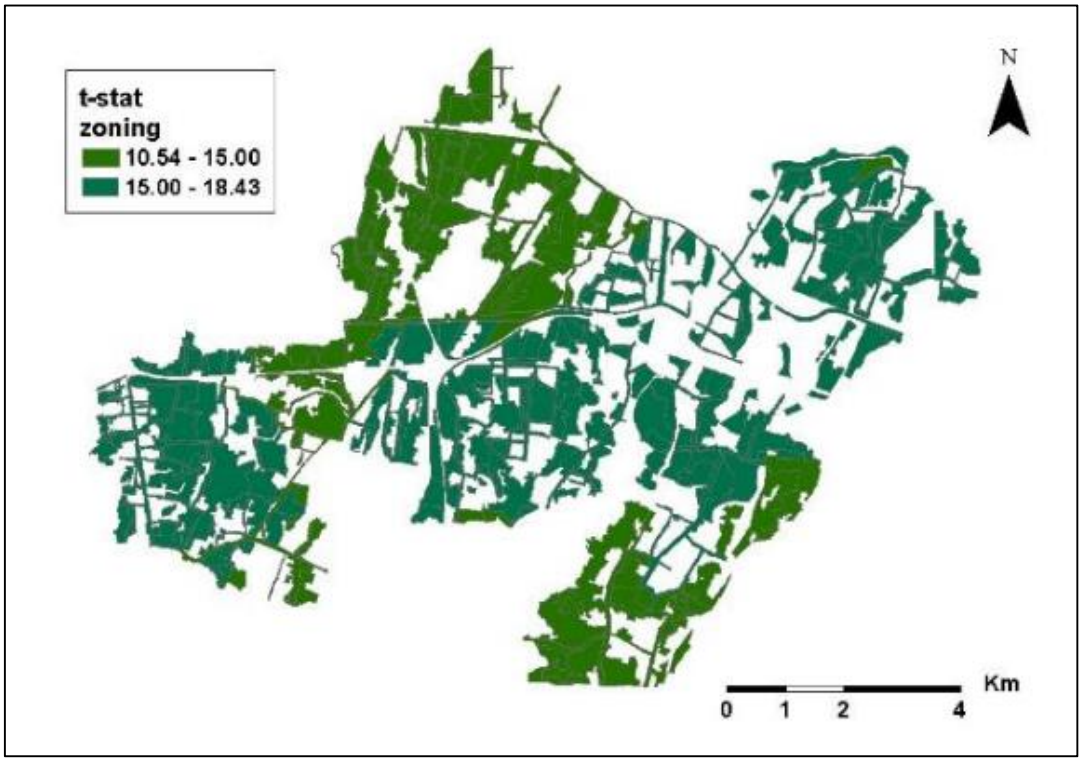


Figure 6.5 Spatial distribution of t values for variable zoning in GWR model using value zones

All zones support the inference that the zoning class is an important explanatory variable for land value, as indicated by large t values for variable zoning across the study area. This inference from the GWR model is in line with the results from the initial data examination using the OLS model and backward elimination stepwise regression.

Travel time to nearest tollgate is expected to have a negative correlation with land value. Variable tollgate had the strongest negative correlation to land value for zones located in the central west (Figure 6.6). In these zones, shorter travel time to nearest tollgate is a significant factor in increasing land value. In part this is because traffic jams in the city centre of Bekasi can be very bad during busy times, so travelling on toll roads which have less traffic is much preferable to using non-toll roads. However, some zones in the north-east and in the mid-south have positive correlations (Figure 6.6). In these zones, shorter travel time to nearest tollgate does not make a positive contribution to increased land value. Zones in north-east Bekasi have a primary arterial road (Juanda Road) as the main access to Jakarta, and this may be preferable to commuters than the toll road.

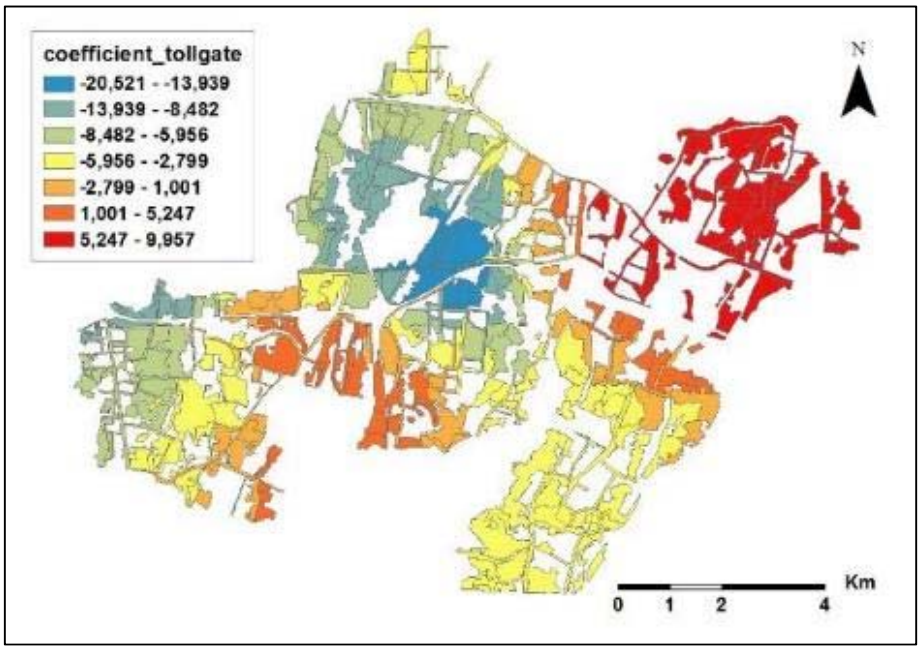


Figure 6.6 Spatial distribution of coefficient estimates for variable tollgate in GWR model using value zones

Zones located in the north-east have the largest positive coefficients and largest positive t-values (Figure 6.7). This adds confidence to the earlier inference that shorter travel time to the nearest tollgate does not lead to increased land values in this area.

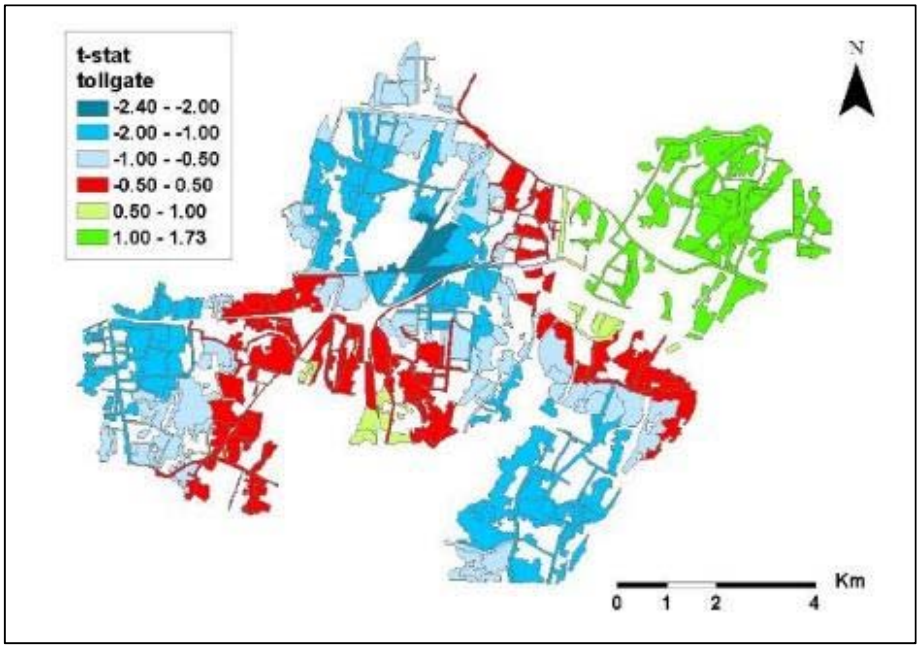


Figure 6.7 Spatial distribution of t value for variable tollgate in GWR model using value zones

Another inference is also supported in the west-centre, where several zones with the largest negative coefficients came up with the largest t-statistic values as well. This also adds confidence to the inference saying that shorter travel time to nearest tollgate can increase land values in this area.

Data examination in Section 6.2.2 showed that travel time to nearest hospital has a positive coefficient (9626.49). This means that a shorter travel time to the nearest hospital does not increase the land value. The GWR model reveals that this inference from the global OLS model does not apply for the whole study area. Coefficients for variable hospital from the GWR model (Figure 6.8) vary significantly across the study area. Around one third of all zones came up with negative coefficients, while around two third of all zones came up with positive coefficients.

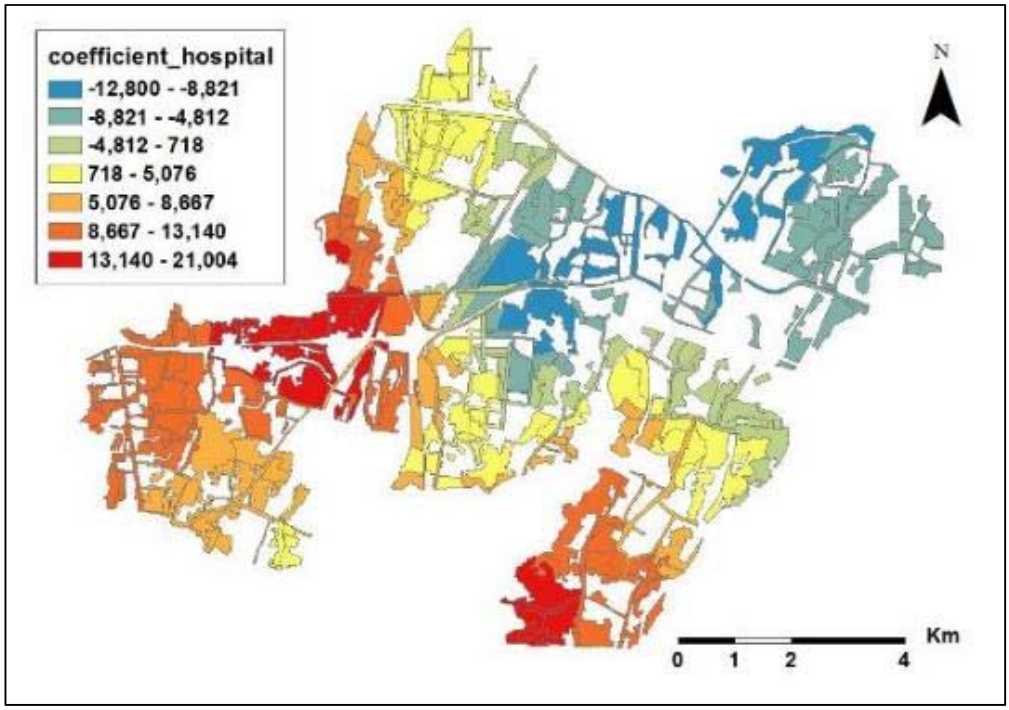


Figure 6.8Spatial distribution of coefficient estimates for variable hospital in GWR model using value zones

The range ofcoefficient's standard errors in relation to variable hospital is nearly half of the range of actual coefficient values. Coefficient's standard errors range from 5,661 to 9,830, while the actual coefficients range from -12,800 to 21,004. The resulting t values range from -1.74 to 2.80 (Figure 6.9), indicating that no zones have high confidence in terms of the

coefficients for variable hospital. Zones with the least reliability on hospital coefficients are in the north, central-south, and the mid-east where the standard errors are more than twice the corresponding actual coefficients (the t statistic values are between -0.5 and 0.5).

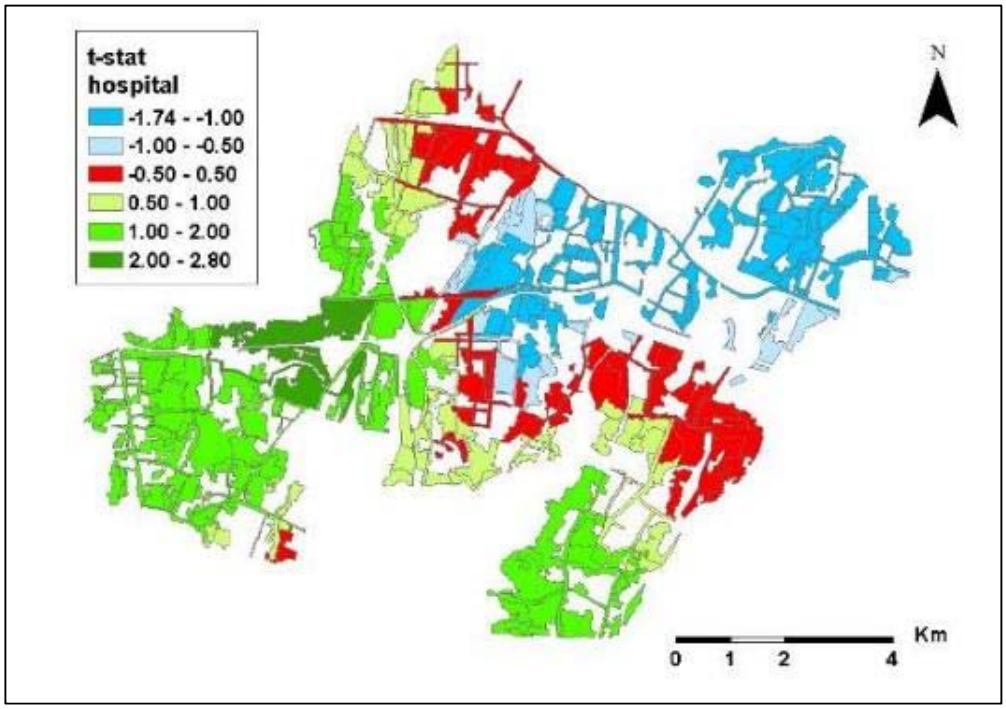


Figure 6.9Spatial distribution of t values for variable hospital in GWR model using value zones

Travel time to the CBD of Bekasi has the strongest positive correlations with land value for zones in the south-east of the city (Figure 3.10). In this area, the shorter travel time to CBD does not increase land value as the area is mostly industrial precincts. An immediate assumption is that there are relatively few personal movements to and from the CBD.

The area in the mid-west has the largest negative parameter estimates (between -792,497 and -532,292), which means shorter travel time to CBD pays its highest contribution to increase land value in this area. This area used to be the business centre until the 1990s before the current CBD was established. The commercial precincts in this area still exist but they are not growing as fast as those in the current CBD. Connection between the commercial precincts in this area to those in the CBD may contribute to the high importance of travel time to the CBD.

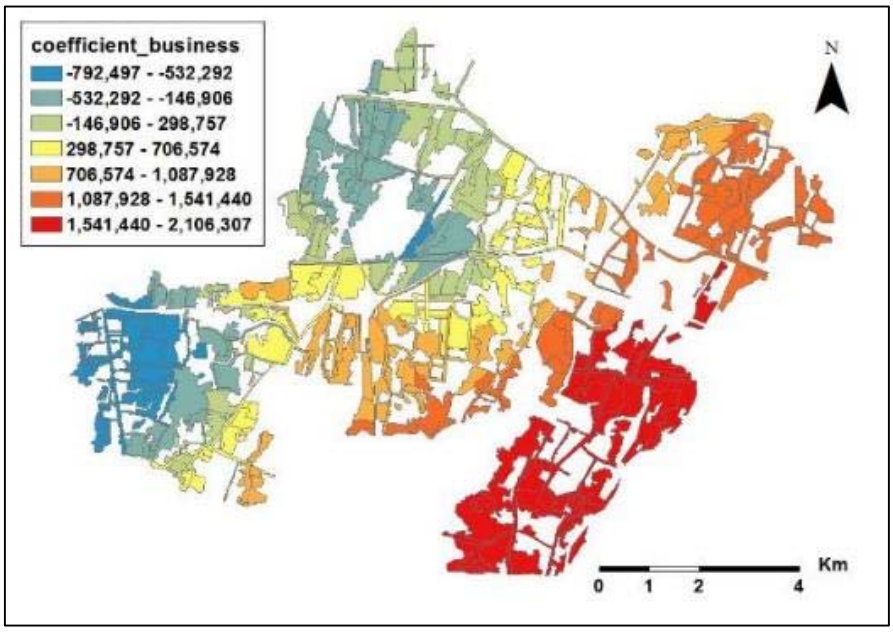


Figure 6.10Spatial distribution of coefficient estimates for variable CBD in GWR model using value zones

An interesting result is found in the case of travel time to CBD. Positive t-statistic values are mostly found in the east (Figure 6.11), and the t value tends to increase by the distance from CBD. The coefficient map (Figure 6.10) suggests that shorter travel time to CBD does not increase land value for zones in the east.

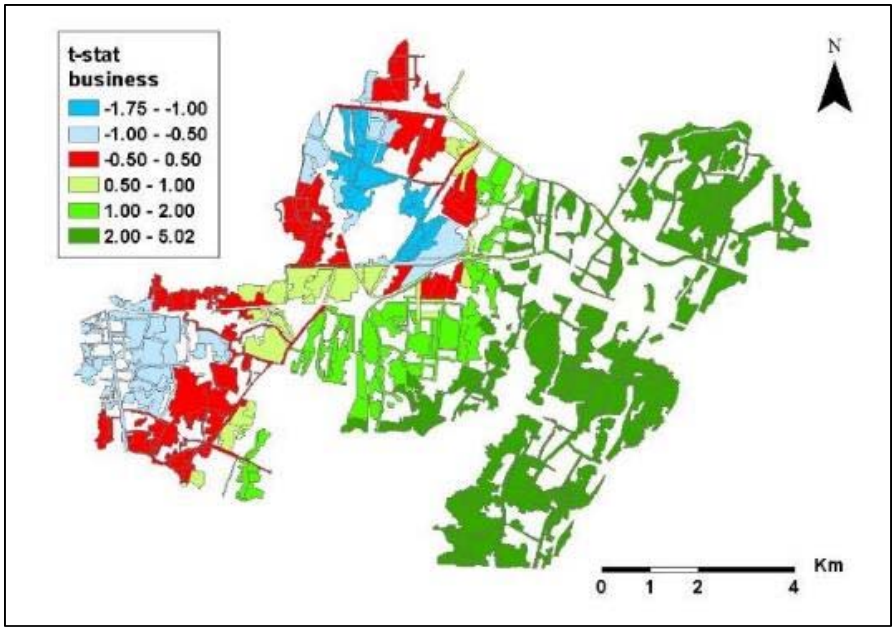


Figure 6.11Spatial distribution of t values for variable CBD in GWR model using value zones

The t-statistic map indicates that the farther from CBD, the more confidence is found on this inference. In the west, zones with moderate to large negative t values are surrounded by zones with small t values. The coefficient map suggests that the shorter travel time to the CBD increases land value in western zones. However, the confidence in parameter estimates in the west is generally lower than in the east.

