COMPARISON OF SPATIAL HEDONIC HOUSE PRICE MODELS: APPLICATION TO REAL ESTATE TRANSACTIONS IN VANCOUVER WEST

By

Wai Man Chan
BSc Statistics, University of British Columbia, 2007

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN FINANCE

In the Master of Science in Finance Program
of the
Faculty
of
Business Administration

SIMON FRASER UNIVERSITY

Summer 2014

Approval

Name: **Wai Man (Raymond) Chan**

Degree: **Master of Science in Finance**

Title of Project: **COMPARISON OF SPATIAL HEDONIC HOUSE PRICE MODELS: APPLICATION TO REAL ESTATE TRANSACTIONS IN VANCOUVER WEST**

Supervisory Committee:

____

Dr. Andrey Pavlov
Senior Supervisor
Professor of Finance

____

Steven Adang
Second Reader
Lecturer

Date Approved: ______________________________________

Abstract

This study compares hedonic house price models for single family properties in Vancouver West, Canada. The real estate literature has shown that traditional hedonic models based on OLS are unable to handle spatial effects inherent in housing markets, prompting the application of spatial econometric methods. This study compares four hedonic house price models: (i) classical OLS model, (ii) OLS model with neighborhood code dummies, (iii) Spatial Durbin Model, and (iv) Geographically Weighted Regression. The latter two models are common spatial econometric techniques that researchers have used. Models are compared based on model R^2 , out-of-sample prediction error, and ability to remove spatial effects from the data. Results indicate that Geographically Weighted Regression is the best performing model. In addition, classical OLS overestimates effects and is unable to address spatial effects. All four models predict a similar impact of property attributes on sale price.

Keywords: Hedonic Model; Spatial Hedonic Model; Spatial Durbin Model; Geographically Weighted Regression; Vancouver West Real Estate

Acknowledgements

I would like to thank my classmates Zarko Kolaric and Lei Yang for their months of hard work in compiling the dataset. Without it, this study would not have materialized. I would also like to thank David Ho and Adam Xu from CBRE Canada for allowing me to use the dataset. Finally, I would like to thank my thesis supervisors Andrey Pavlov and Steven Adang.

Table of Contents

List of Figures

List of Tables

1: Introduction

The real estate market is a very important one in all economies. In addition to providing housing for households, real estate activity is linked to many sectors of the economy, primarily construction, finance, and insurance. For 2012, Statistics Canada reports that the "Real Estate and Rental and Leasing" sector accounted for about 12% of the Canadian economy (GDP), and the construction sector accounted for about 7% of the Canadian economy[1].

Concerns over rapid appreciation in prices, risk from residential mortgages, and record levels of household debt have prompted regulators of the Canadian financial system to pay close attention to the Canadian real estate market. For example, in July of 2013, the Finance Minister reduced the maximum amortization period from 30 years to 25 years for insured mortgages. It marked the fourth time the Finance Minister restricted mortgage lending rules in as many years. The Office of the Superintendent of Financial Institutions (OSFI), Canada's prudential regulator of financial institutions and private pension plans, issued a draft guideline on residential mortgage underwriting practices and procedures in June of 2012[2].

Providing updated fair market value of residential properties is therefore extremely valuable to financial regulators, financial institutions, municipal assessors, housing index compilers, market participants, real estate developers, investors, and many others. Unlike traditional financial assets like stocks and bonds, estimating the fair value of a residential property is non-trivial. A private firm could conduct an appraisal to estimate the fair value of an individual property, but doing so for all properties in a region would be prohibitively costly.

In the real estate literature, hedonic modelling is the most widely used method to estimate the fair market value of real estate properties. The hedonic model from Rosen (1974) postulates that products are sold as a package encompassing its attributes. Each attribute has a price, called the implicit price. The implicit prices of each attribute are revealed from observed prices of differentiated products and the specific amount of each attribute associated with the product. This framework is suitable for real estate properties, as each property is sold as a package with non-separable attributes (square feet, lot size, number of bedrooms, etc.) and the price of each attribute is not known precisely.

[1] https://www.ic.gc.ca/app/scr/sbms/sbb/cis/gdp.html?code=11-91&lang=eng

[2] http://www.osfi-bsif.gc.ca/Eng/fi-if/rg-ro/gdn-ort/gl-ld/Pages/b20.aspx

The traditional hedonic model uses linear regression (OLS) to estimate a relationship between the value of a property and a set of housing characteristics. This relationship can then be used to predict the market value of unsold properties with known characteristics. However, researchers have discovered many drawbacks with the traditional hedonic model, among them the lack of treatment for spatial effects (Dubin, 1992; 1998).

This paper applies recent spatial econometrics techniques towards constructing hedonic pricing models for single family homes in Vancouver West. Using the OLS model as the benchmark, four models are considered; classical OLS, OLS with neighborhood code dummies, Spatial Durbin Model, and Geographically Weighted Regression. The models are estimated with approximately 5,000 sale records collected through the multiple listing service (MLS). The models are compared with respect to model R^2, out-of-sample prediction error, and ability to remove spatial autocorrelation. For this paper, Moran's I is used to assess the degree of spatial autocorrelation remaining in model residuals.

This paper is organized into 6 sections. Section 2 provides a literature review of hedonic models and spatial effects. The dataset with descriptive statistics are discussed in section 3. Section 4 details the spatial econometric techniques used in this paper. Results from the four models are given in Section 5. Section 6 concludes the paper.

2: Literature Review

2.1 Hedonic Pricing Method

Lancaster's (1966) work on consumer theory laid the theoretical basis for hedonic price modelling (Rosen, 1974). Lancaster argued that consumers derive utility from the characteristics of the product, rather than the product itself. However, the market with respect to the product's characteristics is often not explicit and is hidden in the background of product price determination.

The hedonic price model from Rosen asserts that products are sold as a package encompassing its attributes. This is certainly the case for houses, since the housing attributes associated with a home are non-separable. If the marginal or implicit price of each housing attributes can be estimated, then the price of a house would equal to the summation of all its marginal or implicit prices.

Despite its conceptual appeal and simplicity, Pavlov (2000) points out two substantial drawbacks with the hedonic pricing method. The first is the misspecification of the functional form. Rosen's work provides little guidance with respect to the actual functional form relating the price of the product and its attributes. Can and Megbolugbe (1997) finds that an incorrect functional form leads to unreliable and biased estimates. The second drawback is the sensitivity to omitted variables. For a product as complex as a house, it is inevitable that researchers cannot identify and measure all price-determining attributes.

The hedonic pricing method is typically implemented with the classical linear regression model, and estimated using ordinary least squares (OLS). The dependent variable in the model is the sale price of a property, and the independent variables in the model are the attributes of that property. In a broad literature review, Malpezzi (2002) identifies the following types of attributes: structural (living area, lot size, age, number of rooms, etc.), locational (absolute location of the dwelling, proximity to central business district, etc.), neighborhood (availability of public schools, income levels, population density, etc.), contract conditions (appliances/furniture included in sale, time to possession, etc.) and time specific attributes. The study also finds that most researchers use a semi-log or log-log specification. The advantage of these specifications compared to linear form is that implicit prices vary with the quantity of housing attributes and mitigates the problem of heterosecedasticity.

There has been a vast number of published work in the real estate literature that makes use of hedonic regression analysis to explain house prices with housing characteristics. In addition to the Malpezzi (2002) literature review above, Sirmans et al. (2005) examined hedonic pricing models for over 125 empirical studies. The study concludes that both the magnitude and direction of certain characteristics (such as number of rooms) differ across studies.

2.2 Spatial Effects

The hedonic pricing method implemented though OLS accounts for spatial effects through locational attributes (absolute location of the dwelling, proximity to central business district, etc.) and neighborhood attributes (availability of public schools, income levels, population density, etc.). However, it does not account for spatial interaction effects (or “spillover” effects) between properties. In the real estate literature, the spatial interactions is called the “adjacency” effect (Can, 1992). Moreover, OLS assumes a constant relationship between the dependent variable and the independent variables. That is, the relationship is invariant to space and time, which is unlikely the case for the housing market.

Bitter and Krause (2012) identified the increased use of advanced spatial methods in published studies as one of the leading trends in real estate valuation research. The authors state that spatial dependence, spatial heterogeneity, anisotropic phenomena and boundary effects make obsolete the basic monocentric urban economic model, which yields the simple non-linear decline of house values from the central business district. In place of the monocentric urban economic model, cities are characterized by, quoting the authors, “polycentric urban regions complete with localized amenities (or disamenities), geographic heterogeneities, fragmented municipal governments, and complex systems of land use regulations.”

From Anselin (1998), there are two major types of spatial effects: spatial autocorrelation[3] and spatial heterogeneity. Spatial autocorrelation refers to a functional relationship between observations. Spatial heterogeneity, on the other hand, refers to the lack of uniformity arising from space, leading to spatial heterosecedasticity and spatially varying parameters.

2.2.1 Spatial Autocorrelation

The classical OLS method for hedonic modelling relies on several assumptions. One of the assumption is that the error terms are uncorrelated. Moreover, OLS assumes the price of a

[3] In the real estate literature, spatial autocorrelation is also referred to as spatial dependence.

property is related to only its characteristics, but not the characteristics of other properties. But the real estate literature have put forth strong evidence that spillover effects between properties exist, casting doubt on the above two assumptions. In this case, OLS estimation would be inefficient and possibly biased.

Section 2.2 of Elhorst (2014) describes three types of interaction effects that explain the dependency among observations. The first are endogenous interaction effects, where the dependent variable of observation i is related to the dependent variable of another observation j. In the real estate literature, this interaction is referred to as the adjacency effect (Can, 1992). One explanation for the adjacency effect is that buyers consult listing prices of nearby properties prior to making an offer. Similarly, sellers and listing agents use listing prices in the neighborhood to determine listing prices. Therefore, it is reasonable to expect that this effect is present for the housing market.

The second effect is exogenous interaction effect, where the dependent variable of observation i is related to the independent variables of another observation j. At first glance, it seems unlikely that the price of a property depends on the attributes of nearby houses. But Braisington et al. (2005) suggests that houses with characteristics atypical within a block (for example, having the largest or smallest living area) would result in a discounted sale price.

