Automated Valuation Model: An Application in Japanese Rental Markets

Lewen Guo[*] Yuichiro Kawaguchi[†] William Cheung[‡]

October 23, 2019

Abstract

Researchers in real estate and practitioners in mass appraisal industry have developed Automated Valuation Models (AVMs) for estimating housing prices in different housing markets. However, few of them develop AVMs in rental market. By constructing a unique micro-level housing rental dataset, we build two residential rental AVMs for five districts in Tokyo using two different methods -OLS and Ordinary Kriging- in this study. The accuracy metrics from our training and test sets illustrate ambiguous preferences of which method performs better in terms of R-square and RMSE. Besides, we investigate the roles of spatial variables based on our baseline hedonic regression models. Spatial variables -latitudes, longitudes and distances to Tokyo Station- are crucial in determining the housing rents in Tokyo residential market. In addition, we conduct Kriging Error Decomposition analysis based on relative likelihood ratios and find that Ordinary Kriging method, despite its simplicity in interpolation of spatial data, may lead to information losses in modeling the rental functions as the method omits important housing attributes.

JEL classification: R3, R32.
Keywords: Automated Valuation Model, Hedonic Pricing Model, Kriging, Spatial Pricing

1 Introduction

As technology advances, the Automated Valuation Model (shorted as AVM hereinafter) is attracting increasingly attention among researchers in real estate and practitioners in

[*] Independent Researcher
[†] Graduate School of Business and Finance
[‡] Graduate School of Business and Finance

mass appraisal industry. We quote the definition from International Association of Assessing Officers (shorted as IAAO hereinafter) for AVM:

> *An automated valuation model (AVM) is a mathematically based computer software program that produces an estimate of market value based on market analysis of location, market conditions, and real estate characteristics from information that was previously and separately collected. The distinguishing feature of an AVM is that it is an estimate of market value produced through mathematical modeling. Credibility of an AVM is dependent on the data used and the skills of the modeler producing the AVM.*

The origin of the AVM research may date back to late 1980s where most of the researchers and practitioners use traditional models such as cost approach, income approach, comparable sales method as well as income approach as theoretic basis to build their AVMs or to conduct mass appraisal analysis. During that period, the multiple regression analysis (MRA) or the hedonic regression technics are widely used in their research(Foundation, 2003; D'Agostino, 1986). After IAAO set a standard for AVM in the industry in 2003, the AVM is officially distinguished from traditional appraisal method in which an appraiser physically inspects properties and relies more on experience and judgement to analyze the data and develop an estimate of market value. Not until 1990s with the availability of various statistics packages, have a substantial number of empirical works in related to mass appraisals and predictions of housing price been conducted. It is from then, in our understanding, AVM is being recognized as an important independent research field in real estate economics and appraisal society.

Existing literatures in this new field are heavily centered in building house-price-related AVMs (Faishal Ibrahim et al., 2005; García et al., 2008), while existing literature has seldom endeavored to construct rental AVM despite few have investigated and compared different methodologies which have been utilized on modeling the functions of housing rent (Brunauer et al., 2010; Djurdjevic et al., 2008; Löchl and Axhausen, 2010; Seya et al., 2011). The reasons, from our perspectives, which lead to the blank in this rent-related field may lie in the following facts. First, it is difficult to collect rental data since such information is often exclusively owned by big-brand housing brokerages. Without high-quality data to analyze, there's little space for researchers to conduct empirical studies. Second, the fact that no consensus has ever been reached in academia of how to construct a standardized model for modeling housing rents makes benchmarking a related study difficult which in turn disperses research interests among researchers.

Despite this blank in rental market research, there's an urgent need from the public that real estate market should be more transparent since real estate markets exhibit strong evidence that information asymmetry results in biased behaviors among market participants (Garmaise and Moskowitz, 2003). Therefore, one of our objectives in this study is to alleviate information asymmetry by constructing an AVM where the rental of any room in our study areas could be estimated based on our models. Market participants and the general public can refer to AVM-based information before they make decisions.

We select Japan as our study area because of the stability of rental prices after bubble burst (Deng et al., 2017), which provides a temporal-controlled environment for testing cross-sectional data.[1] Japan is a stable developed country which suffered from housing bubbles and bubble burst. However, its housing rents remain almost unchanged after 2008, it is estimated the average of appreciation rate in Tokyo is only 0.7% between 2008 and 2013 (Sta, 2013). The asking rents, in Japan, could also be interpreted as a proxy for the real transactional rents as landlords seldom altered their rents after posting the ads on brokerages (Seya et al., 2011).

Practically, we construct a micro-level dataset which includes both micro-level attributes of individual houses and geographical coordinates resembling famous Boston Housing Price dataset (Harrison Jr and Rubinfeld, 1978). Our dataset is unique and abundant in terms of the attributes of houses and geographic representativeness, which could be extended to a panel dataset if necessary.

We investigate the roles of spatial variables in Japanese rental markets. It is found that latitudes and longitudes in our sample areas are extremely crucial in determining the housing rents. The distance to Tokyo Station can explain part of variations in housing rents but cannot capture the spatial effects completely in our hedonic pricing models. Other structure variables exhibit similar characteristics as many hedonic pricing literatures summarized by Sirmans et al. (2005). We also build AVM based on both OLS and Ordinary Kriging Method, an interpolation method in geostatistics, which only takes geographic coordinates as inputs while could be used to estimate unknown rents. Our AVMs exhibit spatial heterogeneity across different regions in both OLS-based AVM and Kriging-based AVM.

We also compare the accuracy metrics of two different approaches -hedonic regression models (OLS) and Ordinary Kriging- in modeling AVM. The results indicate OLS performs better than Ordinary Kriging in 4 districts out of 5 in our sample. The failure to model rental function in Minato district warns us that simple methodology should be used with caution when constructing AVM. And splitting whole sample into training set and test set is almost mandatory to obtain appropriate interpretation.

This study has the following contributions. First, to our best knowledge, we are among the first to develop the AVM in rental market and the first few to investigate the roles of space in Japanese rental market. This fills the gap between the Japanese AVM research field and the leading AVM research in the world. Second, the AVMs that we have develop provide an important database for researchers and practitioners including appraisers, bankers, portfolio managers, government as well as the public. Third but not least, it is often cited in many papers (Cocco, 2000) that real estate market is incomplete, and information is not as fully transparent as stock market. By providing

[1] In this study, we can only obtain the cross-sectional data due to limitations in data-collecting process.

an open-source AVM to the public, we can to some extend alleviate the Information Asymmetry in real estate market which will lead to a more efficient market in future. In addition, our research is useful to provide more precise estimated imputed rents for owner-occupied dwellings in the Japanese System of National Accounts.

The rest of the thesis is arranged as follows. Section 2 reviews existing literatures which use different methodologies to model housing price (rent) functions. Section 3 introduces data constructions and estimation methods. Section 4 shows the results and discussions. Section 5 concludes our findings and contributions.

2 Literature Review

labelsec2 In terms of methodologies that researchers use to develop AVMs, existing papers can be divided into roughly three mainstreams—regression-based methods, AI-based methods (Zurada et al., 2011) and geostatistic-based methods. Regression-based methods can be further split into traditional assessment based on multiple regression analysis (MRA) methods or hedonic methods (ADAIR and McGREAL, 1988; Mark and Goldberg, 1988; Do and Grudnitski, 1992; Garrod and Willis, 1992; Faishal Ibrahim et al., 2005), spatial regression methods in consideration of spatial autocorrelation (Anselin, 1990; Kelejian and Robinson, 1998) and spatial heterogeneity (Fotheringham et al., 2003, 2015), and semiparametric model (Robinson, 1988; Anglin and Gencay, 1996; Clapp et al., 2002; Fan, 2018) and non-parametric model (McMillen and Redfearn, 2010). AI-based methods including famous Neural-Network(NN) based methods (Do and Grudnitski, 1992; Tay and Ho, 1992; Worzala et al., 1995; McCluskey et al., 1996; Borst and McCluskey, 1997; Rossini, 1999; Limsombunchai, 2004; García et al., 2008; Peterson and Flanagan, 2009; McCluskey et al., 2012), and Fuzzy Logic (Bagnoli and Smith, 1998; Aurélio Stumpf González and Torres Formoso, 2006; Guan et al., 2008) are becoming hotspots in the past two decades. And the popularity raised from keywords like ‘Big Data’, ‘Machine Learning’ among mass media and public makes such research field a cross-discipline of computer science and real estate economics. In contrast with AI-Method, Geostatistics, a branch of statistics which is widely used in spatial datasets, developed by South Africa engineering Krige (1951) and is used for predict probability distributions of ore grades for mining operations. However, it is not until recently, geostatistics-based method is applied in real estate research (Kuntz and Helbich, 2014).

Despite intensive research available with different approaches especially in model specifications, no widely-accepted consensus had ever been reached in academia in terms of how to construct AVM. Crone and Voith (1992) did an overall comparison of five regression models: three parametric and two non-parametric methods. They argue that parametric methods can yield higher accuracy, mean absolute error in this case, than non-parametric methods. In contrast, Brunauer et al. (2010) drew an opposite conclusion in their paper that additive mixed regression models (AMM), non-parametric methods, are shown superior results in contrast to parametric method. The inconsistency of which method performs better appears in the works of AI-based method as well. McCluskey et al. (2012) summarize predictive performance of several works where comparisons between NN methods and hedonic regression are made. However, they find that there is no clear explanations and evidence of why, in most of time, NN methods would have better predictive performances. In early work of Worzala et al. (1995), they argue the results of NN methods will vary depending on packages of software and the long run times of the same package.

The pros and cons of different methods are summarized as following: Hedonic price model and MRA model is easy to estimate, and coefficients are easy to interpret. But naïve OLS model cannot deal with non-linearity and all spatial problems. Spatial econometrics developed by Anselin (1990) and Kelejian and Robinson (1998) can deal with spatial autocorrelation by adding a pre-specified spatial matrix into the regression equations. Empirical works suggest that, if setting spatial weights correctly, predicting power could increase if compared with non-spatial regressions. However, as Martinetti and Geniaux (2017) argue in their paper, few literatures focus on how to address issues such as spatial heterogeneity, spatial autocorrelation, non-linearity and time invariance simultaneously. Geostatistics, which requires only geographical coordinates, is an easy-to-implement interpolation method. However, the kriging process, which excludes other possible explanatory variables may suffer information loss compared with regression methods and AI methods.

Non-linearity methodologies such as local polynomial regression or additive hedonic regression can allow researchers to evaluate data without assuming the functional form in advance between dependent variables and independent variables. By modelling non-parametrically through P-splines, non-linearity and time-invariant effects could be partially wiped out (Brunauer et al., 2010). But many of theoretic properties are difficult to understand and the computation requires a larger time as covariates increase (Opsomer et al., 1997).

Up to now, most of research focus on real estate housing price, few have investigated to construct AVM in rentals given the fact that rental markets are attached equal importance. Existing literatures which set predictive variable as rental or rental related data include the following: Djurdjevic et al. (2008) test hedonic model in Swiss market and find that their multilevel model has better predictive performance than segmented OLS models. Löchl and Axhausen (2010) compared 4 models by using Swiss asking rental data from a publicly available web site. They find that GWR model which intends to solve the spatial heterogeneity, though not perfectly and still suffered from spatial autocorrelation, can provide better predictive accuracy than OLS and SAR models. Seya et al. (2011) are the first to investigate empirical comparisons among spatial econometrics, spatial statistics (kriging) and semiparametric models. Their works provide us an intuition of how suitable Japanese data are for testing rental functions. Vienna paper, however, focuses on more on dealing with non-linearity and models time-trend through P-splines non-parametrically (Brunauer et al., 2010). These literatures provide us some sorts of benchmark to conduct empirical research. However, as far as we are concerned, no existing literature so far has modeled the housing rents by geostatistics, therefore our study is quite experimental and may require a number of tests and trials to obtain optimal results.