The spatial distribution of t value for each of the explanatory variables indicates that the level of reliability of coefficient varies across space. The GWR allows coefficients to vary by location, which enables assessment of a coefficient's reliability to be made by location. This can help to decide which sampled locations should be set aside from the model in order to improve the prediction accuracy. Prediction accuracy was checked at locations with very low t values to find out whether or not the low precision of parameter estimation is related to low accuracy of prediction. To do this, the prediction residual at each location was calculated.

Residuals were measured by comparing the predicted value of a zone to the average value of observed samples in the zone. The distribution of residuals is plotted in Figure 6.12. An immediate concern is that the average observed values vary significantly among value zones, so residuals are measured using various references. Instead of comparing residuals among zones, it wasmore objective to compare each residual to its own observed value. This comparison results in a percentage residual. The distribution of percentage residuals for allvalue zones is plotted in Figure 6.13.

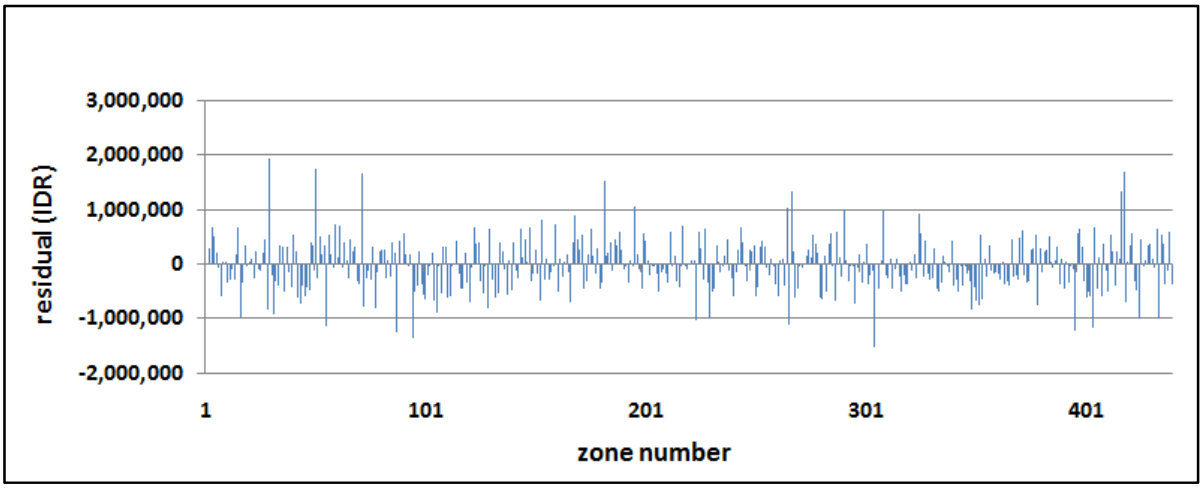


Figure 6.12 Plot of prediction residual in GWR model using value zones

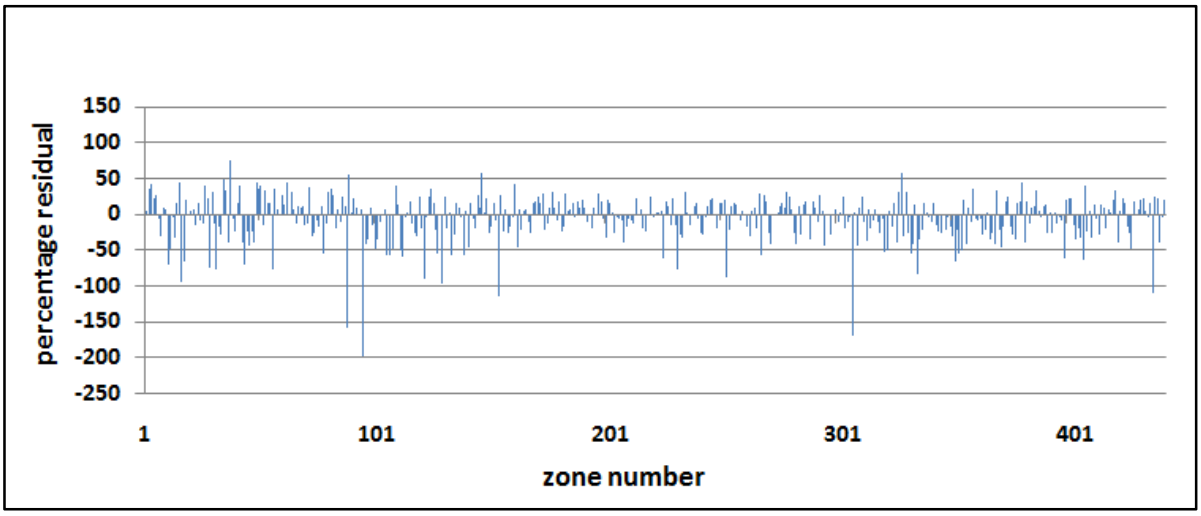


Figure 6.13 Plot of percentage residual in GWR model using value zones

The mean absolute percentage error (MAPE) is 23.33 per cent, and the standard deviation of percentage error is 31.82 per cent. As a whole, the model meets the standards used within the Land Valuation Directorate in BPN. Almost three-quarters (74.09 per cent) of all the zones have percentage residuals lower than 30 per cent. Hence, nearly a quarter of the zones are considered inaccurate and require further investigation.

The graph of percentage residuals (Figure 6.14) looks very different from that of residuals (Figure 6.12). The percentage residual graph gives a more objective assessment of the prediction accuracy because for each zone, the residual was compared to its observed value. From the graph of percentage residuals, it can be seen that the percentage residuals at five zones are over 100 per cent.

The distribution of standardised residuals (Figure 6.14) is displayed in order to examine the interaction between standardised residual and percentage residual at each prediction location.

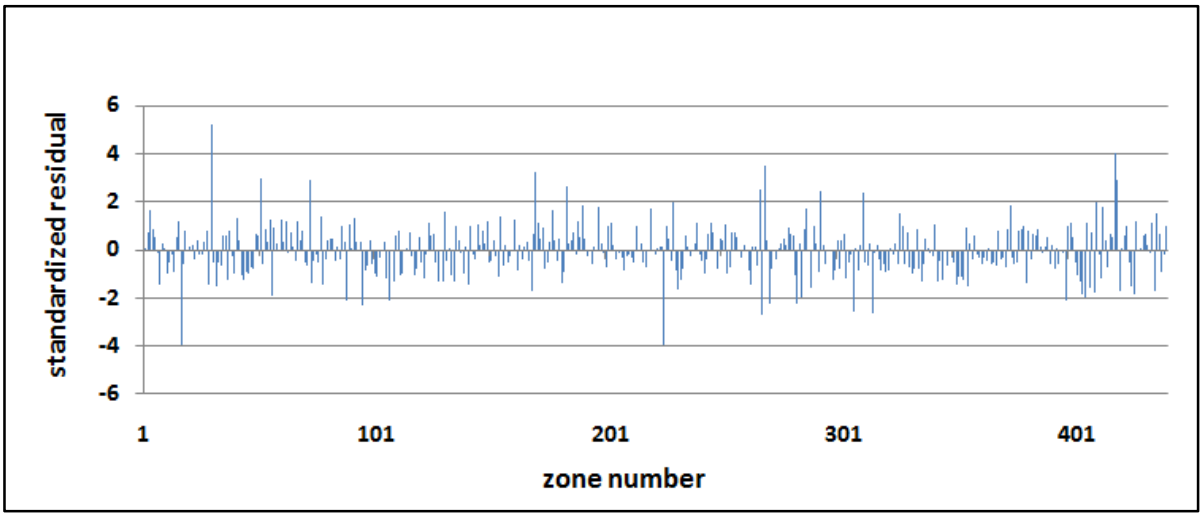


Figure 6.14 Plot of standardised residual in GWR model using value zones

The distribution of standardised residual is quite different from the distribution of percentage residual. The five zones with the largest percentage residuals have relatively small standardised residuals. Therefore, the standardised residuals cannot be used to explain the extremely large percentage residuals in some zones.

The reliability of each local model can be measured by its R-squaredvalue and standard error. However, both measures cannot be used as indicators on predictive performance of the local model: R-squared provides a measure of how well the local model fits the data involved in a local regression, while the standard error provides a measure of precision of the data involved in a local regression when a local model is applied.

Therefore, none of the measures contained in the GWR model output table can be used to explain extremely low prediction accuracy in several zones. The set of parameter estimates at each prediction zone was examined next. The level of reliability of a coefficient in each zone is indicated by the t value. Examination ofzones with extremely low prediction accuracy was undertaken by examining the t values, to find out whether low prediction accuracy is related to the low precision of parameter estimation inthe local model.

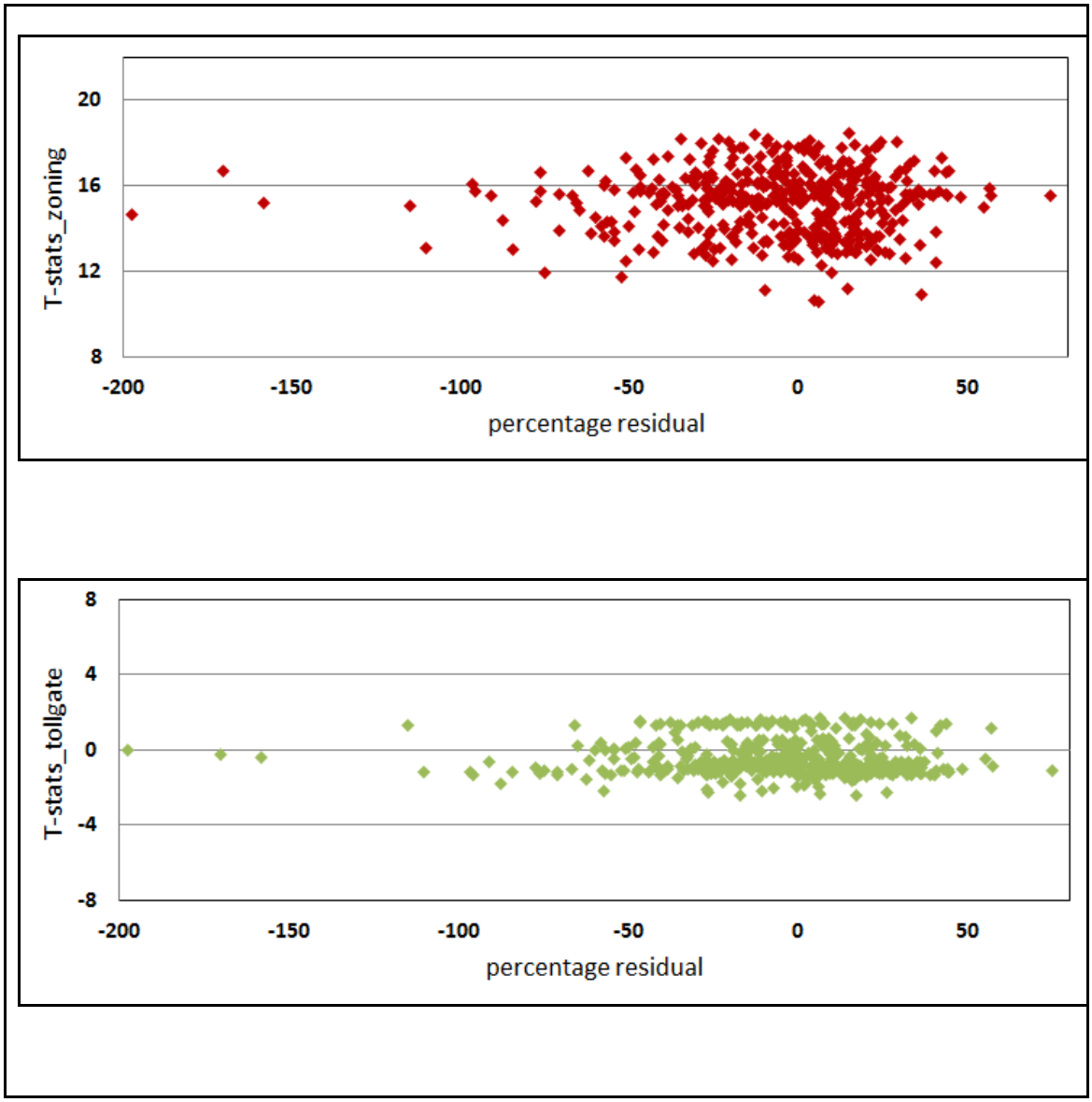
Table 6.9 T-statistic values at locations with the largest percentage residuals in GWR model using value zones

Value zone	Percentage residual	Variable	T-statistic
95	-197.56	zoning	14.68
		tollgate	0.01
		hospital	0.97
		business	0.87
305	-170.20	zoning	16.68
		tollgate	-0.24
		hospital	1.99
		business	1.22
88	-158.24	zoning	15.21
		tollgate	-0.38
		hospital	1.01
		business	0.45
153	-114.87	zoning	15.08
		tollgate	1.30
		hospital	-1.08
		business	4.44
434	-110.30	zoning	13.08
		tollgate	-1.20
		hospital	1.17
		business	3.62

Table 6.9 shows that very small t values exist for variable tollgate atzones 88, 95 and 305 (-0.38, 0.01, and -0.24 respectively), while moderately small values are at zone 153 and 434 (1.30 and -1.20). Variable zoning always has large t values, while variable hospital and business centre have moderately small to moderately large t values in the five zones with the largest residuals. There is no convincing pattern in the relationships between the percentage residual and t values in these zones with the largest percentage residuals: it can be argued that in the case of GWR modelling with zones, low prediction accuracy is not always related with low precision on parameter estimation. An immediate question is

whether or not this inference also applies for all zones in the Bekasi dataset. A scatter plot of percentage residual and t value for each explanatory variableis provided to answer this question (Figure 6.15).

There is no clear pattern that explains the relationships between the prediction residual and t values for each variable (Figure 6.15). This result provides the foundation to infer that prediction accuracy is not related tothe precision of the parameter estimates.



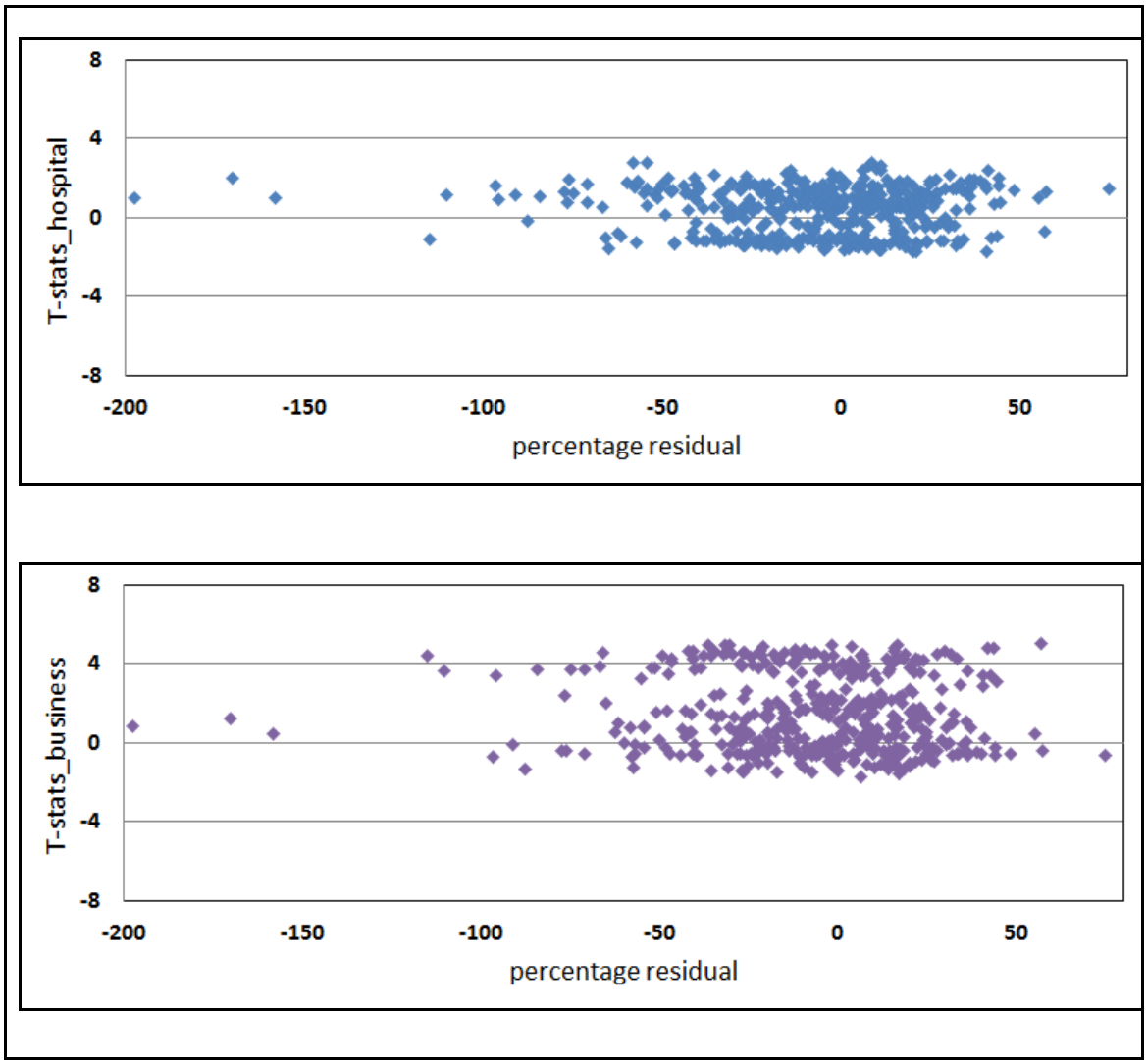
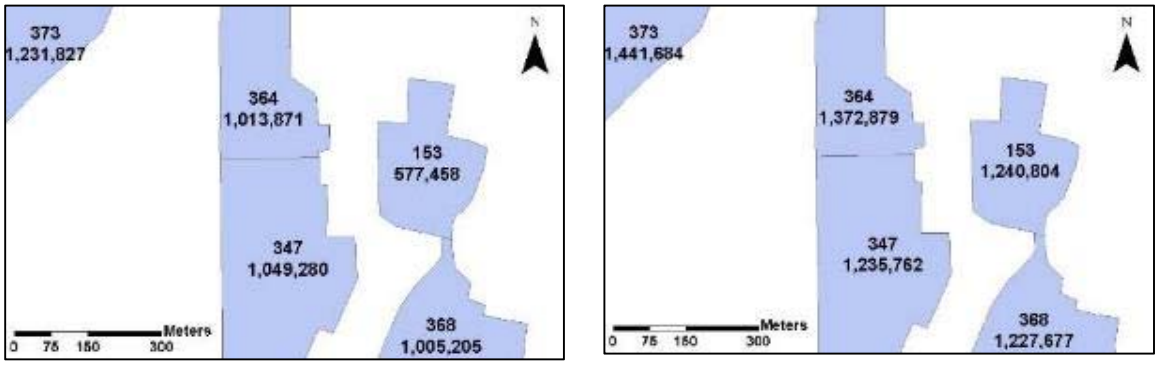


Figure 6.15Scatterplots of T statistic value and percentage residual in GWR model using value zones

A more convincing conclusion can be drawn if the associations between percentage residuals and t valuesfor each explanatory variable is examined at all local regression. The correlation coefficient between percentage residuals and t value of variables zoning, tollgate, hospital, and CBD are 0.018, -0.029, -0.060and -0.042 respectively. These very low correlation coefficients confirm that prediction accuracy is not related to the precision of parameter estimation of any of the explanatory variables.

