SPATIAL ANALYSIS OF HOUSE PRICE DETERMINANTS: A GREATER LONDON CASE STUDY

LAURENT J. L. SANTOS[1]

RUI JIANG[1]

December 2018

[1] Department of Civil, Environmental & Geomatic Engineering,
University College London | UCL, UK
e-mail: laurent.santos.18@ucl.ac.uk and rui.jiang.18@alumni.ucl.ac.uk

How prices are determined by preferences and constrains is central to understand market trends and patterns. This study aims to investigate the relationship between a range of explanatory variables and house prices in greater London. More precisely, the research applies Exploratory Data Analysis and Spatial Regression methods to investigate factors affecting house prices in greater London. To what extent are housing prices determined by better health services and environmental areas? How does distance to underground stations and schools affect property prices? To what extent are housing values shaped by society income areas? To address these questions, a locally based spatial analysis is presented, and local estimates of regression parameters area investigated. The strategy of inquiry places emphasis on the structure of the housing market and the interrelationships of the variables that influence, directly or indirectly the functioning of the market. The study demonstrates that different variables have both positive and negative impacts in house prices across space.

1. INTRODUCTION

The focal point of this study is to investigate the relationship between a range of explanatory variables and house prices in greater London. More precisely, the study stems from Exploratory Data Analysis and Spatial Regression methods to investigate factors affecting house prices in greater London.

Over the last past decades, housing markets as an area of interest has long drawn attention given the causes and consequences of macroeconomics events as the great financial crash, drivers (for instance, labour, capital and material markets), but also for microeconomic theories (e.g. consumer decision analysis) and urban planning studies (e.g. migration and mobility). Housing markets is a broad topic of study with a wide spectrum of applications.

The mathematical modelling of spatial processes is useful in providing information on the determinants of those processes through the estimate of variables. It also provides a framework to predict spatial impacts of various actions and under different scenarios. To this end, a locally based spatial analysis is presented, a technique derived for producing local estimates of regression parameters which can be used to produce maps from which a geography of spatial relationships can be examined.

Spatial analysis is a type of geographical inquiry to explain patterns of human behaviour and its spatial expression in terms of mathematics and geometry. It also can be understood as a process that you can model problems geographically, derive results by computer processing, and then explore and examine results.

Spatial analysis provides a unique set of techniques and methods for analysing events that are located in geographical space. By means of quantitative analysis, it seeks to predict the spatial patterns of some phenomena, as for example the behaviour of property prices within a city.

2. LITERATURE REVIEW

2.1 Housing Markets

In academia or government, housing market is contented upon two broad frameworks: 1. macroeconomics and 2. microeconomics research. The former aims to analyse the drivers that might lead to economic outcomes. The latter is focused on consumer decision by modelling personal preferences and constraints (e.g. income).

Macroeconomics emphasises the context that somehow leads to market behaviours. From one side, the variables of interest are focused on the drivers that may affect pricing mechanisms. In this context, variables as income, taxes and interest rates are associated to housing outcomes (Maclennan, 2012). Such analysis also considers environmental statistics as labour, capital markets and the business of constructing to predict market outcomes.

In microeconomics, consumer preferences and decision mechanisms are the core variables that must be analysed. Housing markets are led by processes of search and matching (Maclennan, 2012). Prices fluctuate in order to bring supply and demand into *economic equilibrium*. Price determination is the outcome of rational choices made by individual agents (i.e. demand) and production (i.e. supply) (Centre for Co-operation with European Economies in Transition, 1993). Figure 1.1 depicts the traditional supply and demand curve model.

In this context, housing is regarded as a composite commodity with attributes that are valued by the decision maker (the purchaser) in order to satisfy a range of preferences. This process is known as consumer-utility function and assumes that buyers make informed decisions based on a matrix of attributes, options and risks throughout a specific spatial search process.

Economics research has been good at establishing the existence of the complexity of housing as a good. House price studies have been used to identify the economic significance of different, distinctive attributes of housing (a good literature review can be found in Maclennan 2012). These studies, that almost invariably have high levels of explanatory power, lead to some important conclusions and confirm that housing prices are influenced by attributes or characteristics such as:

- Size, style, layout and internal amenity (variety).
- The location of the dwelling: households pay not just for size, type, quality but for the characteristics of the location: + the costs of accessibility to the wider spread of locations used by household (such as foot instance, employment), + shopping and leisure locations: the quality and availability of neighbourhood amenity including neighbours, + access to local retail and service facilities.
- The asset importance of their home and possibilities for (relative) gain and loss as well as

quality and maintenance obligations (fixity and durability).

All these theories have a very similar moral and socio-political grounding, namely they assume that “people are rational, utility-maximising decision makers and that economic activity takes place in freely competitive, equilibrium-seeking contexts or settings” (Brown, 1993, p. 186).

Home and neighbourhood are an ineluctably conjoined choice. This jointness not only adds to the variety of attributes that have to be considered but also adds other distinctive aspects to housing choice.

2.2 Geospatial Data Analysis

Geospatial analysis can be defined as the process of examining the locations, attributes, and relationships of features within a spatial scale. We may better refer the techniques covered in this book as spatial rather than geographic because they are applied to data arrayed in any space, not only upon geographic space.

The methods of geospatial analysis provide a better understanding of spatial problems, offering a special perspective to examine events, patterns and processes with spatial variables. Another perspective is offered by De Smith (2007) who content that ultimately, geospatial analysis concerns what happens where and makes use of geographic information that links features and phenomena on the Earth’ s surface to their locations. Any analysis carried out without knowledge of location is definitively not a spatial analysis.

Before diving in a more profound toolset of quantitative geography, it makes sense to introduce the main elements of geospatial analysis.

- *Attributes* are the recorded characteristics or property of a place. The simplest type of attribute, termed nominal, is one that serves to identify one property from another. For example, the number of a property or house serves to identify that instance of a class of entities, and to distinguish it from other members of the same class.
- *Vector data* is the conceptual view of discrete objects. In a geographic scale, the world is represented using points, lines, and polygons. These data are created by digitizing the base data. They store information in (x_i, y_i) coordinates. Vector models are useful for storing data that has discrete boundaries, with country borders, land parcels, and streets.
- *Density estimation* is one of he most useful concepts for spatial analysis. It is a quantitative measure to address a context, to what extent events or properties at some location are related to the location’ s surroundings. The density expresses the number of discrete objects per unit area.
- *Spatial scale* or simply scale refers to the order of magnitude of extent or size of a land area or geographical distance studied or described. It also used in the sense of spatial resolution, or the level of spatial detail in data.

The idea of using models in problem analysis is really not new. For complex problems, visual and mathematical approaches might be used to study the relationships, to describe or represent an object or a phenomenon. A *model* is a simplified representation of reality, a mathematical expression that accurately represents the relevant characteristics of the object or problem being studied. (Ragsdale 2004).

Models allows us to gain insight and understanding examining complex problems under investigation. The modelling techniques in this book use mathematics to describe and explain a spatial pattern. But before dealing with spatial modelling techniques, we have to examine three categories of modeling techniques.

The first category is known as prescriptive models. In some situations, we face a problem involving a very precise, well-defined functional relationship *f()* between the independent variable $x_1, x_2, \ldots x_k$ and the dependent variable *Y*. These types of models are called prescriptive models because their solutions tell the decision maker what action to take.

A second category of problems is one in which the objective is to predict or estimate what value the dependent variable *Y* will take on when the independent variables $x_1, x_2, \ldots x_k$ take on specific values. If the function *f()* relating the dependent and independent variables is known, we might simply enter the specific values for $x_1, x_2, \ldots x_k$ into the function f() and compute the value of *Y*. These types of models are called *predictive models*. For example, a real estate appraiser might know that the value of a house (*Y*) is influenced by several variables as proximity to the underground (x_1) and parks (x_2) but the functional relationship *f()* that relates these variables to one another might be unknown. By analysing the relationship between the dependent variables, the appraiser might be able to identify a function *f()* that relates these variables in a reasonably accurate manner.

The third category of models are called descriptive models. In these situations, a problem has a very precise, well-defined functional relationship *f()* between the independent $x_1, x_2, \ldots x_k$ and the dependent variable *Y*. However, there might be great uncertainty as to the exact values that will be assumed by one or more of the independent variables. In these types of problems, the objective is to describe the outcome or behaviour of a given operation or system.

In this research, various mathematical models are applied to study a specific spatial problem: how housing prices are determined by health, transport, school and income within the greater London authority boundary. Essentially, predictive techniques are applied to tackle a spatial analysis problem.

3. EXPERIMENT DESIGN AND METHODOLOGY

The process of spatial analysis generally follows a number of stages: problem formulation, data gathering, exploratory analysis, modeling and testing and, ultimately, final reporting of the findings.

In this section, the methodologies applied in the study are introduced. The framework of the experiments and the hypotheses that will be tested are first reviewed.