3 Data and Methodology

3.1 Dataset Construction

We construct a unique dataset for analysis from two different sources. One dataset that we use throughout this thesis is “asking-rent” dataset constructed from one of the largest Japanese real estate brokerage company- “Athome”. This dataset is composed of asking rents for residential houses whose owners put their ads on Websites through “Athome”, and micro-level housing attributes of the underlying houses for rent.

Since our interests are focused on constructing AVM for rental markets, which, from our perspectives, involves empirical analysis based on geographical data, we therefore obtain an additional “point” dataset to complement “asking-rent” dataset by adding the geographical coordinates of each house – latitudes and longitudes - into each single piece of data. The “point” dataset provided by Zenrin Corporation can fulfill our tasks in that this dataset collects geographical coordinates of all the residential buildings in a specified area, say, the whole residential houses in one specific district. The “point” dataset is updated until July 2017 and therefore it is suitable for our analysis.

We merge the “point” dataset with “asking-rent” dataset in the following way: First, for each piece of data in “asking-rent” dataset, we query the address from “point” dataset to match the address of “asking-rent” dataset.[2] Second, within a subsample of the same block, we match the floor number of the buildings in two datasets. Third, within the same block and the underlying building with the same floor number, we use the python package “fuzzywuzzy” to match the names of the buildings in two datasets.[3] We get “one-to-one” matching results and select the matched data whose fuzzy ratio is larger than 80. After data cleaning process, the final dataset resembles famous Boston Housing Price dataset (Harrison Jr and Rubinfeld, 1978) where both micro-level of housing attributes and geographical coordinates are available for analysis.

[2] Since the address in “asking-rent” dataset is incomplete, we can only match the address of two datasets in block level. We then obtain a subsample of “point” dataset in the same block.

[3] The package “fuzzywuzzy” is a string matching toolbox in python using Levenshtein Distance to calculate the differences between sequences. It can output a ratio indicating similarity between two strings. A quick example is, if we put two sentences ”fuzzy wuzzy was a bear”, ”wuzzy fuzzy was a bear” into the fuzz.ratio function of “fuzzywuzzy” package, the output ratio is 91. The code for matching the name of the buildings is available upon request.

For the purposes of data visualization, we obtain the mapping data and GIS-related data from Geospatial Information Authority of Japan. The mapping data includes base-map, road-map, and topographic map of district-level in Japan. Figure 1 shows the geographic plot of our final sample data.

[Figure 1]

3.2 Variables Description

We obtain 10892 pieces of data from the period of Jan 2018 to March 2018 within five districts in Tokyo.[4] We have 24 variables for each piece of data in our final dataset. The dependent variables are, *Unitrent* and *Logunitrent* representing rent per square meter and logarithm of rent per square meter, respectively. The numerical attributes for each piece of data are named and defined as the following: latitude and longitude of the house, *Latitude* and *Longitude*; floor area of the house for rent, *Floorarea*; number of rooms, *Room*; number of living rooms, *Living*; number of kitchens, *Kitchen*; number of storage room, *Storage*; number of dinner room, *Dinner*, management fee, *ManagementFee*; deposit fee, *Shikikin*; gratuity fee, *Reikin*; security money, *Hosyoukin*; floor number of the house, *Floornum*; total floor of the building where the underlying house lies in, *Totalfloor*; distance in kilometers to Tokyo Station, *DistToTokyo*; the time (in minutes) to walk to nearest station, *Accessibility*; the age of the building, *Age*; and the floor number of the house relative to the total floor of the underlying building, *RelativeFloor*. Dummy variables in our dataset are *DummyPark*, *DummyBikepark* and *DummyMaterial*, indicating whether the house has a car-parking lot, bike-parking lot and material used for construction is concrete.[5] We also construct district dummy variables for each district and in total we have five dummy variables: *DummyShinjuku*, *DummyKoutou*, *DummyMinato*, *DummySumida*, and *DummySetagaya*. We also include one categorical variable in our dataset. *Orientation* is a variable indicating the positioning of a house in relation to seasonal variations in the sun's path. In our dataset, there are eight different directions which indicate the different positioning of a house. Based on previous survey literatures (Green et al., 2003; Sirmans et al., 2005), we use the following terminologies to describe different types of variables:

[4] In this study, we test the data of five districts in Tokyo which are Shinjuku, Minato, Sumida, Koutou, Setagaya. Our final dataset is cross-sectional since we delete the duplicated observations. Within our collection period, the values of the attributes and the rent for each underlying house are consistent in our data-cleaning process and there are no two different values for the same attributes of each underlying house.

[5] The materials used for construction are roughly divided into two types – concrete and wood- in Japan. According to Statistics Japan Residential House and Land survey conducted in 2013, among 52.1 million houses in Japan 21.99 million (42.2%) are concrete-made houses, while this number for wooden houses is 30.11 million (57.8%).

- (Dummy) Structural Variables: *Floorarea*, *Room*, *Living*, *Kitchen*, *Storage*, *Dinner*, *ManagementFee*, *Shikikin*, *Reikin*, *Hosyoukin*, *Floornum*, *TotalFloor*, *RelativeFloor*, *Accessibility*, *Age*, *DummyPark*, *DummyBikepark* and *DummyMaterial*.
- Spatial Variables: *Latitude*, *Longitude* and *DistToTokyo*.
- District Dummy Variables: *DummyShinjuku*, *DummyKoutou*, *DummyMinato*, *DummySumida*, and *DummySetagaya*.

3.3 TrainTest Splitting and Hypothetical Dataset

We separate the full dataset into training set and test set. By using the functions from scikit-learn, we split the full dataset randomly based on an 80% to 20% split ratio.[6] The small-size sample data – approximately 10000 pieces in total – makes us determine 80% to 20% split ratio as our benchmark which enables us to obtain enough observations to perform accuracy test on test set.

For the purposes of predicting the rental price for all the properties within a specific area based on regression models, it is required that the attributes of unknown properties as the inputs of our AVM. However, actual house-specific data cannot be obtained since "point" dataset only has the building-level attributes but not house-level attributes. We, therefore, construct a hypothetical dataset based on "asking-rent" dataset and "point" dataset. For each piece of data in "point" dataset, we average the attributes of ten nearest-neighbors from "asking-rent" dataset based on their great circle distances and assign the mean value of these attributes, except for latitude and longitude, as the hypothetical values for the attributes of each single piece of data. We must admit here that these hypothetical data are only used for calculation of regression-based AVM since they are smoothed hypothetical data which cannot represent the true values.

3.4 Model Specifications and Methodologies

3.4.1 Hedonic Pricing Model

The origin of hedonic model can be traced back to late 1930s when Court (1939) first developed a hedonic pricing index for the automobile industry. However, it was not until late 1960s and early 1970s after Lancaster (1966) developed a utility-generating microeconomic theory and Rosen (1974) constructed an equilibrium hedonic pricing model based on buyer and seller choices that researchers started to conduct empirical studies for estimating the functions of housing price. The Hedonic Pricing Model views

[6] Scikit-learn is a python package for machine-learning. The function used to split train test data is sklearn.model_selection.train_test_split.

the value of a house is contributed by the satisfaction that users gain from each separate attribute of the house. Due to its straightforward interpretation and simplicity in calculation, the Hedonic Pricing Model attracts attention from people in real estate appraisal as well. However, in real estate appraisal, the terminology used for Hedonic Regression Model is "Multiple Regression Analysis" though the statistics tool -OLS- that both economists and appraisers is the same. Despite inconsistency between two different research areas, the basic regression equation is summarized by Sirmans et al. (2005) and generally takes the following form:

$$Price = F(PhysicalCharacteristics, OtherFactors)$$

Where the Physical Characteristics are typically, the physical attributes of a house, and Other Factors are the external factors which may, affect the housing price such as level of income within the area, GDP per capita, crime rate and so forth.

Following previous hedonic rental literature (Djurdjevic et al., 2008; Löchl and Axhausen, 2010; Seya et al., 2011), we construct the hedonic regression model as our baseline regression and could be written as follows:

$$lnP = \alpha + \beta X + \epsilon \tag{1}$$

Where lnP is the logarithm of a vector of asking rents (N$\times$1, where N is the number of observations), α is the constant, β is a vector of coefficients (N$\times$1) and X is a matrix of house attributes (N$\times$K, where K is the number of attributes). In our baseline hedonic regression models, we regress logarthmns asking rent per square meter, *Logunitrent* on both structural variables, and spatial variables. Structural variables are *Floorarea*, *Room*, *Living*, *Kitchen*, *Storage*, *Dinner*, *ManagementFee*, *Shikikin*, *Reikin*, *Hosyoukin*, *TotalFloor*, *Accessibility*, *Age*, *RelativeFloor*, *DummyPark*, *DummyBikepark*, and *DummyMaterial*, representing the structural characteristics of each house. Spatial variables are *Latitude*, *Longitude*, capturing the geographical location of each house. In addition to the baseline regression, we also include the *DistToTokyo* and district dummy variables in our regression to test what kind of roles can spatial variables play in determining the rental prices.

3.4.2 Ordinary Kriging

Ordinary kriging is used to estimate the unknown true values of points where there are no observable sample available. The values of unknown points are estimated by linear combinations (weights) of the values of known-value points. By constructing a fitted covariance function or a semi-variogram function, the variance of the error, which is the difference between the true values and estimated values of unknown-value points, can be minimized conditioning on the following assumptions: First, the weights for calculating the unknown-value points should be added up to 1. Second, the variance of the values of both known and unknown points should be the same within the study area. Third, for interpolating the values of unknown-value points, the covariance between any pair of the points should decrease as the distance of the pair increases. In mathematics, the ordinary kriging problem can be written as:

$$\begin{gathered}\min \widehat{\sigma}^2_{V_0} = Var(\widehat{V_0} - V_0) \\ Subject\ to\ \widehat{V_0} = \sum_{i=1}^{n} w_i V_i \\ \sum_{i=1}^{n} w_i = 1\end{gathered} \tag{2}$$

Where $\widehat{V_0}$ is the estimated value of the unknown-value point, V_0 is the true value of the unknown-value point, V_i is the value of observed sample i, w_i is the weight of value-observed sample i used to calculate the estimated value of unknown value point.

By rearranging equation (2) with covariance expressions, equation (2) can be written as:

$$\widehat{\sigma}^2_{V_0} = \widehat{\sigma}^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} \widehat{C}_{ij} - 2\sum_{i=1}^{n} w_i \widehat{C}_{i0} \tag{3}$$

Where $\widehat{\sigma}^2$ is the variance of the values of all points, $\widehat{C}_{ij}$ is the estimated covariance between point i and j, $\widehat{C}_{i0}$ is the covariance between point i and unknown-value point 0.

By adding Lagrange Parameter, $2\mu \sum_{i=1}^{n} w_i - 1$ to the right-hand side of equation (3), and taking partial derivatives with respect to the weights and μ, we can obtain n equations with regards to n weights and μ. Together with the constrain of equation (2) the ordinary kriging system (Isaaks and Srivastava, 1989) can be written as the following n+1 equations:

$$\sum_{j=1}^{n} w_j \widehat{C}_{ij} + \mu = \widehat{C}_{i0} \quad \forall i = 1, 2, 3..., n \tag{4}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{5}$$

This system of equations (Isaaks and Srivastava, 1989) could be written in matrix notation as:

$$\boldsymbol{C} \quad \times \quad \boldsymbol{w} \quad = \boldsymbol{D} \tag{6}$$

$$\begin{bmatrix} \widehat{\boldsymbol{C}}_{11} & \cdots & \widehat{\boldsymbol{C}}_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \widehat{\boldsymbol{C}}_{n1} & \cdots & \widehat{\boldsymbol{C}}_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{C}}_{10} \\ \vdots \\ \widehat{\boldsymbol{C}}_{n0} \\ 1 \end{bmatrix} \tag{7}$$

Solving the above matrix problem, we can obtain the solutions for the weights written in matrix notation as:

$$\boldsymbol{w} = \boldsymbol{C}^{-1} \times \boldsymbol{D} \tag{8}$$

So far, we have obtained the solutions for calculating the weights, however, we still need the function for calculating the covariance between points of unknown-value and the sample. Following the (Isaaks and Srivastava, 1989) we define the semi-variogram as follows:

$$\gamma(h) = \frac{1}{2}\mathrm{E}[(V_i - V_j)^2], ||i(x_i, y_i) - j(x_j, y_j) = h|| \tag{9}$$

where semi-variogram, $\gamma(h)$ is a function of distance h, which is calculated by certain kind of distance between point i and point j, typically great circle distance measured by latitude and longitude.