Recalling the zones with large percentage residuals (Figure 6.13), there is no measure from the GWR output table that can be used to explain their very low prediction accuracy. The data at the five locations with largest percentage residuals were checked (Table 6.9), and zone 153 was chosen to give the simplest explanation of the extremely large percentage

residual. Zone 153 is a residential complex. The average observed land values and predicted land values from the neighbouring nearest residential complexes are shown in Figure 6.16.



a. Observed values **b. Predicted values**

Figure 6.16Observed and predicted land values for zone 153 and zones with the same zoning type in its vicinity

Zones with the same zoning type are expected to have similar average land values, but Figure 6.16 shows that zone 153 has a considerably lower average observed land value compared with nearby zones (577,458). This raises questions about the quality of observation data in zone 153. Zone 153 only has one sample within it, sample number 473. There is a chance that the observed sales price of sample 473 is underrated or overrated but unfortunately there is no other sample within the zone to be compared with sample 473.

Although zone 153 is observed to have considerably lower observed land value compared to other zones that share similar characteristics, the predicted value of zone 153 is quite similar to the zones of similar characteristics. The GWR model applies the predominant relationship between explanatory variables and land value for zone 153. The result is thatzone 153 has a considerably large prediction residual (-114.87 per cent).

The GWR using value zones has the same issue with GWR using individual locations. Several locations had prediction residuals over 100 per cent (Section 5.3). Using the cut-off value of 30 per cent, as previously used to assess the result from the previous GWR model, 114 of 440 zones (25.91 per cent) are considered invalid. Despite the serious accuracy issues in around a quarter of all zones, the MAPE of 23.33 per cent is still below the cut-off value.

In the in-sample GWR model, the model was built using only the sampled zones. During the process of minimising the error, the model was optimally fitted to these particular zones. Being only optimally fitted to a set of sampled locations, the prediction model has the potential to overrate its prediction performance (see Hastie *et al.*, 2009). For a convincing assessment on prediction performance, out-of-sample estimation of GWR model with value zones was also run.

6.4. Out-of-sample estimation of GWR model with value zones

A GWR model with out-of-sample estimation was run to give a more objective assessment of prediction ability, and as before the Monte-Carlo Cross Validation (MCCV) was chosen because it was expected to give objective validation as it completely separates the data for training and validation subsets in each iteration. The concept was introduced by Burman (1989), and was originally called Repeated Learning-Testing Method because the dataset was split into training and validation subsets multiple times. Each time, samples for the training subset and validation subsets were randomly selected. The same number of samples in the training subset was maintained in each iteration, the number of possible iterations is calculated as follows.

$$\textbf{Number of iterations} = \frac{\boldsymbol{n}!}{\boldsymbol{t}!\boldsymbol{x}\ \boldsymbol{v}!} \qquad \textbf{(Equation 6.1)}$$

Where: $n!$ is the number of total samples-factorial
$t!$is the number of training samples-factorial
$v!$is the number of validation samples-factorial

There are 440 zones containing at least one sample: 352 zones (80 per cent) were used as a training dataset and 88 zones (20 per cent) as a validation dataset. The number of iterations is determined by the number of combinations that can be made from the number of samples in the training subset and the number of samples in the validation subset. The number of iteration is calculated as follows.

$$\frac{440!}{352!\,x\,88!}$$

The number of possible iterations is extremely large, so it cannot be calculated using normal calculation software. As in the case of GWR for individual locations, 1,000 iterations were run. In each iteration, out-of-sample predictions were made on 88 validation zones. The predicted land values were compared with the observed land values at the corresponding zones.

The numberof predictions at each zone validated using the corresponding observation at the zone, ranges from 163 to 237 (Figure 6.17); the average frequency is 200.

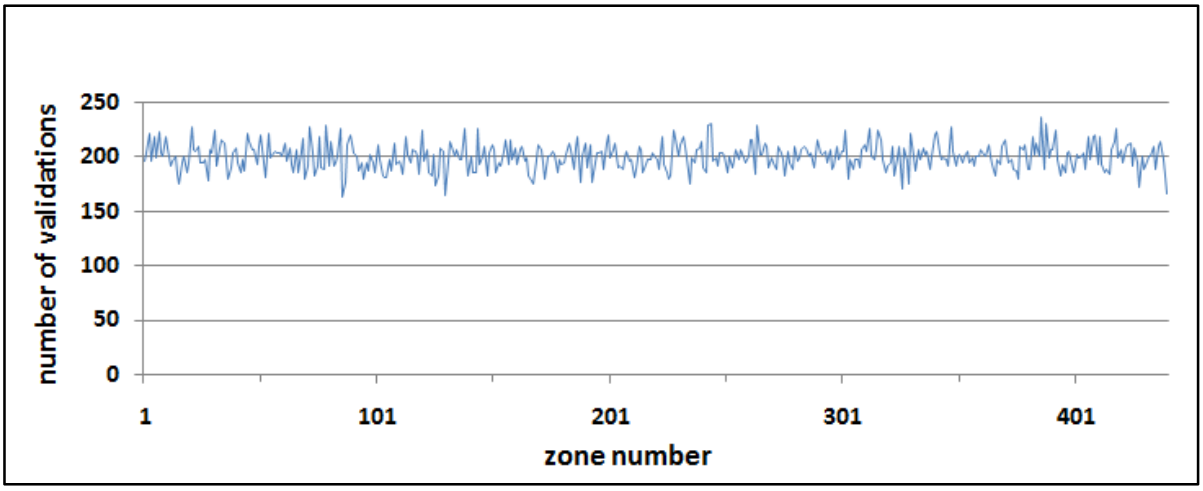


Figure 6.17The number of predictions that were validated for each zone in MCCV on GWR prediction using value zones

The multiple predictions made at each zone were compared with the observed value for the zone. An average percentage residual was calculated for each zone to assess the average prediction accuracy. As discussed in Chapter 5, a positive residual and a negative residual in a zone may counteract each other. In addition to the distribution of average percentage residuals, the distribution of the range of percentage residuals was calculated. Both are shown in Figure 6.18.

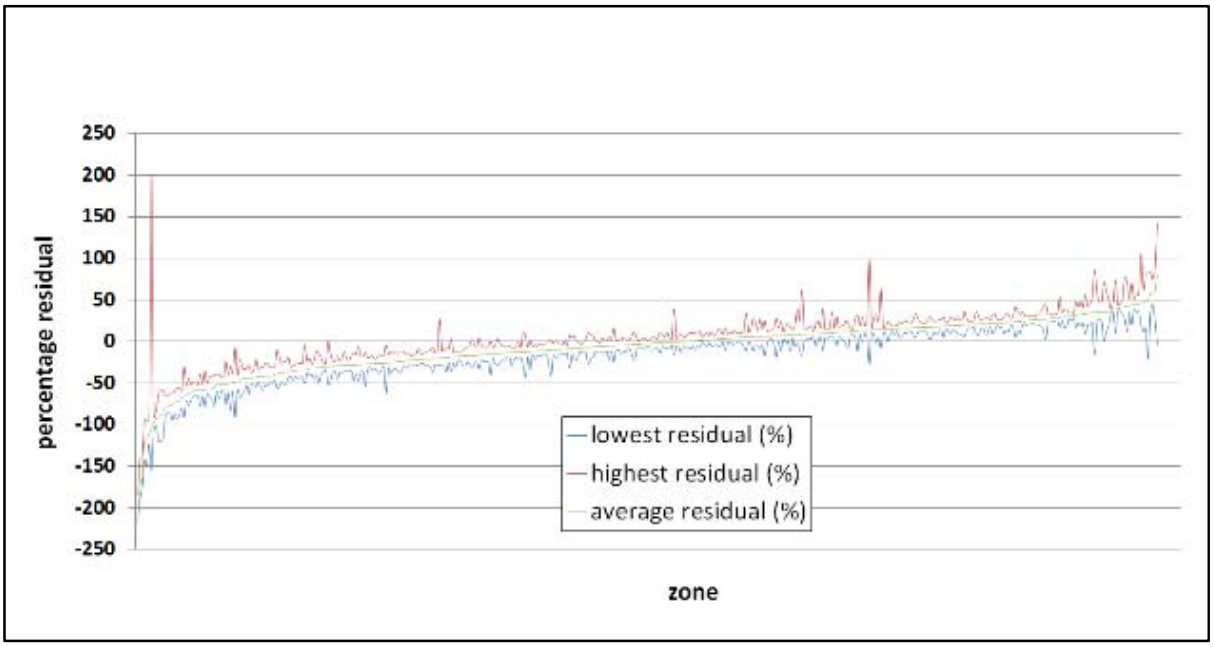


Figure 6.18Range and average of percentage residual in MCCV on GWR prediction using value zones, plotted in increasing value of the average residual

There are a significant number of zones with extremely large average percentage residuals and large ranges of percentage residuals. Prediction from GWR applied to zones has the same issue as with individual samples. For a more thorough assessment, the distribution of standard deviations was also examined (Figure 6.19) as it is a more appropriate measure of prediction precision than the range of residuals.

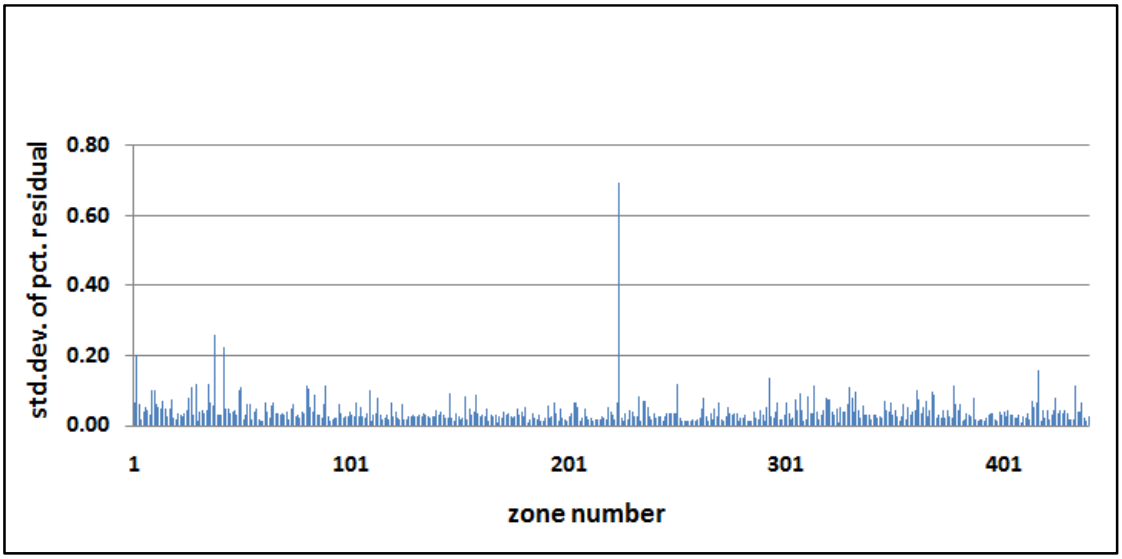


Figure 6.19Plot of standard deviation of percentage residuals in MCCV on GWR prediction using value zones

In the out-of-sample estimation of GWR with value zones, nearly all zones have standard deviations of the percentage residuals less than 30 per cent. Figure 6.19 shows one zone (zone 224) with an extremely large standard deviation of the percentage residual. All of the percentage residuals at zone 224 are plotted in Figure 6.20.

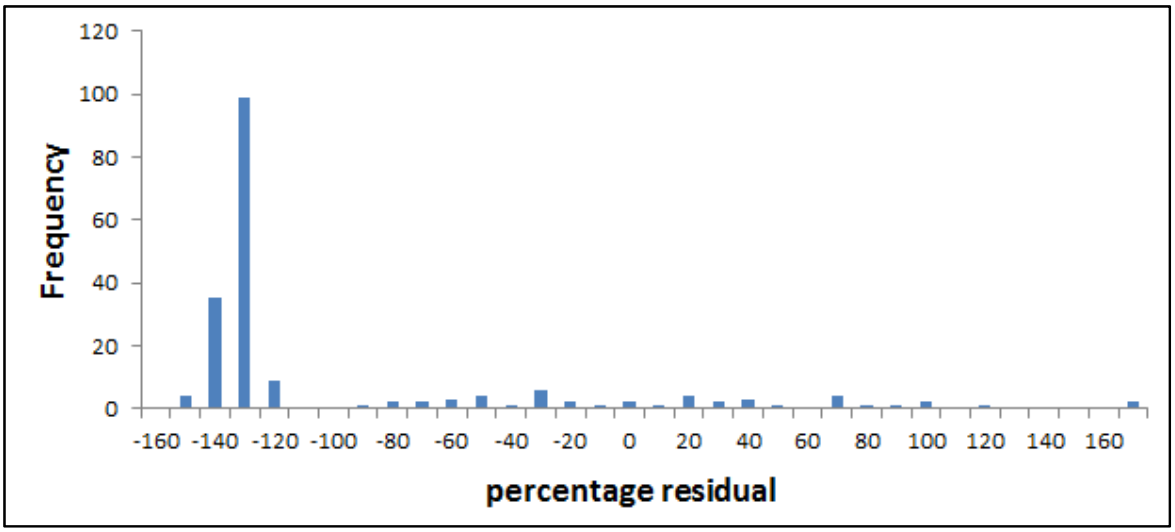


Figure 6.20 Distribution of percentage residuals at an example of zone (zone 224)

The percentage residuals of all predictions for zone 224 are not normally distributed. Over three quartersof the predications, 147 of 193, have percentage residuals between -127 and -156 per cent of the observed value. Forty six (23.83 per cent) predictions have percentage residuals ranging from -98 to 199 per cent of the observed value, making a long tail to the distribution. This situation raises questions about the model's efficacy and data quality.

6.5. Summary

6.5.1. Prediction accuracy

In the in-sample estimation of the GWR model, the MAPE and the standard deviation of percentage residuals were 24.86 per cent and 33.70 per cent respectively. These can be compared with the MAPE and standard deviation of percentage residuals from standard BPN Zonation Method, which are 10.8 per cent and 15.9 per cent respectively. The GWR applied to value zones is less reliable than the BPN Zonation Method in terms of accuracy. However, the prediction model, as a whole, performs moderately well in terms of accuracy;

and importantly the prediction model has resolved the issue which the existing Zonation Method cannot cope with. The GWR model using value zones is able to provide verifiable predictions at every value zone in the study area while the Zonation Method can only provide verifiable predictions in zones where an adequate number of samples exists.

At the individual zone level, 326 out of 440 (74.09 per cent) zones are considered accurate as the prediction residuals are lower than 30 per cent of the corresponding observed values. The result from the out-of-sample estimation is similar; 317 out of 440 (72.05 per cent) zones have average percentage residuals lower than 30 per cent of the corresponding observed values. The distributions of residuals from both approaches are compared in Figure 6.21.

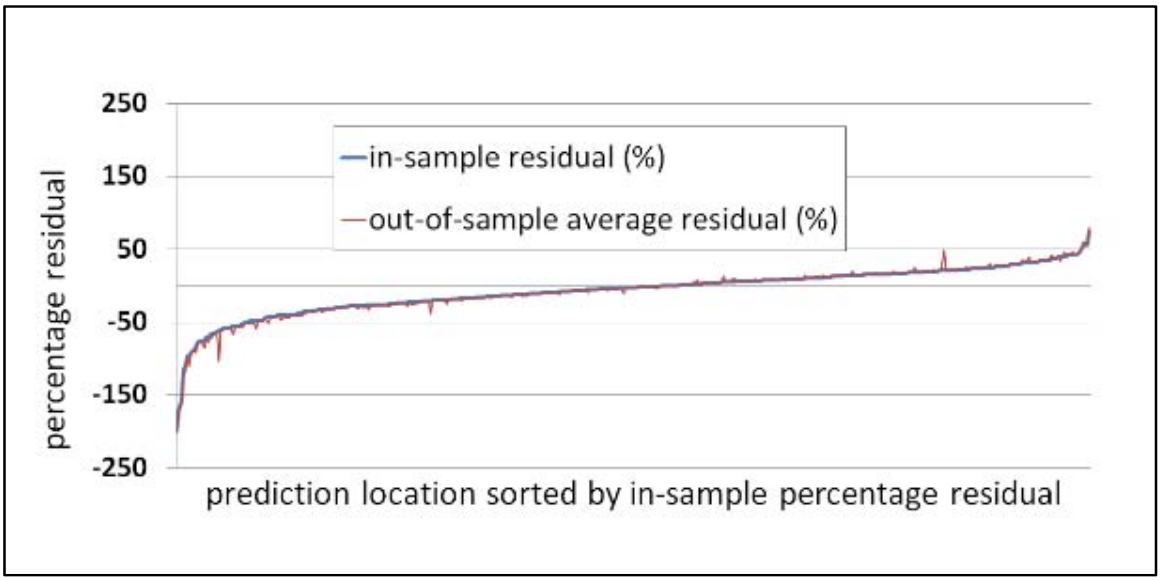


Figure 6.21Average residuals from out-of-sample estimation and residuals from in-sample estimation using value zone data

The distribution of average percentage residuals in the out-of-sample estimation looks similar with the distribution of percentage residuals in the in-sample estimation. This indicates that the prediction accuracy of the model when used to predict non-sampled locations, on average, is quite similar to the prediction accuracy of the model when used to predict sampled locations. The correlation coefficient between the percentage residuals from the in-sample and out-of-sample estimations of 0.996 validates this inference. The very high correlation coefficient also indicates that the GWR prediction model using value zones does not have a significant overfitting issue. With no significant overfitting, the GWR prediction

gives good estimates of the actual prediction accuracy per zone. The prediction accuracy at the non-sampled locations should be similar to the prediction accuracy at the sampled zones.

6.5.2. Dealing with predictions with extremely large percentage residuals

The GWR model using value zones was designed to solve the main issue of the GWR model using individual samples, i.e. large percentage residuals at a number of locations. Anomalous observations are detected by assessing the variation of observed prices among samples in each value zone. Next, a weighting scheme was also applied in each zone so that zones with more samples can contribute more to shape local regressions: the model was then expected to come up with no predictions with extremely large percentage residuals. Unfortunately, the result reveals that there are a number of zones with extremely large percentage residuals. This approach therefore fails the task.