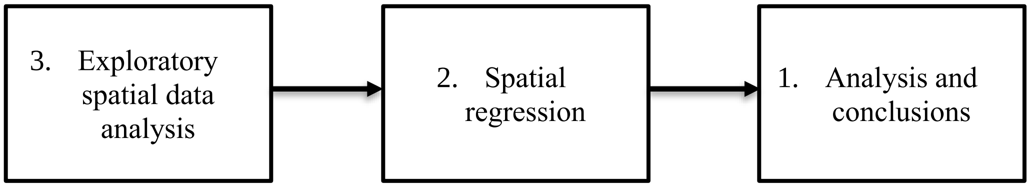
The following subsections are focused on geospatial statistical modeling approaches: exploratory spatial data analysis, spatial regression and testing. The final section of this chapter is dedicated to the description of the data applied in the experiment.

3.1 Analytical Methods

Spatial analysis consists of the analysis of numerical spatial data and the construction and testing of mathematical model of spatial processes. The goal of all these activities is to add to our understanding of spatial processes (Fotheringham, 2000).

The strategy of inquiry places emphasis on the structure of the housing market and the interrelationships of the variables that influence, directly or indirectly the functioning of the market. The strategy of inquiry that will be developed is based on a framework decomposed in three main sections. Figure 3.1 outlines the line of inquiry.

FIGURE 3.1 Outline of Applied Methods


3. Exploratory spatial data analysis
2. Spatial regression
1. Analysis and conclusions

A first step is to examine how data vary in space and explore their general characteristics. Explanatory data analysis aims to examine the relationships between variables to test causal relationships. The autocorrelation indices of spatial heterogeneity in house prices across London have to be calculated and measured.

The second part of the experiment is dedicated to spatial regression modelling. Spatial regression is an extension of linear regression to deal with the properties of spatial data, such as autocorrelation and heterogeneity. A regression model is built and accounts for the autocorrelation of the dependent variable through an additional parameter or set of parameters. With the best regression fit, spatial heterogeneity analysis can then be performed to identify non-uniformity phenomena in the dataset.

3.2 Quantitative Spatial Data Analysis

Exploratory data analysis (EDA) is a strategy to analyse datasets and explore their characteristics and generate hypotheses. The approach has its roots in the work of American mathematician John Tukey in the 1960s (see Tukey 1962). It emphasises graphical methods such as histograms, box-plots and scatterplots.

Exploratory spatial data analysis (ESDA) is the extension of EDA to spatial data. A key aspect of ESDA is the identification of patterns via data visualisation and statistics. It implies using techniques to summarize and visualise the data in search of meaningful patterns and to identify problematic and extraordinary data called *outliners*.

According to T. Cheng (2019), among the objectives of ESTDA, are included:

- maximising insight into a dataset
- uncovering underlying structure
- extracting important variables
- detecting outliers and anomalies
- testing underlying assumptions
- developing parsimonious models (simplest model that can achieve a certain performance level).

The most widely known way to explain a pattern in a dataset is a technique known as *correlation*. In spatial analysis, spatial autocorrelation aims to examine the magnitude of a variable at one location correlates with nearby locations (Cliff, 1981). The magnitude of the effects can be measured using a number of statistics which describes spatial variation in terms of a function that shows how spatial autocorrelation decreases with increasing distance. This core concepts have been captured by Tobler (1970) in what has become known as his First Law of Geography: "*All thinks are related, but nearby things are more related than distant things*" (p.235).

Testing for spatial autocorrelation and incorporating it into a model requires a formal specification of the spatial structure. This is done through a spatial weight matrix and spatial lags. A spatial weight matrix is a nxn square matrix, designated as W, where each element w_{ij} describes the strength of the spatial relationship between each of the n spatial observations. The weight matrix embodies the analyst' s knowledge and assumptions about the study area and data within the model.

A distance weight matrix considers the distance between observations when defining the spatial structure. It works well with point data since precise distances can be captured. The weight matrix captures the strength of the relationships, so two points that are close in space should have a larger weight than those further apart, typically going to zero at some cutoff point (Singleton, 2017).

The spatial weight matrix formalizes the neighbourhood around each observation. A spatial lag is created by multiplying the weight matrix by a variable measured for each observation.

3.3 Spatial Regression

In the study, spatial regression models are used to explore the effect of several explanatory variables on house price data in London by taking spatial autocorrelation and heterogeneity into account. Spatial regression is an extension of linear regression to deal with the properties of spatial data, such as autocorrelation and heterogeneity.

Through the regression approaches, including linear regression and spatial regression methods, a series of explanatory variables are applied used in the analyses:

- retail environment
- health services accessibility
- physical environment
- distance to undergrounds
- distance to schools
- annual household income.

These explanatory variables are utilised to explain and predict the dependent variable, the mean value of the house price in London. Four variables have been log transformed in all models. Replacing the data with the log aims to remove the skew and normalise the data, making the pattern more visible.

When dealing with spatial data, conventional multiple linear regression is often not enough to account for the presence of spatial regression. After the residual autocorrelation has been tested, the spatial regression models are used to fit the model.

Three widely used spatial regression models (Anselin, 2010) will be presented:

- spatial lag model

- spatial error model
- spatial Durbin model

These techniques are going to be applied by considering a spatial lag term, a spatial error term and a lagged term of the independent variables.

3.3.1 Multiple Linear Regression

As one of the most widely used and simplest statistical methods, multiple linear regression is conducted to summarize and explain the relationships between dependent variable and independent variables based on their statistical relationships.

Scatterplots between each independent variable and the dependent variable are created to verify associations between the variables which may indicate significant relationships between house prices and other factors.

Also, prior to linear regression, multicollinearity must be tested to ensure the reliability of the model. *Variance inflation factor* (VIF) and *Kappa* statistic are used to diagnose multicollinearity. Both statistics aimed to explore the correlation among independent variables. VIF, the reciprocal of the tolerance, quantifies the severity of multicollinearity by measuring the variance in a model in the presence of multiple terms and one single term[1] while Kappa, the condition number, explains the multicollinearity by calculating the eigenvalues of the variables (Belsley, 2004).

In general, a value of VIF less than 5 or a Kappa statistic less than 1000 indicates that it is safe to ignore multicollinearity. When a VIF goes above 5 and Kappa goes above 1000, it can be assumed that the regression coefficients are modelled incorrectly and requiring correction. In this case, the independent variable with a high VIF/Kappa should be removed from the regression model, or *Partial Least Square Regression* would be used instead.

3.3.2 Spatial Autocorrelation in Regression Errors

When working with spatial data, it is important to test for spatial autocorrelation in the residuals of linear models. There are a number of tests that can be used. We examine some of them in the following sections.

3.3.2.1 Spatial Autocorrelation in Ordinary Least Squares (OLS) Regression Errors

Although the association between covariates must be examined and excluded by conducting multicollinearity tests, spatial correlations among the variable itself may still exist. When dealing with spatial variables, it is essential to test the regression errors after the model been initially fitted in

[1] For more details, see James, 2013.

the linear model due to the independency of the explanatory variables in terms of their spatial correlations.

Therefore, to avoid spatial model misspecification, it is critically necessary to calculate the spatial autocorrelation in the residuals of multiple linear regression model by employing spatial test models.

Based on the assumption that the residual errors are normally distributed, different approaches of diagnostics are adopted to quantify spatial dependence in the presence of covariates (Darmofal, 2015). Generally, there are two types of tests, unfocused diagnostics and focused diagnostics. Unfocused diagnostics include *Moran's I* and *Kelejian-Robinson* (KR) diagnostics, which simply measure the presence of spatial dependence. On the contrary, focused diagnostics apply the effect of spatial lag or spatial error dependence on the present dependence as well as LM test for spatial lag dependence and LM test for spatial error dependence.

3.3.2.2 Moran's I

Moran's I stands as the most common diagnostic approach for regression residuals.

Unlike the usual Moran's I test which measures autocorrelation in spatial data, Moran's I for residual spatial autocorrelation assumes that all the variables are the residuals from a linear regression.

3.3.2.3 Lagrange Multiplier Test for Spatial Lag Dependence and Error Dependence

Lagrange Multiplier (LM) test is usually used for situation where several sources of misspecification are considered. Specifically, the test is applied for situations in which spatial dependence in the form of an omitted spatially lagged variable or spatial autocorrelation are presented.

3.3.2.4 Spatial Lag Model

Spatial lag model assumes that direct interactions between observations at near locations exist (Ward, 2008). The model is purely autoregressive and posits that autocorrelated dependent variable causes the presence of spatial autocorrelated residuals. The model can be expressed as the following equation:

$$\mathrm{y} = \rho W y + X\beta + \epsilon$$

where *ρ* represents the spatial autocorrelation parameter and *W* represents the row-sum standardized spatial weight matrix.

3.3.2.5 Spatial Error Model

The *spatial error model* assumes that autocorrelation exists in the errors as a result of unobserved variables. The model can be expressed as the following equation:

$$\mathrm{y} = X\beta + \gamma W\epsilon + \delta$$

where *γ* is the spatial autocorrelation parameter, *W* is the row-sum standardized spatial weight matrix and *δ* is a vector of the error terms.

3.3.2.6 Spatial Durbin Model

Based on spatial lag model, the spatial Durbin model utilises both the explanatory variables and predictor variable as its autoregressive terms.

$$\mathrm{y} = \rho W y + X\beta + W X\theta + \epsilon$$

where *θ* is a vector of parameters.