The third effect relates to the dependency among the disturbance terms. That is, ε_i is positively (or negatively) related to ε_j for distinct observations i and j. A few examples that lead to this are minor misspecifications, incorrect spatial delineations of geographical variables, and omitted neighborhood characteristics (Osland, 2010). For instance, suppose house i is located near an airport, and this information is not captured in the hedonic model. Then we would expect the model to overestimate its market price, because buyers demand a discount to compensate for increased noise level. In terms of the hedonic model, this translates to a large negative value for ε_i. Similarly, the model would also overestimate market prices of houses in the proximity of house i, leading to spatial dependency among the disturbance terms.

2.2.2 Spatial Heterogeneity

In the OLS hedonic pricing model, the regression parameters represent the implicit prices of housing attributes. They are assumed constant for all observations. An implication is that the implicit prices are assumed constant through space. Spatial heterogeneity refers to the case where this assumption is invalid. In the real estate literature, there has been strong evidence to support spatial heterogeneity.

For example, Pavlov (2000) suggests that omitted variables influence not only the intercept of the model, but also the implicit prices of attributes. Consider some omitted variables that are related to construction quality. In the OLS model without the omitted variables, an additional square feet of marble floors (high construction quality) would be valued equally as an additional square feet of carpeted floor (lower construction quality). It appears reasonable that the implicit price of an additional square feet (and other physical characteristics) should be related to construction quality. If we further conjecture that construction quality is spatially correlated (construction quality of a house is similar to that of its neighbors), then the implicit price for an additional square feet exhibits spatial heterogeneity.

Localized supply and demand imbalances within a large metropolitan real estate market also lead to spatial heterogeneity (Michaels et al. 1990). With respect to supply, it is often the case that housing characteristics exhibit a high degree of spatial correlation; homes near the central business district are typically older, smaller, and lack new features like multiple garages and air conditioning. On the other hand, suburban homes are generally newer, larger, and include newer features.

If there is a shift in household preference to a certain housing attribute (eg. air conditioning), then competition for those attributes in an area where houses lack that attribute should result in higher implicit prices, compared to an area where houses with that attribute are plentiful.

2.3 Spatial Econometric Models

Analogous to the approach in Farber and Yeates (2006), this paper classifies spatial econometric models into two types: ‘Global’ regression models and ‘Local’ regression models. The global regression models considered in this paper are the standard hedonic house price model (OLS) and linear spatial dependence models. The local regression model used in this paper is geographically weighted regression (GWR).

2.3.1 Linear Spatial Dependence Models

Section 2.2 of Elhorst (2014) details the taxonomy of linear spatial dependence models commonly applied in empirical studies. The three types of interaction effects considered in these models are (i) endogenous interaction effects, where the dependent variable of unit i interacts with the dependent variable of another unit j, (ii) exogenous interaction effects, where the dependent variable of unit i depends on independent explanatory variables of another unit j, and (iii) interactions among the error terms, where the error term of unit i interacts with the error term of another unit j.

A full model with all three types of interaction effects is expressed as:

$$\begin{aligned} Y &= \delta WY + \alpha 1_N + X\beta + WX\theta + u \\ u &= \lambda Wu + \varepsilon \end{aligned} \tag{2.1}$$

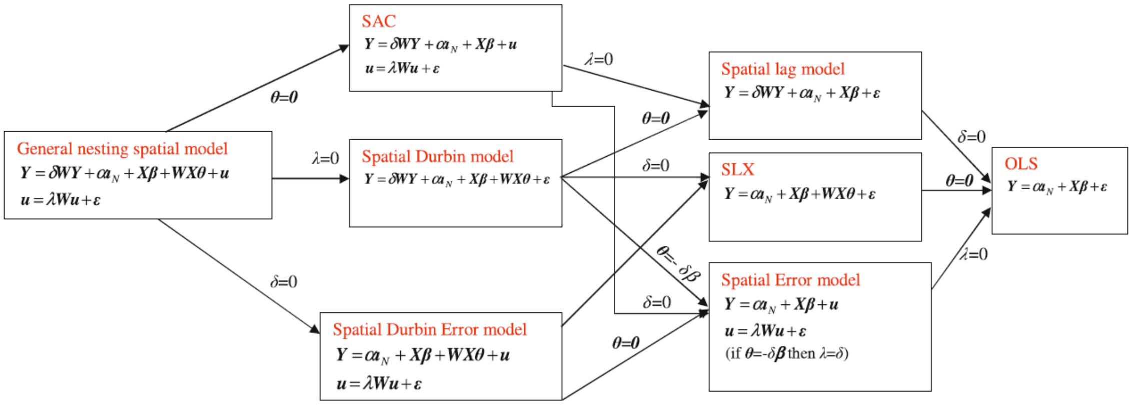
where

- Y is an (Nx1) vector of the dependent variable
- δ is the spatial autoregressive coefficient
- W is an (NxN) spatial weight matrix
- WY is the endogenous interaction effects among the dependent variable
- α is the constant term parameter, 1_N is an (Nx1) vector of one's
- X is an (NxK) matrix of explanatory variables
- β is a (Kx1) vector of fixed but unknown coefficients
- WX is the exogenous interaction effects
- θ is a (Kx1) vector of fixed but unknown coefficients
- λ is the spatial autocorrelation coefficient
- Wu is the interaction effects among the disturbance terms
- ε is an (Nx1) vector of uncorrelated errors with zero mean and constant variance

By placing restrictions on δ, θ and λ, a family of spatial models is obtained. For example, the classical OLS model is achieved by setting the above parameters to zero. Figure 2.1 shows the family of spatial dependence models obtained with parameter restrictions.

Care must be exercised in choosing among the available spatial models. LeSage (2014) states that practitioners of spatial regression models should first determine whether the phenomena being modelled are likely to produce local or global spatial spillovers. After this determination, only two models need to be considered; the spatial Durbin error model (for local spillovers) or the spatial Durbin model (for global spillovers).

Figure 2.1 The relationship between different spatial dependence models for cross-section data. Source: Vega and Elhorst (2012)


SAC
$Y = \delta WY + \alpha\iota_N + X\beta + u$
$u = \lambda Wu + \varepsilon$
General nesting spatial model
$Y = \delta WY + \alpha\iota_N + X\beta + WX\theta + u$
$u = \lambda Wu + \varepsilon$
Spatial Durbin model
$Y = \delta WY + \alpha\iota_N + X\beta + WX\theta + \varepsilon$
Spatial Durbin Error model
$Y = \alpha\iota_N + X\beta + WX\theta + u$
$u = \lambda Wu + \varepsilon$
Spatial lag model
$Y = \delta WY + \alpha\iota_N + X\beta + \varepsilon$
SLX
$Y = \alpha\iota_N + X\beta + WX\theta + \varepsilon$
Spatial Error model
$Y = \alpha\iota_N + X\beta + u$
$u = \lambda Wu + \varepsilon$
(if θ=-$\delta\beta$ then λ=δ)
OLS
$Y = \alpha\iota_N + X\beta + \varepsilon$
θ=0
λ=0
δ=0
λ=0
θ=0
δ=0
θ=- $\delta\beta$
δ=0
θ=0
δ=0
θ=0
λ=0

A spatial spillover is present when a causal relationship exists between an independent variable of a unit i and the dependent variable of another unit j. A mathematical definition is $\partial y_j / \partial X_i^r \neq 0$, implying a spillover from the r^{th} characteristic of unit i on to the dependent variable of unit j. If the non-zero cross-partial derivative implies an impact on neighboring units that do not instigate endogenous feedback effects, then the spillover is referred to as a local spillover.

Global spillover, on the other hand, refers to spillover effects where the non-zero cross-partial derivative implies an impact on neighboring units, plus neighbors to the neighboring units, and so on. This chain of impacts results in endogenous interaction and feedback effects. Endogenous interaction is when changes in one unit triggers a sequence of changes in potentially all other units, leading to a new long-run steady state equilibrium.

In a study of the economic impact of sports facilities on residential property values in Columbus, Ohio, Feng and Humphreys (2008) suggest that shared neighborhood amenities lead to neighboring spillover effects among properties. Therefore, the price of each property affects all other properties in the neighborhood, implying a global range of spillovers. The study modelled the housing data with the spatial lag model, which is a special case of the spatial Durbin model with θ restricted to zero.

2.3.2 Geographically Weighed Regression (GWR)

GWR is implemented through a series of local linear regression for each unit in the sample (Fotheringham et al., 1998). Mathematically, GWR is expressed as:

$$y_i = X_i \beta_i + \varepsilon_i \qquad (2.2)$$

where the subscript i indicates that parameter estimates are specific to unit i. In the context of hedonic house price modelling, this means the marginal prices of housing attributes varies across space in a continuous fashion. The weighted least squares method is used to estimate β_i:

$$\beta_i = (X^T W_i X)^{-1} X^T W_i y \qquad (2.3)$$

where W_i is an NxN diagonal matrix with diagonal entries reflecting the weighting of each unit with respect to unit i .

The most commonly used kernel to compute the weights is the Gaussian distance decay function, which specifies $W_i(j, j) = \exp(-\frac{d_{ij}^2}{h^2})$, where d_{ij} is the distance between unit i and unit j, and $h > 0$ is the bandwidth parameter. The bandwidth parameter controls the rate at which the weighting function declines with distance. The bi-square function and the tri-cube kernel function are also weighting functions that are commonly used.

The weighting functions mentioned above have a fixed bandwidth. A potential problem arises for units that are located in a sparsely populated area. For these units, β_i is estimated based on data from very few neighbours ($W_i(j, j)$ would be very small for most units j), resulting in large estimation variance. To remedy this, a specific bandwidth parameter h for each observation is used to create adaptive weighting functions. A small bandwidth is used for units located in densely populated area, while a larger bandwidth is used for units located in a sparsely populated area.

2.4 Spatial Hedonic Models Applied to Canadian Housing Markets

This section briefly introduces five studies that have applied spatial methods to the Canadian housing market. The studies are Leblond (2004), Boxall et al. (2005), Farber and Yeates (2006), Kestens et al. (2006), and Huang et al. (2010).