If we follow our assumptions that the estimated variance of the values of both known and unknown points should be the same, we can rearrange the equation (9) as follows:

$$\gamma(h) = \widehat{\sigma}^2 - C(h), \quad C(h) = \widehat{C}_{ij} \tag{10}$$

where $\widehat{\sigma}^2$ is the estimated variance of the values of both known and unknown points, C(h) is the covariance between point i and point j where the distance h is calculated by the distance function in equation (9).

In many geostatistics books and literature (Isaaks and Srivastava, 1989; Cressie, 1992), equation (10) which is derived from equation (9) is often written together as follows:

$$\gamma(h) = \sigma^2 - C(h) \tag{11}$$

$$C(h) = \mathrm{Cov}(V_i, V_j), \quad ||i(x_i, y_i) - j(x_j, y_j) = h|| \tag{12}$$

In geostatistics, equation (11) is often called as “Theoretical Semi-variogram” and usually takes a variety of different functions including spherical function, gaussian function, exponential function, power function, linear function and so forth (Isaaks and Srivastava, 1989; Cressie, 1992; Liu and Maghsoodloo, 2009), Figure 2 is an illustration of semi-variogram and co-variogram of spherical model which is often used in literature as a standard model for empirical testing (Cressie, 1992; Chica Olmo, 1995; Basu and Thibodeau, 1998; Gillen et al., 2001; Kuntz and Helbich, 2014). The semi-variogram function for spherical model can be written as:

$$\gamma(h; a, s, r) = \begin{cases} a + (s - a)(\frac{3h}{2r} - \frac{h^3}{2r^3}), & 0 \leq h \leq r \\ s, & h > r \end{cases} \tag{13}$$

[Figure 2]

Following the methods of Schabenberger and Gotway (2004); Kuntz and Helbich (2014), we set the empirical semi-variogram as:

$$\widehat{\gamma}(h_k) = \frac{1}{2N(h_k)} \sum_{S_i,S_j \in N(h_k)} [\widehat{V}(S_i) - \widehat{V}(S_j)]^2 \tag{14}$$

Where $N(h_k)$ is the number of pairs for interval h_k, .$k = 1, 2, 3, \vdots n$ is the number of lags (or bins) for determining the number of intervals in empirical semi-variogram. h_k is the lag distance by taking the average of distances of all pairs within the lag k. The $\widehat{\gamma}(h_k)$ is determined by taking the average of squared-difference pairs, $[\widehat{V}(S_i) - \widehat{V}(S_j)]^2$.

For empirical fitting strategies, we choose weighted least squares method proposed by Cressie (1992); Kuntz and Helbich (2014) and the general weighted problem can be written as:

$$\min_{(r,s,a)} \sum_{k=1}^{n} W_k[\widehat{\gamma}(h_k) - \gamma(h_k; a, s, r)]^2 \tag{15}$$

By putting more weights on those bins which have more observations and less weights on bins with less observations, we can write the weights W_k as follows:

$$W_k = \frac{N(h_k)}{\gamma(h_k; a, s, r)}, \quad k = 1, 2, 3, \ldots n \tag{16}$$

After we obtain the best fitted parameters $\widehat{a}$, $\widehat{s}$, $\widehat{r}$, we then get back to equation (7) where covariance between any unknow-value point and value-observed sample can be calculated by . Since the parameter of lags, is a parameter pre-specified by researchers, in this thesis, we use the grid search algorithm to find the best lag which produces the highest accuracy metrics when using our training set.[7]

3.5 Estimation Metrics for Model Accuracy

We use two regression metrics -Coefficient of Determination (R-Squared) and Root Mean Squared Error (RMSE) for estimating our model accuracy. The two metrics are defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{17}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2}{n}} \tag{18}$$

[7] Scikit-learn provides grid search algorithms where you can specify parameters as input and return you a score based on different accuracy metrics such as R-Squared, Mean Squared Error and so forth.

Where $\widehat{Y_i}$ is the predicted values for observation i, Y_i is the true value for observation i, $\overline{Y}$ is the mean of all the observations, and n is the number of the observations in the model.

As criticized by Alexander et al. (2015), other accuracy measurements should be reported since R-Squared sometimes can lead to inappropriate interpretation of model fitness.[8] We, therefore, following many AVM and model-fitting-related literatures (Limsombunchai, 2004; Faishal Ibrahim et al., 2005; McCluskey et al., 2012; Bency et al., 2017), include Root Mean Squared Error as alternative accuracy metric other than R-Squared.

4 Results and Discussions

4.1 Summary Statistics

Table I summarizes the statistics for the variables in our whole sample data except for variable Orientation since it is a categorical variable. To illustrate our sample representativeness, we follow Djurdjevic et al. (2008) and summarize geographic representativity in Table II. Table II reports the geographic differences between our "point" dataset and sample dataset. Table II illustrates that Setagaya district is under-represented in our sample compared with "point" dataset while Minato district is over-represented. The discrepancy in sample representativeness between the "point" dataset and sample dataset may arise from the fact that Setagaya district has more single-family and owner-occupied houses than Minato district which are all included in "point" dataset but never appear in sample dataset. Our sample data is overall sufficiently representative for the rental market in these five districts.

[Table I]

[Table II]

Figure 3 illustrates the kernel distributions of *Unitrent* for five different districts. Among these five districts, Minato district has the highest *Unitrent* compared with other four districts.

[Figure 3]

[8] Many statistics softwares including R, statsmodels in Python and excel report a different R-Square ratio based on the following formula: $R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(\overline{Y} - 0)^2}$.

4.2 Results and Interpretations

4.2.1 Regression Results

Table III reports OLS results of five rental hedonic models for different districts in Tokyo. In Table III, we regress logarithms of asking-rent per square meter, *Logunitrent* on structure variables: *Floorarea* (in logarithm), *Room*, *Living*, *Kitchen*, *Storage*, *Dinner*, *ManagementFee*, *Shikikin*, *Hosyoukin*, *Reikin*, *TotalFloor*, *Accessibility*, *Age*, and *RelativeFloor*; dummy structural variables: *DummyMaterial*, *DummyPark*, and *DummyBikePark*; and spatial variables: *Latitude* and *Longitude* (Details of variable constructions are discussed in Section 3.2). Column (1) to Column (5) illustrate regression results for different districts. Consistent with the results of previous studies summarized by Sirmans, Macpherson and Zietz (2005) and rental-related hedonic pricing model literature (Djurdjevic et al., 2008; Seya et al., 2011), the coefficients of *Age* in our regression models are all negative and significant at 1% level across all the districts. Among other (dummy) variables, the coefficients of structural variables such as *Room*, *Living*, *Storage*, *Dinner* are positively significant in some districts but exhibit no significance in some districts as well. These internal features of houses in our sample area resemble similar characteristics of previous studies where the coefficients of these structural variables exhibit ambiguous signs and levels of significance (Sirmans et al., 2005). In terms of fee-related variables, the coefficients of *ManagementFee*, *Hosyoukin* and *Reikin* are inconsistent across different districts while the coefficients of *Shikikin* (amount of deposit) are all positive within our sample areas. These finding are intuitive as higher deposits are generally associated with higher total rent and these effects also exist with *Logunitrent*, rent per square meter as well. The coefficients of *Floorarea*, are all negative and significant at 1% across all the districts. These findings are also consistent with our intuitions that the rent per square meter should be negatively correlated with the floor area of a house. Since our approach is to model the rent per square meter instead of total rent, our findings are inconsistent with previous studies, most of which regressing total rents or prices on different housing attributes (Sirmans et al., 2005). Perhaps the most interesting findings, which may probably only be observed in Japan, are that both *TotalFloor* and *RelativeFloor* show positively significant coefficients in our baseline regression implying that buildings with higher total floor numbers and the houses with higher floor numbers in the building will have larger unit rents. These findings are intuitive in that tenants could gain more satisfactions from broader viewshed and therefore higher rent will be requested when living in a skyscraper compared with living in a low-rise building.

The external feature variables -*DummyPark* and *DummyBikePark*- have different characters in our regression models where *DummyPark* exhibits strong and positive correlation with *Logunitrent* while the coefficients of *DummyBikePark* are almost not significant across different districts. Spatial variables, *Latitude* and *Longitude*, which capture the relative position on space are all significant across five study districts. Our findings are showing evidence against previous study in Japan (Deng et al., 2017) in that both latitude and longitude in our model are significant while the results of Deng et al. (2017) illustrate that only latitude matters in his hedonic pricing model.

[Table III]

As an interpretation of our spatial variables in baseline regression results, we illustrate the geographical plot of our sample data including the point of Tokyo Station in Figure 4. The coefficients of latitude and longitude in our baseline regression results, except for the latitude of Minato district, are indicating that the rents per square meter are related to the relative positions to Tokyo station. Take the coefficients of Shinjuku as an example. If our assumption that the relative positions to Tokyo Station matter in determining the housing rents holds, the coefficient of latitude in Shinjuku regression model should be negative while the coefficient of longitude should be positive since Shinjuku district is located at upper left (Northwest) relative to Tokyo Station. Table IV compares the empirical signs of coefficients of latitude and longitude from regression models and hypothetical signs of the coefficients if our assumption holds true.

[Figure 4]

[Table IV]

To test how the distance to Tokyo Station can affect the housing rent in these five districts, we add the variable *DistToTokyo* to our baseline results. Table V shows that after adding the structural variable *DistToTokyo*, coefficients of some of the spatial variables in our 5 study districts, *Latitude* and *Longitude* are becoming not significant in contrast with baseline regression results as reported in Table III. However, on the other hand, the coefficients for *DistToTokyo* are all negatively significant except for Shinjuku district indicating that the distance to Tokyo Station explains some variations of rent per square meter in our study districts. However, this variable, *DistToTokyo*, alone cannot fully explain the true spatial effects between the rents and the model since the coefficients of latitude and longitude exhibit ambiguous level of significance. One possible reason that the coefficient of *DistToTokyo* is not significant may result from the fact that Shinjuku Station is possible another functional-equivalent point for Shinjuku district as it is the transportation center and commercial center especially for the west part of Tokyo. The Shinjuku StWWation may alleviate the needs for accessibility to Tokyo Station which possibly weakens the level of significance of the variable *DistToTokyo*.

[Table V]

We further documented regression results in Table VI adding district dummies in our regression equations. We regress *Logunitrent* on the same structural variables, spatial variables as used in Table V but adding district dummies into regression models. After adding district dummies and testing the whole training set, as shown in Table VI Column (1), the coefficients for *Latitude* and *Longitude* are still significant at 1% level and the coefficient of distance to Tokyo Station, *DisToTokyo* is still negatively significant. Spatial variables: *Latitude*, *Longitude* and *DistToTokyo* are all capturing the effects of space in this model. In addition, the coefficients of district dummies indicate the price level of different districts. From Column (1) and the Row of DummyMinato in Table VI, if a house is located at Minato district, it would probably have the highest rent per square meter compared with a similar house in other four districts. The results are consistent with the kernel distributions of *Unitrent* in Figure 4.

[Table VI]

4.2.2 Discussions of OLS Methodology

As discussed in Section 2, hedonic pricing model (OLS) is often criticized by a variety of researchers and academicians since the method assumes many unrealistic assumptions which can never hold true when fitting real data. In our case, the Tobler's first law of geography "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970) will be violated if we use OLS method as it assumes homoskedasticity in error terms while spatial-related data usually suffer from spatial autocorrelation (Anselin, 1990; Basile et al., 2014). However, the reason we choose to report OLS results in this thesis is to investigate our research questions that what roles can spatial variables play in rental market and provide evidence of how important geographic coordinates are in determining the housing rents. The results of naïve OLS could provide insights of what kinds of other methodologies which may yield more appropriate results should be used for modeling a more precise AVM.