3.4 Geographically Weighted Regression

The global regression model (GWR) estimates a constant parameter for the relationship between each independent variable and the dependable variable across the study area.

GWR is a technique that allows local, rather than global, parameters to be estimated. It recognizes that spatial variations might exits and should be measured**.** That is, observed data near to point *i* have more influence than those located farther than *i* *(Fotheringham, 2000)*.

In this model, the equation measures the relationship around each point *i*. Weighted least square estimation provides the basis to understanding how GWR operates. It attempts to model heterogeneity using geographically varying regression coefficients. This enables maps of the coefficients to be produced, providing a better understanding of the relationship between the dependent and independent variables across space.

The general form of the GWR model is:

$$\mathrm{y_i} = \beta_{i0} + \sum_{p=1}^{m} \beta_{ip} x_{ip} + \varepsilon_i$$

Where: β_{i0} is the intercept term at location *i*, $i=1,2,...,N_i=1,2,...,N$; *N* is the number of spatial locations; β_{ip} is the value of the p^{th} parameter at location *i*, $p=1,2,...,m$; *m* is the number of independent variables; ϵ_i is a random error.

3.5 Data Collection and Description

Economic studies content the existence of a broad range of attributes for housing. In his work, Maclennan (2012) recognizes this and suggests that they have high levels of “explanatory power” (p. 7), confirming that prices are influenced by a spectrum of characteristics and factors dictated by housing characteristics, location and access to local retail and service facilities. Quite in line with this, Malpezzi (2002) asserts that the value of houses correlate with different attributes.

That said, it can be suggested that house pricing is a direct application of the economic significance of different factors, ranging from internal attributes as well as locational characteristics (e.g. services facilities, public transport, neighborhood, etc.). Maclennan (2012) points to the fact that space effects interact with variety in shaping the dynamics of housing markets: "households pay not just for size, type, quality but for the characteristics of the location" (p. 7).

A fundamental feature of housing is the availability of underground station nearby. Indeed, many studies suggest a positive relationship between house prices and proximity to rail station. In their work, Cervero et al. (2004) found a 6.4% to 45% increase in the house prices near to rail station. This claim is also supported by Debrezion et al. (2007) who found that for every 250m closer to the station is worth 2.4% of the house price.

By the same extent, living close to school is an asset for couples with children. Their life course is particularly salient when living nearby school facilities in the sense that it turns to be an important attribute for housing search. In this context, Clark and Davies Withers (2007) argue that there is higher demand for schools and facilities when young families move into a community.

The access to hospitals, has also been pointed out as an important component for house prices. Clark (2012) argues that housing prices is seen as a function of demands for associated services. The quality and availability of health services might lead to a positive relationship to house prices.

Additionally, Zhang and Yi (2018) argue that house prices increase significantly as a result of a willingness to pay for a pleasing environment, that is, zones with low traffic congestion, environmental pollution and availability of leisure areas. In this line of reasoning, environmental health might also present a positive effect in housing markets.

A further component affecting housing choices is the association with neighbourhood attributes. In this sense, Maclennan suggests that as neighbourhood attributes are often associated with "social situational good" (2012, p. 9). Butler and Hamnett (2012) explain that residential patterning presents strong differences in the geography of income and social class. Income and housing prices might also be positively correlated.

Taking into account this brief review, Table 3.1 details the description and sources of the data applied to the study.

TABLE 3.1 Data Description

Category	**Data base**	**Description**	**Type**	**Source**	**Geographical scale**	**Maintainer**	**Data range**
Pricing	House price data	Price Paid Data includes information on all property sales in England and Wales that are sold for full market value and are lodged with us for registration.	Continuous data	https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads	Point data	HM Land Registry	Mar-17
Education	Education	List of London schools.	Location data	https://data.london.gov.uk/dataset/london-schools-atlas	Point data	Consumer Data Research Centre	2017
Environmental health	Environmental domain	London environmental domains	Discrete data	https://data.cdrc.ac.uk/dataset/access-to-healthy-assets-and-hazards-ahah/resource/dd146935-8181-4cef-95cf-68e11b47c1ae	Lower layer super output areas (LSOA)	Consumer Data Research Centre	2017
Health Services	Health Services domain	London health services.	Location data	https://data.cdrc.ac.uk/dataset/access-to-healthy-assets-and-hazards-ahah/resource/dd146935-8181-4cef-95cf-68e11b47c1ae	Lower layer super output areas (LSOA)	Consumer Data Research Centre	2017
Transport	Underground stations	List of London stations with their Ordnance Survey coordinates. Include all tube stations and mainline stations.	Location data	https://github.com/oobrien/vis	Point data	OpenStreetMap	November, 2018
Income	CDRC Individual Income Estimates (PAYE)	Individual income from Pay As You Earn (PAYE) and benefits.	Discrete data	https://data.cdrc.ac.uk/dataset/cdrc-2016-individual-income-estimates-greater-london-area-geodata-pack-london-e12000007	Lower layer super output areas (LSOA)	Consumer Data Research Centre	2016

4. Exploratory Spatial Data Analysis

To explore the characteristics of the variables, exploratory spatial data analysis (ESDA) was applied prior to the spatial regression models.

4.1 Non-spatial Characteristics

Table 4.1 summarizes the statistics of variables. From the table, a wide range of values of variables can be observed. The minimum value of house price is 190722, while the maximum value is more than four times greater than that. The wide range can also be observed in the distance to undergrounds, with LSOA area containing undergrounds having a value of 0, while for some LSOA areas, the distance to the nearest underground is as far as 18847.6 meters. Compared with undergrounds, schools seem to be more evenly distributed. The maximum distance from a LSOA area to the nearest school is 753.88 meters, which is about 1% of the maximum distance from LSOA to the underground. The indexes of retail outlets, health service and physical environments vary from 0 to 100. The mean annual household income ranges from 25020 to 88090.

TABLE 4.1 Summary of Variables

	House price	**Retail**	**Health**	**Environment**	**Underground**	**School**	**Income**
Min.	190722	4.116	0.0049	4.898	0	0	25020
1st Qu	375081	26.962	1.3618	28.692	327.2	0	34130
Median	483722	40.418	3.4774	44.501	970.5	14.52	38440
Mean	671684	42.915	5.5587	47.221	2370.9	62.7	39565
3rd Qu	671684	56.956	7.5203	63.93	3329.7	98.77	43545
Max.	8061563	99.255	46.1394	100	18847.6	753.88	88090

4.2 Spatial Characteristics

Figure 4.1 uncovers wide range values of different variables. Different spatial patterns can be observed between the central of London and the peripheral areas. In general, variables appear to be spatially correlated since adjacent LSOA areas tend to have similar values than distant LSOA areas.

Figure 4.1 also demonstrates that houses with higher property prices cluster in the central area of London. Together with figure 4.2, it seems that houses with a higher price appear to have a better accessibility to retail outlets and a shorter distance to the nearest underground. However, those houses tend to be surrounded by worse physical environments, and they appear to have less accessibility to health services. Also, the distance to schools has less spatial patterns and does not present strong correlation with house prices. As expected, annual household income appears to have a positive correlation with the house prices.