In the Leblond (2004) study, four hedonic models are used for automated mass valuation for the housing market of Montreal. The four models are (i) a simple hedonic model with no spatial effects, (ii) a model with spatially lagged independent variable, (iii) a model with spatially lagged dependent variable, and (iv) a model with spatially lagged errors. Multiple listing service (MLS) transaction data for about 5,000 sales from single family homes during the period January 1999 to September 2003 are used in the estimation. Sales from the fourth quarter of 2003 are withheld for the purpose of out-of-sample prediction. With respect to root mean squared prediction error, the model with spatially lagged independent variable performed best.

Boxall et al. (2005) investigated the impact of oil and gas facilities on rural residential property values in 36 townships (6-mile by 6-mile block) near Calgary. MLS data on 532 sales from residential properties during the period January 2004 to March 2001 were analyzed. The sample was restricted to properties that ranged in size from 1 to 40 acres and priced from $150,000 to $450,000[4]. The spatial lag model and the spatial error model were considered. When variables related to oil and gas facilities were included, LM tests and robust LM tests supported the spatial error model over the spatial lag model. The results of the study concluded that the presence of oil and gas facilities have significant negative impacts on the values of nearby rural residential properties.

The Farber and Yeates (2006) study examined the performance of four hedonic price models for the city of Toronto. The study classifies the four models into two ‘global’ models and two ‘local’ models. The two global models are (i) standard hedonic model and (ii) hedonic model with spatially lagged sales price. The two local models are (i) geographically weighted regression (GWR) and (ii) moving window regression. The data set used consists of 19,007 freehold housing sales between July 2000 and June 2001 in the City of Toronto. The study found that GWR achieved the highest coefficient of determination R^2 at 91.9%[5]. In addition, the model residuals from GWR had the lowest level of spatial autocorrelation.

Kestens et al. (2006) assessed the hypothesis that variability of implicit prices of certain housing attributes is linked to individual preferences. The study used two spatial hedonic pricing models: the Casetti expansion method and GWR. The dataset is based on a survey of 761 households that acquired property in Quebec City between 1993 and 2001. The study concluded

[4] The restriction in size ensures the property was rural but with no commercial agriculture value. The restriction in price mitigates the impact of abnormally low or high priced properties.

[5] For the two local models, R^2 is approximated by the pseudo-R^2, defined as the squared correlation coefficient between the observed and predicted values.

that some characteristics of the buyer’s household (household income, previous tenure status, and age) have a direct impact on transaction prices.

Huang et al. (2010) applied local regression models to residential housing sales from 2002 to 2004 in Calgary. The study developed the geographically and temporally weighted regression (GTWR) model, an extension of the GWR model to include time effects. When compared to a global OLS model, the study found that GWR reduced absolute prediction error by 31.5%, and GTWR reduced absolute prediction error by 46.4%. Also, model R^2 increased from 76.31% in global OLS model to 88.97% in GWR and 92.82% in GTWR.

3: Data

In this paper, the dataset consists of transactions data pertaining to single family homes in Vancouver West that occurred in 2011, 2012, or 2013. In addition to transaction price and housing characteristics variables used in traditional hedonic price modelling, the dataset includes the latitude and longitude of each property. The latitude and longitude information allows the calculation of distances between properties, which is an essential step for the two spatial models considered in this study. In summary, the following is a list of variables used in this paper:

- Most recent sale date and sale price
- Living area of property
- Lot size
- Age of property at sale date
- Number of bedrooms and bathrooms[6]
- Longitude and latitude of the property
- Neighborhood code of the property

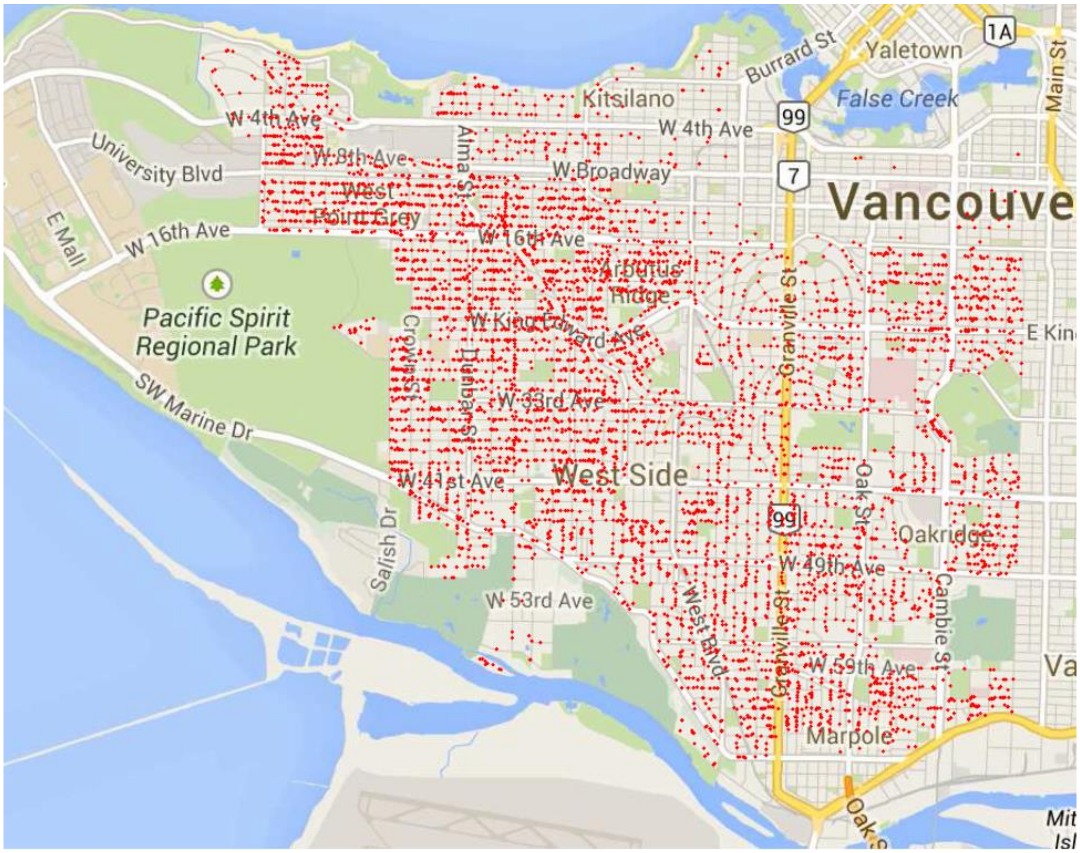
Initially, the dataset contained 6,045 properties. Records with missing data in any of the fields are removed. To ensure the analysis is restricted to arms-length sales and eliminate possible data entry errors, records with sale price less than 50% of the city assessment (2013) are removed. Finally, several sales in the downtown west end are deleted to facilitate spatial analysis. In summary, 767 records are removed[7] and 5,278 properties remain in the data set.

Figure 3.1 displays the spatial distribution of the sale transactions used in this study. It can be seen that there is a widespread distribution. However, it is also apparent that some areas, in particular the north east region of the map, have very few or no transactions. This is attributed to those areas having predominately condo, townhouse, or commercial developments. Another point to note is that highway 99 depicted on the map is actually a major street (Granville Street) but not a highway.

[6] Half bathrooms are included as well as 0.5 bathrooms.

[7] Of the 767 deletion, 341 are attributed to the sale price less than 50% of city assessment.

Figure 3.1 Property Sale Locations[8]



3.1 Treatment of Time Effects

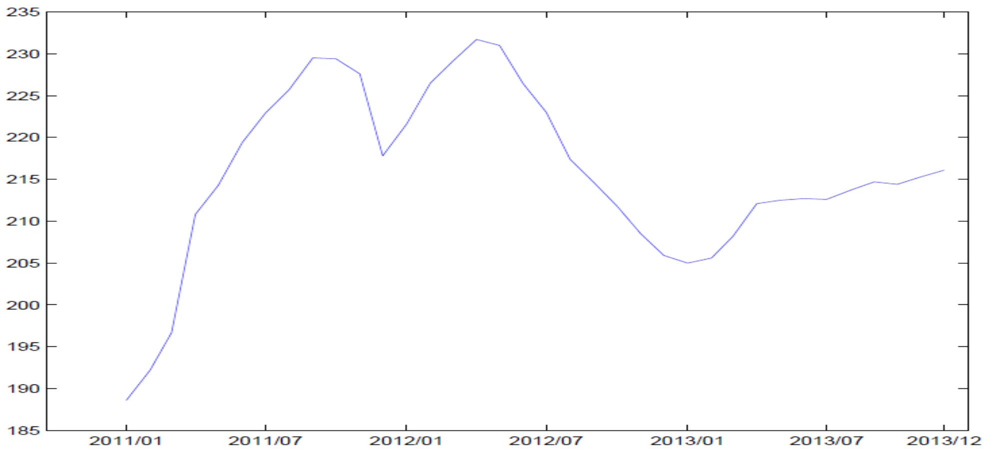
Of the 5,278 properties in the dataset, 1,993 (37.8%) transacted in 2011, 1,491 (28.2%) transacted in 2012, and 1,794 (34.0%) transacted in 2013. Hedonic pricing models in the real estate literature account for time effects in a number of ways. For example, Farber and Yeates (2006) included a variable to capture the sale date in the 12-month study period to account for time trends. Another approach is to add time dummy variables to capture the year and month at which the property was sold.

Figure 3.2 charts the Home Price Index[9] (HPI) published monthly by the Multiple Listing Service (MLS) for Vancouver West detached homes, over the years 2011, 2012, and 2013. Based on this index, prices for detached homes in Vancouver West exhibited considerable volatility over this period.

[8] The Matlab function used to draw the Google Map is courtesy of Zohar Bar-Yehuda. http://www.mathworks.com/matlabcentral/fileexchange/27627-plot-google-map

[9] http://www.rebgv.org/home-price-index

Figure 3.2 MLS Home Price Index – Vancouver West Detached (2005/01 = 100)



This paper handles time effects by adjusting the sale price of each transaction with the HPI. Specifically, sale price are indexed to December 2013. The adjusted sale price are subsequently used in the models, and henceforth abbreviated as the "sale price". As an example, a property that sold for $1,000,000 in June 2012 would have an adjusted price of

$$\$1{,}000{,}000 * \frac{HPI(Dec/2013)}{HPI(June/2012)} = \$1{,}000{,}000 * (216.1/226.4) = \$954{,}505.$$

3.2 Descriptive Statistics

This section reports summary statistics for the variables used in this study: sale price, living area, lot size, age, number of rooms, number of bathrooms, and number of half bathrooms. For sale price, living area, and lot size, Moran's I[10] is computed to assess the degree of spatial autocorrelation. Similar to the correlation coefficient, the Moran Coefficient varies between -1 to +1, with positive values indicating positive autocorrelation.