4.3 Ordinary Kriging Results and Discussions

4.3.1 Ordinary Kriging Results

As discussed in detail in Section 2 and Section 3.4, Ordinary Kriging is a statistic tool often used by geologists to predict the values of coordinated points where the true values are unknown. Therefore, we only take the variables of latitudes, longitudes and the corresponding rents per square meter out of our training sample to build Ordinary Kriging models. Since the Ordinary Kriging models are sensitive to their parameters -sill, nugget, range and lags- in our spherical models, we select the best-fitted model for each district based on the following procedures:

① Run a for loop on the number of lags k
② Calculate the fitted parameters based on equation (15) and equation (16) for empirical semi-variograms: $\widehat{s}$ (sill), $\widehat{r}$ (range) and $\widehat{a}$ (nugget)
③ Execute the Ordinary Kriging Systems based on training set and test set, separately
④ Check the accuracy metrics of both training set and test set
⑤ Check whether the predicted values are clustered or not [9]
⑥ Select the appropriate models based on comprehensive fitness according to procedure ④ and ⑤

The selections of the number of loops to run in terms of lags k are quite subjective and difficult to determine. In our case, we set loop number as 1000 for one time to search the best-fitted model. Fortunately, except for Koutou district, we figure out the optimal models for the other four districts within one loop. Table VII shows the best-fitted parameters for Ordinary Kriging models of five different districts. Figure 5 illustrate the semi-variogram plots for these five districts as well. From these semi-variogram figures, we find the parameters for varied a lot across different districts. These findings provide the evidence that strong spatial heterogeneity exist in rent prices across these five districts. Our findings are consistent with previous studies in real estate that housing-related data exhibit spatial heterogeneity (Anselin, 2003; Goodman and Thibodeau, 2003; Khalid, 2015). The model fitness for different districts are distinctive as the parameters are varied.

[Table VII]

[Figure 5]

4.3.2 Ordinary Kriging Versus OLS

Table VIII] reports the accuracy metrics from both methods - Ordinary Kriging and OLS - for different districts. Both training set and test set are included in Table VIII. Ordinary Kriging Method yields higher accuracy in training set than test set indicating potential overfitting problems while OLS yields similar accuracy in both training set and test set. Comparing both methods, OLS performs better in test set for Shinjuku

[9] The defects of using spherical semi-variogram for ordinary kriging is that the value of $\widehat{r}$(range) is very important in determining the correctness of model fitting. If the range r is too small, most of the values of semi-variogram will be s rather than $a + (s - a)(\frac{3h}{2r} - \frac{h^3}{2r^3})$. This will lead to a terrible situation where matrix $\boldsymbol{C}$ and $\boldsymbol{D}$ from equation (7) would be the same for most of distance (h), consequently, resulting in the kriged values clustered in one specific value. To monitor the parameter tuning process, we print the mode of the kriged values both on training set and test set in our program to detect the problems of too small range.

district, Koutou District, Sumida district and Setagaya district but fails to capture the relationship between attributes of house and housing rents in Minato district. On the other hand, Ordinary Kriging method which only takes geographic coordinates as model inputs, can explain 50% to 70% variations for output, in our case, rent per square meter. Our findings suggest both methods could be used for predictions while it is difficult to identify which method is superior to the other.

[Table VIII]

4.4 AVM Constructions

4.4.1 OLS-based AVM

Figure 6 illustrates the OLS-based AVM. This map shows the rent per square meter of any building in five districts of Tokyo. The colors of heatmap indicate the values of rent per square meter, which are simulated by baseline OLS regression models in Table III. The data used for predictions is our hypothetical data (details of construction are in Section 3.3).

[Figure 6]

4.4.2 Kriging-based AVM

Figure 7 illustrates the Kriging-based AVM. This map is the same as the map in Figure 6 except for the values of rent per square meter are calculated by Ordinary Kriging method. The difference between the calculation processes of Ordinary Kriging and OLS is that only geographic coordinates are selected as inputs for Ordinary Kriging model while OLS requires exact the same housing attributes as inputs as the models in Table III. Compared with the map in Figure 6, the map in Figure 7 exhibits stronger effects of spatial heterogeneity in some of the districts -especially in Koutou district and Shinjuku district. However, on the other hand, two methods yield similar heatmaps in Minato district where rents per square meter are clustered at relatively high values. Overall, there are no apparent discrepancies in terms of general trends between two methods. To illustrate more precisely of our AVMs, we show details of Minato district in Figure 8.

[Figure 7]

[Figure 8]

5 Conclusions

By constructing a unique micro-level housing rental dataset from two different datasets, we build two different AVMs for five districts in Tokyo based on two different methods and compare the accuracy metrics of these two methods. Strong spatial heterogeneities across different districts have been detected from both OLS-based and Kriging-based AVMs.

We investigate simple hedonic pricing models in rental market of our study areas-five districts in Tokyo. Structural variables such as age of the house, number of rooms, whether the house has a parking lot and so forth exhibit similar characteristics as previous studies. However, inconsistent with existing literatures spatial variables such as latitude, longitude and distance to Tokyo station, show explanatory power in determining housing rents in our study areas. In our baseline hedonic regressions, the coefficients of latitude and longitude are all significant across different districts. However, if adding the distance to Tokyo station to our baseline regression results, the marginal impact on rent per square from latitude and longitude would decrease while the distance to Tokyo station is negatively correlated with housing rents. These results indicate both geographical coordinates and distance to Tokyo station could marginally capture roles of space and spatial variables are of great importance to determine the housing rents.

Furthermore, we conduct Kriging Error Decomposition analysis based on relative likelihood ratios to investigate how much information loss would be if using Ordinary Kriging method instead of OLS to estimate the housing rents of our test set. Our results indicate that Ordinary Kriging method will lead to information loss since this method omits important housing attributes which could possible contain rental-determinable information.

Table I. Summary Statistics for Whole Sample

This table summarizes the statistics of variables (without locational dummy variables) in our dataset. All the variables are in raw form without any transformation. *Unitrent* is the rental price per square meter, *Floorarea* is the floor area (in square meter) of the house for rent, *Room* is the number of rooms, *Living* is the number of living rooms, *Kitchen* is the number of kitchens, *Storage* is the number of storage room, *Dinner* is the number of dinner room, *ManagementFee* is the management fee (in yen) charged on the tenant if the house is rented, *Shikikin* is the deposit fee (in thousand yen) charged on the tenant if the house is rented, *Hosyoukin* is the security money (in thousand yen) charged on the tenant if the house is rented, *Reikin* is the gratuity fee (in thousand yen) charged on the tenant if the house is rented, *Totalfloor* total floor of the building where the underlying house lies in, *Floornum* is the floor number of the house, *Accessibility* is the time (in minutes) to walk to nearest station, *DummyMaterial* is 1 if the building material of the house is concrete-based 0 if the building material is wooden-based; *Age* the age (in years) of the building where the underlying house lies in, *DummyPark* is 1 if the house has a parking lot 0 if no parking lot available, *DummyBikePark* is 1 if the house has a bike-parking lot 0 if no bike-parking lot available, *RelativeFloor* is the floor number of the house relative to the total floor of the underlying building, *Latitude* and *Longitude* are the latitude and longitude of the house, and *DisttoTokyo* is the distance in kilometers to Tokyo station.

Statistic	N	Mean	St. Dev.	Min	Max
Unitrent	10892	3376.1	757.31	1556.1	10874
Floorarea	10892	33.296	18.314	5.3000	266.19
Room	10892	1.2530	0.5530	1.0000	5.0000
Living	10892	0.2280	0.4190	0.0000	1.0000
Kitchen	10892	0.8510	0.3560	0.0000	1.0000
Storage	10892	0.0150	0.1230	0.0000	1.0000
Dinner	10892	0.3670	0.4820	0.0000	1.0000
ManagementFee	10892	5779.6	4261.0	0.0000	40000
Shikikin	10892	108.73	136.60	0.0000	3000.0
Hosyoukin	10892	2.5763	18.583	0.0000	460.00
Reikin	10892	85.552	80.153	0.0000	2200.0
TotalFloor	10892	7.1360	5.6760	2.0000	56.000
Floornum	10892	4.0520	3.5920	1.0000	50.000

Accessibility	10892	14.389	25.388	1.0000	211.04
DummyMaterial	10892	0.8950	0.3070	0.0000	1.0000
Age	10892	18.835	12.016	0.0000	77.137
DummyPark	10892	0.3670	0.4820	0.0000	1.0000
DummyBikePark	10892	0.3560	0.4790	0.0000	1.0000
RelativeFloor	10892	0.6170	0.2670	0.0540	1.0000
Longitude	10892	139.73	0.0740	139.58	139.85
Latitude	10892	35.673	0.0300	35.594	35.737
DisttoTokyo	10892	7.2520	3.9180	1.9080	16.963

Table II. Geographic Representativeness

This table shows the representativeness of our sample dataset and point dataset where the latter contains all the residential buildings within one district.

District	Point Dataset(N)	Point Dataset in %	Sample(N)	Sample in %
Shinjuku	41123	13.9%	1904	17.5%
Koutou	39651	13.4%	2110	19.4%
Minato	18066	6.1%	1513	13.9%
Sumida	38450	13.0%	1936	17.8%
Setagaya	159279	53.7%	3429	31.5%

Baseline Hedonic Regression Results

This table reports the baseline hedonic regression results of five districts (Column (1) to Column (5)) in Tokyo. We regress logarithms of rent per square meter, *Logunitrent*, on Spatial Variables: *Longitude* and *Latitude*, and Structure Variables: *Floorarea*, *Room*, *Living*, *Kitchen*, *Storage*, *Dinner*, *ManagementFee*, *Shikikin*, *Hosyoukin*, *Reikin*, *TotalFloor*, *RelativeFloor*, *Accessibility*, *Age*, *MaterialDummy*, *DummyPark*, and *DummyBikePark*. The details of variable constructions are discussed in Section 3.2. We report Variance Inflation Factor (VIF) for each variable in our five models. The table shows the coefficients and heteroscedasticity consistent standard-errors (in parentheses) obtained from five predictive OLS models. The statistics significance at the 10%, 5%, and 1% levels are indicated by *, **, and ***. Ψ indicates the variable is in logarithm.