FIGURE 4.1 Map of House Prices

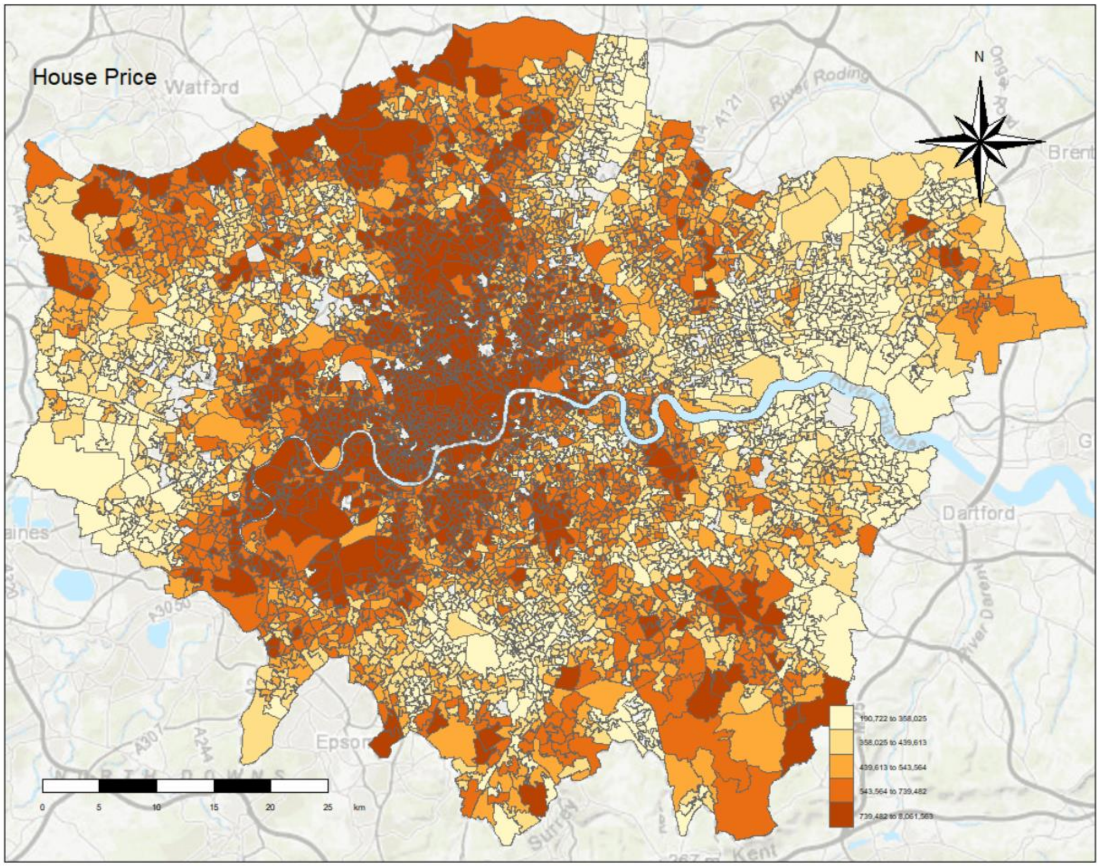
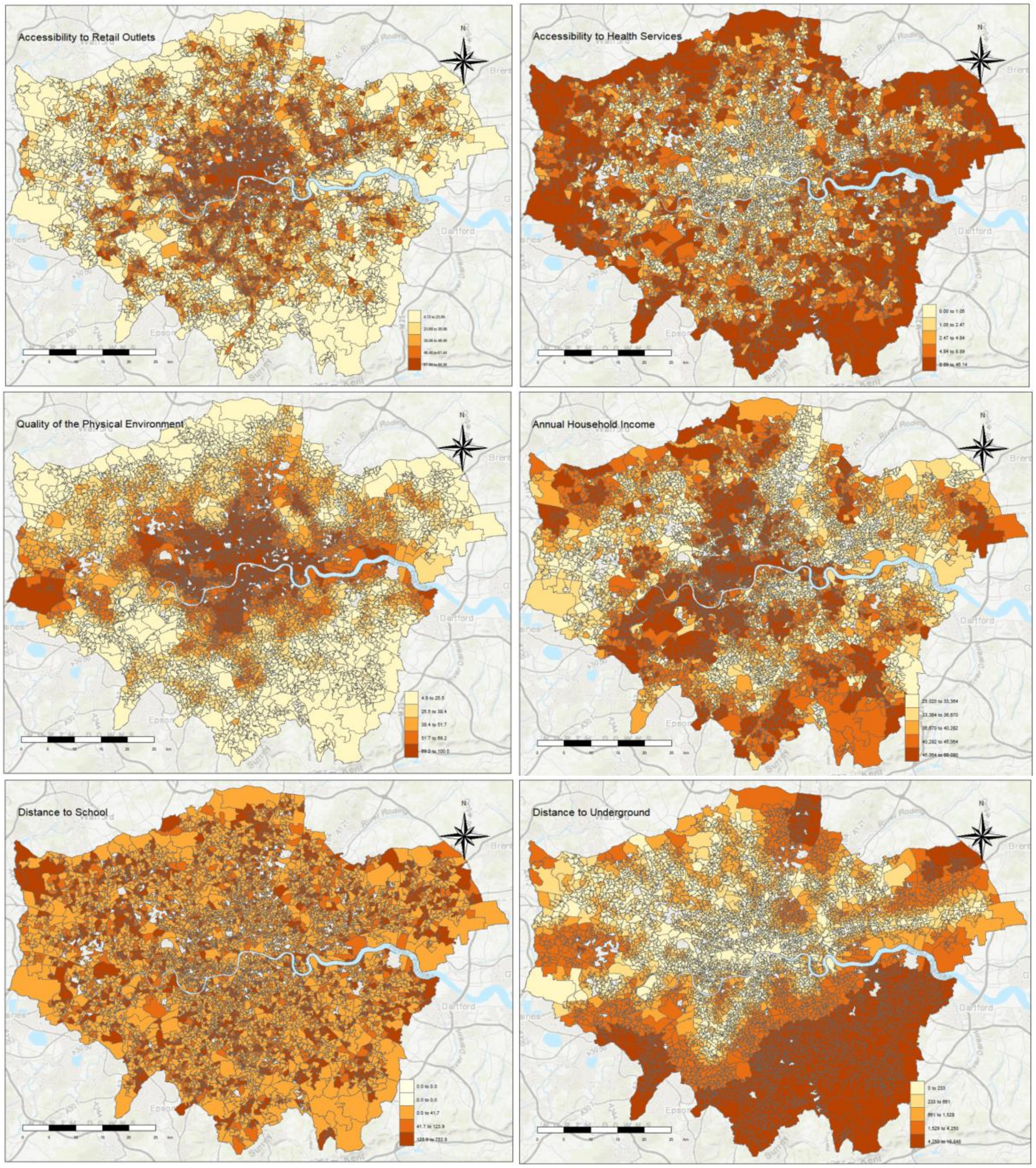


FIGURE 4.2 Maps of (from top to bottom, from left to right): 1. Accessibility to Retail Outlets and 2. health service, 3. Quality of Physical Environment; 4. Annual Household Income; 5. Distance to School and 6. To Underground stations.



4.3 Global Spatial Autocorrelation

As Tobler (1970) mentioned, everything tends to be related to everything else, but things close to each other tend to be more related.

In the study, Moran’s I have been employed to uncover the presence of global spatial autocorrelation by measuring a single value of autocorrelation level across the study area for each variable.

TABLE 4.2 Outputs of Monte-Carlo Simulation

Monte-Carlo simulation	Moran’s I	observed rank	p-value
House Price	0.681	1000	0.001
Retail	0.705	1000	0.001
Health	0.642	1000	0.001
Environment	0.943	1000	0.001
Underground	0.987	1000	0.001
School	0.150	1000	0.001
Income	0.733	1000	0.001

Table 4.2 summarizes the values of calculated Moran’s I, observed rank and p-value for each variable. The positive values of Moran’s I suggest that all variables are positively correlated to nearby variables: distance to underground and physical environment present a strong correlation whereas distance to school has a weaker correlation. The observed ranks are all 100 and p-values are all less than 0.001, which suggests that the probability of the observed Moran’s I being due to chance is less than 0.1%. Therefore, it can be concluded that all variables are significantly positively autocorrelated.

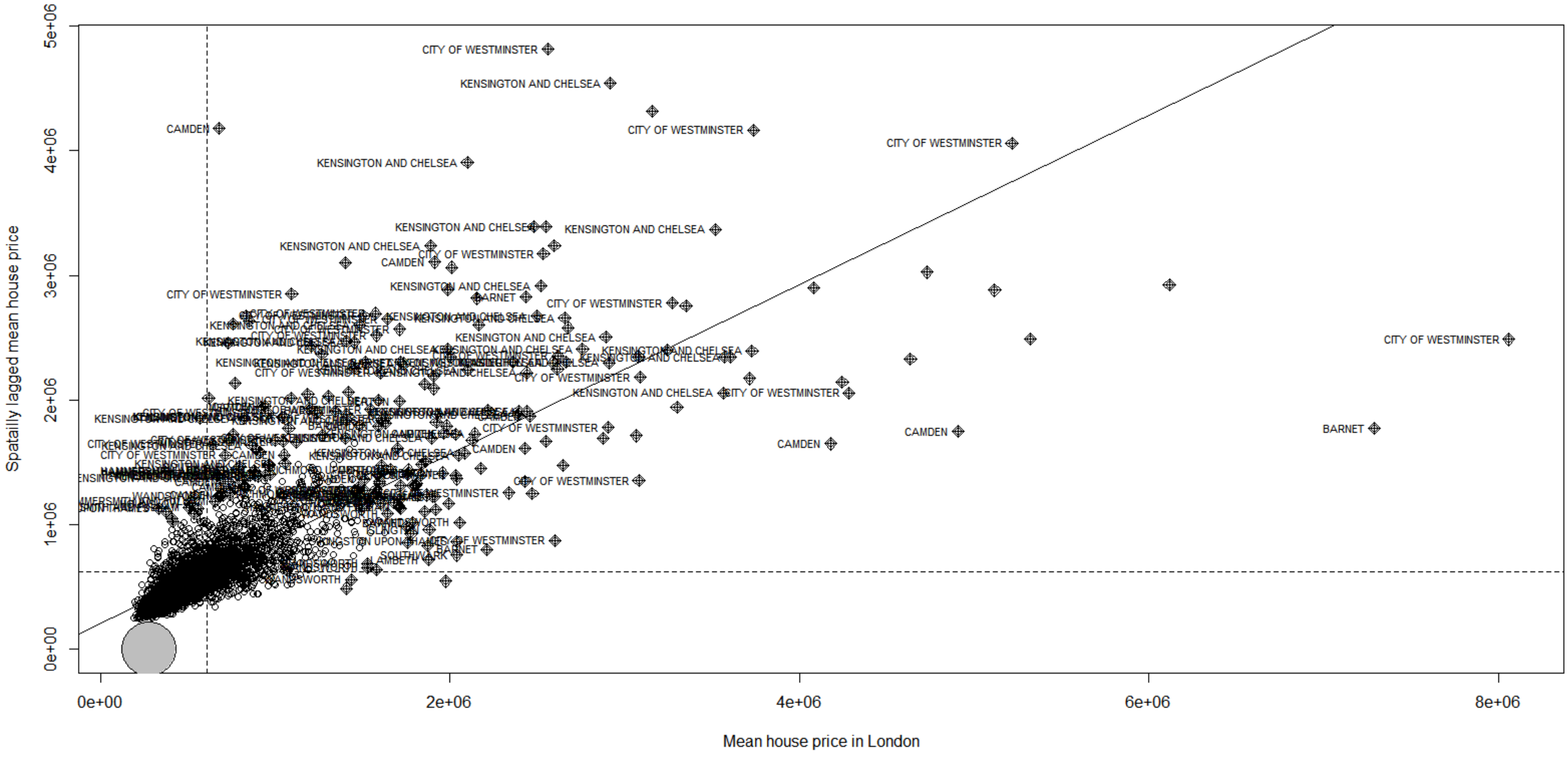
4.4 Local autocorrelation

Due to the presence of spatial heterogeneity, the level of autocorrelation could vary. In this case, a single value is not enough to explain the variance in the variables. Therefore, local autocorrelation methods were employed in the study as well.