Two limitations are most likely to be the main reasons behind this failure. First, there are only around ten per cent of the zones used in the model have three or more samples. Although each zone is given weight based on the number of samples, this small number of 'valid' zones raises question about the reliance of the data. Second, the use of zones centroids could also contribute for inaccuracy as the geographical shapes of zones vary considerably from a very compact geometry to elongated linear. Centroids from elongated zones could significantly affect the distance variables used in the model.

The biggest challenge of applying the GWR model using the Bekasi dataset remains the same – how to deal with the extremely largepercentage residuals at a number of predictions. Controlling the input for the model (by running the model using value zones) did not solve the problem. An alternative approach to undertake is to control the output of the prediction model. Measures to control the output of GWR prediction model will be discussed in the next chapter (Chapter Seven) where all of the predictions, especially the ones with extremely large percentage residuals, will be carefully examined.

7. VERIFYING PREDICTIONS

7.1. Introduction

This chapter will discuss the measures adopted to verify the predictions of the GWR model. Checking the predictions of the model becomes a very important task because the results from GWR models outlined in Chapter Five and Chapter Six reveal that the GWR models have the same major issue, i.e. extremely large residuals at a number of locations or zones. The potentially inaccurate predictions at several locations requires further examination. Before the examination, these anomalous predictions must be detected. Detection of potentially inaccurate predictions at the sampled locations can be done using the prediction residuals. This is possible because each prediction has the corresponding observation as a reference point to calculate the prediction residual.

At the non-sampled locations, another measure must be arranged to detect potentially inaccurate predictions because no observations are available as references to calculate the prediction residuals. Value zone is again employed for this task. The idea is that neighbouring land parcels in a value zone tend to have similar prices. Following this logic, predicted prices of neighbouring land parcels in a value zone should also be similar to one another. Discussion on employing value zone to control predictions at the non-sampled locations is the main functionof this chapter.

7.2. Why value zone is useful to verify predictions

For non-sampled locations, detection of potentially inaccurate predictions is done by comparing each prediction to its neighbouring predictions. A rational measure in selecting comparable neighbours that is applicable for each location across the study area is required.Making use of value zones may be feasible in this regard, as the most comparable locations for a target location are those located in the same value zone. This is based on the premise that neighbouring land parcels in a value zone share common characteristics. An excerpt from the Bekasi dataset highlights this (Figure 7.1).

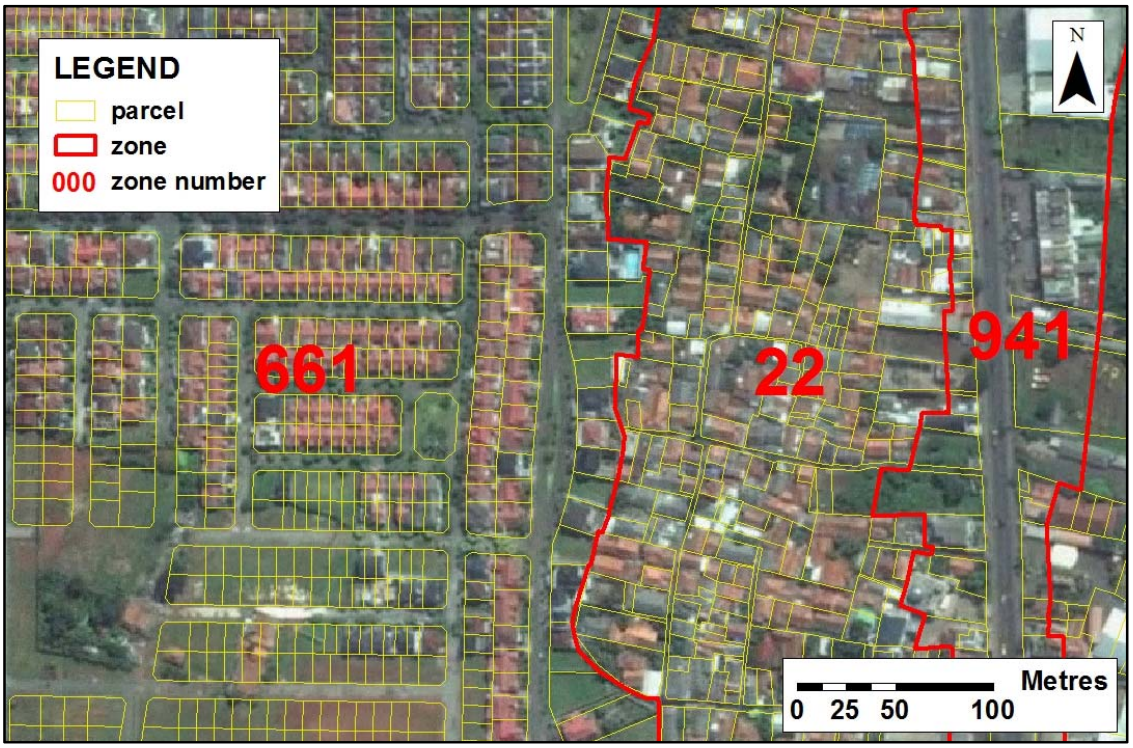


Figure 7.1 A cross section through Value Zones 22, 641, and 941 in South-central Bekasi

Source: Land Office of Bekasi (2012b)

It is clear from Figure 7.1 that the spatialcharacteristics of zone 661 (a residential complex) are quite different from zones 22 (a residential area with irregular land parcels) and 941 (a buffer zone of primary arterial road). The zoning type actually acts as a label that represents the characteristics of a zone. A residential complex like zone 661 is usually a very exclusive residential area with a 'one-gate system' to limit access. In such complexes, land parcels and other features are well arranged and road segments are wide enough for two-way traffic.

Although the zoning type represents the characteristics of the zone, it does not represent the dynamics within the zone. In residential complexes, for instance, high-priced land parcels will have some particular qualities such as being located near a park. Private sellers and brokers take this factor into account when determining the sale prices. So while proximity to the park contributes to the price, it is not listed currently on the BPN data collection form and therefore is not used in prediction. Proximity to the park is an example of an 'unmeasured variable'. A crucial point is that this particular 'unmeasured variable' only exists in a residential complex type of zoning. In Figure 7.1, only land parcels in zone 661 are affected by the proximity to the park. The distance to the park (in the bottom left of the image) is not important in shaping prices of land parcels in zone 22 because the residents in zone 22 do not have access to the park located in zone 661.

An irregular residential area, e.g., zone 22, is defined as having land parcels that are not regularly arranged and road segments that are not regular either. The road widths in zone 22 are much smaller than in zone 661 (Figure 7.1). Many segments are so narrow that only small motorbikes can pass through them, and in busy times traveling through these narrow lanes is challenging. For parcels located on these narrow lanes in areas like zone 22, a land parcel with a short distance to the nearest wider street is preferable and has a higher value. The BPN survey form acquires information on the distances to the nearest collector road and the nearest arterial road. However, these variables were removed from the prediction model because the OLS model and Backward Elimination Stepwise Regression determined these variables were not statistically significant within the whole dataset. Therefore, distance to collector and arterial roads are 'unmeasured variables' in zones like 22. In residential complexes, these variables do not come into play in determining prices.

A buffer zone of primary arterial road, e.g., zone 941, consists of land parcels (e.g., shops and commercial precincts) located on the primary arterial road. For these parcels, the length of frontage to the road is crucial, the longer the frontage the higher the value. However, the ratio between the frontage and the length of a parcel is not taken into account in the prediction model and is an 'unmeasured variable' in buffer zones along streets with commercial premises.

By running local regressions and applying a weighting scheme, GWR tries to capture local variations, and in doing so, GWR also tries to incorporate the 'unmeasured variables' (see Fotheringham *et al.*, 1997). But there is no measure of how well the effects of 'unmeasured variables' are incorporated into GWR models, so it is difficult to infer whether or not the effects are optimally represented in the prediction. Although the effects of 'unmeasured variables' are not well measured, they can be expected to be similar for neighbouring locations in a value zone. This is because neighbouring locations in a value zone are affected by the same set of 'unmeasured variables', and the closeness of locations in a zone increases the probability of similar effects. Hence, for example, a predicted value in zone 661 can be compared with other predictions in zone 661, but cannot be compared with predictions in zones 22 and 941 because of the different effects from 'unmeasured variables' influencing those zones.

It can be concluded that land parcels in a value zone not only have similar data related to the listed explanatory variables (road width, zoning, and travel time to nearest tollgate) but also have similar effects of 'unmeasured variables'. Following this logic, predictions in one value zone are expected to be similar to another. This is a solid justification for using neighbouring predictions in one value zone as the most suitable comparable locations for each prediction.

The next task is to identify the most effective measure to detect anomalous predictions among all predictions in one value zone.

7.3. Detecting anomalous predictions in a value zone

Predictions in one value zone are expected to be similar to one another, so predictions are expected to form a homogenous pattern in each value zone. A prediction that stands out among all other predictions in a value zone is a potential candidate to be an anomalous prediction. Analysing the spatial patterns of predicted prices in a zone can be an effective measure to detect anomalous predictions.

7.3.1. Spatial patterns among predictions in one value zone

Zones 22, 661 and 941, which were compared in Section 7.2 are again used to explain this. As shownin Figure 7.1 from west to east the three zones are a residential complex, an irregular residential area, and the buffer zone of a primary arterial road. Predicted price distribution (Figure 7.2) confirms the idea that the predicted price of a land parcel tends to be much similar to neighbouring predictions in one zone than to neighbouring predictions from other zones.

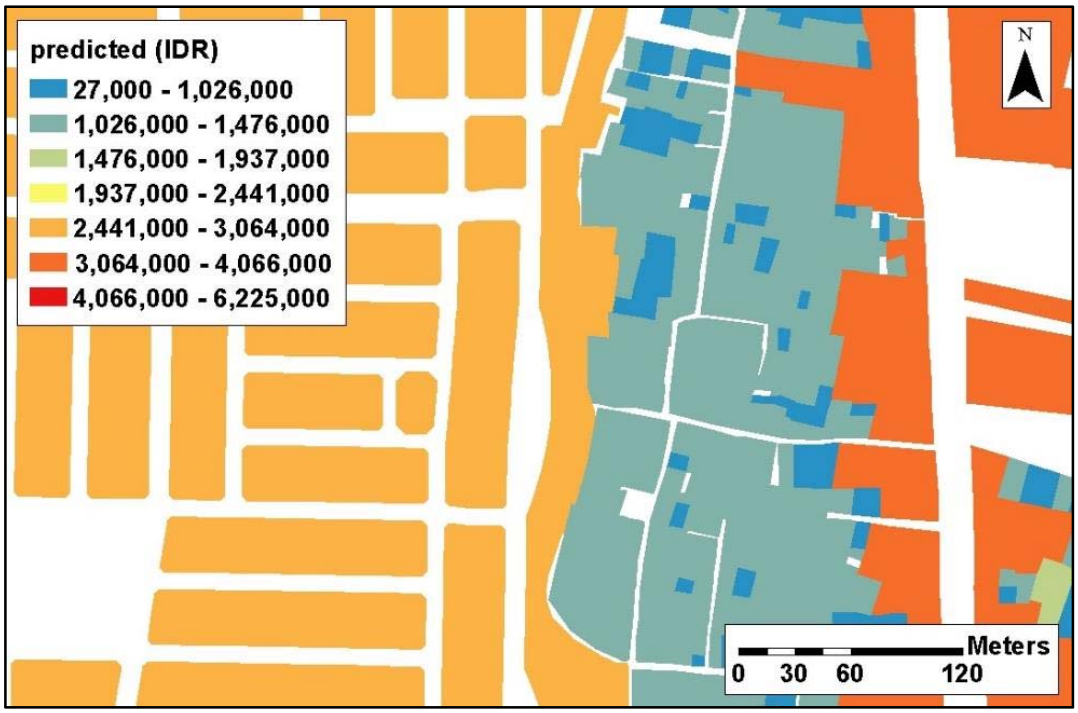


Figure 7.2 Predicted land values in an extract of zones 22, 641, and 941

Modification of the price intervals can be useful in analysing the patterns of predicted prices. However, visual analysis of the price patterns may lead to inconsistent assessment. The Local Moran's I test can be employed to detect outliers among predictions in each zone. This test uses the procedures of local indicators of spatial association (LISA) developed by Anselin (1995). The Local Moran's Index was calculated at each parcel based prediction, using Equation 7.1.

$$I_i = \frac{x_i - \bar{x}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j} \left(x_j - \bar{x}\right) \quad \textbf{(Equation 7.1)}$$

Where: x_i is prediction at parcel i

$\bar{x}$ is the average value of neighbouring predictions

$w_{i,j}$ is the spatial weight between parcel i and parcel j

S_i^2 is the variance of neighbouring predictions of parcel i, which is calculated as follow.

$$S_i^2 = \frac{\sum_{j=1,\ j \neq i}^{n} \left(x_j - \bar{x}\right)^2}{n-1} \quad \textbf{(Equation 7.2)}$$

A significant positive index means that a number of parcels form a clustering pattern because they have similar values which are significantly higher or lower compared with other parcels surrounding them. A significant negative index means that a parcel of a low prediction value has neighbours of high prediction values or a high prediction value has neighbours of low prediction values.

Detection of anomalous predictions using Local Moran's I test was run on predictions for each value zone. In a value zone, adjacent land parcels are not expected to have extremely different values. The idea is that the value of one land parcel should be more similar to its nearest neighbours than to parcels farther away in the same zone. If a prediction is significantly higher or lower than the surrounding predictions, the predicted value will be considered an anomalous prediction.

Zone 209 (Figure 7.3) is an example of a zone in which the Local Moran's I test is an effective tool to detect anomalous predictions.

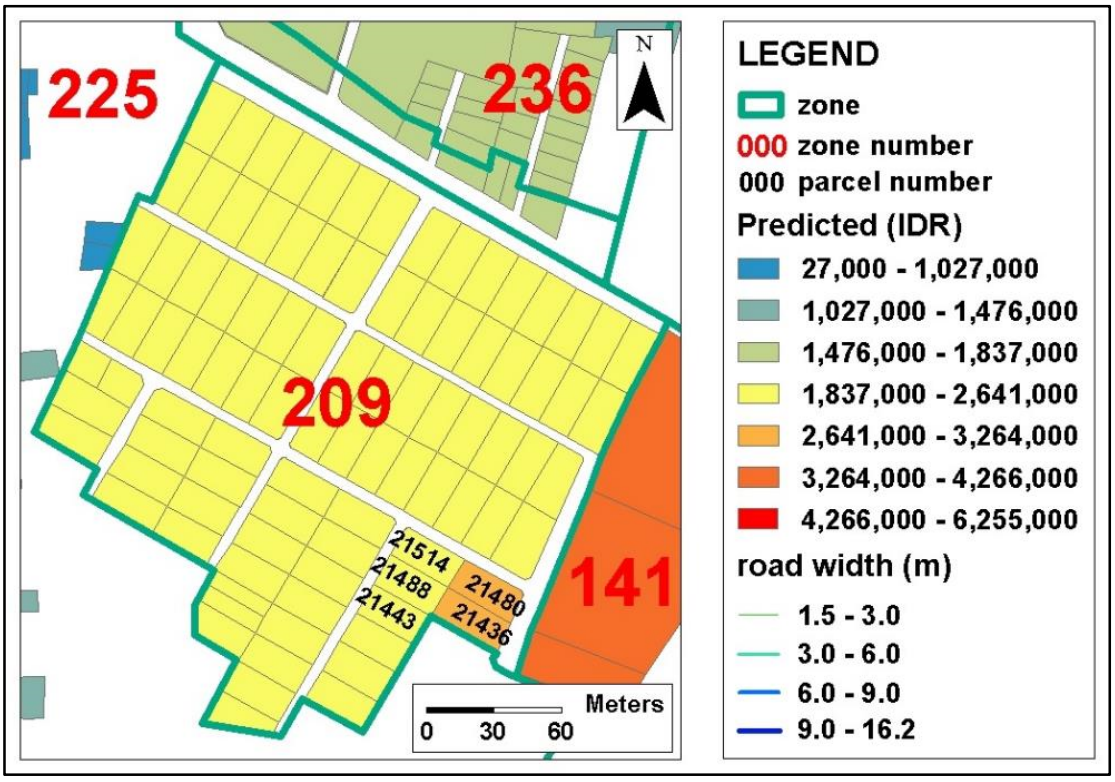


Figure 7.3Predicted land values in zone 209

Zone 209 is a small residential complex of 95 land parcels. Predictions at Locations 21436 and 21480 are much higher than adjacent predictions. The predicted values are IDR 2,869,000 (21480) and IDR 1,894,000 (21514). The prediction for Parcel 21480 is 51.48 per cent higher than adjacent parcel 21514; while 21436 is 52.12 per cent higher than 21488. Predictions at locations 21480 and 21436 are very likely to be outliers. Manual detection using symbology schemes is practical and convenient (see Figure 7.3). However, determination of whether a prediction is an outlier must be supported by robust examination. Local Moran's I test was applied on all 95 predictions in the zone. At each target prediction, all of the 94 neighbouring parcels in zone 209 were involved in the test. The contribution of each neighbour in determining the Local Moran's Index at any target prediction was specified by using the contiguity weighting scheme. This scheme was chosen because parcels in each block are adjacent to one another, they have similar sizes and are regularly arranged.

A worked example is provided for the prediction at parcel number 21436. The contiguous parcels are 21480, 21433, 21488, and 21514 (Figure 7.3). The predicted value at Location 21436 (x_i) is IDR 2,872,000, and the predicted values at the neighbouring (x_j) Locations

21480, 21433, 21488, and 21514 are IDR 2,869,000; 1,884,000; 1,888,000; and 1,894,000 respectively. The average prediction value in zone 209 was calculated from all 95 locations, while the variance of prediction at parcel number 21436 was calculated using all 94 neighbours. The average value and the variance of neighbours for Location 21436 are IDR 1,975,700 and 15,699,373,760 respectively. Next, the weights for neighbouring parcels ($W_{i,j}$) were set. There are 90 parcels in zone 209 not contiguous to 21436, so deviations from the average prediction value for these parcels were given weights of 0. A weighting of 1.0 was given to the contiguous parcels, i.e., 21480, 21443, 21488, and 21514. Each weight was then standardised by dividing it by the sum of contiguity weights at the target parcel. The difference between the average prediction and each adjacent prediction was then calculated. Each difference was given the standardised weight, so there will be weighted difference at each contiguous neighbour. The weighted differences from all four neighbours were then summed (Table 7.1).