Table 3.1 displays summary statistics of the study variables in the data set. The mean and median sale price are close to the Dec/2013 HPI benchmark price of $2.1 million for a typical Vancouver West detached home. The mean and median lot sizes are both considerably larger than the standard Vancouver lot size (33ft x 122ft = 4026 sqft.).

[10] The calculation of Moran's I requires an (NxN) spatial weight matrix W. For this section, $W(i,j) = 1 - \frac{d_{ij}}{0.4km}$ if $d_{ij} \le 0.4km$ and $W(i,j) = 0$ otherwise. d_{ij} is the distance between properties i and j. W is subsequently row-normalized so each row sums to 1.

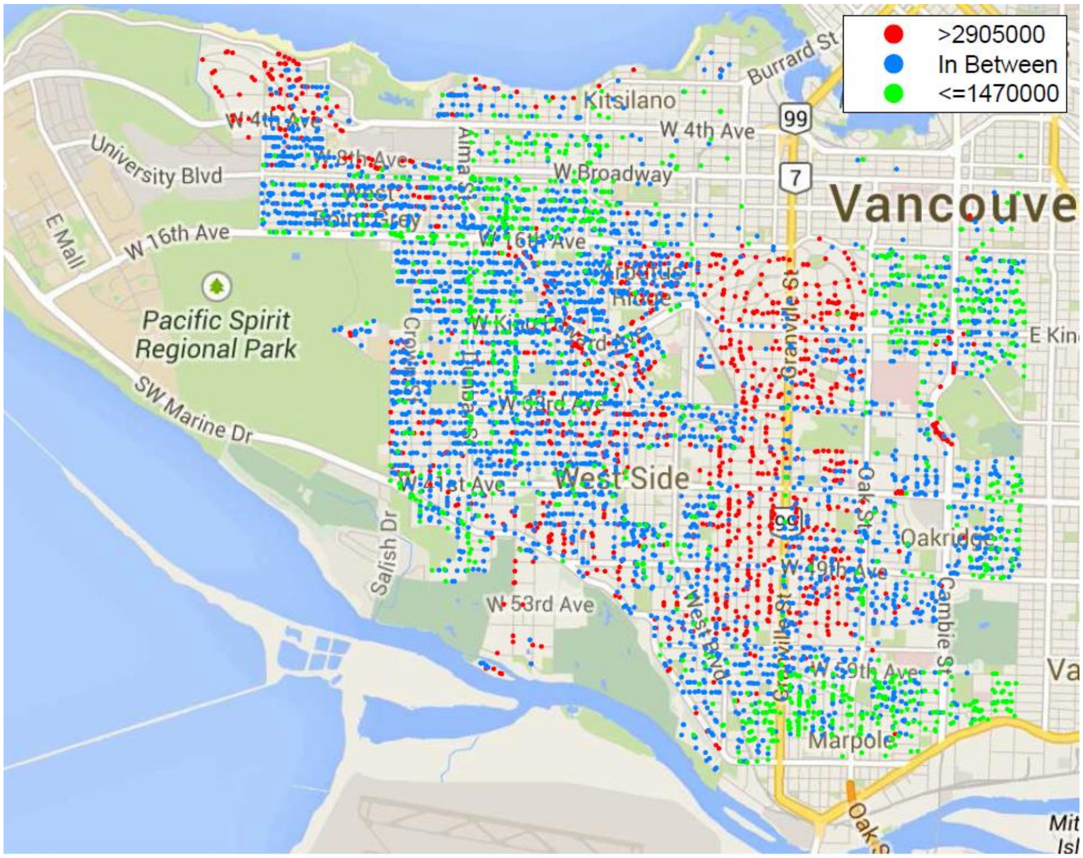
There is significant variation in the age of the properties in the dataset. Further analysis shows evidence of several clusters; 17.6% of the properties are aged between 0 to 5 years, 18% between 15 to 26 years, 15.1% between 63 to 76 years, and 18.7% between 81 to 91 years.

Moran's I for sale price, living area, and lot size are all positive, suggesting positive spatial autocorrelation in these variables. Figure 3.3 maps the spatial distribution of sale prices in the dataset.

Table 3.1 Summary statistics for sale price, living area, and lot size

Variable	Mean	Median	Standard Deviation	20th percentile	80th percentile	Moran's I
Sale Price	$2,335,000	$2,018,000	$1,356,000	$1,470,000	$2,905,000	0.471
Living Area (sqft.)	3032	2742	1320	2052	3978	0.324
Lot Size (sqft.)	6806	6041	4650	4026	8040	0.622
Age	46.65	51	33.85	9	83	
Number of Rooms	4.43	4	1.32	3	5	
Number of Bathrooms	3.51	3.5	1.63	2	5	

Figure 3.3 Spatial distribution of sale prices (adjusted to Dec/2013 HPI). The green, blue, and red points correspond to, respectively, properties with sale price in the bottom 20th percentile, 20th to 80th percentile, and top 20th percentile.



4: Hedonic Pricing Models

This section describes the three hedonic models considered in this paper: Ordinary Least Squares (OLS), Spatial Durbin Model (SDM), and Geographically Weighted Regression (GWR).

4.1 Ordinary Least Squares

In addition to variable selection, selecting an appropriate functional form is essential for an OLS hedonic pricing model. There are predominately three functional forms in the hedonic house price literature: linear, semi-log, and log-log. Malpezzi (2002) found that the semi-log and log-log specifications are common in empirical studies. Both the semi-log and log-log form express the marginal implicit price of attributes in terms of percentage change in sale price. In contrast, the linear form assumes a constant additive effect of each attribute. So in the linear form, the addition of a bedroom to a ten bedroom home would have the same effect as the addition of a bedroom to a two bedroom home.

Following the methodology used in other Canadian studies (Leblond, 2004; Farber and Yeates, 2006), this paper uses the semi-log specification for the OLS model:

$$\begin{aligned}\log(sale\mathrm{Pr}ice_i) = \alpha_1 + \beta_1 \log(livingArea_i) + \beta_2 \log(lotSize_i) + \beta_3 NumberOfBedrooms_i \\ + \beta_4 NumberOfBathrooms_i + \beta_5 Age_i + \beta_6 Age_i^2 + \varepsilon_i\end{aligned} \qquad (4.1)$$

In this model, β_1 (and similarly for β_2) is interpreted as the elasticity of sale price with respect to living area. If $\beta_1 = 0.4$, then a 1% increase in living area (holding everything constant) would lead to an increase of 0.4*1% = 0.4% in the sales price. β_3 and β_4 are interpreted as the percent change in sales price for an extra bedroom and bathroom, respectively. If $\beta_3 = 0.02$, then an extra bedroom (holding everything else constant) would yield an increase of 0.02% in the sales price. The interpretation of β_5 and β_6 is less straightforward because of the Age_i^2 variable. Taking the partial derivative of $sale\mathrm{Pr}ice_i$ with respect to Age_i in equation (4.1) gives:

$$\frac{\partial \log(sale\mathrm{Pr}ice_i)}{\partial Age_i} = \frac{\frac{\partial sale\mathrm{Pr}ice_i}{sale\mathrm{Pr}ice_i}}{\partial Age_i} = \beta_5 + 2\beta_6 Age_i \qquad (4.2)$$

So an extra year in age of a property would imply a percentage change in sale price by $[\beta_5 + 2\beta_6 Age_i]*100\%$.

The OLS model in equation (4.1) lacks locational attributes. To facilitate comparison with SDM and GWR, a variable with the neighborhood code of each property is incorporated. A total of eleven neighborhoods are identified[11] in the dataset. This variable is added to the OLS model through a set of dummy variables. This model is estimated separately and referred to as "OLS model with neighborhood codes".

4.2 Spatial Durbin Model

Figure 2.1 depicts the family of spatial econometric models that are used in spatial econometric analysis. As LeSage (2014) points out, only two should be considered for applied work: (i) the Spatial Durbin Error model (SDEM) or (ii) the Spatial Durbin model (SDM). SDEM is appropriate if the relationship being investigated generates local spillover effects, whereas SDM is applicable in the case of global spillover effects.

Feng and Humphreys (2008) suggests that the housing market generates a global range of spillovers. Shared neighborhood amenities lead to neighboring spillover effects among properties. As a result, the price of each house affects all other houses in the neighborhood, with the effect diminishing with distance.

In another study, Montero et al. (2011) selected the SDM to model the impact of noise on housing prices in Madrid, Spain. The authors cites that SDM is quite general and robust, and that the commonly used spatial autoregressive model and spatial error model are special cases of the SDM. And for the majority of spatially correlated data generating processes, SDM is able to provide consistent estimates.

With the above considerations, this paper uses the SDM. The SDM is expressed by a matrix equation:

$$y = \alpha 1_N + \rho W y + X\beta + WX\gamma + \varepsilon \qquad (4.3)$$

where

- y is an (Nx1) vector of the logarithms of the sale prices

[11] The neighborhoods identified in the dataset are *Arbutus/Mackenzie Heights*, *Cambie/Fairview/Mount Pleasant*, *Dunbar*, *Kerrisdale*, *Kitsilano*, *Marpole*, *Oakridge*, *Point Grey*, *Shaughnessy*, *South Granville*, and *Southlands/Marine Drive*.

- α is the constant term parameter, 1_N is an (Nx1) vector of one's
- ρ is the spatial autoregressive coefficient
- W is an (NxN) spatial weight matrix
- X is an (Nx6) matrix of housing attributes; $\log(livingArea)$, $\log(lotSize)$, $NumberOfBedrooms_i$, $NumberOfBathrooms$, Age, Age^2
- β is an (6x1) vector of parameters associated with X
- γ is an (6x1) vector of parameters associated with WX
- ε is an (Nx1) vector of uncorrelated errors with zero mean and constant variance

4.2.1 Direct and Indirect Effects in SDM

Unlike OLS, the interpretation of parameters is more involved in SDM. This is attributed to spatial spillover effects. Following the analysis from Montero et al. (2011), equation (4.3) can be re-written as (LeSage and Pace, 2009):

$$y = (I_N - \rho W)^{-1}[\alpha 1_N + X\beta + WX\gamma] + (I_N - \rho W)^{-1}\varepsilon \tag{4.4}$$

Since $(I_N - \rho W)^{-1}$ is in general a non-sparse matrix, a change in the housing characteristic of property j has a non-zero impact on the sale price of property i.