					Logunitrent					
	Shinjuku		Koutou		Minato		Sumida		Setagaya	
	(1)		(2)		(3)		(4)		(5)	
Longitude	0.732^{***}	VIF	-3.715^{***}	VIF	-5.146^{***}	VIF	-2.248^{***}	VIF	2.535^{***}	VIF
	(0.159)	1.134	(0.177)	1.514	(0.381)	1.357	(0.299)	1.837	(0.083)	1.373
Latitude	-3.649^{***}	VIF	0.702^{***}	VIF	5.410^{***}	VIF	-3.288^{***}	VIF	0.232^{**}	VIF
	(0.262)	1.193	(0.188)	1.371	(0.357)	1.142	(0.256)	1.884	(0.096)	1.187
Floorarea$^{\Psi}$	-0.446^{***}	VIF	-0.481^{***}	VIF	-0.133^{***}	VIF	-0.551^{***}	VIF	-0.460^{***}	VIF
	(0.017)	5.116	(0.014)	6.532	(0.027)	5.845	(0.020)	6.597	(0.012)	6.786
Room	0.055^{***}	VIF	-0.005	VIF	0.092^{***}	VIF	0.035^{***}	VIF	0.030^{***}	VIF
	(0.010)	2.030	(0.007)	3.160	(0.015)	2.183	(0.010)	2.983	(0.006)	2.708
Living	0.103^{***}	VIF	0.040^{***}	VIF	0.033^{*}	VIF	0.081^{***}	VIF	0.042^{***}	VIF
	(0.011)	2.759	(0.009)	2.442	(0.020)	5.551	(0.010)	2.361	(0.007)	2.437
Kitchen	-0.026^{***}	VIF	-0.026^{***}	VIF	0.004	VIF	0.001	VIF	-0.003	VIF
	(0.007)	1.278	(0.007)	1.179	(0.010)	1.489	(0.007)	1.137	(0.006)	1.231
Storage	0.032	VIF	0.080^{**}	VIF	-0.031	VIF	0.038^{**}	VIF	0.011	VIF
	(0.020)	1.066	(0.033)	1.103	(0.032)	1.103	(0.015)	1.101	(0.013)	1.028

Dinner	0.060***	VIF	0.052***	VIF	−0.040**	VIF	0.075***	VIF	0.039***	VIF
	(0.010)	3.056	(0.009)	3.285	(0.019)	5.636	(0.009)	3.566	(0.008)	3.289
ManagementFee	−0.006	VIF	−0.005	VIF	−0.045***	VIF	−0.024***	VIF	0.005	VIF
	(0.008)	1.535	(0.007)	1.623	(0.010)	1.111	(0.009)	1.803	(0.007)	1.515
Shikikin	0.004***	VIF	0.005***	VIF	0.001***	VIF	0.002***	VIF	0.006***	VIF
	(0.0004)	2.283	(0.001)	2.253	(0.0003)	1.864	(0.0005)	1.927	(0.0004)	1.942
Hosyoukin	0.003***	VIF	0.002***	VIF	−0.003***	VIF	−0.0003	VIF	0.005***	VIF
	(0.001)	1.125	(0.001)	1.156	(0.001)	1.049	(0.001)	1.138	(0.002)	1.060
Reikin	0.001**	VIF	0.003***	VIF	−0.0001	VIF	0.003***	VIF	0.004***	VIF
	(0.001)	1.731	(0.0004)	1.568	(0.001)	1.341	(0.001)	1.546	(0.0004)	1.445
TotalFloor	0.008***	VIF	0.006***	VIF	0.003***	VIF	0.010***	VIF	0.014***	VIF
	(0.001)	1.514	(0.001)	1.571	(0.001)	1.570	(0.001)	1.549	(0.001)	1.464
RelativeFloor	0.046***	VIF	0.029***	VIF	0.064***	VIF	0.050***	VIF	0.025***	VIF
	(0.010)	1.072	(0.008)	1.072	(0.015)	1.066	(0.009)	1.108	(0.007)	1.096
Accessibility	−0.001	VIF	−0.0002***	VIF	−0.003***	VIF	−0.004***	VIF	−0.00001***	VIF
	(0.001)	1.059	(0.0001)	1.015	(0.001)	1.057	(0.001)	1.086	(0.0000)	1.011
Age	−0.006***	VIF	−0.006***	VIF	−0.008***	VIF	−0.007***	VIF	−0.005***	VIF
	(0.0003)	1.571	(0.0003)	1.859	(0.0004)	1.318	(0.0003)	1.855	(0.0002)	1.322
MaterialDummy	0.043***	VIF	0.020	VIF	0.029	VIF	0.035**	VIF	0.033***	VIF
	(0.012)	1.315	(0.018)	1.080	(0.106)	1.032	(0.015)	1.206	(0.005)	1.359
DummyPark	0.022***	VIF	0.015***	VIF	0.035***	VIF	0.019***	VIF	0.008**	VIF
	(0.006)	1.121	(0.004)	1.066	(0.008)	1.176	(0.004)	1.066	(0.004)	1.176
DummyBikePark	0.0002	VIF	0.004	VIF	−0.0002	VIF	0.011**	VIF	−0.002	VIF
	(0.005)	1.044	(0.004)	1.027	(0.008)	1.125	(0.005)	1.065	(0.004)	1.097

Constant	37.535	503.993***	534.955***	441.436***	-352.795^{***}
	(26.252)	(22.383)	(53.099)	(37.533)	(13.233)
N	1507	1715	1187	1576	2728
Adjusted R^2	0.684	0.833	0.506	0.803	0.744

Table IV. The Signs of Coefficients

This table shows the empirical signs (OLS columns) of coefficients of Longitude and Latitude and hypothetical signs (Hypo) of coefficients if our assumption in Section 4.2.1 holds true. The red sign in OLS column means the empirical sign of corresponding coefficient is opposite to the hypothetical sign.

	Shinjuku		Koutou		Minato		Sumida		Setagaya	
	Hypo	OLS	Hypo	OLS	Hypo	OLS	Hypo	OLS	Hypo	OLS
Longitude	+	+	-	-	+	-	-	-	+	+
Latitude	-	-	+	+	+	+	-	-	+	+

Table V. OLS Results with Distance to Tokyo Station

This table reports the regression results when adding the distance to Tokyo station, *DistToTokyo* to baseline hedonic regression models as in Table III. Structure Variables are those structure variables which are included in baseline hedonic regressions but are not reported here. The table shows the coefficients and heteroscedasticity consistent standard-errors (in parentheses) obtained from five predictive OLS models. The statistics significance at the 10%, 5%, and 1% levels are indicated by *, **, and ***.

			Logunitrent		
	Shinjuku	Koutou	Minato	Sumida	Setagaya
	(1)	(2)	(3)	(4)	(5)
Longitude	-0.269	-1.011^{*}	-8.519^{***}	6.790^{***}	1.402^{**}
	(2.116)	(0.583)	(1.054)	(2.483)	(0.569)
Latitude	-3.135^{***}	0.198	0.082	3.879^{**}	-0.352
	(1.117)	(0.198)	(1.522)	(1.969)	(0.293)
DisttoTokyo	-0.012	-0.029^{***}	-0.065^{***}	-0.117^{***}	-0.013^{**}
	(0.026)	(0.006)	(0.018)	(0.032)	(0.0060)
Structural Variables	Yes	Yes	Yes	Yes	Yes
Constant	159.154	144.021^{*}	1196.540^{***}	-1077.481^{***}	-173.577^{*}
	(257.490)	(78.048)	(197.565)	(415.846)	(89.509)
N	1507	1715	1187	1576	2728
Adjusted R^2	0.684	0.835	0.513	0.805	0.744

Table VI. OLS Results with District-Dummy Variable

This table reports regression results with district dummy variables. Details of variable construction are discussed in Section 3.2. The table shows the coefficients and heteroscedasticity consistent standard-errors (in parentheses) obtained from five predictive OLS models. The statistics significance at the 10%, 5%, and 1% levels are indicated by *, **, and ***. Ψ indicates the variable is in logarithm.

	Logunitrent				
	(1)	(2)	(3)	(4)	(5)
Longitude	-0.839^{***} (0.078)	Yes	Yes	Yes	Yes
Latitude	-0.932^{***} (0.082)	Yes	Yes	Yes	Yes
DisttoTokyo	-0.038^{***} (0.001)	Yes	Yes	Yes	Yes
Floorarea$^{\Psi}$	-0.385^{***} (0.007)	Yes	Yes	Yes	Yes
Room	0.027^{***} (0.003)	Yes	Yes	Yes	Yes
Living	0.062^{***} (0.005)	Yes	Yes	Yes	Yes
Kitchen	-0.020^{***} (0.004)	Yes	Yes	Yes	Yes
Storage	0.012 (0.010)	Yes	Yes	Yes	Yes
Dinner	0.020^{***} (0.004)	Yes	Yes	Yes	Yes
ManagementFee	-0.008^{**} (0.003)	Yes	Yes	Yes	Yes
Shikikin	0.004^{***} (0.0001)	Yes	Yes	Yes	Yes
Hosyoukin	0.002^{***} (0.001)	Yes	Yes	Yes	Yes
Reikin	0.002^{***} (0.0002)	Yes	Yes	Yes	Yes

TotalFloor	0.004***	Yes	Yes	Yes	Yes
	(0.0003)				
RelativeFloor	0.031***	Yes	Yes	Yes	Yes
	(0.005)				
Accessibility	−0.00001*	Yes	Yes	Yes	Yes
	(0.00000)				
Age	−0.006***	Yes	Yes	Yes	Yes
	(0.0001)				
MaterialDummy	0.044***	Yes	Yes	Yes	Yes
	(0.004)				
DummyPark	0.012***	Yes	Yes	Yes	Yes
	(0.003)				
DummyBikePark	−0.001	Yes	Yes	Yes	Yes
	(0.003)				
DummyShinjuku	−0.089***	0.112***	−0.033***	0.116***	
	(0.006)	(0.008)	(0.009)	(0.009)	
DummyKoutou	−0.205***	−0.004	−0.149***		−0.116***
	(0.008)	(0.004)	(0.013)		(0.009)
DummySetagaya	−0.056***	0.145***		0.149***	0.033***
	(0.009)	(0.014)		(0.013)	(0.009)
DummySumida	−0.201***		−0.145***	0.004	−0.112***
	(0.009)		(0.014)	(0.004)	(0.008)
DummyMinato		0.201***	0.056***	0.205***	0.089***
		(0.009)	(0.009)	(0.008)	(0.006)
Constant	160.183***	159.981***	160.126***	159.977***	160.094***
	(11.556)	(11.563)	(11.552)	(11.563)	(11.556)
N	8713	8713	8713	8713	8713
Adjusted R^2	0.741	0.741	0.741	0.741	0.741

Table VII. Parameters of Ordinary Kriging Model for Each District

This table reports parameters, partial sill, sill, range, nugget and nlags for each district.

District	$\widehat{s}-\widehat{a}$ (Partial Sill)	$\widehat{s}$ (Sill)	$\widehat{r}$ (Range)	$\widehat{a}$ (Nugget)	k (nlags)
Shinjuku	335369.5	392880.8	0.00002	57511.3	490
Koutou	324831.1	347837.3	0.00004	23006.2	2047
Minato	471035.0	585621.2	0.00007	114586.2	623
Sumida	458277.5	462721.8	0.00075	4444.3	109
Setagaya	243626.4	295918.0	0.00014	52291.6	715

Table VIII. Comparisons of Accuracy Metrics for Ordinary Kriging and OLS

This table reports the accuracy metrics of different methods for each district. OK represents Ordinary Kriging and OLS represents Ordinary Least Squared. R-Squared is defined as equation (17) and RMSE is defined as equation (18). Results of both test set and training set are reported.

District	Method	R-Squared (Test)	RMSE (Test)	R-Squared (Training)	RMSE (Training)
Shinjuku	OK	33.70%	557.39	93.50%	165.07
	OLS	54.70%	379.83	51.94%	353.25
Koutou	OK	64.44%	343.57	91.60%	172.39
	OLS	77.73%	250.63	78.50%	249.43
Minato	OK	56.55%	565.57	78.56%	363.66
	OLS	−65.52%	690.82	−10.60 %	574.09
Sumida	OK	63.06%	338.85	83.24%	224.44
	OLS	75.79%	249.14	73.36%	253.78
Setagaya	OK	52.95%	392.63	85.83%	213.07
	OLS	62.16%	308.12	63.86%	290.42