4.4.1 Moran Scatterplot

The Moran scatterplot on figure 4.3 identifies clusters of high and low values in house price data. Indicated by the labels on the plot, there are some clusters of extremely high price houses in city of Westminster and Camden town, while most data are within the ‘Low-Low’ quadrant. The missing district information is due to the combination of two datasets, with one dataset has less data.

FIGURE 4.3. Moran Scatterplot of House Prices



4.4.2 Getis and Ord’s Gi and Gi *

The figures below are the outputs of Getis and Ord’s Gi and Gi*. The figure 4.4 indicates the distribution of raw Gi* values across the study area. Central London has a higher value of Gi* compared with peripheral area, which suggests that the house price in central London is more spatially autocorrelated.

Figure 4.5 displays the distribution of hotspots and coldspots. As expected, the central area of London stands as a cluster of high property prices while some parts of peripheral areas are identified as a cluster of relatively low property price, especially east part. No significant clusters, presented by red colour, are found in most of the areas.

FIGURE 4.4. The Output of Getis and Ord’s Gi*

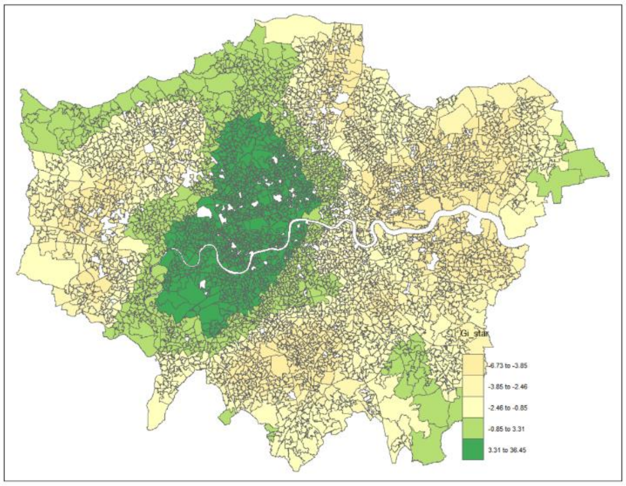
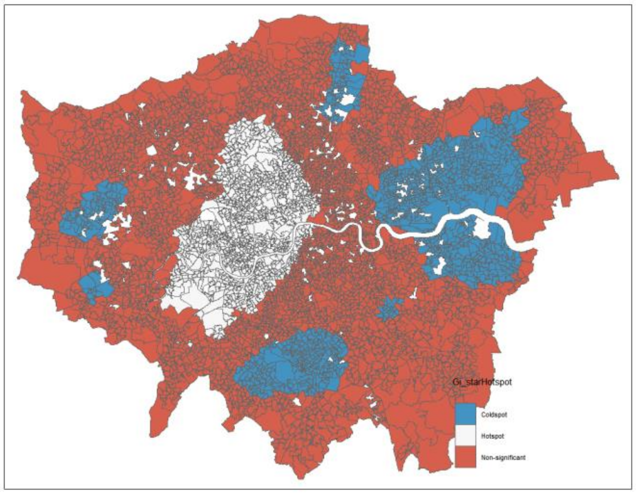


FIGURE 4.5. Hotspots and Coldspots



4.4.3 Local Moran’s I

Areas with significant clusters have been uncovered by the local Moran approach. Consistent with the results calculated by Getis and Ord’s Gi and Gi *, there is a small area in the central London with high-high clusters. No low-low clusters have been found in this study area by local Moran’s I (see figures 4.6 and 4.7).

FIGURE 4.6 Local Moran's I

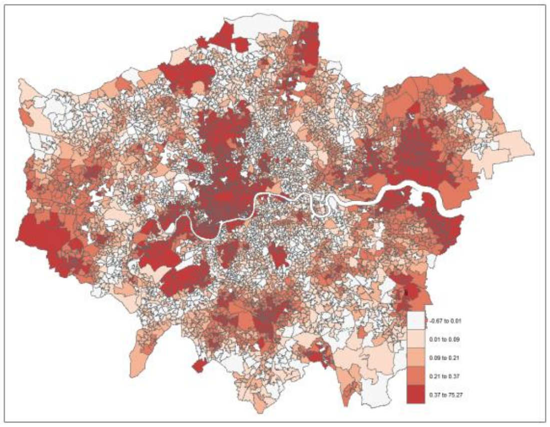
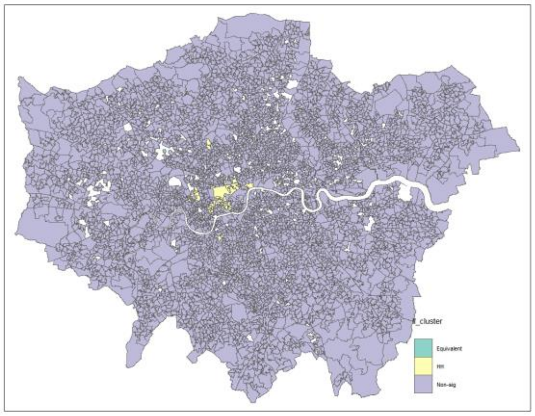


FIGURE 4.7 Map of Significant Clusters



5. RESULTS

5.1 Multiple Linear Regression

To acquire an initial overview of the data, a scatter plot of matrices was created. It helps to visualise the relationships among each variable. In figure 5.1, the diagonal, bivariate scatter plots with the linear regression fits (red lines) reveals the statistical relationships between independent and dependent variables. On the diagonal, histograms are presented, which indicates most of the variable data are normal distributed after log transformation. Pearson correlation coefficients are presented in the area above the diagonal. Overall positive associations were found between house price and retail, environment and income variables, which negative associations were found between house price and health service, distance to the underground and school.