Now let $S_r = (I_N - \rho W)^{-1}(I_N \beta_r + W\gamma_r)$ be the (NxN) matrix associated with a change in characteristic r. The direct and indirect effects, respectively, of a change in characteristic r on the sale price of property i are given by:

$$\frac{\partial y_i}{\partial x_{ir}} = S_r(i,i) \qquad \text{and} \qquad \frac{\partial y_i}{\partial x_{jr}} = S_r(i,j) \tag{4.5}$$

The Average Direct Impact (ADI), Average Total Impact (ATI), and Average Indirect Impact (AII) of characteristic r are defined as:

$$\begin{aligned} ADI &= \frac{1}{N}\sum_{i=1}^{N} S_r(i,i) \\ ATI &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} S_r(i,j) \\ AII &= ATI - ADI \end{aligned} \tag{4.6}$$

4.2.2 Specification of Spatial Weight Matrix

An important consideration in applying spatial econometric method is the specification of a spatial weight matrix W. A spatial matrix is an (NXN) non-negative matrix that specifies a set of neighbors for each observation in the data set. It captures all the spatial interaction in the data and is specified a priori by the researcher. If property i and property j are neighbors, then $W_{ij} > 0$, and $W_{ij} = 0$ otherwise. Conventionally, a property is not a neighbor to itself, so that $W_{ii} = 0$. W is usually row-normalized for interpretation and estimation purposes.

The spatial statistics literature have proposed many techniques to specify W. Three common approaches are based on border contiguity, distance contiguity, and k-nearest neighbors. This study constructs the spatial weight matrix based on distance contiguity from Pace and Gilley (1997):

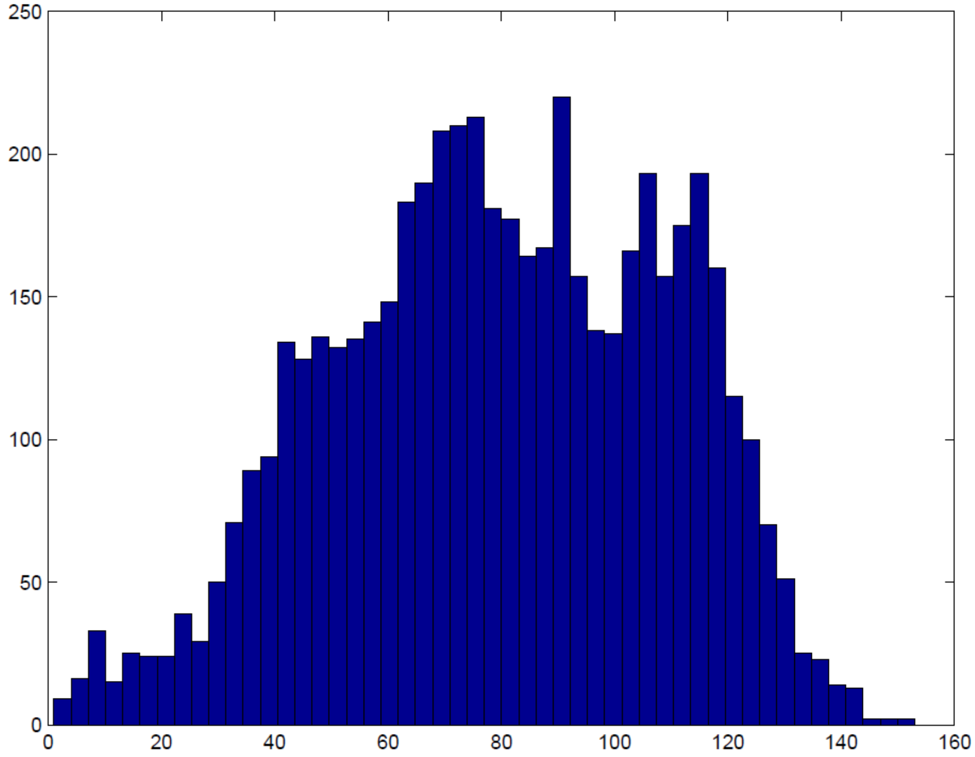
$$W_{ij} = \max(1 - (d_{ij} / d_{\max}), 0) \qquad (4.7)$$

where d_{ij} is the straight-line Euclidian distance between property i and property j, and $d_{\max}$ is a predetermined cutoff. In this formulation, properties with distance within $d_{\max}$ apart are considered neighbors, with the weight declining linearly with distance.

There are three appealing properties with the spatial matrix built from equation (4.7). The first is that W is symmetric, so that if property j is a neighbor of property i, then necessarily property i is a neighbor of property j. Spatial matrices based on border contiguity and k-nearest neighbors in general would not have this property. The second ideal property is that the sparseness of W can be controlled by adjusting $d_{\max}$. LeSage (2014) suggests that the weight matrix should be sparse. Thirdly, the weights are linearly declining with distance, so that neighbors that are closer together have larger weights than neighbors that are further apart.

In this paper, following the approach from Feng and Humphreys (2008), $d_{\max}$ is chosen such that every property has at least one neighbor. For the 5,278 properties in the data set, it turns out that $d_{\max} = 0.4km$. For this cut-off distance, the average number of neighbors is 79.66 with a standard deviation of 29.29. The minimum number of neighbors is one (by construction), and the maximum number of neighbors is 153. There are a total of 420,444 links (non-zero elements) in the spatial weight matrix, or 1.51% of the maximum possible number of links (5278*5278 – 5278). The distribution of the number of neighbors is shown in Figure 4.1.

Figure 4.1 Distribution of the number of neighbors with spatial weight matrix based on equation (4.6)



4.3 Geographically Weighed Regression

GWR extends the classical OLS model by estimating the parameters locally. For this study, GWR is an extension of the OLS model in equation (4.1). In vector notation, GWR is expressed as:

$$Y_i = \alpha_i + X_i \beta_i + \varepsilon_i \qquad (4.8)$$

where

- Y_i is the logarithm of sale price of property i
- α_i is the model intercept for property i
- X_i is a (1x6) vector of housing attributes for property i analogous to equation (4.2)
- β_i is a (6x1) vector of parameters for property i
- ε_i is the error term for property i

Now define an (NxN) diagonal weight matrix W_i with diagonal entries reflecting the weighting of properties with respect to property i [12]. The estimation of β_i then follows from weighted least squares:

$$\beta_i = (X^T W_i X)^{-1} X^T W_i Y \qquad (4.9)$$

In this paper, the weights are calculated using a k-nearest-neighbor weighting scheme from Pavlov (2000). First, the k nearest neighbors of property i are identified. Let this set be denoted by $N_k(i)$. Then weights are assigned to the k nearest neighbors using a parabolic shape function:

$$\begin{aligned} W_i(j,j) &= 1 - u_j^2 && \text{for } j \in N_k(i) \\ &= 0 && \text{for } j \notin N_k(i) \qquad (4.10) \\ u_j &= \frac{d_{ij}}{\max\limits_{k \in N_k(i)} d_{ik}} \end{aligned}$$

An important property of this weighting scheme is that a constant number of neighbors is admitted to estimate β_i from equation (4.8). Therefore, the relevant neighborhood of each property varies with the density of observations. The parameter k is a smoothing parameter. A small k will restrict estimation to only nearby properties, whereas a large k will allow distant observations to enter in the estimation. In terms of model complexity, the larger the k, the more complex the model.

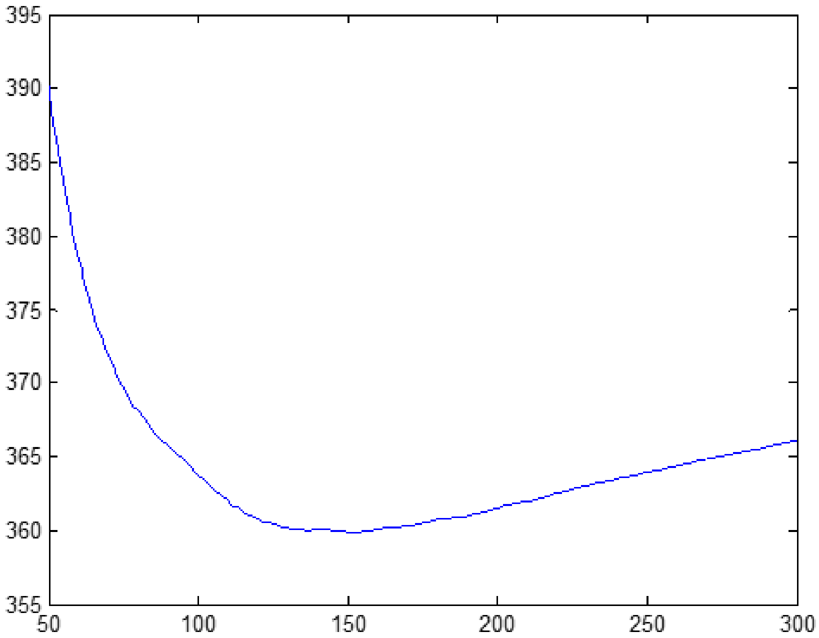
In this paper, k is determined by cross-validation. In this procedure, k is varied over a range of values. For each k, the series of model parameters are estimated from equation (4.8), then the predicted logarithm of the sale price is calculated for each property. The squared prediction error is subsequently calculated. The k that minimizes the total sum of squared prediction error is selected. To avoid unstable parameter estimates and singularity issues in local regressions, k is varied over the range $[50, 300]$.

Figure 4.2 shows that the sum of squared prediction error decreases rapidly as k is increased from the lowest value of 50. It reaches a minimum at $k = 153$ and gradually increases

[12] The weight matrix used for GWR is distinct from the spatial weight matrix used for SDM.

afterwards. This behaviour is typical for smoothing parameters in non-parametric methods, a phenomenon known as the "bias-variance tradeoff". For comparison, Farber and Yeates (2006) found $k = 274$ using cross-validation for 19,007 housing sales in the city of Toronto.