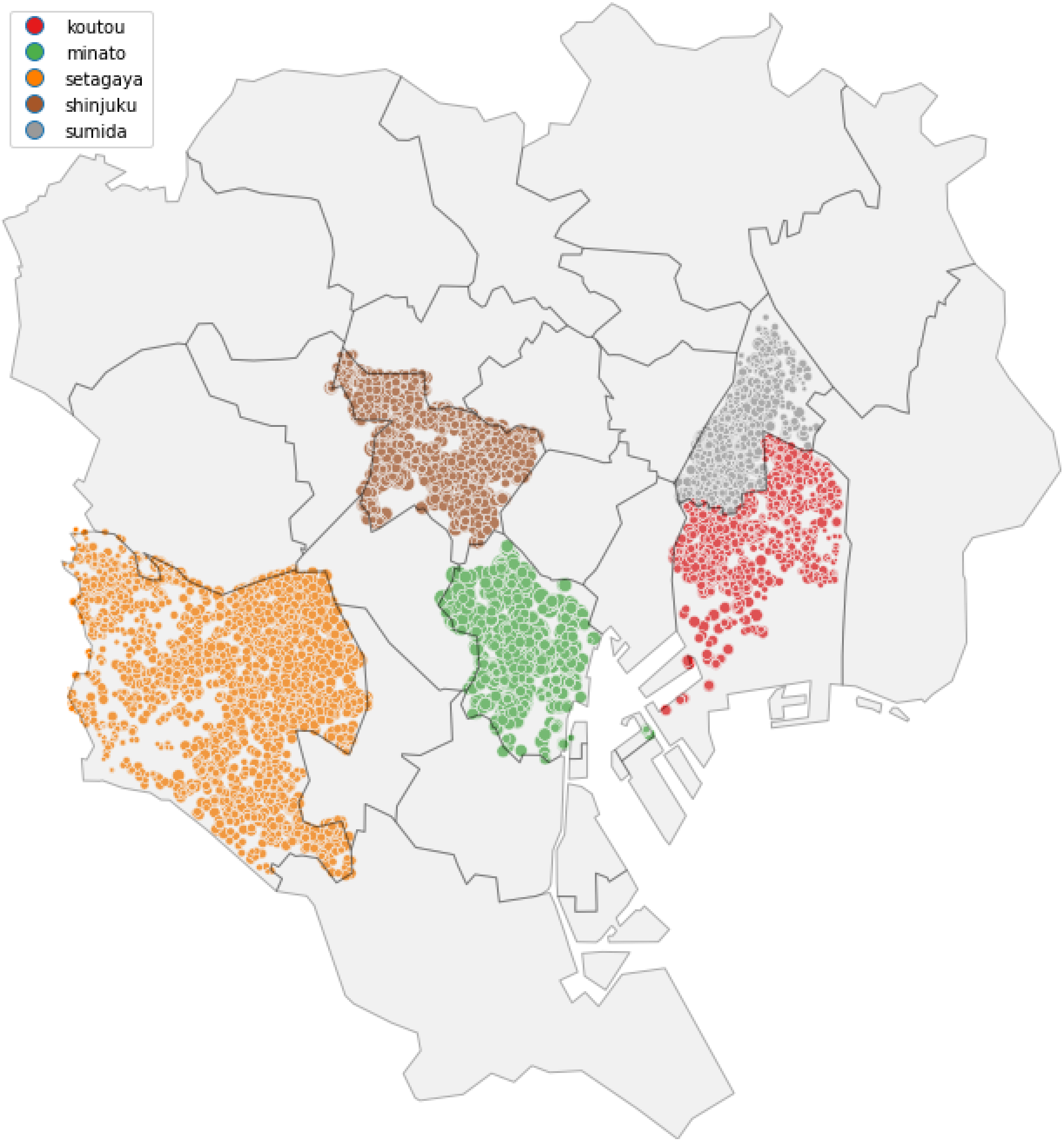


Figure 1: Geographic Plot of Sample Data.
Notes: The figure shows the geographic plot of sample data in our final dataset. Each point represents the geographic location of one observation. The size of each point represents the relative value (higher value, larger size) of rent per square meter of this observation. Different color represents different district.

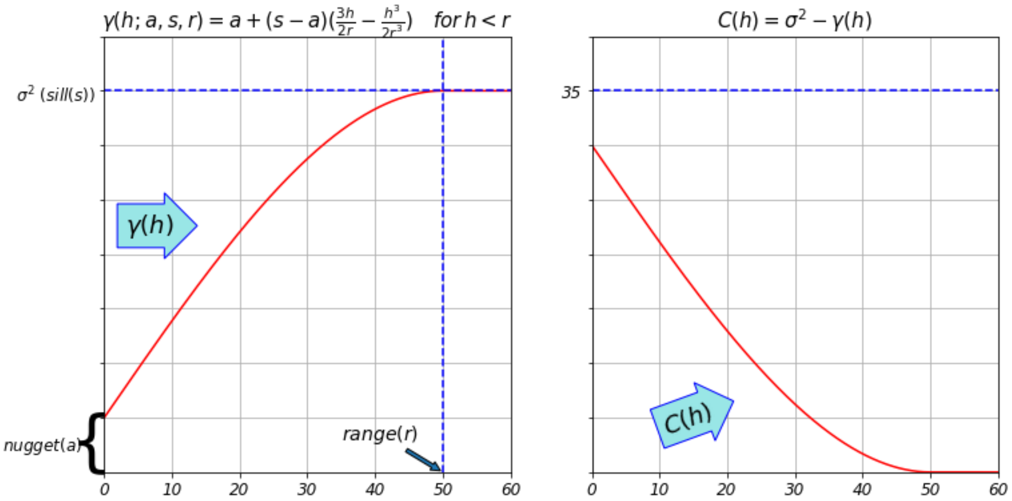


Figure 2: Theoretical Semi-Variogram (Left) and Co-Variogram (Right) for spherical model.

Notes: The figure shows the theoretical Semi-Variogram function and Co-Variogram function with parameters nugget (a) equals to 5, sill (s) equals to 35, and range (r) equals to 50.

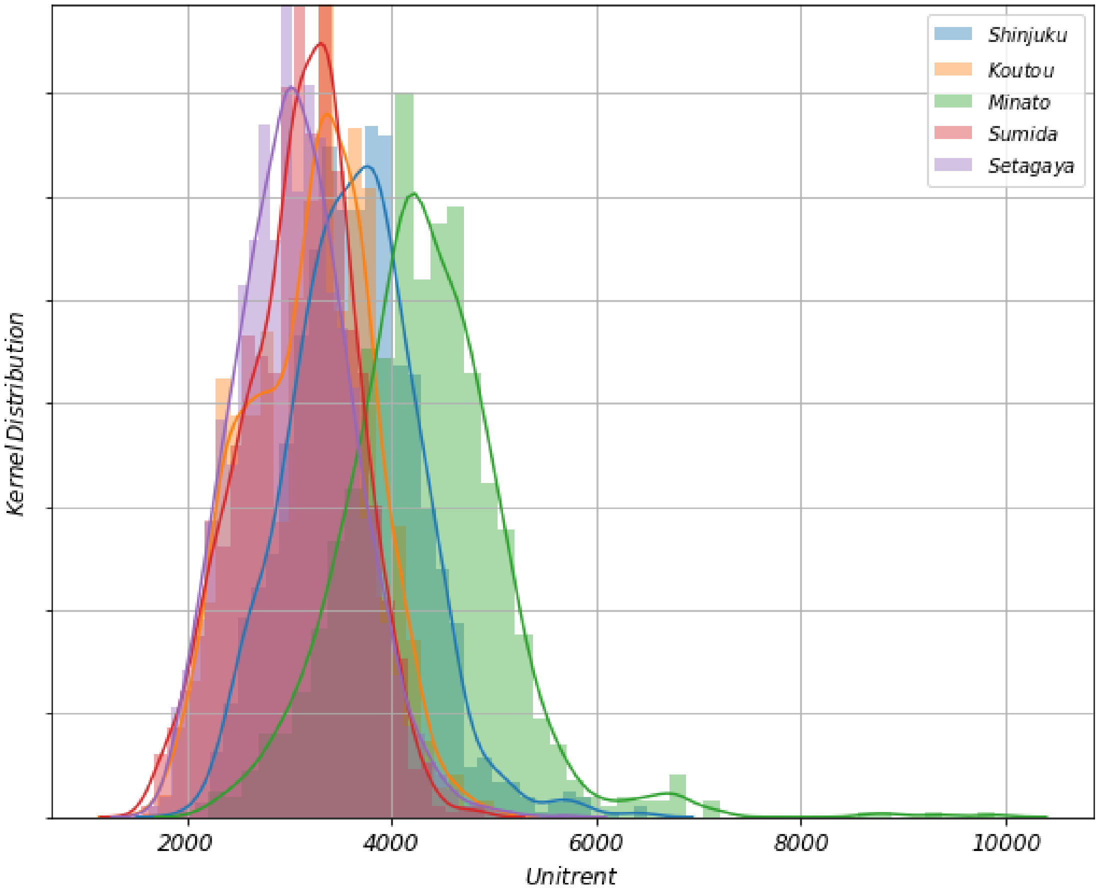


Figure 3: Kernel Distributions of rent per square meter, *Unitrent*, for five districts. *Notes*: The figure shows Kernel Distributions of rent per square meter for five districts. X-axis indicates the Kernel Distribution Probability and y-axis indicates the rent per square meter. Different color represents different district.

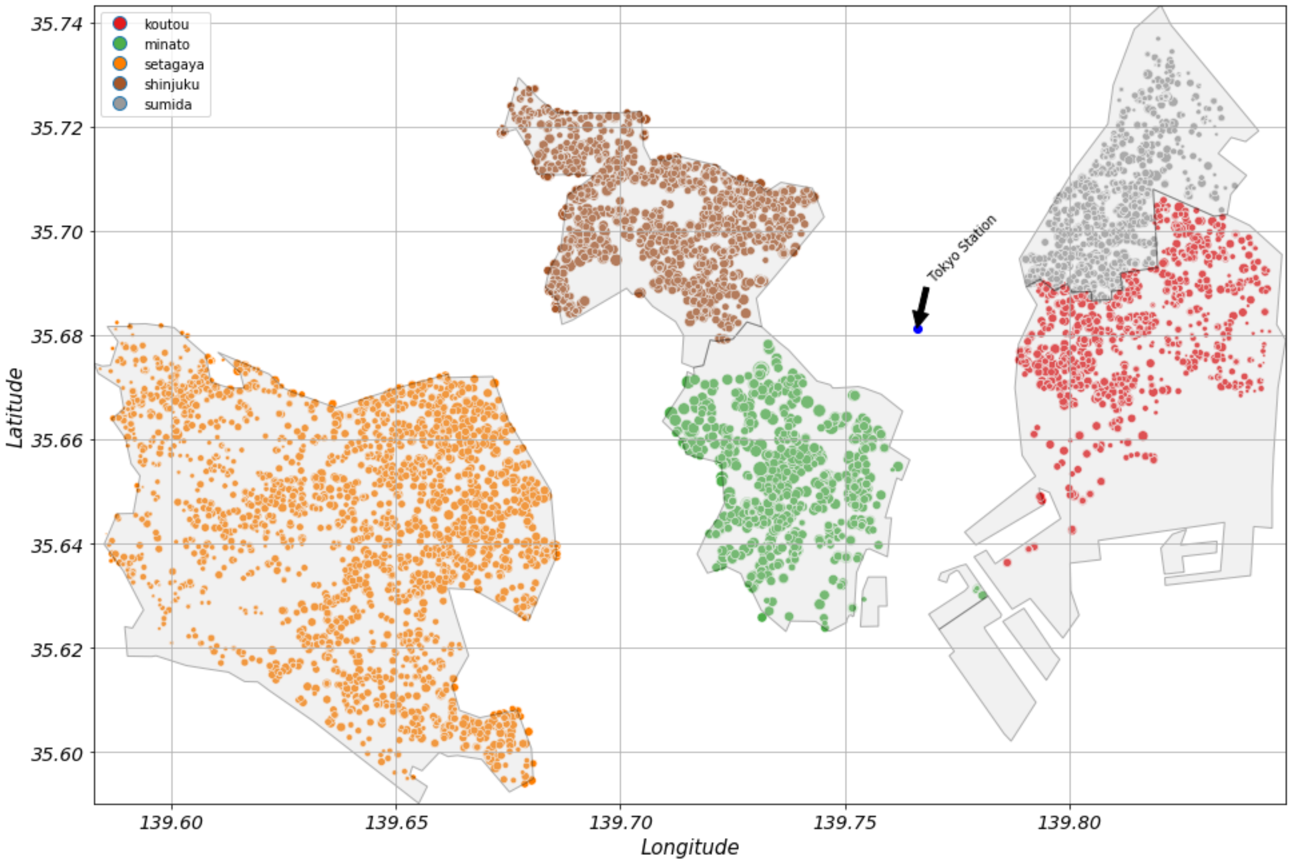
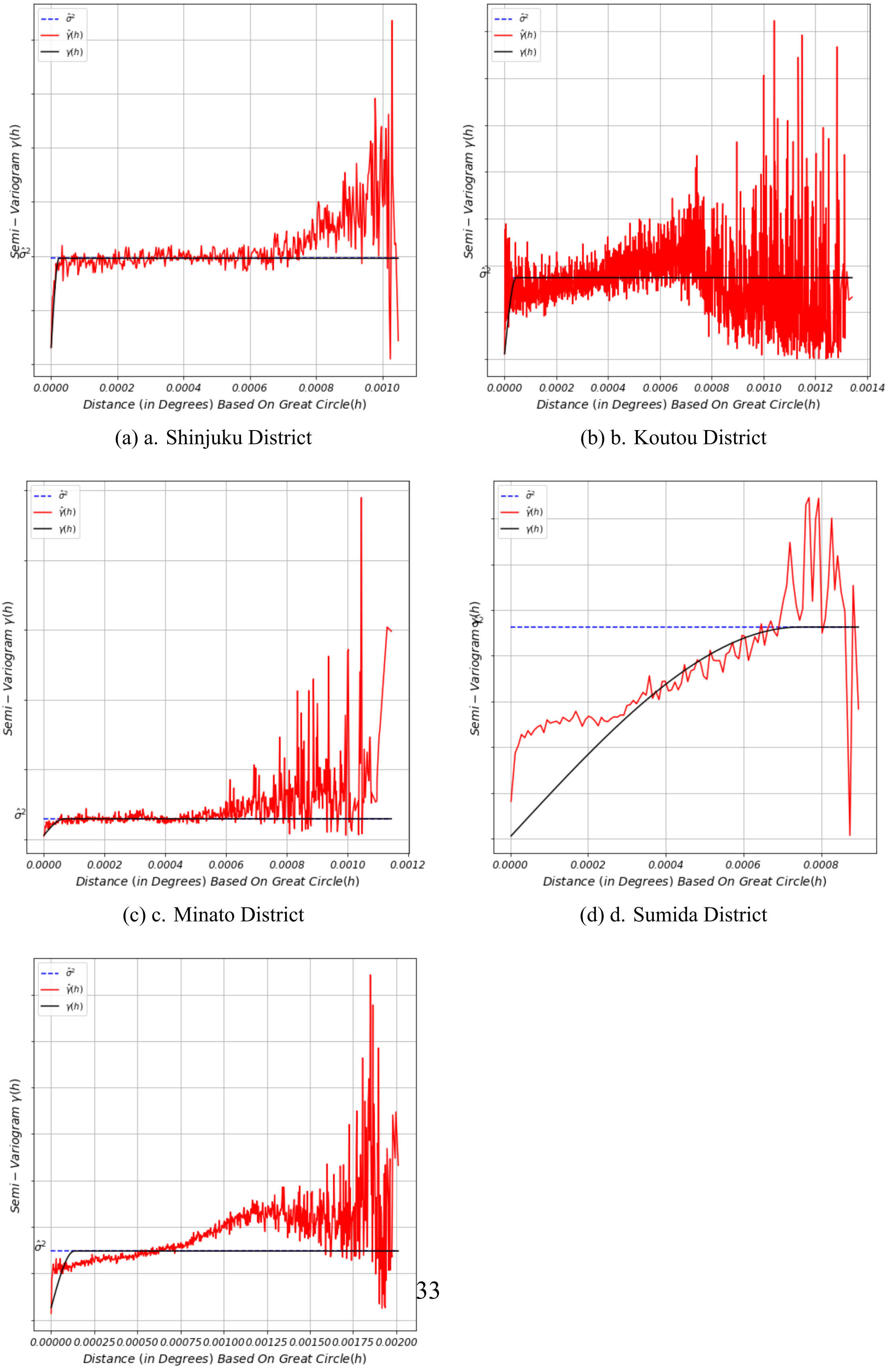