FIGURE 5.1 A Scatter Plot of Matrices

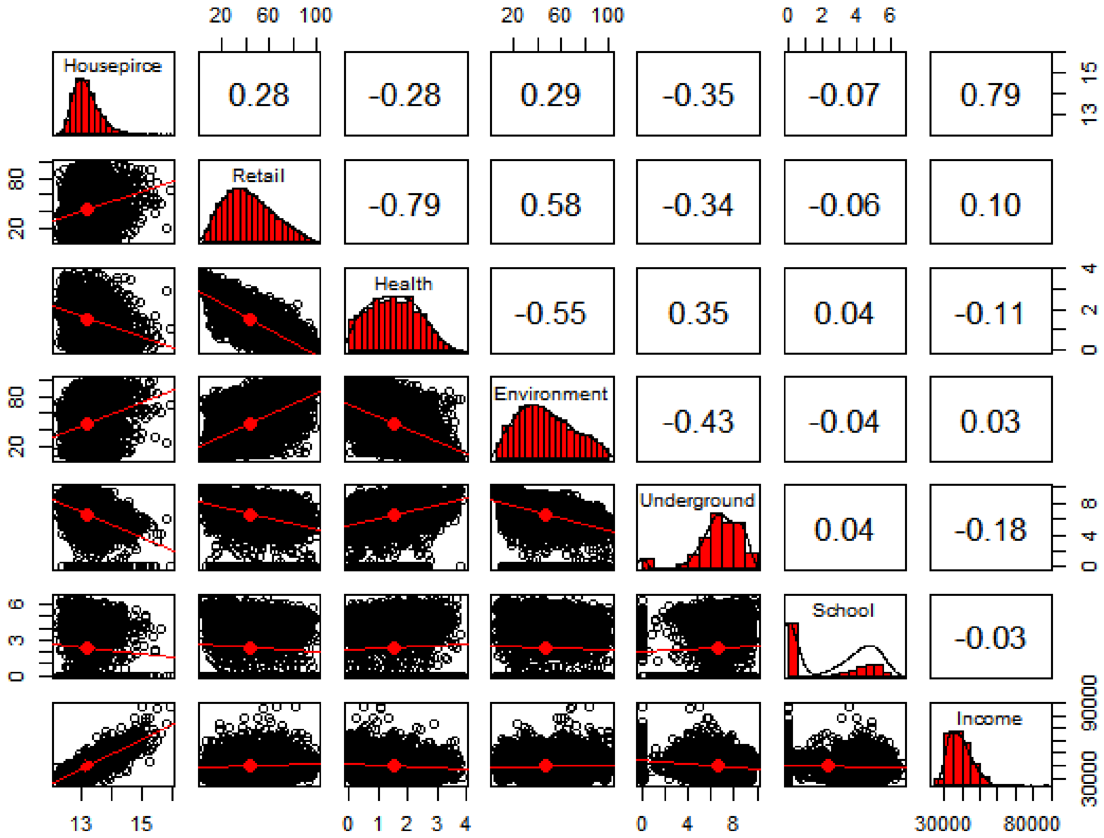


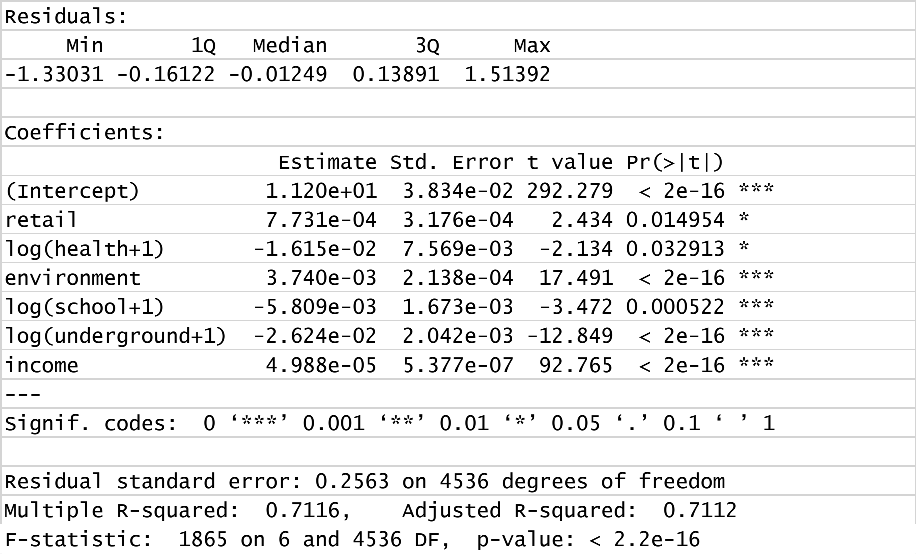
TABLE 5.1 Summary of Multicollinearity Analysis

Variables	retail	log(health+1)	environment	log(school+1)	log(underground+1)	income
VIF	2.96	2.80	1.72	1.01	1.30	1.04
Mean VIF	1.80					
Kappa	16.23					

Then, multicollinearity analysis was employed on the independent variables (see table 5.1). The relative low values of VIF and Kappa statistic indicate that there is no severe multicollinearity among the explanatory variables.

After that, the multiple linear regression was conducted. From the result below, it turns that all explanatory variables are statistically significant because their p-values are less than the common alpha value 0.05. The value of multiple R-squared (0.7116) indicates that the multiple linear regression explains 71.16% of the variance in the data. The statistically significant p-value in F-test also explains a significant amount of the variance in house price data.

FIGURE 5.2 Result of Multiple Linear Regression

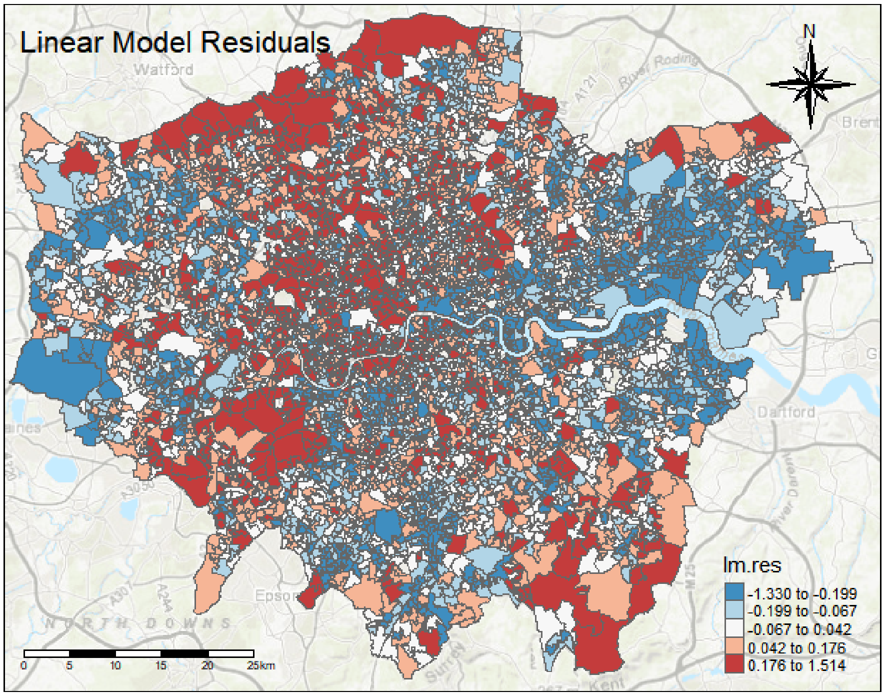


```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.33031 -0.16122 -0.01249  0.13891  1.51392 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.120e+01  3.834e-02 292.279  < 2e-16 ***
retail               7.731e-04  3.176e-04   2.434 0.014954 *  
log(health+1)       -1.615e-02  7.569e-03  -2.134 0.032913 *  
environment          3.740e-03  2.138e-04  17.491  < 2e-16 ***
log(school+1)       -5.809e-03  1.673e-03  -3.472 0.000522 ***
log(underground+1)  -2.624e-02  2.042e-03 -12.849  < 2e-16 ***
income               4.988e-05  5.377e-07  92.765  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2563 on 4536 degrees of freedom
Multiple R-squared:  0.7116,	Adjusted R-squared:  0.7112 
F-statistic:  1865 on 6 and 4536 DF,  p-value: < 2.2e-16
```

5.2 Spatial Autocorrelation in OLS Regression Errors

Figure 5.3 displays the spatial distribution of the residuals of the linear regression model. Spatial patterns can be observed as positive residuals (in red) tend to cluster together in the north and south-east where the house prices are overestimated by the fitted model, while negative residuals (in blue) are clustered together in the east and west London where the house prices are underestimated.

FIGURE 5.3 Residuals of Multiple Linear Regression Model



5.2.1 Moran’s I

To test the spatial autocorrelation in the residuals, the Moran test has been applied. It returns a value of 0.432 (expected value -0.001), with a p-value of 0.000, which shows that the residuals are significantly autocorrelated.

5.2.2 LM test for Spatial lag and spatial error dependence

The lag test returns a chi-squared statistic of 104.12 with a p-value of 0.000, while the error test returns a chi-squared statistic of 1219.8 with a p-value of 0.000. The high values of chi-square statistic along with the small p-values indicate that both regression models can be used as the regression models.

Indicated by the higher value of chi-squared statistic in spatial error test, spatial error model might be used as a better interpretation of the spatial autocorrelation.

TABLE 5.2 Lagrange Multiplier (LM) test for Spatial Dependence.

Test	Value	DF	p-Value
LM lag	104.12	1	0.0000
LM error	1219.8	1	0.0000

5.3 Spatial Regression Models

5.3.1 Global Spatial Regression Models

TABLE 5.3 Summary of Spatial Regression Models

	ρ (rho) / Lambda	Log-likelihood	AIC	LM-test
Spatial Lag Model	0.34791***	253.8128	-489.63	474.71***
Spatial Error Model	0.70226***	603.4894	**-1189**	
Spatial Durbin Model	0.38217***	396.0785	-762.16	

Note: *** represents the significance level is at 0.001.

Table 5.3 displays the summarized results of three spatial regression models. Four spatial parameters including rho, log-likelihood, AIC and LM-test were selected to compare the performance of different models.

In all models, the lag factor (rho/Lambda) is always statistically significant with a p-value less than 0.001, which indicates a high level of spatial autocorrelation presents in the house price data. Therefore, the result of multiple linear regression would be misleading. The LM test for spatial lag model reveals that significant spatial autocorrelation remains after the model has been fitted.

The log-likelihood and Akaike Information Criterion (AIC) were used to compare the suitability of each model. The higher value of log-likelihood and the lower negative value of AIC, the better the model. In the study, a larger log-likelihood and a larger negative AIC reveals that spatial error model has the best fitness compared with other two models. However, although this model performs best among three models, spatial autocorrelation still exists and cannot be fully explained by the model.

TABLE 5.4 Significant coefficients in spatial Durbin model

Variable	Non-lagged	Lagged	Significance
Retail	No	Yes	positive
log(Health+1)	Yes	Yes	positive/negative
Environment	No	No	
log(school+1)	Yes	Yes	negative/negative
log(underground+1)	No	No	
Income	Yes	Yes	positive/negative

Figure 5.4 shows the residuals of three spatial models, spatial lag model, spatial error model and spatial Durbin model respectively. In general, the house price in north London has been overestimated while east and west London have been underestimated.