Figure 4.2 GWR Sum of Squared Prediction Error vs k



5: Estimation Results

5.1 Ordinary Least Squares

The estimation results of the two OLS models are presented in Table 5.1. The first OLS model does not incorporate any locational attributes. The second OLS model includes the neighborhood code of each property through dummy variables. The parameter estimates for the dummy variables are presented in Table 5.2. The Age^2 variable is divided by 1,000 to help show this variable's parameter estimate.

The inclusion of neighborhood codes significant improved the OLS model. R^2 increased from 55.4% to 63.0%, and Moran's I of the residuals is reduced from 0.252 to 0.109[13]. The reduction in Moran's I suggests that neighborhood codes are effective at removing spatial autocorrelation from the data.

For both models, all estimated parameters take the expected sign and are strongly statistically significant. Living area and lot size both have a large positive effect on sale price. Using the OLS estimates, a 1% increase in living area leads to a 0.36% increase in sale price, and a 1% increase in lot size yields a 0.47% increase in sale price.

An extra bedroom reduces sale price in both models. An explanation is that, for a fixed living area, more bedrooms reduces space available for desirable features such as a larger kitchen, larger living room, home theatre room, and bathrooms. A property with many bedrooms is also more likely to be a practical home rather than a luxurious home.

Both models predict an extra bathroom have a positive influence on sale price. The OLS estimate implies an additional bathroom increases sale price by 2.5%. In this study, the number of bathrooms also includes the number of half bathrooms. Since half bathrooms are a popular feature in newer luxury homes, an additional bathroom may also be an indication that the property is a luxury home.

[13] The calculation of Moran's I requires an (NxN) spatial weight matrix W. For Moran's I calculations, $W(i, j) = 1 - \frac{d_{ij}}{0.4km}$ if $d_{ij} \le 0.4km$ and $W(i, j) = 0$ otherwise. d_{ij} is the distance between properties i and j. W is subsequently row-normalized so each row sums to one.

Table 5.1 Estimation Results for Ordinary Least Squares Models (N=5,278). Dependent variable is the logarithm of sale price. Parameter t-values in brackets.

	OLS	OLS – With Neighborhood Codes
R^2	0.5535	0.6298
Moran's I - residuals	0.2520	0.1091
Constant	7.7636 (63.92)	N/A
Log(Living Area)	0.3591 (16.85)	0.2920 (14.85)
Log(Lot Size)	0.4742 (34.74)	0.5054 (35.76)
# of Bedrooms	- 0.0363 (- 9.60)	- 0.0242 (- 6.94)
# of Bathrooms	0.0251 (5.04)	0.0221 (4.85)
Age	- 0.0078 (-12.48)	- 0.0062 (-10.71)
$Age^2/1000$	0.0733 (12.61)	0.0500 (9.13)

As expected, age is estimated to have a negative impact on sale price. However, the square of age is estimated to have a positive impact on sale price, suggesting a positive vintage effect for older properties. Taken together, the model predicts a U-shaped effect of age on sale price.

Another finding is that the magnitude of the parameter estimates are smaller (except lot size) in the OLS model with neighborhood codes. In other words, after adjusting for location effects with neighborhood codes, the effect of each attribute (except lot size) on sale price is mitigated.

From Table 5.2, OLS model predicts that the Point Grey neighborhood is the most expensive, and the Marpole neighborhood is least expensive. The relative price of a home in Point Grey compared to a home in Marpole is approximately $e^{8.1627-7.7322} = 153.8\%$. That is, a property with fixed characteristics located in the Marpole neighborhood will sell for 53.8% more if it re-located to the Point Grey neighborhood.

Table 5.2 Dummy variable estimates for the Ordinary Least Squares Model with neighbourhood codes (N = 5,278)

Neighborhood	Dummy Variable Estimate
Arbutus	8.0851
Cambie/Fairview/Mount Pleasant	7.8658
Dunbar	8.0091
Kerrisdale	8.0116
Kitsilano	7.9981
Marpole	7.7322
Oakridge	7.8201
Point Grey	8.1627
Shahghnessy	8.1116
South Granville	7.9894
Southlands	7.7702

5.2 Spatial Durbin Model

Recall from Section 4.2 that the SDM is expressed in matrix notation as:

$$y = \alpha 1_N + \rho W y + X\beta + WX\gamma + \varepsilon \qquad (5.1)$$

James LeSage's Econometric Toolbox[14] is used to estimate the SDM. In particular, the *sdm* function is used with $[1_N \;\; X]$ as the independent variables. Table 5.3 presents the estimation

[14] http://www.spatial-econometrics.com/

results of the SDM, using the row-normalized spatial weight matrix based on equation (4.7)[15]. The first column shows the parameter estimates for the six explanatory variables (β). The second column displays the parameter estimates for the six explanatory variables with a spatial lag (γ). The third, fourth, and fifth columns calculates, respectively, the average direct impact, average indirect impact, and average total impact of each housing characteristic (see Section 4.2.1). The estimate for the constant parameter α is shown in the first column.

The estimated value for ρ , the spatial autoregressive parameter, is 0.78, indicating a high level of spatial dependency. Model R^2 is 60.3%, which is higher than the OLS model but lower than the OLS model with neighborhood codes. However, the use of R^2 for spatial econometric models is not appropriate and should be interpreted with caution (Anselin, 1988). Moran's I is 0.18, suggesting spatial autocorrelation is not entirely removed by SDM.

For SDM, the effect of explanatory variables is best measured by the average direct effect (see Section 4.2.1). Referring to the third column of Table 5.3, the average direct effects of the housing attributes are similar to those from the OLS model with neighborhood codes.

Living area and lot size are estimated to have a large positive direct effect on sale price, with price elasticities of 0.25 and 0.48, respectively. An extra bedroom has a negative direct effect of -2.1% on sale price. An extra bathroom adds 2.4% to sale price. Similar to the two OLS models, SDM estimates a negative direct effect with age but a positive direct effect with age squared, suggesting a U-shaped relationship.

Except for lot size, the average indirect effect takes the same sign as the average direct effect. In addition, the average indirect effect is generally much larger than the average direct effect. For example, a 1% increase in living area of a property leads to a 2.5% increase in its sale price, but it also increases sale price of neighboring properties by an average of 1.1%. An extra bathroom adds 2.5% to a property's sale price, and on average 10.5% to neighboring property's sale price.

The average indirect effect and average total effect estimates are sensitive to the connectedness of the spatial weight matrix W . It was discovered that if the cutoff distance in the weight function is increased (so that each property has more neighbors), the average indirect

[15] $W_{ij} = \max(1-(d_{ij}/0.4km),0)$

effect and average total effect increases rapidly. However, the average direct effect remain relatively constant.

Table 5.3 *Estimation results for Spatial Durbin Model (N=5,278). Dependent variable is the logarithm of sale price. t-values in brackets. Note: t-values may be inaccurate because of negative variances from numerical hessian.*

	Coeff X (β)	Coeff W*X (γ)	Average Direct Effect	Average Indirect Effect	Average Total Effect
Constant	1.0564 (3.16)	N/A	N/A	N/A	N/A
Log(Living Area)	0.2400 (12.72)	0.0461 (0.76)	0.2509 (12.75)	1.056 (3.29)	1.3071 (4.01)
Log(Lot Size)	0.4869 (28.41)	- 0.4758 (- 13.85)	0.4825 (27.88)	- 0.4308 (- 2.96)	0.0517 (0.35)
# of Bedrooms	- 0.01789 (- 5.45)	- 0.0521 (- 3.51)	- 0.0210 (- 6.03)	- 0.3017 (- 4.42)	- 0.3227 (- 4.67)
# of Bathrooms	0.02367 (5.54)	0.0035 (0.16)	0.0247 (5.41)	0.1054 (1.04)	0.1302 (1.26)
Age	- 0.0060 (- 10.82)	0.0010 (0.42)	- 0.0061 (- 10.89)	- 0.0166 (- 1.55)	- 0.0227 (- 2.09)
$\text{Age}^2/1000$	0.0452 (8.49)	0.0087 (0.45)	0.0472 (8.80)	0.2008 (2.28)	0.2480 (2.79)
ρ	0.7829 (130.53)				
R^2	0.6034				
Moran's I - residuals	0.1849				

Another observation is that the average direct effect estimates are similar to the parameter estimates (β) of the explanatory variables. This finding is consistent with LeSage and Pace (2009).

5.3 Geographically Weighted Regression

In this study, the weight matrix for GWR is constructed from a k-nearest-neighbor weighting scheme using a parabolic shape function. A cross-validation procedure selected $k = 153$ neighbors. For more details, see Section 4.3.

The estimation results for GWR is presented in Table 5.4. The mean, median, standard deviation, 25th percentile, 75th percentile, and Moran's I of the parameter estimates for each explanatory variable is presented. The same statistics for the local R^2 values are included as well. Following Farber and Yeates (2008), the global R^2 for GWR is calculated as the squared correlation coefficient between the observed and predicted values.

Although the mean and median local R^2 values are lower than those of the two OLS models and SDM, GWR achieved the highest global R^2 value of 65.9%. GWR also achieved the lowest Moran's I value of 0.0265, indicating GWR is most effective at removing spatial autocorrelation.

The mean and median of each parameter has the same sign as those from the two OLS models. The parameter magnitudes for most variables are similar to those from the OLS model with neighborhood codes. This is not a surprising finding because GWR estimates parameters locally, and the OLS model with neighborhood codes accounts for the neighborhood of each property.

Moran's I exceeds 0.8 for each attribute's coefficient. This suggests spatial heterogeneity in the effect of attributes on sale price. Figure 5.1 plots the spatial distribution of the parameter estimates for $\log(LivingArea)$, $\log(LotSize)$, number of bedrooms, and number of bathrooms. The four plots groups the properties by the 25th and 75th percentile of the respective parameter estimates.