Table 7.1Calculation of weighted differences from the average prediction for contiguous neighbours of the target prediction

Step 1	target parcel	weight for contiguous neighbour				Sum
		21480	21443	21488	21514	
	21436	$W_{21436,21480}$	$W_{21436,21443}$	$W_{21436,21488}$	$W_{21436,21514}$	**4**
		(1)	**(1)**	**(1)**	**(1)**	
step 2	target parcel	standardized weight for contiguous neighbour				Sum
		21480	21443	21488	21514	
	21436	$W_{21436,21480}$	$W_{21436,21443}$	$W_{21436,21488}$	$W_{21436,21514}$	**1**
		(1/4)	**(1/4)**	**(1/4)**	**(1/4)**	
step 3	target parcel	weighted difference from average prediction value for contiguous neighbour				Sum
		21480	21443	21488	21514	
	21436	$\frac{1}{4} * (x_{21480} - \bar{x})$	$\frac{1}{4} * (x_{21443} - \bar{x})$	$\frac{1}{4} * (x_{21488} - \bar{x})$	$\frac{1}{4} * (x_{21514} - \bar{x})$	**157,938**
		(223,249)	**(-22,860)**	**(-21,940)**	**(-20,511)**	

The average prediction in zone 209 ($\bar{x}$), the variance of all 94 neighbouring predictions from parcel 21436 (S_i^2), and sum of weighted differences from the average prediction of the four

contiguous neighbours $\left(\sum_{j=1,j\neq i}^{n} w_{i,j}\,(x_j - \bar{x})\right)$ were already calculated. The Local Moran's Index for prediction at parcel 21436 was calculated using Equation 7.1, and the result was 5.85.

Fiaschi *et al.* (2015) used a Local Directional Moran Scatter Plot (LDMS) to explain the dispersion and clustering issues among data. A similar tool is utilised for this study (Figure 7.4).

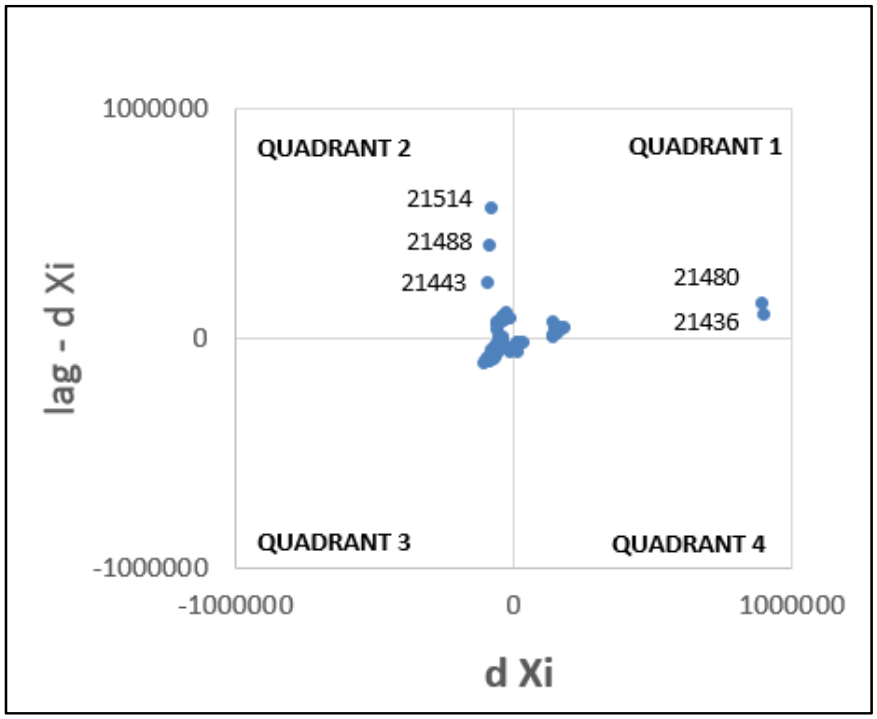


Figure 7.4Local Directional Moran Scatter Plot for zone 209

The x axis, d Xi is the difference between prediction at location i and the average prediction from all predictions. The y axis, lag-d Xi is the difference between the weighted average of predictions contiguous to location i and the average prediction from all predictions in the zone. In Figure 7.4 quadrant 1, a prediction and the weighted average of predictions contiguous with it are both larger than the average prediction in the zone. Quadrant 3 includes prediction and its contiguous neighbours that are significantly smaller than the average prediction in the zone. In quadrant 2, a prediction is smaller than the average prediction in the zone but the weighted average of predictions contiguous to it is larger than the average prediction in the zone. In quadrant 4, a prediction is larger than the average prediction in the zone but the weighted average of predictions contiguous to it is smaller than

the average prediction in the zone. In quadrants 2 and 4, a prediction tends to become an outlier when there is a large contrast between the prediction and its contiguous neighbours.

Predictions 21443, 21488, and 21514 are very likely to be outliers and the large negative Local Moran's indexes confirm this. Interestingly, the above scatter diagram (Figure 7.4) shows that these predictions are within 1.04 to 1.05 per cent of the average prediction. Prediction 21514 has the largest negative Local Moran's index, so it is discussed further. Prediction 21514 only has three contiguous neighbours, i.e. prediction 21488, 21436, and 21480. While prediction 21488 is very close to the average prediction, predictions 21436 and 21480 are 45.37 per cent and 45.20 per cent larger. These extremely large neighbours are the dominant cause for the contrast between prediction 21514 and its contiguous neighbours. The extremely large predictions at 21436 and 21480 overwhelm the influence of the other contiguous neighbours when calculating Local Moran's indexes at parcels 21488 and 21433 resulting in them being detected as outliers even though they are similar to the average prediction for the zone.

Beside causing other 'average predictions' detected to be outliers, the large predictions at parcels 21436 and 21480 also form a cluster of large predictions and are behind the dispersion and clustering issues in zone 209. Dispersion patterns at locations 21443, 21488, and 21514 are marked by Local Moran's indexes of -0.90, -1.47, and -1.93 respectively, while clustering patterns at locations 21436 and 21480 are marked by Local Moran's indexes of 5.85 and 8.95 respectively. Putting aside those five locations with dispersion and clustering patterns, the other 90 locations in zone 209 have Local Moran's indexes between -0.19 and 0.54.

The finding from zone 209 suggests that not only the predictions forming dispersion (outlier) patterns require examination. In this zone, predictions which form a cluster of high predictions were the ones causing the dispersion pattern at other predictions. For this reason, the predictions forming clustering pattern were examined further. Data related to the explanatory variables and parameter estimates at locations 21436 and 21480 were evaluated.

Variation in road width is most likely to be the main reason behind the significantly high predictions at locations 21436 and 21480. Both parcels are located on a 7.84 metre wide road segment, while the average width of all other road segments in the zone is only 3.57 metres. There is no variation in zoning and only very small variation in travel time to the nearest tollgate. In order to come up with solid conclusion, variation of model parameters is also examined (Figure 7.5).

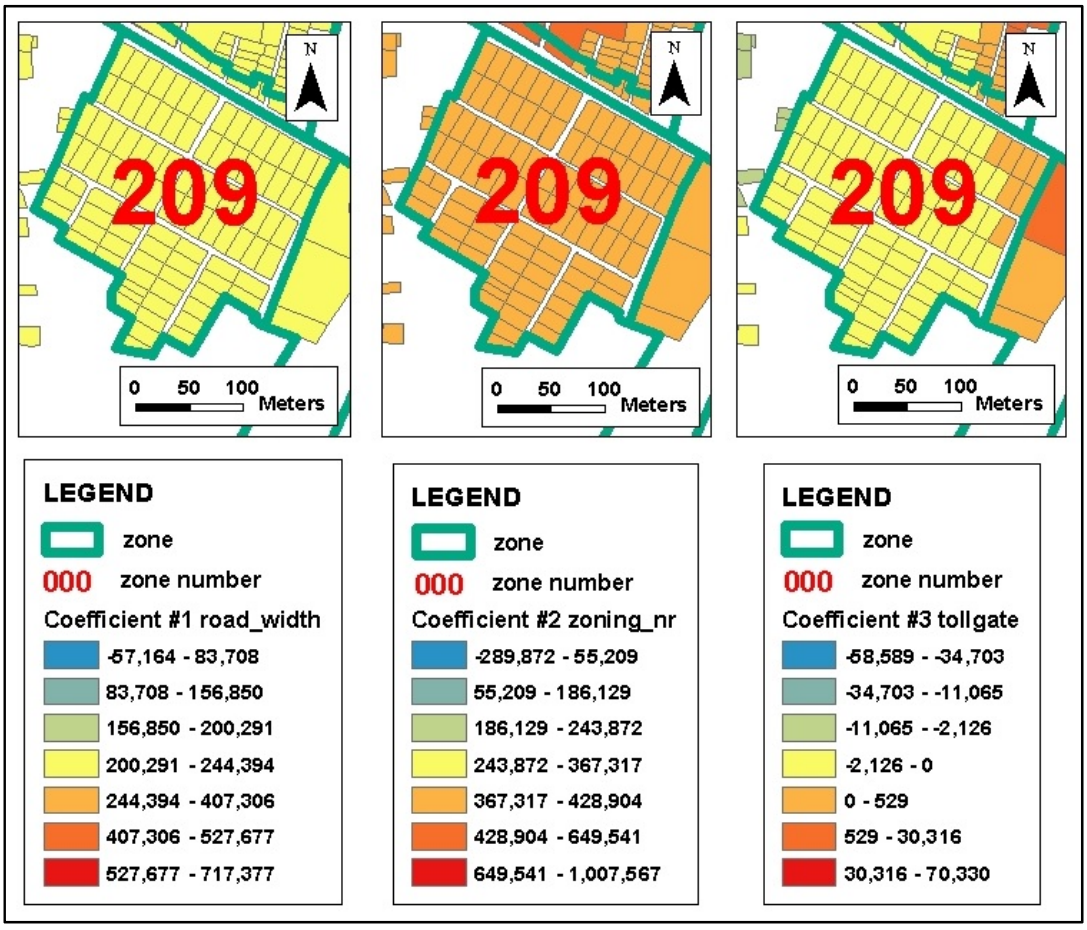


Figure 7.5Parameter estimates in zone 209

At parcel number 21480 and 21436, the parameters on variables road width, zoning and tollgate do not stand out. At location 21480, for instance, the parameters are 223,078, 387,996, and -143 respectively which are in all cases close to the medians of their ranges. Therefore, the model parameters at these locations do not seem to have a significant contribution to the high predictions, making it clear that the significant difference in road width is the main cause for the significantly higher predictions. However, the price differences of 51.48 per cent and 52.12 per cent between these parcels and their adjacent neighbours in a small residential complex like zone 209 is not normal, as prices of land parcels in residential complexes are usually very similar. The two most reasonable possibilities are the following. First, the width of the road segment adjacent to 21480 and 21436 may be wrong due to errors introduced during the process of delineating blocks. Inaccurate gaps between blocks create inaccurate road segment widths. Second, the road width is accurate but the actual prices may be only slightly or moderately higher than their neighbours. A field check would be the best solution to examine these two potential sources of error.

Zone 209 is an example in which analysing the spatial pattern of predictions is quite effective for detecting anomalous predictions. There are cases in which the spatial pattern does not indicate an anomaly among predictions but the variation among predictions in a value zone is quite high. The coefficient of variation among predicted prices in one value zone can also be employed to detect anomalous predictions.

7.3.2. Coefficient of variation among predictions in one value zone

There are zones, like 448, in which no prediction appears markedly different from its neighbours but the overall variation among predictions is quite large. Zone 448 is covered by agricultural land parcels, with 33 of these being registered (Figure 7.6). The coefficient of variation amongst predictions is 35.12 per cent, and this moderately large variation is clearly shown in the map of the zone. Variation in road width is again most likely to be the main reason for this variation.

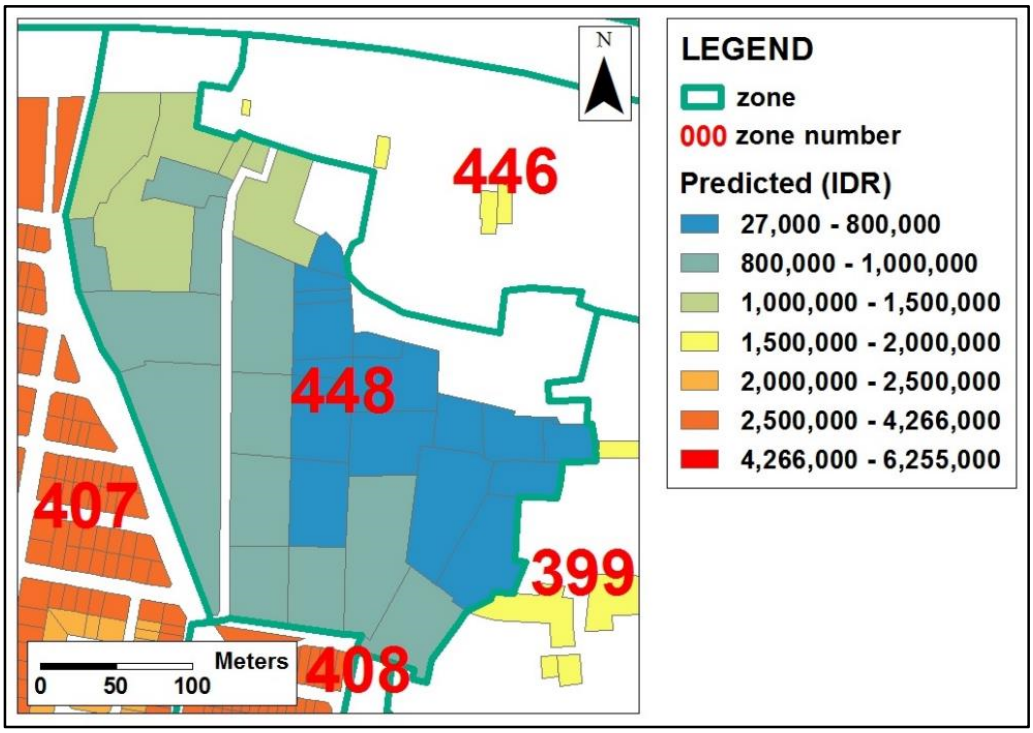


Figure 7.6Predicted land values in zone 448

Spearman's rank correlation was run between predictions and each of the explanatory variables to understand how prices change with each data input, except zoning as there was no variation. The correlation coefficients for price for road width and travel time to nearest

tollgate are 0.975 and -0.484 respectively. Road width data is highly correlated with price, which the inference that it is the dominant factor shaping price in this zone. The value predictions in zone 448 occur in three price bands. Land parcels with values between IDR 1,000,000 and 1,500,000 are located on road segments wider than four metres; those between IDR 800,000 and 1,000,000 are located on road segments of three to four metre wide; while those less than IDR 800,000 are located on lanes less than three metres wide.

Although it can be well explained by the input data, the moderately large prediction variation in zone 448 should be considered as a warning that predictions in the zone may require further examination. Comparing the predictions (Figure 7.7) with observed actual prices is an effective examination procedure. It is an advantage to have one sampled location located in zone 448.

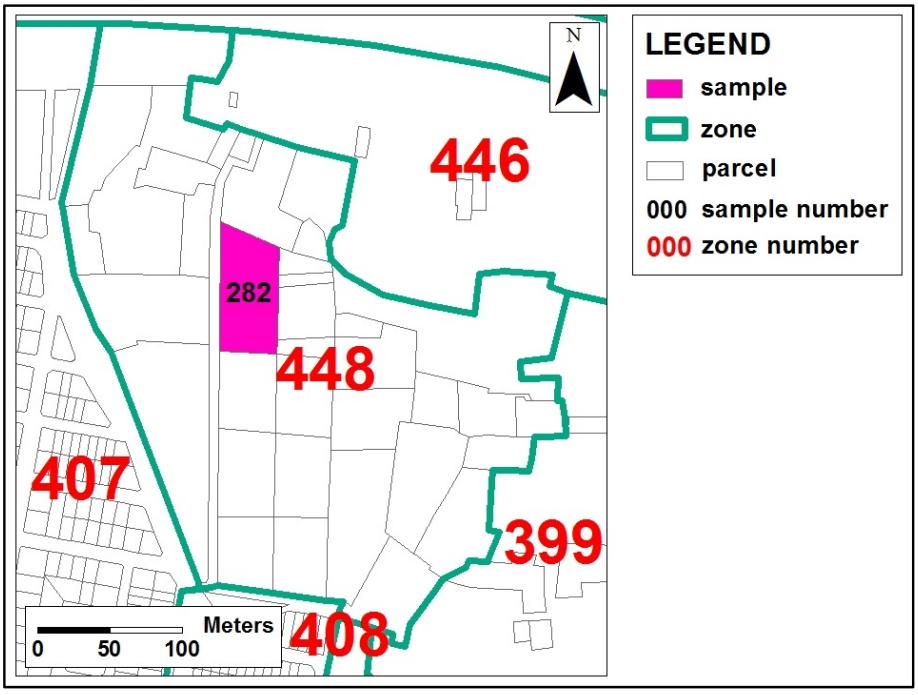


Figure 7.7 Sampled location in zone 448

The observed price at location number 282 is IDR 1,087,000, while the predicted price at the same location (parcel number 74170) is IDR 910,000. The predicted price is smaller than the observed price by 16.26 per cent. Recalling the cut-off value of 30 per cent, this residual is moderate and still acceptable. It is always useful to have observation data in a zone with largevariation of prediction like zone 448. Predictions in the zone are considered to be less reliable, and the observation is used to ensure that the predictions are not too inaccurate.

7.4. Summary

7.4.1. Verifying predictions using spatial pattern of predictions in one zone

Analysing the spatial pattern of predictions is an effective measure to detect anomalous prediction, and it becomes more important when the coefficient of variation among predictions is low. Zone 209 is a good example of this situation. With a low coefficient of variation (7.87 per cent), predictions within zone 209 are expected to be similar to one another. However, significant clustering and dispersion patterns were found when the Local Moran's I test was run on each prediction.

Two predictions deviate significantly from the average prediction value, while 93 other predictions in zone 209 are quite similar to one another. Although the anomalous predictions deviate considerably from the average value, the small number of anomalous predictions only results in a small variation among predictions. In this kind of situation, the coefficient of variation does not indicate a serious issue of anomaly among predictions but the spatial pattern of predictions does.

7.4.2. Verifying predictions using the coefficient of variation among predictions in one zone

The coefficient of variation is an effective tool to verify predictions which differ gradually across space in a value zone. Zone 448 is a good representation of this type of situation. The Local Moran's I test does not detect any anomaly in the spatial pattern of predictions because predictions change gradually across space.

Although there is no anomalous prediction detected in zone 448, the variation among predictions is moderately high(35.11 per cent). It is true that this significant variation can be explained by the significant variation in data related to variable road width, but the high variation should be taken as an indication of potentially inaccurate predictions. Whether or not the variation in data related to variable road width should cause such high variation on land values of neighbouring agricultural lands, should be verified through a field check.

8. DISCUSSION

8.1. Introduction

The Zonation Method, the mass valuation technique currently employed in BPN RI, requires at least three samples of valid data for each value zone. Unfortunately, every mass valuation project administered by BPN RI has failed to provide a sufficient number of samples to comply with this standard. Valid property sales data are scarce because the parties involved in a property transaction tend to give false statements about the transaction price to lessen the transaction tax.