FIGURE 5.4 Residuals from Spatial Lag Model, Spatial Error Model and Spatial Durbin Model

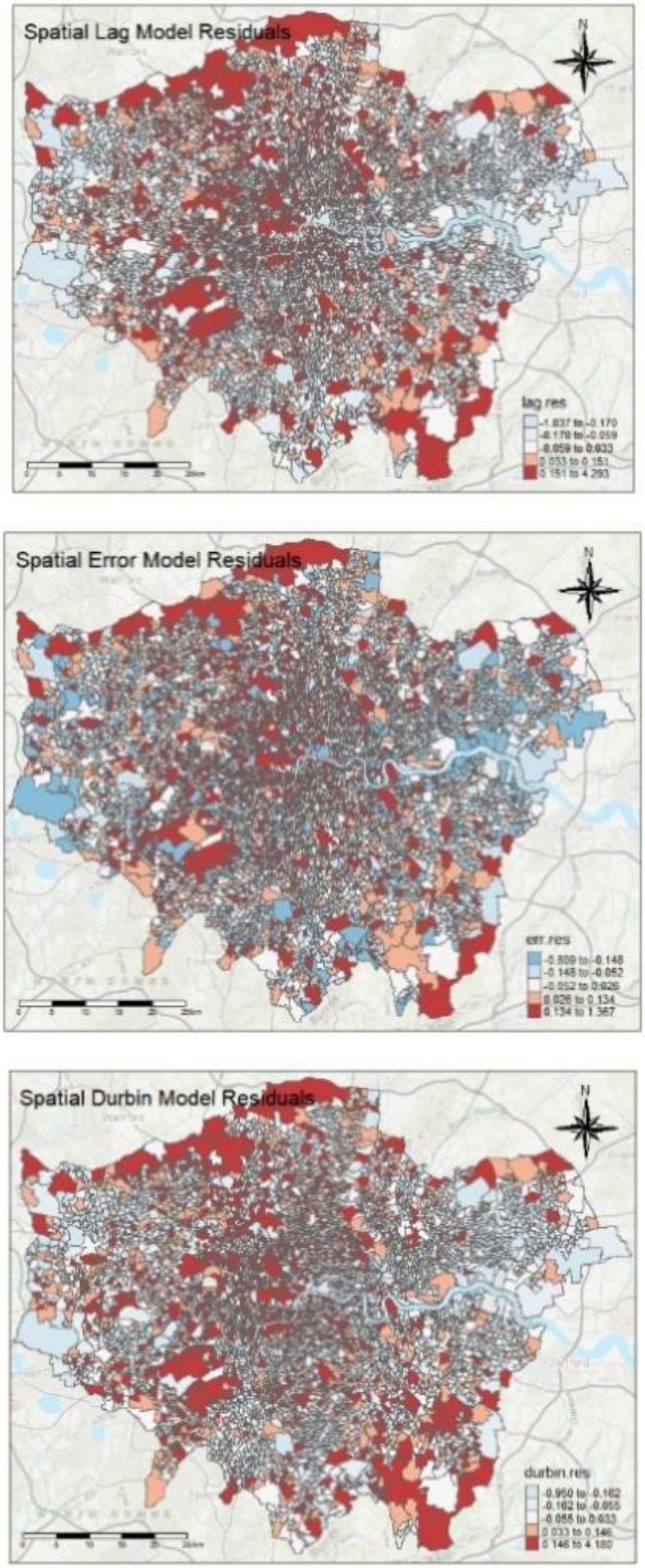
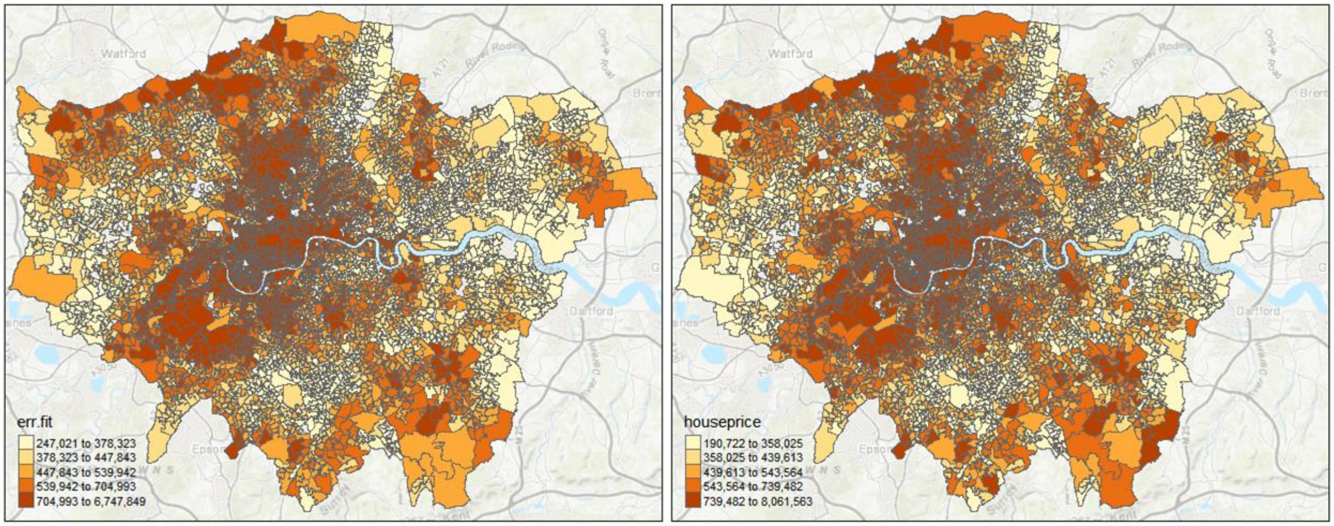


Figure 5.5 presents the fitted house price values calculated by spatial Durbin model and the actual house price values. By comparing two figures, it can be concluded that spatial Durbin model predicts the house price well. However, since the model is only a simplified version of the real world, residuals remain at some extent.

FIGURE 5.5 Fitted values from Spatial Error Model and Observed House Prices in London



5.3.2 Geographically Weighted Regression

In the study, a Gaussian kernel and a Bisquare kernel have been applied when fit the GWR model. Fixed bandwidths are chosen by leave-one-out cross-validation method for Gaussian and Bisquare kernel respectively.

Summarized in table 5.5, the values of optimal bandwidth are estimated to be 1243.81m and 3822.40m for the Gaussian kernel and the Bisquare kernel, which in turn result in different effective numbers of parameters (776.57 and 561.72). Also, a larger negative AIC (-2703.02) and a larger positive R2 (0.88) in Gaussian kernel indicates that Gaussian kernel performs better compared with Bisquare kernel.

The result of Moran test of GWR model quantifies the residuals for autocorrelation. A Moran's I value of 0.02 and a significant p-value (p<0.000) demonstrate significant spatial autocorrelation remains in the house price after fitted by a GWR model.

TABLE 5.5 Comparison between Gaussian kernel and Bisquare Kernel

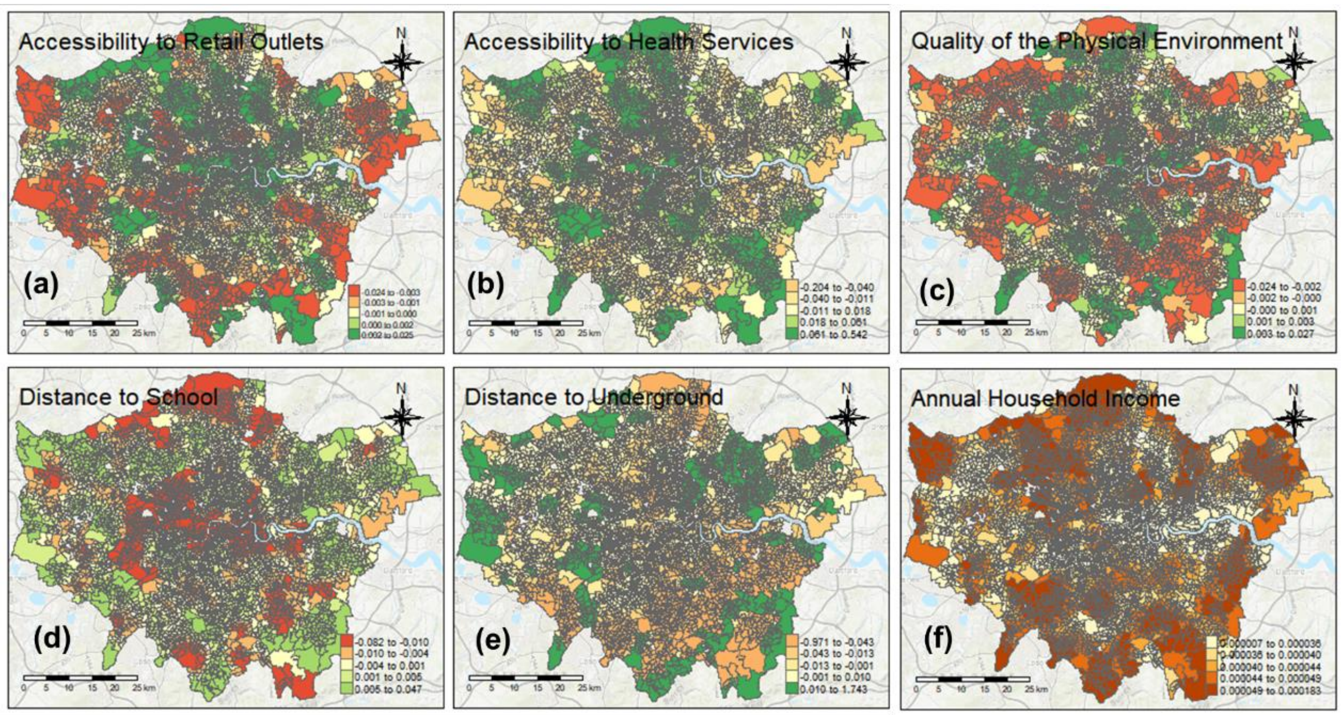
		Gaussian	**Bisquare**
GWR	Bandwidth	1243.81	3822.40
	Effective number of parameters	776.57	561.72
	AIC	-2703.02	-2388.83
	Quasi-global R2	0.88	0.87
Moran test	Residuals	37864.00	33159.00
	df	41483.00	37002.00
	p-value	< 2.2e-16	< 2.2e-16
	I	0.02	0.06

TABLE 5.6 Summary of GWR Coefficient Estimates at Data Points by Gauss Kernel.