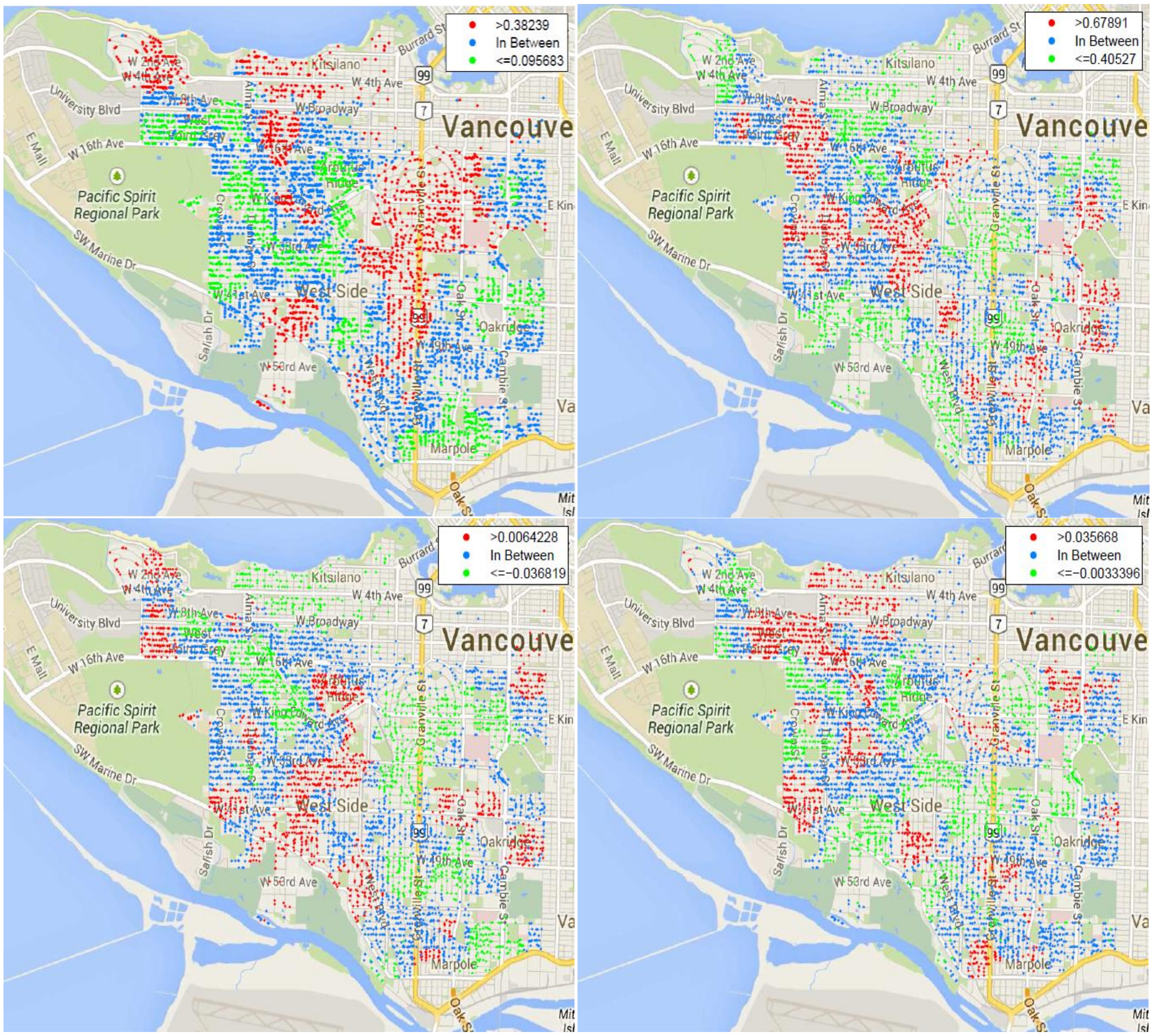
Note that for each of the four parameter estimates, properties for each percentile group appear in patches. That is, if a property's parameter estimate for an attribute is in the top 25th percentile, then nearby property's parameter estimate for that attribute will likely be in the top 25th percentile as well. This is a consequence of the GWR model and the weight function adopted in this study; nearby properties admit similar neighbors with similar weightings towards the estimation of the parameters in equation (4.9), resulting in similar parameter estimates.

An interesting finding from Figure 5.1 is that the effects of living area and lot size appear negatively correlated. Further analysis shows that only 59 properties, or 1.1% of the dataset, have parameter estimates of both $\log(LivingArea)$ and $\log(LotSize)$ in the top 25th percentile. Similarly, only 45 properties, or 0.85% of the dataset, have parameter estimates of both $\log(LivingArea)$ and $\log(LotSize)$ in the bottom 25th percentile. If the effects of these two attributes are independent, we would expect to observe 25%*25% = 6.25% in both cases.

Table 5.4 Estimation result for GWR (N=5,278).

	Mean	Median	Standard Deviation	25th Percentile	75th Percentile	Moran's I
Local R^2	0.4970	0.4948	0.1178	0.4228	0.5727	0.8371
Constant	7.9878	7.9697	1.5901	7.0007	9.0567	0.7613
Log(Living Area)	0.2519	0.2283	0.2074	0.0957	0.3824	0.8810
Log(Lot Size)	0.5425	0.5368	0.1853	0.4053	0.6789	0.8079
# of Bedrooms	- 0.0167	- 0.0145	0.0336	- 0.0368	0.0064	0.8833
# of Bathrooms	0.0185	0.0147	0.0380	- 0.0033	0.0357	0.8403
Age	- 0.0074	- 0.0079	0.0063	- 0.0116	- 0.0027	0.8969
$Age^2/1000$	0.0577	0.0583	0.0624	0.0148	0.0914	0.8778
R^2	0.6606					
Moran's I - residuals	0.0277					

Figure 5.1 Spatial distribution of GWR parameters: log(livingArea) (top left), log(lotSize) (top right), number of bedrooms (bottom left), and number of bathrooms (bottom right). The green, blue, and red points correspond to, respectively, properties with parameter estimates in the bottom 25th percentile, 25th to 75th percentile, and top 25th percentile.



5.4 Prediction Error

This section calculates the out-of-sample prediction error for the four models in this study. For the two OLS models and SDM, the calculation of out-of-sample prediction error involves the following procedure: (i) Randomly divide the observations into training set (50%) and a testing set (50%). (ii) Estimate the model parameters using the training set. For SDM, a spatial weight matrix based only on the training set is required. (iii) Use the estimated parameters to predict the logarithm of sale price in the testing set. For SDM, equation (4.4) is used with ε set to zero, and W is the spatial weight matrix based only on the testing set. (iv) Compute the prediction error for each observation in the testing set. Out-of-sample prediction error for each model is then equal to the squared prediction error averaged over the observations in the testing set. To reduce sampling variability, this procedure is repeated for 100 trials.

For GWR, no training set or testing set are required because a local regression model is built for each property using data from its neighbors. The predicted logarithm of sale price can readily be calculated with each observation's estimated local regression model. GWR out-of-sample prediction error is equal to the squared prediction error averaged over the observations in the dataset.

Table 5.4 shows the prediction error of the four models. Model R^2 and Moran's I of model residuals are included for comparison. GWR achieved the lowest prediction error, followed by the OLS with neighborhood codes, SDM, and OLS. For reference, the mean logarithm of sale price in the dataset is 14.55, and the median is 14.52.

GWR also had the highest model R^2, while OLS had the lowest. With respect to addressing spatial autocorrelation, GWR is most effective, having by far the lowest Moran's I of model residuals.

The results corroborate that of Farber and Yeates (2006). In that study, 19,007 records of housing sales over a 12-month period in Toronto were analyzed. The authors found that spatial autoregressive model (a special case of SDM) outperformed OLS, but that GWR outperformed both spatial autoregressive model and OLS.

Finally, Table 5.6 summarizes the effects of each attribute for the four models. For SDM, the average direct effects are reported. For GWR, the mean effects are reported.

Table 5.5 Prediction error for log(salePrice), R^2, and Moran's I of model residuals for the four hedonic pricing models in this study. Prediction error is calculated out-of-sample and based on 100 independent trials.

Model	Prediction Error	R^2	Moran's I - Residuals
OLS	0.0898	0.5535	0.2520
OLS – With Neighborhood Codes	0.0746	0.6298	0.1091
SDM	0.0823	0.6034	0.1849
GWR	0.0682	0.6606	0.0277

Table 5.6 Parameter estimates of attributes for the four hedonic models in this study.

	OLS	OLS – with Neighborhood Codes	SDM - Average Direct Effect	GWR - Mean estimate
Log(Living Area)	0.3591	0.2920	0.2509	0.2519
Log(Lot Size)	0.4742	0.5054	0.4825	0.5425
# of Bedrooms	- 0.0363	- 0.0242	- 0.0210	- 0.0167
# of Bathrooms	0.0251	0.0221	0.0247	0.0185
Age	- 0.0078	- 0.0062	- 0.0061	- 0.0074
$\text{Age}^2/1000$	0.0733	0.0500	0.0472	0.0577

5.5 Monthly Price Index Estimate

This section estimates a monthly price index to track price changes in the study period. Because no transactions occurred in December 2013, the index is estimated from January 2011 to December 2013. For this section, the GWR model from Section 4.3 is used but with Y_i set to the logarithm of the unadjusted sale price and α_i replaced by time dummy variables. The dummy variables are added to capture the year and month at which each property was transacted.

Table 5.7 shows the estimates of the time dummy variables for the 35 months from January 2011 to November 2013. From these estimates, the relative price of properties (adjusted for characteristics) between months may be calculated. For example, the relative price between January 2011 to July 2011 is $e^{7.78370-7.7552} = 102.89\%$. Computing the relative prices for every month relative to January 2011 yields an estimate of the monthly price index.

Figure 5.2 charts the estimated index, with the base period January 2011 set to 188.6 to coincide with the HPI index. Compared to the HPI index, the estimated index is much more volatile, showing significantly more month to month changes. The estimated index showed a decline of about 10% from January 2011 to February 2011, whereas the HPI showed a moderate increase. The HPI increased at a moderate for 2013, but the estimated index showed a large increase for September (6.35%) and decreases for October (- 2.15%) and November (- 1.01%).