A mass valuation method that uses a limited amount of data is required for mass valuation in Indonesia. The analysis of existing methods (Chapter Two) identified Geographically Weighted Regression (GWR) as the method with the most potential to replace the Zonation Method. GWR was applied to an existing dataset from the 2012 Bekasi City Mass Valuation Project, even though it had a limited number of samples. This dataset was chosen for this research because the sampling densityin this dataset is the highest amongst all datasets from all of the mass valuation projects administered by BPN RI. Moreover, the samples are evenly distributed in the Bekasi dataset.

GWR was run using in-sample and out-of-sample estimation approaches. The out-of-sample approach was run using the Monte Carlo Cross Validation method (Section 5.3.2). The key finding of this study is that the results from both approaches show a very strong association (Section 5.4.1). This strong association indicates that the GWR prediction model does not have overfitting issues, which in turn means that the prediction accuracy at the non-sampled locations is expected to be similar to the prediction accuracy at the sampled locations.

However, when the GWR model is applied to the sampled locations, a number of predictions are potentially inaccurate because they come up with very large percentage residuals (Section 5.3.1). Given the strong association between the in-sample and out-of-sample estimations in the GWR model, the issue of potentially inaccurate prediction is very likely to also occur at non-sampled locations. For the non-sampled locations, detection of potentially inaccurate predictions was conducted per value zone because predictions in one value zone were expected to be similar to one another (Section 7.2). In each value zone, the coefficient of variation was effective in indicating extremely high variation among predictions and the Local Moran's I test was also found to be effective in detecting anomalous predictions. These measures worked well to verify GWR predictions in Bekasi, and they should also work

well to verify mass valuation predictions in other administrations where land parcels can be aggregated into value zones.

8.2. Discussion on the results

8.2.1. Developing a dataset for geospatial modelling from the BPN RI dataset

A key finding from this research is that existing BPN RI-datasetsare not produced in a format that is suitable for use in geospatial modelling of land transaction values in the context of mass valuation. While they are suitable for use with BPN RI's Zonation Method (Section 1.5), significant amounts of research effort were required to convert the dataset from Bekasi into a format suitable for GWR modelling (Chapters Three and Four).

The data collection forms used by BPN RI for mass valuation work record 12 variables. Ten of the 12 variables are related to road class, road width, and travel times to amenities. After the location related to property sales data is defined in the Land Parcel Map, a road network dataset is required to provide data related to these ten explanatory variables for that sample. The road network dataset was derived from the Land Parcel Map because the road network data from other sources (OpenStreetMap and the Topographic Map of Indonesia) had severe mismatches with the Land Parcel Map (Section 3.4). Similar work deriving road network data from a cadastral map was conducted by Haunert and Sester (2008) in Hildesheim, Germany. The main issue in the work was to reconstruct the junctions correctly. Nonetheless, 89.80 per cent of all junctions were appropriately represented. In research conducted by Zhang *et al.* (2010) in Barcelona, Spain, the authors also found reconstructing unique intersections was the main issue but overall, 97.00 per cent of reconstructions were reasonable. For this current research project, the percentage of well reconstructed intersections was not calculated because the exact representation of road segments at intersections was not a crucial requirement. A more important matter was to ensure that all road segments were connected to one another to allow the proper calculation of travel distances, and the topological validation was successful, to ensure that all road segments were appropriately connected.

From the road network dataset, travel distances from each parcel to amenities were calculated. Unfortunately, travel distance is not an effective measure of accessibility because it regards similar contribution of each road class in relation to ease of travelling to amenities. In order to allow a unique contribution of each road class, travel distances to amenities are transformed into travel times to amenities (Section 3.6). This is actually in line with the

guidelines in the Internal Standards for Land Valuation in BPN RI[9] which advise the use of travel time as a measure of accessibility. Optimisation of travel time in this work is a basic application of the Djikstra's algorithm on Shortest Path Optimisation, so the optimisation is rather standard compared with advanced optimisations in other works. For instance, Sun *et al.* (2014) modelled the driver's route choice and used road usage as the weight for route optimisation in Shenzhen, China. They found that incorporating the modelled driver's preference resulted in a high match between the predicted optimum route and the actual travel route usually chosen by drivers. Yiannakoulias *et al.* (2013) incorporated turn penalties and traffic congestion to optimise travel time in Edmonton, Canada. The results revealed that travel times when turn penalties and traffic congestion were incorporated in the estimation were generally over twice the free-flow travel times. It is true that incorporating the driver's preference, turn penalties, and the traffic congestion can result in a more realistic travel time optimisation but calculating this comes with significant technical challenges. Yiannakoulias *et al.* (2013) highlighted that the schemes of free-flow travel time and congested travel time ended up telling the same story; the areas of low or high accessibility were likely to remain in the same classes for both schemes. This lends weight to the decision to apply the basic route optimisation algorithm to calculate travel times to amenities in this study.

In terms of travel times from each land parcel to amenities, the research revealed that only travel time to the nearest tollgate is statistically significant with coefficient correlation to price of -2.680. However, this association with price is moderately lower than those of road width, zoning type and road classwhich have correlation coefficients with price of 0.710, 0.663 and 0.393 respectively. Although road class has a statistically significant association with price, the Backward Elimination Regression and OLS models show that it is not statistically significant within the model. With only road width, zoning type, and travel time to nearest tollgate as the explanatory variables in the model, the prediction model is free from spatial autocorrelation issue. The absence of spatial autocorrelation indicates that the model does not have issues of 'missing variables', model misspecification, redundant information, failure to capture the spatial process mechanism, and areal unit problem(see Griffith, 2009). Two of the five causes of spatial autocorrelation were tackled in this study. The first issue causing spatial autocorrelation addressed was model misspecification. Griffith (2009) observed that specifying linear relationship for non-linear relationship is a common example of model misspecification. OLS and GWR assume linear relationships between the dependent variable and each explanatory variable. Therefore, instead of imposing linear relationships in the OLS or GWR models, the explanatory variables are transformed in order to improve the

[9]Standar Operasional Prosedur Internal Survei Potensi Tanah, BPN RI(2013).

linearity of the price variables (Section 4.6). The second issue causing spatial autocorrelation tackled in this study was the failure to capture spatial process mechanisms. A spatial process mechanism is very likely to cause local variations, and applying a global model to a spatially varying process causes spatial autocorrelation, therefore calibration of local regressions will remove spatial autocorrelation caused by the failure to capture spatially varying process (see Fotheringham, 2009).

While road width, zoning type, and travel time to the nearest tollgate were found to be the most effective predictors of land price in Bekasi, this does not mean that they would be the most important explanatory variables in other administrations. Road width, zoning type, and travel time to nearest tollgate (as part of the 12 variables listed in the data collection forms of BPN RI) are predetermined by appraisers within BPN RI. The appraisers in other administrations may predetermine different sets of explanatory variables for mass valuation. Next, the uniqueness of a dataset from an administration gives a unique relationship between price and each explanatory variable. If a set of explanatory variables are applied to two datasets from different administrations, the relationships between price and each of the explanatory variables will vary between the two administrations. For example, in Bekasi where only 11.42 per cent of commuters travel by public transport, the travel time to nearest station is not an important variable but travel time to nearest tollgate is. A contrasting situation is found in Tokyo, where 48 per cent of workers commute by public transport, and accessibility to the nearest station is an important variable for mass valuation but travel time to nearest tollgate is not. Nevertheless, both mass valuations in Tokyo (see Shimizu and Nishimura, 2007) and in Bekasi agree that road width is an important predictor for land price. It is difficult to find mass valuation works that use variable zoning as an explanatory variable but in mass valuation research in Madrid (see Morillo *et al.*, 2017), zoning was considered to be an important predictor of price because a value zone indicates housing-parameters (typology, age, surface area, price per square metre, etc.) for houses within it. For similar reasons, land zoning was used in Bekasi.

To sum up, modifying the original Bekasi dataset into a format that is compatible with geospatial modelling was a huge task, and this applies to other Indonesian cities as they all have mass valuation datasetssimilar to the Bekasi dataset. Developing a road network dataset from the Land Parcel Map was the main part of data preparation, but it was valuable because the data for ten of the 12 explanatory variables were provided from the road network dataset. Data examination (Chapter Four) concluded that only three of the 12 explanatory variables were statistically significant in Bekasi but results may vary in other cities.

8.2.2. Geospatial modelling of the Bekasi dataset

The second research question focuses on how the application of a geospatial technique that has been used for mass valuation in other administrations performs with the Bekasi dataset. The first step in answering this question was to select the geospatial technique that is the most fit for purpose when applied to the Bekasi dataset. First, the selected technique must be applicable in the circumstances of the Bekasi dataset and the results must be understandable. Next, the selected technique must also be able to provide accurate predictions for all the land parcels in Bekasi.

An extensive literature review was conducted to identify quantitative methods that have been used worldwide for mass valuation. Only a few of the techniques discussed in Chapter Two have been employed in mass valuation since 2010. It was very difficult to find mass valuation work using Hierarchical Trend Modelling (HTM), Logistic Regression, Generalized Additive Modelling (GAM), Piecewise Parabolic Multiple Regression Analysis (PPMRA), Spatial Expansion Model (SEM), or Case-Based Reasoning (CBR) since 2010. There have been only a few researcherswho have usedCokriging, Rule-Based Expert System, and Genetic Algorithms (GA) in mass valuation since 2010. The analysis of research literature shows that artificial neural network-based (ANN-based) techniques and Geographically Weighted Regression dominate most recent mass appraisal research.

ANN techniques can provide high prediction accuracies but the processes and the results are not easy to explain, and this is why Kauko and d'Amato (2008) refer the nature of ANN as a 'black box'. This is a big issue because the output of the mass valuation work is predominantly used for taxation purposes, and mass valuation in BPN RI is no different. Having a clear explanation on how the amounts of tariffs and taxes are determined is a big concern in this area. Moreover, model-free estimation techniques like ANN-based techniques appear to require more data than regression-based techniques because inferences are made from the data themselves (see Lin and Mohan, 2011; Nguyen and Cripps, 2001; Peterson and Flanagan, 2009). With the very limited amount of data in the Bekasi dataset, ANN-based techniques are very likely to give unreliable predictions.

GWR has been widely used in the area of mass appraisal, because it isolates and combines spatial dependency and heterogeneity, accounting for locational or adjacency effects and market segmentation (McCluskey *et al.*, 2013). Because of the ability to capture spatial heterogeneity which in turn results in accurate predictions, GWR is also widely used in other fields. Wang *et al.* (2012) mapped the distribution of soil organic matter in Longyan, China using GWR and Regression Kriging (RK). For both methods, the root-mean-square error (RMSE) values are 2.748 and 7.576 respectively, while the adjusted R-square values are

0.909 and 0.699 respectively. These results indicate that GWR not only comes up with more accurate predictions than RK but it also fits the data better than RK. Liu *et al.* (2017) employed GWR, Ordinary Kriging, Inverse Distance Weighted Interpolation, Multiple Linear Regression Model, and the Linear Mixed-Effect Model to map the spatial distribution of soil organic carbon density (SOCD) in Jinjing Town, China. The standardised mean square error (SMSE) values are 3.92, 4.22, 14.91, 14.82, and 9.12 respectively, while the R values are 0.66, 0.57, 0.58, 0.56, and 0.64 respectively. These results suggest that GWR generated more accurate spatial distribution of SOCD than the other techniques.

In the field of mass valuation, GWR also shows its superiority. McCluskey *et al.* (2013) applied a number of techniques (Multiple Regression Model – MRM, Artificial Neural Network – ANN, Spatial Simultaneous Autoregressive – SAR, and GWR) for mass appraisal in Northern Ireland, UK. GWR came up with the highest prediction accuracy as indicated by the lowest mean absolute percentage error (MAPE) value; 10.40 per cent, while MRM, ANN, andSAR came up with MAPE values of 12.27, 11.97, and 13.69 per cent respectively. Chrostek and Kopczewska (2014) compared the performances of a number of techniques (Ordinary Least Squares – OLS, Spatial Expansion Model – SEM, Spatial Lag, Spatial Error Model, and GWR) in mass appraisal work in Wroclaw, Poland. GWR was also the most accurate technique with the MAPE value of 12.96 per cent, while OLS, SEM, Spatial Error Model, and the Spatial Lag Model came up with MAPE values of 13.405, 36.476, 13.186, and 14.191 per cent respectively The prediction accuracy of the GWR model for mass valuation using the Bekasi dataset is also measured using MAPE value, and the MAPE value of 19.40 per cent suggests that the GWR model is moderately accurate. Compared with the research conducted by McCluskey *et al.* (2013) and Chrostek and Kopczewska (2014), the prediction accuracy of the GWR model using the Bekasi dataset is moderately lower. The quality of the Bekasi dataset is probably lower than the datasets from the other two pieces of research, or the set of explanatory variables used in Bekasi is probably less effective.

The mean absolute percentage error (MAPE) value is again used to compare the prediction accuracy of the proposed GWR method with the Zonation Method currently employed in BPN RI. The Zonation Method came up with MAPE values of 10.8 per cent. In terms of prediction accuracy, the Zonation Method outperforms the GWR model. Nevertheless, the GWR model is a potential candidate to replace the Zonation Method because the GWR model solves the main problem of theZonation Method, i.e. non-verifiable prediction for zones with fewer than three samples. It is true that the GWR model offers lower accuracy

than the Zonation Method but the MAPE value of 19.40 per cent is way below the cut-off value of 30 per cent currently used in BPN RI.

The main issue with the GWR model using the Bekasi dataset is that there are a number of locations with extremely large prediction residuals. Six locations have percentage residuals larger than 100 per cent, and ten locations have percentage residuals between 75 to 100 per cent. In order to address the issue of extremely large residuals at a number of locations, GWR was run using value zones. The idea is to control the input data for the model by detecting potentially inaccurate observations. A lesson learned from the Zonation Method is that neighbouring land parcels in one value zone tend to have similar prices, so the observed prices at the sampled locations in one zone are expected to be similar. Controlling the observation data (input data for the model) is expected to remove or reduce the extremely large residuals at a number of locations. Unfortunately, this new approach fails the task as GWR using value zones also comes up with large residuals at a number of zones. The prediction accuracy of the GWR model using value zones is lower than the GWR model using individual locations, as indicated by the MAPE value of 24.86 per cent. Compared with the cut-off value of 30 per cent currently used in BPN RI, the prediction accuracies of the GWR model using individual locations and the GWR model using value zones are actually moderately good.

There is a possibility of the performance of a model being overestimated when run in an in-sample validation. Excessive bending to minimise the training error at the sampled locations can cause an issue of overfitting for the model. When validated using the sampled locations (which were previously used to shape the model), the model shows optimum performance. A model with an overfitting issue will potentially show a quite different performance when validated using non-sampled locations. In order to obtain a good estimate on the model's performance at the non-sampled locations, out-of-sample validation was run using the Monte Carlo Cross Validation (MCCV) technique.

When 1,000 iterations of the GWR model are run using MCCV, the distribution of average percentage residuals resembles the distribution of percentage residuals from the in-sample GWR model. The correlation coefficient between out-of-sample average percentage residuals and in-sample percentage residuals is surprisingly high; 0.987. This indicates that the GWR prediction model does not have an issue of overfitting. With no issue of overfitting in the GWR prediction model, the prediction accuracy at the non-sampled locations is expected to be similar with the prediction accuracy at the sampled locations. In other research, GWR was also revealed not to have the issue of overfitting, and therefore resulted in accurate out-of-sample predictions. Páez *et al.* (2008) compared the performance of

moving window techniques (Moving Window Regression – MWR, Moving Window Kriging – MWK, and GWR), and found out that GWR and MWR resulted in more accurate out-of-sample predictions than MWK.Helbich and Griffith (2016)ran out-of-sample estimations to examine the performance of a number of prediction models (Spatial Expansion Method – SEM, Moving Window Regression – MWR, Genetic Algorithm-Based Eigenvector Spatial Filtering – ESF, and GWR) using Leave-One-Out Cross Validation (LOOCV) and Hold-Out (Monte Carlo Cross Validation – MCCV) methods. The results from 100 iterations suggest that the SEM, MWR, and GWR do not have an overfitting issue, while ESF does.

8.2.3. Verifying predictions at non sampled locations

If the GWR model using individual locations is to be employed in BPN RI, adjustments must be organised to cope with the main issue of the model, i.e. the few large residuals. Measures to verify predictions need to be specified to detect the potentially inaccurate predictions, and these measures must suit the circumstances of Bekasi as well as other Indonesian cities. In Indonesia, one neighbourhood can have significantly different characteristics from its surrounding neighbourhoods. For instance, a residential complex is usually a well-looked-after neighbourhood with well-arranged land parcels and road segments, while an irregular residential area is usually a modest neighbourhood in which land parcels and road segments are not regularly arranged. This situation allows delineation of neighbourhoods to split land parcels in a city or district into a number of neighbourhoods. Land parcels in one neighbourhood tend to have similar qualities, and in turn tend to have similar value. For this reason, the polygons of neighbourhood have been called polygons of value zones. Delineation of value zones is undertaken by mass valuation analyst at each local Land Office in Indonesia.

Delineation of value zones for mass appraisal is also conducted in other administrations. Value zone is utilised to aggregate the similarities among neighbouring properties in urban areas in the Ukraine because the available number of land sales data is inadequate to work with the mass appraisal techniques commonly employed worldwide (see Kryvobokov, 2004). In Madrid, Spain, Morillo *et al.* (2017) proposed the use of value zones to represent housing-parameters (typology, age, surface area, price per square metre, etc.) for houses within each zone.

Value zones can be employed to verify predictions of the GWR model for two main reasons. First, the GWR model uses explanatory variables (road width, zoning, and tollgate) for prediction. Land parcels in one value zone share similar values related to these explanatory

variables. Second, by running local regressions and applying a weighting scheme, GWR tries to capture local variations, and in doing so, GWR also tries to incorporate the 'unmeasured variables' (see Fotheringham *et al.*, 1997). Being located close to one another in one neighbourhood of distinguishable characteristics, land parcels in one value zone are expected to receive similar effects of 'unmeasured variables'. For these two reasons, predictions in one value zone are expected to be similar to one another. In order to judge whether or not one prediction is potentially inaccurate, that particular prediction must be compared with the other predictions in the zone where it is located.

Analysing the spatial pattern of predictions in each value zone can be an effective measure to detect anomalous predictions. The Local Moran's I test was run per value zone, so all predictions in one value zone are involved in each test. Because predictions in one value zone are expected to be similar to one another, each value zone is expected to have a homogenous pattern. An anomalous prediction, which is extremely larger or smaller than its nearest neighbours, will stand out among other predictions in a zone. There are cases in which predictions differ gradually across space in a value zone, so the result from the Local Moran's I test does not detect any anomaly on the spatial pattern. Coefficient of variation (COV) can be an additional tool to verify predictions in zones with this kind of situation. A high variation of predictions in one value zone should be taken as an indication for careful use of the predictions in the zone because predictions in one value zone are expected to be similar to one another. Kim and Kim (2016) demonstrated that COV is an effective tool to assess the horizontal equity of predictions of spatial statistic models. Examples in Chapter Seven also show that the use of Local Moran's Index and COV to verify predictions in one value zone is effective to detect anomalies or outliers in the zone.