	Min.	**1st Qu.**	**Median**	**3rd Qu.**	**Max.**	**Global**
X.Intercept.	-5.55e+00	1.13e+01	1.16e+01	1.20e+01	2.10e+01	11.20
retail	-2.40e-02	-2.51e-03	-6.15e-04	1.07e-03	2.51e-02	0.00
log.health	-2.04e-01	-3.17e-02	2.67e-03	4.77e-02	5.42e-01	-0.02
environment	-2.37e-02	-1.74e-03	4.54e-04	2.49e-03	2.67e-02	0.00
log.school.	-8.24e-02	-8.50e-03	-1.49e-03	3.92e-03	4.67e-02	-0.01
log.underground.	-9.71e-01	-3.03e-02	-6.90e-03	6.10e-03	1.74e+00	-0.03
income	7.30e-06	3.69e-05	4.18e-05	4.71e-05	1.83e-04	0.00

Table 5.6 indicates the distribution of the explanatory variables. The effects of different explanatory variables on the house price are also visualized in the distribution map in Figure 5.5.

FIGURE 5.6 Coefficient Maps



In addition, maps of the coefficients have been produced. As the figures shown, different dependent variables have different influences on the dependent variable across space. Expect for household income, both positive and negative impacts are found between different variables and house price data across space. Figure 4.6a indicates that the accessibility to retail outlets has a positive effect on the house price in east and west London, while it has a negative impact on the house price in centre and north the London. Similarly, map of the accessibility to health service (figure 5.6b) reveals a similar spatial pattern on the house price. In terms of the physical environment (figure 5.6c), it has a greater positive influence in the centre London while has a relatively negative influence in the peripheral area. Distances to both schools (figure 5.6d) and undergrounds (figure 5.6e) appear to have a greater influence on the house price in the peripheral area, while a less influence in the central area.

On the contrary, the coefficient map of mean household income (figure 5.6f) tells a different story. An overall positive impact on house price has been visualized by the map, with larger positive values in the central area of London and along the Times river.

DISCUSSIONS AND CONCLUSION

6.1 Spatial autocorrelation

The result of global spatial autocorrelation reveals that geospatial clustering exists in all the variables. The highest level of autocorrelation was observed for distance to underground (I =

0.987) and physical environment (I = 0.943), followed by income level and accessibility to retail outlets.

The high correlation in underground distance can be explained by the high density of undergrounds in central London, while low density in the peripheral area driven by the requirement of residences. Surprisingly, relatively weak autocorrelation has been found in distance to school. The reason might be explained from the distribution map, where schools are relatively evenly distributed across London.

Hotspots in central London have been identified through local spatial autocorrelation methods, while no coldspots have been observed. It is aligned with the traditional economic theory of bid-rent (Alonso, 1964) which refers to the relationship between house price and real estate demand and the distance from the central business district.

6.2 Multiple linear regression

Visualized by the scatter plot, strong associations exist between independent variables and dependent variables. As expected, household income is positively correlated with house price and distances to the underground and school are negatively correlated with house prices. The accessibly to the retail outlet and health service has opposite impacts on house prices. It is understandable that the higher accessibility to the retail outlet, the higher the house price is. However, a negative association has been found between health service accessibility and house price, which could be explained by the cost of health service infrastructure. In central London, fewer but larger health service institutes would be found, while more but smaller size institutes might be found in peripheral areas. In addition, a R-squared value of 0.7116 indicates that the linear model alone can explain most of the variability.

6.3 Spatial Regression Models

Global regression models have been employed to explain the overall spatial autocorrelation across the area. Among three models, spatial error model performs best in the study by comparing AIC and log-likelihood. However, spatial autocorrelation remains between fitted values and observed values.

To further explain the spatial heterogeneity across the study area, a local spatial regression model, GWR model, has been applied. A Quasi-global R2 of 0.88, compared with the aspatial multiple linear regression (R2=0.71), suggests that the GWR model performs well by taking spatial autocorrelation into account. The chosen explanatory variables can therefore explain most of the variability in house prices. However, from the coefficient map, it could be concluded that although most variability being explained, the presence of heterogeneity in variables affect its predictivity level of dependent variable and should be taken into consideration.

6.4 Limitations and Future Works

The pros and cons in the ‘local’ and the ‘global’ in spatial regression must be understood in the context of searching for broad generalisations versus local distinctions.

On the other hand, a major concern that must be raised when applying techniques to spatial data analysis refers to the issue of the geographical scale applied in the study. Spatially aggregated data as LSOA could draw major issues that should be noted.

The modifiable areal unit problem (MAUP) refers to questions whether the correlation coefficient is an appropriate statistic for use with aggregate spatial data. That is, some relationships can be relatively stable to data aggregation while others can be rather sensitive. Fotheringham et al (2000) outline two components of the MAUP problem:

1. The scale effect: different results can be obtained from the same analysis at different levels of spatial resolution.
2. The zoning effect: different results can be obtained through the regrouping of zones.

These issues suggest the importance of applying different aggregation to demonstrate the sensitivity of the results for different scales. This would provide a way to increase the robustness of the results and improve the confidence of the conclusions.

Finally, other statistical techniques other than regression could provide good opportunities for future works as for instance principal components analysis and discriminant analysis. The awareness that aspatial techniques could be transplanted to spatial problem would makes room for new spatial analysis possibilities.

REFERENCES

Alonso, W. (1964). *Location and land use: Toward a general theory of land rents*. Harvard University Press, Cambridge, MA.

Anselin, L. (2010). Spatial econometrics. Dordrecht: Kluwer Academic.

Belsley, D., Kuh, E., & Welsch, R. (2004). *Regression diagnostics*. Hoboken, N.J.: Wiley.

Brown, S., 1993. Retail location theory: evolution and evaluation. The International Review of Retail, Distribution and Consumer Research, 3(2), pp.185–229.

Butler, T., & Hamnett, C. (2012). Social geographic interpretations of housing spaces. *Sage handbook of housing studies,* 147-162.

Centre for Co-operation with European Economies in Transition (1993*). Glossary of industrial organisation economics and competition law*. Organization for Economic.

Cervero, R., Murphy, S., Ferrell, Christopher, et al., 2004. *Transit-oriented Development in the United States: Experiences, Challenges, and Prospects (TCRP 102)*. Transportation Research Board, Washington, DC.

Clark, W. A.V. and Davies Withers, S. (2007). Family migration and mobility sequences in the United States: Spatial mobility in the context of the life course. *Demographic Research (Max Planck Institute for Demographic Research)*17(20): 591–622.

Clark, W. A. (2012). Residential mobility and the housing market. *The Sage handbook of housing studies*, 66-83.

Darmofal, D. (2015). *Spatial analysis for the social sciences* (pp. 68-95). Cambridge: Cambridge University Press.

Debrezion, G., Pels, E., Rietveld, P., 2007. The impact of railway stations on residential and commercial property value: a meta-analysis. *J. Real Estate Financ. Econ*. 35, 161–180.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography: perspectives on spatial data analysis*. Sage.

Haworth, James (2018). *Spatial Analysis and GeoComputation: A tutorial guide*. Unpublished manuscript.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Krugman, P., & Wells, R. (2010). *Microeconomics*. New York: Worth Publishers.

Maclennan, D. (2012). *Understanding housing markets: Real progress or stalled agendas*. The SAGE Handbook of Housing Studies, 5-26.

Malpezzi S (2003)'Hedonic Pricing Models: A selective and applied review'. In T. In TO'Sullivan , S.and K.Gibb. *Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan*. Blackwell Science. Oxford, pp. 67–85.

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234. doi: 10.2307/143141

Tukey, J. (1977). *Exploratory data analysis*. Reading, Mass.: Addison-Wesley Pub. Co.

Ward, M., & Gleditsch, K. (2008). *Spatial regression models*. London: SAGE.

Zhang, L., & Yi, Y. (2018). What contributes to the rising house prices in Beijing? A decomposition approach. *Journal of Housing Economics*, 41, 72-84.