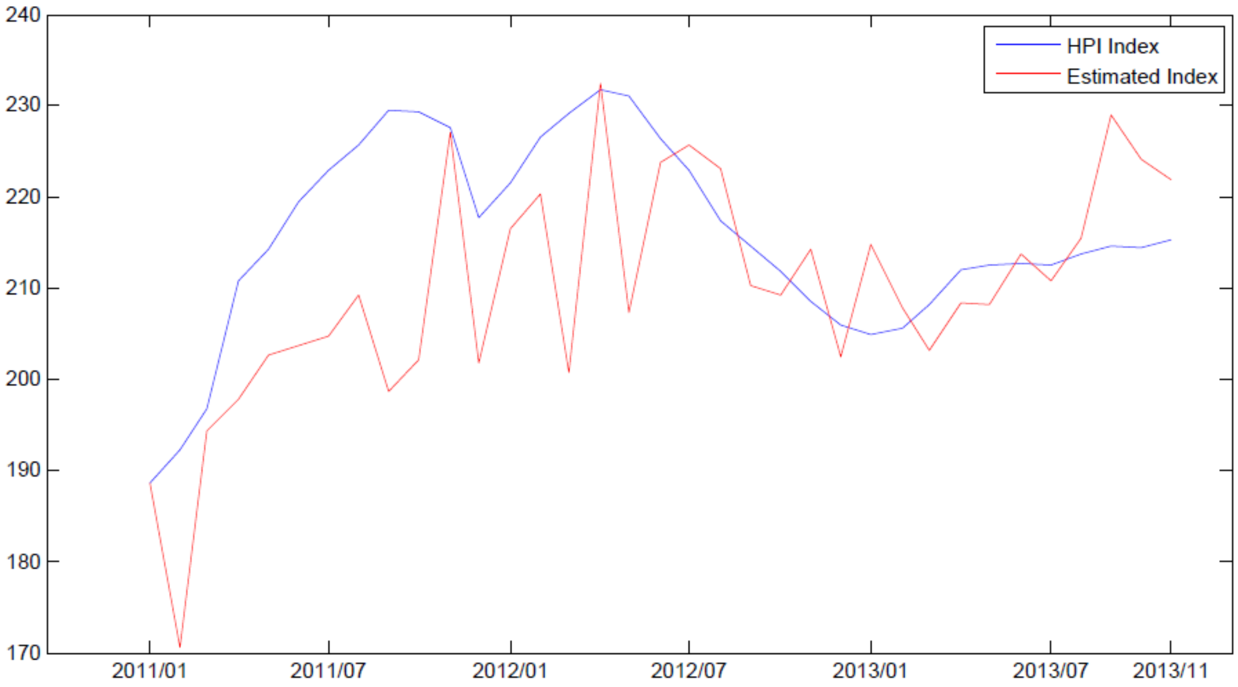
A possible explanation for the volatility of the estimated index is that in the GWR model, $k = 153$ are neighbors are used in the estimation. As a result, some properties might have very few (or zero) neighbors that was sold each month. This would lead to unreliable estimates for the time dummy variables.

Table 5.7 Dummy variable estimates using GWR model (N = 5,278)

Year/Month	Dummy Variable Estimate
2011 - Jan	7.7552
2011 - Feb	7.6549
2011 - Mar	7.7853
2011 - Apr	7.8028
2011 - May	7.8271
2011 - Jun	7.8323
2011 - Jul	7.8370
2011 - Aug	7.8589
2011 - Sep	7.8072
2011 - Oct	7.8248
2011 - Nov	7.9409
2011 - Dec	7.8227
2012 - Jan	7.8935
2012 - Feb	7.9105
2012 - Mar	7.8179
2012 - Apr	7.9647
2012 - May	7.8500
2012 - Jun	7.9266

Year/Month	Dummy Variable Estimate
2012 - Jul	7.9350
2012 - Aug	7.9234
2012 - Sep	7.8645
2012 - Oct	7.8591
2012 - Nov	7.8829
2012 - Dec	7.8266
2013 - Jan	7.8852
2013 - Feb	7.8522
2013 - Mar	7.8298
2013 - Apr	7.8547
2013 - May	7.8539
2013 - Jun	7.8804
2013 - Jul	7.8663
2013 - Aug	7.8882
2013 - Sep	7.9497
2013 - Oct	7.9280
2013 - Nov	7.9178

Figure 5.2 Estimated price index using time dummy variable estimates from GWR. HPI index shown for comparison



6: Conclusion

It is well documented in the real estate literature that hedonic price models based on OLS are inadequate to handle spatial effects inherent in housing data. This study compares the performance of four hedonic housing price models for single family homes in Vancouver West, Canada. The models range from classical ordinary least squares to more sophisticated spatial econometric models that account for spatial autocorrelation and spatial heterogeneity. In total, four models are considered: (i) Ordinary Least Squares (OLS), (ii) OLS with neighborhood code dummies, (iii) Spatial Durbin Model (SDM), and (iv) Geographically Weighted Regression (GWR).

The dataset consists of 5,278 sale records for single-family homes in Vancouver West from 2011 to 2013. A parsimonious set of six property attributes along with geographic coordinates of the properties are used in this study.

The models are compared based on model R^2, out-of-sample predictive power, and effectiveness at addressing spatial autocorrelation. For all three criteria, GWR performs the best, followed by OLS with neighborhood codes, SDM, and OLS. Referring to the Moran's I of model residuals in Table 5.4, GWR appears to be the only model capable of removing spatial autocorrelation from the data.

Encouragingly, all four models predict a similar impact of property attributes on sale price. Both living area (square feet) and lot size have a large positive effect on sale price. The number of bedrooms have a negative effect. Positive effects are estimated for the number of bathrooms and number of half bathrooms. A U-shaped relationship is estimated between the age of the property and sale price. That is, newer homes are more expensive than middle-aged homes, but middle-aged homes are cheaper than very old homes.

In conclusion, this study demonstrates that accounting for spatial effects is essential to building a reliable hedonic pricing model. The classical OLS model without location information over-estimates the impact of property attributes on sale price. In addition, model residuals from OLS exhibit a high degree of spatial autocorrelation. However, a classical OLS model with neighborhood codes outperforms the Spatial Durbin Model.

Future studies could investigate many issues. A few examples are: (i) adding explanatory variables related to structure (garage, pool, etc.), neighborhood (average income, percentage of immigrants, etc.), location (distance to central business district, schools, hospitals, etc.), and economy (stock market returns, mortgage rates, etc.); (ii) specification of spatial weight in SDM, as well as interpretation of indirect effects; (iii) specification of weight matrix and weight function for GWR, possibly including time effects (Huang et al., 2010); (iv) applying the methodology to other municipalities in the Greater Vancouver area; (v) applying the methodology to the apartment and townhouse market.

Reference List

Works Cited

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Bitter, C., and Krause, A. (2012). "Spatial Econometrics, Land Values and Sustainability: Trends in Real Estate Valuation Research". *Current Research on Cities*, 29: 19-25.

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

Boxall, P.C., Chan, W.H., and McMillan, M.L. (2005). "The Impact of Oil and Natural Gas Facilities on Rural Residential Property Values: A Spatial Hedonic Analysis". *Resource and Energy Economics*, 27: 248-269.

Brasington, D.M., and Hite, D. (2005). "Demand for Environmental Quality: A Spatial Hedonic Analysis. *Regional Science and Urban Economics*, 35: 57-82.

Can, A. (1992). "Specification and Estimation of Hedonic Price Models". *Regional Science and Urban Economics*, 1992, 22: 453-74.

Can, A., and Megbolugbe, I. (1997). "Spatial dependence and house price index construction". *The Journal of Real Estate Finance and Economics*, 14: 203-222.

Dubin, R. (1992). "Spatial Autocorrelation and Neighborhood Quality". *Regional Science and Urban Economics*, 22: 433-452.

Dubin, R. (1998). "Spatial Autocorrelation: A Primer". *Journal of Housing Economics*, 7: 304-327.

Elhorst, JP. (2014). *Spatial Econometrics From Cross-Sectional Data to Spatial Panels*. SpringerBriefs in Regional Science.

Farber, S., and Yeates, M. (2006). "A Comparison of Localized Regression Models in a Hedonic House Price Context". *Canadian Journal of Regional Science*, 29: No. 3.

Feng, X., and Humphreys, B.R. (2008). "Assessing the Economic Impact of Sports Facilities on Residential Property Values: A Spatial Hedonic Approach". *Working Paper Series, International Association of Sports Economists/North American Association of Sports Economists*, No. 08-12.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E. (2002). *Geographically Weighted Regression*. Chichester, UK: John Wiley and Sons.

Fotheringham, A.S., Brunsdon, C., and Charlton, M.E. (1998). "Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis". *Environment and Planning*, A, 30: 1905-1927.

Huang, B., Wu, B., and Barry, M. (2010). "Geographically and Temporally Weighted Regression for Modeling Spatio-temporal Variation in House Prices". *International Journal of Geographical Information Science*, 24 (3): 383-401.

Kestens, Y., Thériault, M., and Rosiers, F.D. (2006). "Heterogeneity in Hedonic Modelling of House Prices: Looking at Buyer's Household Profiles". *Journal of Geographical Systems*, 8: 61-96.

Leblond, S.P. (2004). "Comparing Predictive Accuracy of Real Estate Pricing Models: An Applied Study for the City of Montreal". M.Sc. thesis, University of Montreal.

LeSage, J.P., and Pace, R.K. (2009). Introduction to Spatial Econometrics, CRC Press, Boca Ratón.

LeSage, J.P. (2014). "What Regional Scientists Need to Know About Spatial Econometrics". Working paper series, Texas State University.

Malpezzi, S. (2003). "Hedonic pricing models: A selective and applied review", in T. O'Sullivan and K. Gibb, *Housing Economics and Public Policy*, Blackwell Science, Oxford.

Montero, J.M., Fernández-Avilés, G., and Mínguez, R. (2011). "Spatial Hedonic Pricing Models for Testing the Adequacy of Acoustic Areas in Madrid, Spain". *Investigaciones Regionales*, 21: 157-181.

Pace, R.K., Gilley, O.W. (1997). "Using the Spatial Configuration of the Data to Improve Estimation". *Journal of Real Estate Finance and Economics*, 14 (3): 333-340.

Pavlov, A.D. (2000). "Space-Varying Regression Coefficients: A Semi-parametric Approach Applied to Real Estate Markets". *Real Estate Economics*, 28: 249-283.

Sirmans, G. S., Macpherson, D. A., and Zietz, E. N. (2005). "The composition of hedonic pricing models". *Journal of Real Estate Literature*, 13 (1): 3-46

Vega, S.H., and Elhorst, J.P. (2013). "On Spatial Econometric Models, Spillover Effects, and W". *University of Groningen*, Working paper