8.3. Summary

Road width, zoning, and travel time to nearest tollgate were found to be the most effective explanatory variables in a GWR model for land price prediction in Bekasi. When using these three variables, the GWR model does not display multicollinearity or spatial autocorrelation, neither does it have an overfitting issue, so the prediction accuracy at non-sampled locations is expected to be similar to the prediction accuracy at the sampled locations.

In terms of accuracy, the GWR model is not as good as the currently employed Zonation Method. Nevertheless, it is a suitable candidate to replace the Zonation Method because the

GWR model solves the main problem of the Zonation Method, i.e. that of non-verifiable predictions for zones with fewer than three samples; and the average residual of the GWR model is well below the cut-off value of 30 per cent currently used in BPN.

The main issue in applying the GWR model is the large residuals that are generated at a number of locations. Controlling the inputs to the model GWR using value zones did not solve this problem. A more feasible measure was to control the output of the model, i.e., the predictions. Land parcels in a single value zone have similar data related to explanatory variables, and they are very likely to receive similar effects from 'unmeasured variables'. Therefore, predictions in one value zone are expected to be similar to one another. Analysing spatial patterns among predictions in one zone was found to be an effective measure to detect potentially inaccurate predictions. Because Bekasi is typical of Indonesian cities, the prediction model and the verification measures specified in this study should be able to be applied to most other Indonesian cities.

9. CONCLUSIONS AND RECOMMENDATIONS

9.1. Introduction

The main objective of this study was to improve current mass valuation practicesin the National Land Agency of Indonesia (BPN RI) given the existing data scarcity issue introduced in this thesis (Section 1.6). The Zonation Method which is currently employed for mass valuation in BPN RI, gives accurate predictions but only for zones where there are at least three samples available. Collection of actual price of land transfer data has been a longstanding issue in mass valuation practice in Indonesia, and because of the gaps in this information, mass valuation as practised at present in Indonesian cites has never been successful in providing a complete prediction of land values for an entire city. As stated in Chapter One, a new method is required to resolve this situation. Moreover, whatever method is chosen must be able to utilise the limited number of samples available in any particular city.

Among the methods that have been developed worldwide for mass valuation, GWR has been employed frequently because of its ability to capture local variations. In order for the Bekasi dataset to be run using GWR, the dataset required a significant number of adaptations. It is argued that these and similar adaptations would be required for any dataset from an Indonesian city. Statistical examination of the data revealed that only road width, zoning type, and travel time to the nearest tollgate were statistically effective explanatory variables. The prediction performance of the GWR model using these three variables is reasonable, as indicated by the mean absolute percentage error (MAPE) of 19.40 per cent. In order to estimate the prediction performance at the non-sampled locations, an out-of-sample estimation was run using repeated random sub-sampling validation (Monte Carlo Cross Validation – MCCV). The results from GWR modelling using all the sampled locations and the out-of-sample estimations are very similar, as indicated by the correlation coefficient of 0.987. This means that there is a high probability that the accuracy of the GWR model when used to predict the non-sampled locations will be similar to the accuracy of the model when used to predict the sampled locations.

The main issue that was discovered after running the GWR model using individual locations and the GWR model using value zones was large prediction residuals at a number of locations. For the GWR model using individual locations, six of the 706 locations had percentage residuals greater than 100 per cent and ten locations had percentage residuals between 75 to 100 per cent. Recalling the very high correlation coefficient between the results from the GWR model and the MCCV on GWR model, potentially inaccurate

predictions are likely to be found at the non-sampled locations. In order to detect potentially inaccurate predictions at the non-sampled locations, all predictions were verified within the context of the value zone that the predictions occurred in. To do that, the spatial patterns and the coefficients of variation of predictions in each value zone were examined so that anomalous predictions could be identified. In this case, an anomalous prediction is a prediction which is much larger or smaller (defined in terms of percentage difference) than its nearest neighbours in a value zone. Predictions in one value zone are expected to be similar to one another, so these predictions are expected to have small coefficients of variation. A large coefficient of variation and the existence of anomalies among predictions in one value zone were taken as an indication of the existence of potentially inaccurate predictions in the zone.

9.2. Meeting the objectives and answering the research questions

In the process of meeting the main objective of the study (Section 9.1), the research questions listed in Section 1.6. were answered. These are summarised individually below.

9.2.1. Research question 1: Converting an existing BPN RI-dataset into a format that can be used in geospatial modelling of land transaction values

The dataset for Bekasi comprised original property sales data, a land parcel map, and zoning data. These data were obtained from the local (Bekasi) Land Office. The Land Parcel Map was used to derive a road network dataset. Deriving this road network dataset from the Land Parcel Map was very time consuming because of the large number of corrections that were required. This effort was essential because the road network dataset, in turn, allowed data related to road classes, road widths and travel times to amenities to be extracted: without this step, the modelling could not have taken place.

The data extracted from the road network dataset generated data for ten of the 12 variables listed in the mass valuation standards of BPN RI. Developing a road network dataset is, therefore, a major requirement in terms of person effort in order to run geospatial modelling for mass valuation in Indonesian cities. This element of the research project is reported in Chapters Three and Four.

9.2.2. Research question 2: Evaluating the performance of the selected model to predict land values in Bekasi, Indonesia

The mean absolute percentage error (MAPE) was used as a measure of prediction performance for the GWR model. The GWR models using individual locations and value zones came up with MAPE values of 19.40 per cent and 24.86 per cent respectively, indicating that GWR modelling using individual locations is a better candidate than GWR using value zones to replace the Zonation Method. Although the GWR model using individual locations has lower prediction accuracy than the Zonation Method, which has MAPE value of 10.80 per cent, the GWR model solves the main problem of theZonation Method, i.e. non-verifiable prediction for zones with fewer than three samples.

The main issue of GWR model using individual locations is that a number of locations had considerably large prediction residuals. Six locations have percentage residuals larger than 100 per cent, and ten locations have percentage residuals between 75 to 100 per cent. Adopting the cut-off value of 30 per cent from BPN RI, 137 of 706 predictions at the sampled locations (19.41 per cent) do not comply with this cut-off value.

In order to estimate the prediction performance at the non-sampled locations, out-of-sample prediction was conducted. One thousand iterations of the GWR model using the Monte Carlo Cross Validation revealed that at each location, the average percentage residual from out-of-sample prediction was not statistically different from the percentage residual from the in-sample GWR model. The coefficient correlation between the average percentage residuals from the out-of-sample and in-sample analyses was 0.987. This very high correlation indicates that the performance of the GWR model when predicting the non-sampled locations had a high probability of being similar to the performance of the GWR model when predicting the sampled locations. As a consequence, predictions at non-sampled locations can be expected to have low to medium MAPE and very large residuals in a small number of locations. This element of the research project is reported on in Chapters Five and Six.

9.2.3. *Research question 3: Identifying* adjustments to improve BPN's mass valuation practices for Indonesian urban areas

If the GWR model using individual locations is to be employed in BPN RI, the issue of large residuals must be addressed. The first step to deal with potentially inaccurate predictions is to identify them. At the sampled locations, potentially inaccurate predictions can be detected using the prediction residuals. Effective measures must also be employed to detect potentially inaccurate predictionsfor non-sampled locations.

Value zones were utilised to verify predictions because predictions in one value zone are expected to be similar to one another. In order to judge whether or not one prediction is potentially inaccurate, that particular prediction must be compared with the nearest predictions in the value zone in which it is located. Examination of the spatial patterns and the coefficients of variationof the predictions in a single value zone, were found to be effective in detecting potentially inaccurate predictions. This element of the research project is reported on in Chapter Seven.

9.3. Recommendations

9.3.1. Recommendations for BPN RI

The biggest challenge in applying a regression-based prediction model for mass valuation using the Bekasi dataset was in deriving the road network dataset from the Land Parcel Map. As discussed in Section 3.4, there were a large number of drawing errors in the Land Parcel Map. Batch fixing of these errors was not appropriate because each error was found to be unique. A series of recommendations emanate from this issue.

(1) In the current practice of maintaining and updating digital land parcel maps within BPN RI, changes to these maps are saved after editing even though there can be drawing errors in the maps. The Bekasi dataset is a clear example of a map in which errors have been ignored and, therefore, errors have accumulated as more drawing and editing is done and saved. This is a bad practice in terms of spatial data management and in terms of costs in ultimately producing accurate land parcel maps, whether they are used in GWR modelling or not. This would need to be addressed and replaced by a system that requires the drawing errors to be corrected before the map is updated and further drawing can occur. This mechanism will ensure that a drafter cannot ignore drawing errors and also prevent errors from accumulating, so the data will always be accurate and ready-to-use.

(2) Road network layers should be included in land parcel maps. A key lesson learnt from this study is that a ‘corrected’ land parcel map still has the potential to cause errors when deriving a road network dataset. For instance, a complicated intersection in a land parcel map can cause an unresolved intersection in a road network dataset. Presenting the road network layer in such a map would allow early detection of errors in the road network dataset.

The result of this study could be the best case scenario in a national context because of the extensive work to fix errors in the datasets. This raises questions on the viability of applying the GWR approach presented in this thesis in Indonesia. Fixing the geometrical errors in order to create a reliable road network dataset is not practical with current condition of Land Parcel Map at local Land Offices. A more feasible option is to use Euclidean distance rather than road network distance for variable distance to amenities. The Euclidian distance does not mimic the real world process as good as the road network distance does, which in turn may reduce prediction accuracy, but using Euclidian distance will greatly save time and effort in data preparation. It is a trade off between quality and cost.

There are also significant issues in current mass valuation practices in Indonesia that are out of the scope of this research. Potential solutions to these issues were not incorporated in the objectives of this thesis due to the time that would have been involved in researching them. Nevertheless, it is important that these are addressed in trying to achieve better mass valuation practice in Indonesia. By far, the most important of these issues is the scarcity of valid data.

While it has been shown that a GWR model is able to provide accurate predictions in a scare data environment, a larger number of records is always beneficial for any prediction model. Property sales data are compiled by each local Land Office, but the parties involved in transactions often tend to under report the actual transaction price in order to reduce their transaction tax burden, which currently is set at five per cent of the sale price for each party (see Tamtomo *et al.*, 2008).

An effective mechanism is required to encourage the parties involved in property transactions to state the actual sales price on the sale deed orin the sale contract. If the sales prices stated in the sales deeds reflected the real situation, the property sales records compiled in the local land office will be ready-to-use for mass valuation analysis.

When the author started this research project, the author had hoped to be able to research this issue as well. But after discussions with supervisor and other academics from the Mathematical Sciences, Psychology and Policy Studies units at Flinders University it became clear than none of the conceptual and methodological frameworks from these disciplines could be applied within the time frame of a doctoral study in addition to the spatial modelling reported in this thesis. Moreover, none of the academics in these units felt that inputs from their disciplines would resolve the false transaction reporting issue in a timely fashion, arguing that fundamental social, fiscal and legal changes are required in Indonesia to overcome this problem.

9.3.2. Recommendations for future studies

In addition to the explanatory variables listed in this study, there may be other variables that contribute significantly to shaping land prices in particular cities or urban districts in Indonesia. In Jakarta, for instance, whether or not a property is located in a flood-prone area is a crucial factor in determining transaction price. In instances like that, vulnerability to flooding should be added to the list of explanatory variables, either by gathering data from property owners on how often they have been flooded, or incorporating flood vulnerability maps in spatial modelling. Exposure to other natural hazards that occur frequently in the country, e.g., earthquakes, volcanic activity and landslides, could be dealt with in a similar manner. Martı´nez-Cuevas *et al.* (2017) determined the vulnerability scores related to earthquakes for buildings in Lorca, Spain, and recommended adaptations to the urban zoning regulations to reduce the vulnerability of buildings in these high-risk zones. In the context of mass valuation, a vulnerability score of each property to the particular hazard is an important explanatory variable in predicting property prices in cities. Therefore, I argue that a key task of future studies is to examine the most appropriate ways to incorporate hazard vulnerability as an explanatory variable in prediction modelling. This would best be accomplished through a series of Masters or PhD thesis in Indonesia or overseas.

10. REFERENCES

Anguita, D., Ghio, A., Oneto, L., & Ridella, S. (2012). In-Sample and Out-of-Sample Model Selection and Error Estimation for Support Vector Machines. *Neural Networks and Learning Systems, 23*(9), 1390-1406.

Anselin, L. (1995). Local Indicators of Spatial Association – LISA *Geographical Analysis, 27*(2), 93-115.

Bagnoli, C., & Smith, H. C. (1998). The Theory of Fuzzy Logic and Its Application to Real Estate Valuation. *Journal Of Real Estate Research, 16*(2), 169–199.

Balmann, A., & Happe, K. (2000). *Applying Parallel Genetic Algorithms to Economic Problems: The Case of Agricultural Land Markets.* Paper presented at the International Institute of Fisheries Economics and Trade (IIFET) Conference 2000 , Oregon, USA.

Bidanset, P. E., & Lombard, J. R. (2014). The Effect of Kernel and Bandwidth Specification in Geographically Weighted Regression Models on the Accuracy and Uniformity of Mass Real Estate Appraisal. *Journal of Property Tax Assessment & Administration, 10*(3), 5-14.

BIG [Geospatial Information Agency of Indonesia]. (2013). *Peta Rupa Bumi Indonesia Skala 1: 250,000 [Indonesian Topographic Map scale 1:250,000]* [Digital Topographic Map].

Bingham, N. H., & Fry, J. M. (2010). *Regression: Linear Models in Statistics*. London, UK: Springer-Verlag.

Bitter, C., Mulligan, G. F., & Dall'erba , S. (2007). Incorporating Spatial Variation in Housing Attribute Prices: A Comparison of Geographically Weighted Regression and the Spatial Expansion Method. *Geographical Systems, 9*, 7-27.

Bolen, F., Yirmibesoglu, F., Turkoglu, H., & Korca, P. (1999). *Determinants of Land Prices in Istanbul: A Case Study*. Paper presented at the European Regional Science Association (ERSA) 39th European Congress, Dublin, Ireland.

Booth, A. (1974). Ipeda – Indonesia's Land Tax. *Bulletin of Indonesian Economic Studies, 10*(1), 55-81.

BPN RI [National Land Agency of Indonesia]. (2013). Standar Operasional Internal Survei Penilaian Tanah [Internal Operational Standards for Land Valuation Survey]. Jakarta, Indonesia: BPN RI

BPS [Indonesian Bureau of Statistics]. (2015). Statistik Komuter Kota Bekasi 2014 [Commuter Statistics of Bekasi City in 2014]. Jakarta, Indonesia: BPS

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis, 28*, 281–298.

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and The Repeated Learning-Testing Methods. *Biometrika, 76*(3), 503-514.

Caruso, J. C., & Cliff, N. (1997). Empirical Size, Coverage, and Power of Confidence Intervals for Spearman's Rho. *Educational and Psychological Measurement, 57*(4), 637-654.

Chan, N., & Chen, F.-y. (2010). A Comparison of Property Taxes and Fees in Sydney and Taipei. *Property Management, 29*(2), 146-159.

Chan, Y. (2005). *Location, Transport and Land-Use: Modelling Spatial-Temporal Information*. Berlin, Germany: Springer-Verlag.

Charlton, M., & Fotheringham, A. S. (2009). Geographically Weighted Regression: White Paper. Kildare, Ireland: National Centre for Geocomputation, National University of Ireland Maynooth.

Charlton, M., Fotheringham, A. S., & Brunsdon, C. (2006). Geographically Weighted Regression (Vol. NCRM/006). Southampton, UK: ESRC National Centre for Research Methods.

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures*. California, USA: Sage Publications, Inc.

Chica-Olmo, J. (2007). Prediction of Housing Location Price by a Multivariate Spatial Method: Cokriging. *The Journal of Real Estate Research, 29*(1), 91-114.

Chica-Olmo, J., Cano-Guervos, R., & Chica-Olmo, M. (2013). A Coregionalized Model to Predict Housing Prices. *Urban Geography, 34*(3), 395-412.

Chrostek, K., & Kopczewska, K. (2014). Spatial Prediction Models for Real Estate Market Analysis. *Ekonomia Journal, 34*, 27-45.

Colwell, P. F. (1998). A Primer on Piecewise Parabolic Multiple Regression Analysis via Estimations of Chicago CBD Land Prices. *Journal of Real Estate Finance and Economics, 17*(1), 87–97.

Colwell, P. F., & Munneke, H. J. (2003). Estimating a Price Surface for Vacant Land in an Urban Area. *Land Economics, 79*(1), 15-28.

d'Amato, M. (2002). Appraising Property with Rough Set Theory. *Journal of Property Investment and Finance, 20*(4), 406–418.

d'Amato, M. (2008). Rough Set Theory as Property Valuation Methodology: The Whole Story. In T. Kauko & M. d'Amato (Eds.), *Mass Appraisal Methods: An International Perspective for Property Valuers* (pp. 220-258). West Sussex, UK: Blackwell Publishing.

Delluchi, M., & Murphy, J. (2005). Motor-Vehicle Infrastructure and Services Provided by the Public Sector *Report #7 in the series of The Annualized Social Cost of Motor-Vehicle Use in the United States*. California, USA: Institute of Transportation Studies, UC Davis.

Efron, B. (1979). Bootstrap Method: Another Look at The JackKnife. *The Annals of Statistics, 7*(1), 1-26.

Farber, S., & Pa´ez, A. (2007). A Systematic Investigation Of Cross-Validation in GWR Model Estimation: Empirical Analysis and Monte Carlo Simulations. *Geographical Systems, 9*, 371–396.

Fiaschi, D., Gianmoena, L., & Parenti, A. (2015). *Local Directional Moran Scatter Plot - LDMS*. Discussion Paper. Dipartimenti di Economia e Managament - University of Pisa. Retrieved from http://www.ec.unipi.it/ricerca/discussion-papers.html.

Fotheringham, A. S. (2009). The Problem of Spatial Autocorrelation and Local Spatial Statistics. *Geographical Analysis, 41*, 398-403.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. West Sussex, UK: John Wiley & Sons Ltd.

Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. (1997). Measuring Spatial Variations in Relationships with Geographically Weighted Regression. In M. M. Fischer & A. Getis (Eds.), *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling, and Computational Intelligence* (pp. 60-82). Heidelberg, Germany: Springer.

Francke, M. K. (2008). The Hierarchical Trend Model. In T. Kauko & M. d'Amato (Eds.), *Mass Appraisal Methods: An International Perspective for Property Valuers*. West Sussex, UK: Blackwell

Francke, M. K., & Vos, G. A. (2004). The Hierarchical Trend Model for Property Valuation and Local Price Indices. *Journal of Real Estate Finance and Economics, 28*(2/3), 179-208.