Figure 4: Geographic interpretation of baseline regression results.
Notes: The figure shows the relative position between each district and Tokyo Station. The size of the circle indicates the relative value of rent per square meter (larger circle larger *Unitrent*).



(a) a. Shinjuku District

(b) b. Koutou District

(c) c. Minato District

(d) d. Sumida District

(e) e. Setagaya District

Figure 5: Empirical Semi-Variograms for five districts.
Notes: The red line represents actual binned semi-variogram, the black line represents fitted theoretical semi-variogram and the dashed blue line represents sill.

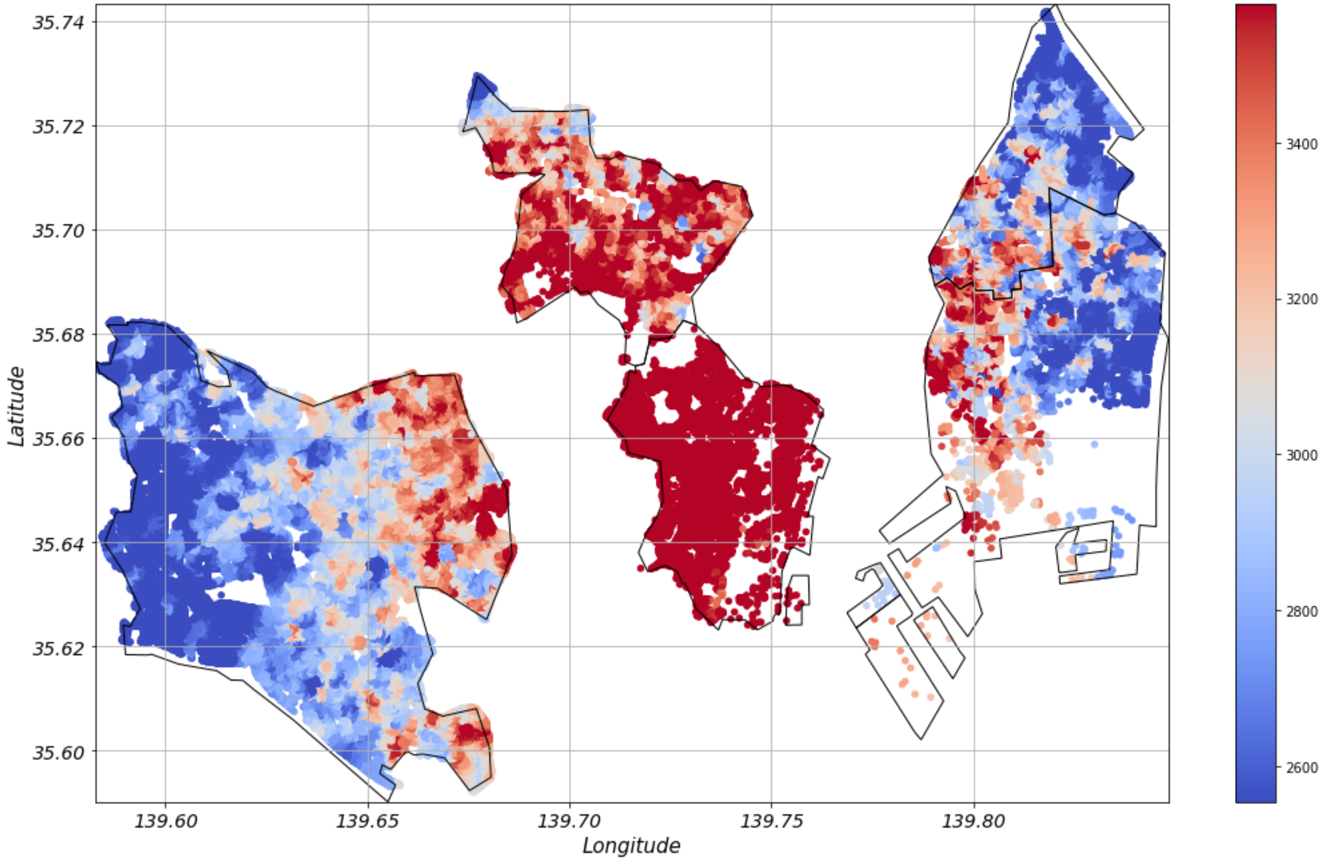


Figure 6: Automated Valuation Model (AVM) based on OLS for five districts in Tokyo.
Notes: The figure shows geographic plot of our Automated Valuation Model based on OLS method. All the residential buildings in our "point" dataset are plotted as color dots on this figure. The red-blue heatmap captures the relative value of rent per square meter for each dot. The value-range of heat map is located at the right of the map. X-axis and y-axis represents longitude and latitude respectively.

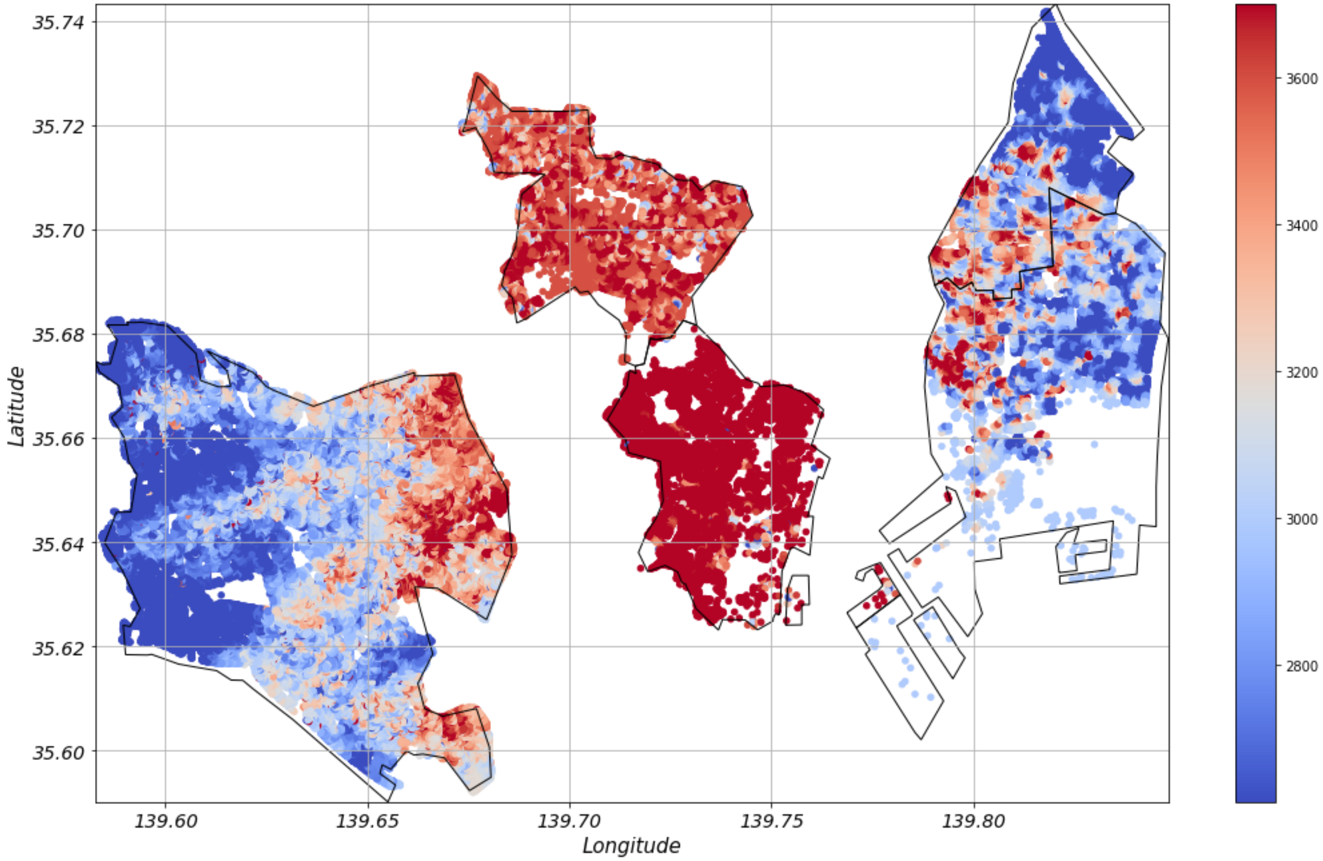


Figure 7: Automated Valuation Model (AVM) based on Ordinary Kriging for five districts in Tokyo.
Notes: The figure shows geographic plot of our Automated Valuation Model based on Orinary Kriging method. All the residential buildings in our “point” dataset are plotted as color dots on this figure. The red-blue heatmap captures the relative value of rent per square meter for each dot. The value-range of heat map is located at the right of the map. X-axis and y-axis represents longitude and latitude respectively.