Geoghegan, J., Wainger, L. A., & Bockstael, N. E. (1997). Spatial Landscape Indices in a Hedonic Framework: An Ecological Economics Analysis Using GIS. *Ecological Economics, 23*, 251–264.

Gonzalez, A. J., & Laureano-Ortiz, R. (1992). A Case-Based Reasoning Approach to Real Estate Property Appraisal. *Expert Systems with Applications, 4*, 229–246.

Griffith, D. A. (2009). Spatial Autocorrelation. University of Texas at Dallas, TX, USA: Elsevier Inc.

Griffith, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering*. Berlin: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Model Assessment and Selection *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). New York, USA: Springer Science+Business Media.

Haunert, J.-H., & Sester, M. (2008). Area Collapse and Road Centerlines based on Straight Skeletons. *Geoinformatica, 12*(2), 169-191.

Helbich, M., & Griffith, D. A. (2016). Spatially Varying Coefficient Models in Real Estate: Eigenvector Spatial Filtering and Alternative Approaches. *Computers, Environment and Urban Systems, 57*, 1-11.

Heryani, E., & Grant, C. (2004). *Land Administration in Indonesia*. Paper presented at the International Federation of Surveyors (FIG) 3rd Regional Conference: Developing Asia and the Pacific, Jakarta, Indonesia.

Hibberts, M., Johnson, R. B., & Hudson, K. (2012). Common Survey Sampling Techniques In L. Gideon (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 53-74). New York, USA: Springer.

Horne, R., & Felsenstein, D. (2010). Is Property Assessment Really Essential for Taxation? Evaluating the Performance of an 'Alternative Assessment' Method. *Land Use Policy, 27*, 1181-1189.

Hui, E. C.-M., Ho, V. S.-M., & Ho, D. K.-H. (2004). Land Value Capture Mechanisms in Hong Kong and Singapore. *Journal of Property Investment and Finance, 22*(1), 76-100.

Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *Royal Statistical Society, 60*(2), 271 - 293.

IAAO [International Association of Assessing Officers]. (2014). Guidance on International Mass Appraisal and Related Tax Policy. *Journal of Property Tax Assessment and Administration, 11*(1), 5-33.

IBM Corp. (2015). SPSS Statistics 22.0.0: Curve Estimation Models. Retrieved January 8, 2017, from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/base/curve_estimation_models.htm

IVSC [International Valuation Standards Council]. (2011). International Valuation Standards 2011. London, UK: International Valuation Standards Council.

Jirong, G., Mingcang, Z., & Liuguangyan, J. (2011). Housing Price Forecasting Based On Genetic Algorithm and Support Vector Machine. *Expert Systems with Applications, 38*, 3383-3386.

Kauko, T., & d'Amato, M. (2008). Suitability Issues in Mass Appraisal Methodology. In T. Kauko & M. d'Amato (Eds.), *Mass Appraisal Methods: An International Perspective for Property Valuers*. West Sussex, UK: Blackwell

Kelly, R. (2004). Property Taxation in Indonesia. In R. M. Bird & E. Slack (Eds.), *International Handbook of Land and Property Taxation* (pp. 117-128). Cheltenham, UK: Edward Elgar.

Kelly, R. (2003). *Property Taxation in Indonesia: Challenges from Decentralization*. Working Paper. Lincoln Institute of Land Policy. Massachusetts, USA.

Kementrian Dalam Negeri Republik Indonesia [Ministry of Home Affairs of Indonesia]. (2014). *Pembentukan Daerah-Daerah Otonom Di Indonesia Sampai Dengan 2014 [Establishment of Autonomous Regions in Indonesia until 2014]*. Jakarta, Indonesia.

Kestens, Y., The´ riault, M., & Des Rosiers, F. (2006). Heterogeneity in Hedonic Modelling of House Prices: Looking at Buyers' Household Profiles. *Journal of Geographical Systems, 8*(1), 61-96.

Kilpatrick, J. (2011). Expert Systems and Mass Appraisal. *Journal of Property Investment & Finance, 29*(4/5), 529-550.

Kim, B., & Kim, T. (2016). A Study on Estimation of Land Value Using Spatial Statistics: Focusing on Real Transaction Land Prices in Korea. *Sustainability, 8(3), 203*

Kryvobokov, M. (2004). Urban Land Zoning for Taxation Purposes in Ukraine. *Property Management, 22*(3), 214–229.

Land Office of Bekasi. (2012a). *Peta Pendaftaran Tanah Kota Bekasi* [Digital Cadastral Map].

Land Office of Bekasi. (2012b). *Peta Zona Nilai Tanah* [Digital Map of Land Value Zones].

Land Office of Bekasi. (2012c). *Survey Penilaian Tanah Kota Bekasi Tahun 2012* [Property Sales Data Compilation].

Land Office of Jakarta Selatan. (2012). *Survey Penilaian Tanah Kota Jakarta Selatan Tahun 2012* [Property Sales Data Compilation].

Lewis, B. D. (2003). Property Tax in Indonesia: Measuring and Explaining Administrative (Under-) Performance. *Public Administration And Development, 23*, 227-239.

Li, F., & Klette, R. (2011). *Euclidean Shortest Paths*. London, UK: Springer-Verlag.

Liao, S.-H. (2005). Expert System Methodologies and Applications—A Decade Review from 1995 to 2004. *Expert Systems with Applications, 28*, 93–103.

Lin, C. C., & Mohan, S. B. (2011). Effectiveness Comparison of The Residential Property Mass Appraisal Methodologies in The USA. *Journal of Housing Markets and Analysis, 4*(3), 224-243.

Liu, H., Zhou, J., Feng, Q., Li, Y., Li, Y., & Wu, J. (2017). Effects of Land Use and Topography on Spatial Variety of Soil Organic Carbon Density in a Hilly, Subtropical Catchment of China. *Soil Research, 55*(2), 134-144.

Man, K. F., Tang, K. S., & Kwong, S. (1996). Genetic Algorithms: Concepts and Applications. *Institute of Electrical and Electronics Engineers (IEEE): Transactions on Industrial Electronics, 43*(5), 519-534.

Martı´nez-Cuevas, S., Benito, M. B., Cervera, J., Morillo, M. C., & Luna, M. (2017). Urban Modifiers of Seismic Vulnerability Aimed at Urban Zoning Regulations. *Bulletin of Earthquake Engineering, 15*(11), 4719-4750.

McCluskey, W., & Anand, S. (1999). The Application of Intelligent Hybrid Techniques for the Mass Appraisal of Residential Properties. *Journal of Property Investment and Finance, 17*(3), 218-238.

McCluskey, W., Davis, P., Haran, M., McCord, M., & McIlhatton, D. (2012a). The Potential of Artificial Neural Networks in Mass Appraisal: The Case Revisited. *Journal of Financial Management of Property and Construction, 17*(3), 274-292.

McCluskey, W., Davis, P., McCord, M., McIlhatton, D., & Haran, M. (2012b). Computer Assisted Mass Appraisal and the Property Tax. In W. McCluskey, G. C. Cornia & L. C. Walters (Eds.), *A Primer on Property Tax: Administration and Policy* (pp. 307-338). West Sussex, UK: Wiley and Sons.

McCluskey, W., Deddis, W., Mannis, A., McBurney, D., & Borst, R. (1997). Interactive Application of Computer Assisted Mass Appraisal and Geographic Information Systems. *Journal of Property Valuation and Investment, 15*(5), 448-465.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction Accuracy in Mass Appraisal: A Comparison of Modern Approaches. *Journal of Property Research, 30*(4), 239-265.

Miller, W. (2013). *Statistics and Measurement Concepts with OpenStat*. New York: Springer

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction Error Estimation: A Comparison of Resampling Methods. *Bioinformatics, 21*(15), 3301-3307.

Moore, W. (2009). A History of Appraisal Theory and Practice: Looking Back from IAAO's 75th Year. *Journal of Property Tax Assessment and Administration, 6*(3), 23-49.

Morillo, M. C., Cepeda, F. G., & Martínez-Cuevas, S. (2017). The Application of Spatial Analysis to Cadastral Zoning of Urban Areas: An Example in The City of Madrid. *Survey Review, 49*(353), 83-92.

Nguyen, N., & Cripps, A. (2001). Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal Of Real Estate Research, 22*(3), 313-336.

O'Roarty, B., McGreal, S., Adair, A., & Patterson, D. (1997). Case-Based Reasoning and Retail Rent Determination. *Journal of Property Research, 14*(4), 309–328.

Oliveira, C. A. S., & Pardalos, P. M. (2011). *Mathematical Aspects of Network Routing Optimization*. New York, USA: Springer Science+Business Media.

OpenStreetMap. (2015). OpenStreetMap. Retrieved 12 February 2015, from OpenStreetMap.org https://www.openstreetmap.org/#map=13/-6.2393/107.0092

Pace, R. K. (1998). Appraisal Using Generalized Additive Models. *Journal Of Real Estate Research, 15*(1/2), 77-99.

Pace, R. K., Sirmans, C. F., & Slawson, V. C. (2002). Automated Valuation Models. In K. Wang & M. L. Wolverton (Eds.), *Real Estate Valuation Theory* (pp. 133-156). New York: Springer.

Páez, A., Farber, S., & Wheeler, D. (2011). A Simulation-Based Study of Geographically Weighted Regression as A Method For Investigating Spatially Varying Relationships. *Environment and Planning, 43*, 2992 - 3010.

Páez, A., Long, F., & Farber, S. (2008). Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques. *Urban Studies, 45*(8), 1565-1581.

Páez, A., & Wheeler, D. C. (2009). Geographically Weighted Regression *International Encyclopedia of Human Geography* (pp. 407-414). Heidelberg, Germany: Elsevier Ltd.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real Estate Appraisal: A Review of Valuation Methods. *Journal of Property Investment and Finance, 21*(4), 383-401.

Pawlak, Z. (1997). Rough Set Approach to Knowledge-Based Decision Support. *European Journal of Operational Research, 99*, 48-57.

Peterson, S., & Flanagan, A. B. (2009). Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. *Journal Of Real Estate Research, 31*(2), 147-164.

Portnov, B. A., McCluskey, W. J., & Deddis, W. G. (2001). Property Taxation in Israel: A Non Ad Valorem Approach. *Land Use Policy, 18*, 351-364.

Prasad, D. (2015). Shortest Path using Dijkstra's Algorithm. Retrieved June 16, 2016, from http://techieme.in/shortest-path-using-dijkstras-algorithm

Republic of Indonesia. (2012). *Undand-Undang Nomor 2 tahun 2012 Tentang Pengadaan Tanah Bagi Pembangunan Untuk Kepentingan Umum [Law Number 2 of 2012 on Land Acquisition for Infrastructure Development in the Public Interests]*. Jakarta, Indonesia.

Şen, Z. (2009). *Spatial Modeling Principles in Earth Sciences*. New York: Springer

Shimizu, C., & Nishimura, K. G. (2007). Pricing Structure in Tokyo Metropolitan Land Markets and its Structural Changes: Pre-bubble, Bubble, and Post-bubble Periods. *Journal of Real Estate Finance and Economics, 35*(4), 475-496.

Sun, D. J., Zhang, C., Zhang, L., Chen, F., & Peng, Z.-R. (2014). Urban Travel Behavior Analyses and Route Prediction Based on Floating Car Data. *The International Journal of Transportation Research, 6*(3), 118-125.

Tamtomo, J. P., Erestajaya, V., Färnkvist, O., & Roos, H. (2008). *Land Valuation Survey in Indonesia*. Paper presented at the International Federation of Surveyors (FIG) Working Week 2008: Intergrating Generations, Stockholm, Sweden.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography, 46*, 234-240.

Wang, K., Zhang, C., & Li, W. (2012). Comparison of Geographically Weighted Regression and Regression Kriging for Estimating the Spatial Distribution. *GIScience & Remote Sensing, 49*(6), 915-932.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. *Neural Computation, 1*, 425- 464.

Yacim, J. A., Boshoff, D. G. B., & Khan, A. (2016). Hybridizing Cuckoo Search with Levenberg-Marquardt Algorithms in Optimization and Training of ANNs for Mass Appraisal of Properties. *Journal Of Real Estate Literature, 24*(2), 473-492.

Yiannakoulias, N., Bland, W., & Svenson, L. W. (2013). Estimating The Effect Of Turn Penalties and Traffic Congestion on Measuring Spatial Accessibility to Primary Health Care. *Applied Geography, 39*, 172-182.

Zhang, J., Zhu, Y., Krisp, J., & Meng, L. (2010). *Derivation of Road Network from Land Parcels*. Paper presented at the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, California, USA.

11. APPENDICES

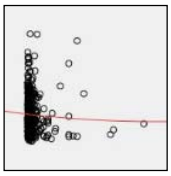
Appendix A: Correlation coefficients between variables

	parcel size sqm	value per sqm	road width	road class nr	zoning nr	tollgate	primary artery	Secondary artery	primary collector	business centre	Market-place	hospital	school
parcel size sqm	1.000	-.115**	-.037	.102**	-.299**	-.089*	.061	-.035	-.018	.038	.049	.014	.139**
value per sqm	-.115**	1.000	.710**	.393**	.663**	-.268**	-.194**	-.003	.024	-.176**	.020	.122**	-.096*
road width	-.037	.710**	1.000	.610**	.651**	-.069	-.198**	.038	-.099**	-.184**	-.044	.030	-.128**
road class nr	.102**	.393**	.610**	1.000	.344**	-.094*	-.209**	.090*	-.222**	-.116**	-.125**	.046	-.026
zoning nr	-.299**	.663**	.651**	.344**	1.000	-.041	-.161**	.047	-.089*	-.143**	-.071	.041	-.178**
tollgate	-.089*	-.268**	-.069	-.094*	-.041	1.000	.324**	.065	.008	.342**	-.078*	-.448**	-.009
primary artery	.061	-.194**	-.198**	-.209**	-.161**	.324**	1.000	.082*	.082*	.617**	.378**	.055	.146**
secondary artery	-.035	-.003	.038	.090*	.047	.065	.082*	1.000	-.272**	.392**	.309**	.595**	.201**
primary collector	-.018	.024	-.099**	-.222**	-.089*	.008	.082*	-.272**	1.000	.028	.179**	-.078*	.131**
business centre	.038	-.176**	-.184**	-.116**	-.143**	.342**	.617**	.392**	.028	1.000	.119**	.238**	.288*
market place	.049	.020	-.044	-.125**	-.071	-.078*	.378**	.309**	.179**	.119**	1.000	.209**	.029
hospital	.014	.122**	.030	.046	.041	-.448**	.055	.595**	-.078*	.238**	.209**	1.000	.257**
school	.139**	-.096*	-.128**	-.026	-.178**	-.009	.146**	.201**	.131**	.288**	.029	.257**	1.000

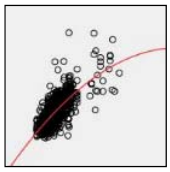
**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

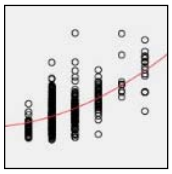
Appendix B: Lines fitted on the scatter plots between price and each explanatory variable



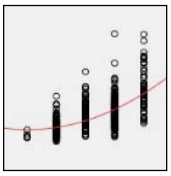
Parcel size



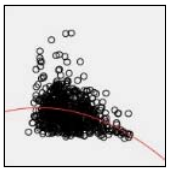
Road width



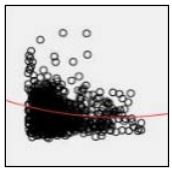
Road class



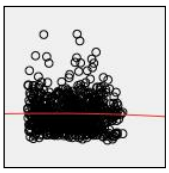
Zoning



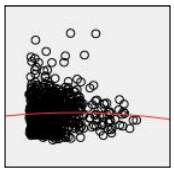
Tollgate



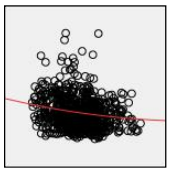
Primary artery



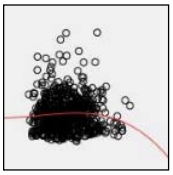
Secondary artery



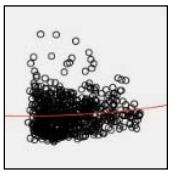
Primary collector



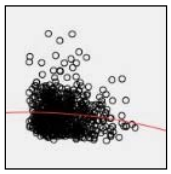
Business centre



Marketplace



Hospital



School

Appendix C: Survey Form

Formulir SPT. 111

FOMULIR PENDATAAN
OBYEK PENILAIAN, PEMBANDING, *SAMPLE*
PENILAIAN TANAH NON PERTANIAN
(untuk zonasi nilai tanah)

1. Nomor Identifikasi : _____
2. Nomor Surat Tugas/ Tanggal : _____
3. Nama Surveyor/Tanggal pelaksanaan : _____ /

A. Data Administrasi/Harga Tanah

4. Alamat : ..... Kode Pos : .....
   ..... Koordinat ☐ (X) : ..... ☐ Lintang : .....
   (Y) : ..... Bujur : .....

5. Status Kepemilikan : ☐ HM ☐ HGB ☐ HP ☐ TMA

6. Jenis data :
   ☐ Transaksi ☐ Penawaran Tanggal transaksi /penawaran .....

7. Harga : ☐ Jual beli ☐ Sewa/kontrak Rp.....
   (.....)

8. Responden : ☐ Pemilik tanah ☐ Real estate/broker ☐ Developer ☐ Notaris/PPAT
   ☐ Penyewa ☐ Lurah/Kepala Desa
   Nama :
   .....
   Alamat /No telp. :
   .....

9. Keterangan :
   .....
   .....
   .....
   .....

B. Data Fisik Tanah

10. Luas tanah = ...m^2

11. Lebar depan = ...m, panjang kebelakang = ...m

12. Bentuk tanah: ☐ Persegi/Normal ☐ Tidak beraturan ☐ Lain-lain .....

13. Elevasi dari jalan : ☐ Lebih Tinggi ☐ Sama ☐ Lebih Rendah

14. Letak Tanah: ☐ Normal ☐ Tusuk Sate ☐ Hadap Taman ☐ Huk ☐ Lain-lain ....

15. Keterangan:
.....
.....
.....
.....

C. Data Lingkungan

16. Kelas Jalan : ☐ Arteri ☐ Kolektor ☐ Lokal ☐ Setapak
Lebar : ..... meter

17. Aksesibilitas : ☐ Sangat Baik ☐ Baik ☐ Cukup ☐ Kurang

18. Drainase : ☐ Sangat Baik ☐ Baik ☐ Cukup ☐ Kurang

19. Utilitas : ☐ Listrik ☐ Air Bersih ☐ Telepon
☐ Gas ☐ TV kabel ☐ Lain-lain .....

20. Fasilitas : ☐ Sekolah ☐ Tempat Ibadah ☐ Rumah Sakit ☐ Pasar ☐ Lain-lain .....

21. Zoning/Peruntukan Kawasan : ☐ Permukiman ☐ Komersial ☐ Industri ☐ Lain-lain .....

22. Keterangan :
.....
.....
.....
.....