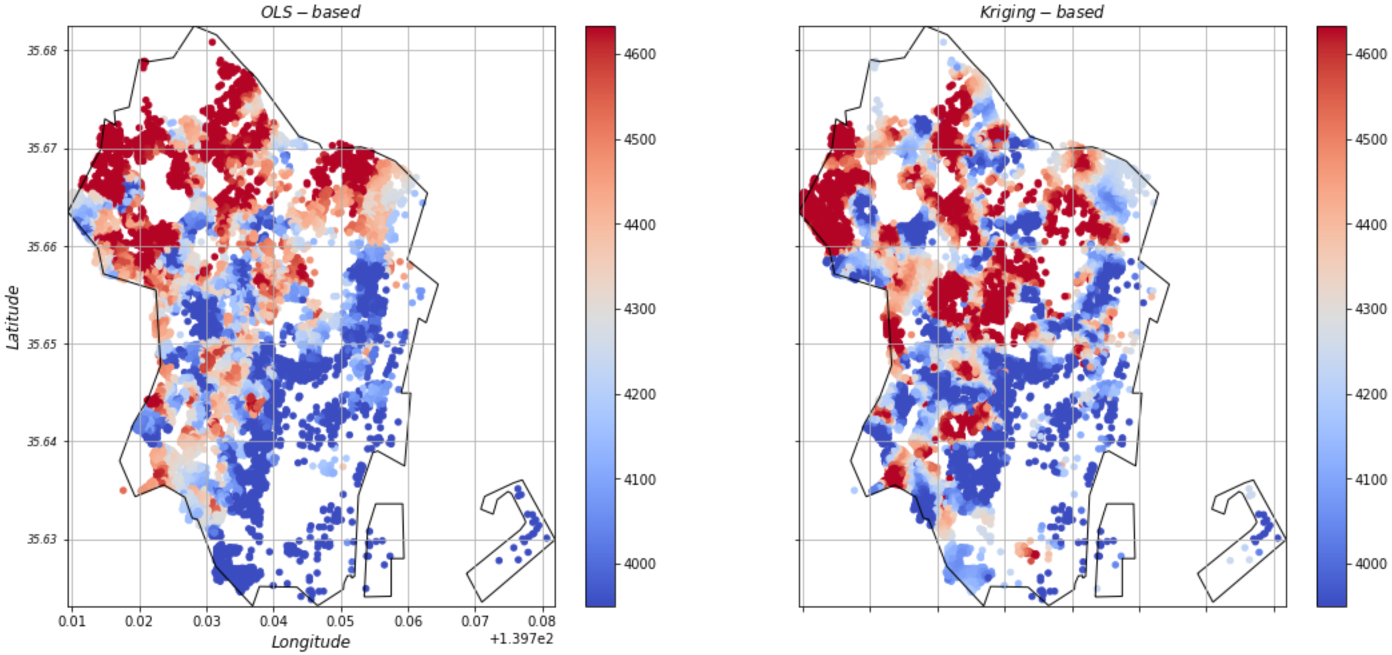


Figure 8: Automated Valuation Model for Minato districts based on OLS (Left) and Ordinary Kriging (Right).
Notes: The figure shows geographic plot of our Automated Valuation Model based on OLS (Left) and Ordinary Kriging (Right). All the residential buildings in our "point" dataset are plotted as color dots on this figure. The red-blue heatmap captures the relative value of rent per square meter for each dot. The value-range of heat map is located at the right of the map. X-axis and y-axis represents longitude and latitude respectively.

References

Statistics Japan Residential House and Land survey 2013. Statistics Bureau, Director-General for Policy Planning (Statistical Standards) and Statistical Research and Training Institute, 2013.

ALISTAIR ADAIR and STANLEY McGREAL. The application of multiple regression analysis in property valuation. *Journal of Valuation*, 6(1):57–67, 1988.

David LJ Alexander, Alexander Tropsha, and David A Winkler. Beware of r 2: simple, unambiguous assessment of the prediction accuracy of qsar and qspr models. *Journal of chemical information and modeling*, 55(7):1316–1322, 2015.

Paul M Anglin and Ramazan Gencay. Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, 11(6):633–648, 1996.

Luc Anselin. Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics*, 20(2):141–163, 1990.

Luc Anselin. Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, 26(2):153–166, 2003.

Marco Aurélio Stumpf González and Carlos Torres Formoso. Mass appraisal with genetic fuzzy rule-based systems. *Property Management*, 24(1):20–30, 2006.

Carlo Bagnoli and Halbert Smith. The theory of fuzz logic and its application to real estate valuation. *Journal of Real Estate Research*, 16(2):169–200, 1998.

Roberto Basile, María Durbán, Román Mínguez, Jose María Montero, and Jesús Mur. Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245, 2014.

Sabyasachi Basu and Thomas G Thibodeau. Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 17(1):61–85, 1998.

Archith J Bency, Swati Rallapalli, Raghu K Ganti, Mudhakar Srivatsa, and BS Manjunath. Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 320–329. IEEE, 2017.

RA Borst and WJ McCluskey. An evaluation of mra, comparable sales analysis and anns for the mass appraisal of residential properties in northern ireland. *Assessment Journal*, 4(1):47–55, 1997.

Wolfgang A Brunauer, Stefan Lang, Peter Wechselberger, and Sven Bienert. Additive hedonic regression models with spatial scaling factors: An application for rents in vienna. *The Journal of Real Estate Finance and Economics*, 41(4):390–411, 2010.

Jorge Chica Olmo. Spatial estimation of housing prices and locational rents. *Urban studies*, 32(8):1331–1344, 1995.

John M Clapp, Hyon-Jung Kim, and Alan E Gelfand. Predicting spatial patterns of house prices using lpr and bayesian smoothing. *Real Estate Economics*, 30(4):505–532, 2002.

Joao F Cocco. Hedging house price risk with incomplete markets. In *AFA 2001 New Orleans Meetings*, 2000.

A.T. Court. *The Dynamics of Automobile Demand.* General Motor, New York, 1939.

Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

Theodore M Crone and Richard P Voith. Estimating house price appreciation: a comparison of methods. *Journal of Housing Economics*, 2(4):324–338, 1992.

Ralph B D'Agostino. *Goodness-of-fit-techniques*, volume 68. CRC press, 1986.

Yonghe Deng, Xiangyu Guo, and Chihiro Shimizu. Change in the distribution of sale rental prices: Comparison of beijing and tokyo. In *Hitotsubashi-RIETI International Workshop on Real Estate and the Macro Economy Presenter4*. Hitotsubashi Project on Real Estate, Financial Crisis, and Economics Dynamics, 2017.

Dragana Djurdjevic, Christine Eugster, and Ronny Haase. Estimation of hedonic models using a multilevel approach: An application for the swiss rental market. *Swiss Journal of Economics and Statistics*, 144(4):679–701, 2008.

A Quang Do and Gary Grudnitski. A neural network approach to residential property appraisal. *The Real Estate Appraiser*, 58(3):38–45, 1992.

Muhammad Faishal Ibrahim, Fook Jam Cheng, and Kheng How Eng. Automated valuation model: an application to the public housing resale market in singapore. *Property Management*, 23(5):357–373, 2005.

Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66.* Routledge, 2018.

A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

A Stewart Fotheringham, Ricardo Crespo, and Jing Yao. Geographical and temporal weighted regression (gtwr). *Geographical Analysis*, 47(4):431–452, 2015.

The Apprasial Foundation. *USPAP 2003 Uniform Standards of Professional Appraisal Practice*. The Apprasial Foundation, 2003.

Noelia García, Matías Gámez, and Esteban Alfaro. Ann+gis: An automated system for property valuation. *Neurocomputing*, 71(4-6):733–742, 2008.

Mark J Garmaise and Tobias J Moskowitz. Confronting information asymmetries: Evidence from real estate markets. *The Review of Financial Studies*, 17(2):405–437, 2003.

Guy D Garrod and Kenneth G Willis. Valuing goods' characteristics: an application of the hedonic price method to environmental attributes. *Journal of Environmental management*, 34(1):59–76, 1992.

Kevin Gillen, Thomas Thibodeau, and Susan Wachter. Anisotropic autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, 23(1):5–30, 2001.

Allen C Goodman and Thomas G Thibodeau. Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, 12(3):181–201, 2003.

Richard K Green, Richard Green, and Stephen Malpezzi. *A primer on US housing markets and housing policy*. The Urban Insitute, 2003.

Jian Guan, Jozef Zurada, and Alan Levitan. An adaptive neuro-fuzzy inference system based approach to real estate property assessment. *Journal of Real Estate Research*, 30(4):395–422, 2008.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Edward H Isaaks and RM Srivastava. An introduction to geostatistics, 1989.

Harry H Kelejian and Dennis P Robinson. A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding monte carlo results. *Regional Science and Urban Economics*, 28(4):389–417, 1998.

Haniza Khalid. Spatial heterogeneity and spatial bias analyses in hedonic price models: some practical considerations. *Bulletin of geography. Socio-economic series*, 28(28): 113–128, 2015.

Daniel G Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

Michael Kuntz and Marco Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science*, 28(9):1904–1921, 2014.

Kelvin J Lancaster. A new approach to consumer theory. *Journal of political economy*, 74(2):132–157, 1966.

Visit Limsombunchai. House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference*, pages 25–26, 2004.

HP Liu and S Maghsoodloo. Taylor kriging for simulation metamodeling. *Auburn, Auburn University, Dissertation*, pages 99–124, 2009.

Michael Löchl and Kay Axhausen. Modelling hedonic residential rents for land use and transport simulation while considering spatial effects. 2010.

Jonathan Mark and Michael A Goldberg. Multiple regression analysis and mass assessment: A review. *The Appraisal Journal*, 56(1):89, 1988.

Davide Martinetti and Ghislain Geniaux. Approximate likelihood estimation of spatial probit models. *Regional Science and Urban Economics*, 64:30–45, 2017.

W McCluskey, K Dyson, D McFall, and SS Anand. Mass appraisal for property taxation: an artificial intelligence approach. *Australian Land Economics Review*, 2(1), 1996.

William McCluskey, Peadar Davis, Martin Haran, Michael McCord, and David McIlhatton. The potential of artificial neural networks in mass appraisal: the case revisited. *Journal of Financial Management of Property and Construction*, 17(3):274–292, 2012.

Daniel P McMillen and Christian L Redfearn. Estimation and hypothesis testing for nonparametric hedonic house price functions. *Journal of Regional Science*, 50(3): 712–733, 2010.

Jean D Opsomer, David Ruppert, et al. Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, 25(1):186–211, 1997.

Steven Peterson and Albert Flanagan. Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, 31(2):147–164, 2009.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55, 1974.

PETER Rossini. Accuracy issues for automated and artificial intelligent residential valuation systems. In *International Real Estate Society Conference*, 1999.

O Schabenberger and CA Gotway. Statistical methods for spatial data analysis vol. chapman & hall/crc texts in statistical science. 2004.

Hajime Seya, Morito Tsutsumi, Yasushi Yoshida, and Yuichiro Kawaguchi. Empirical comparison of the various spatial prediction models: in spatial econometrics, spatial statistics, and semiparametric statistics. *Procedia-Social and Behavioral Sciences*, 21:120–129, 2011.

Stacy Sirmans, David Macpherson, and Emily Zietz. The composition of hedonic pricing models. *Journal of real estate literature*, 13(1):1–44, 2005.

Danny PH Tay and David KH Ho. Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, 10(2):525–540, 1992.

Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

Elaine Worzala, Margarita Lenk, and Ana Silva. An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, 10(2):185–201, 1995.

Jozef Zurada, Alan Levitan, and Jian Guan. A comparison of regression and artificial intelligence methods in a mass appraisal context. *Journal of Real Estate Research*, 33(3):349–387, 2011.