Swarm-Inspired Data-Driven Approach for Housing Market Segmentation: A Case Study of Taipei City

Abstract

Data-driven housing-market segmentation has been given increasing prominence for its objectiveness in identifying submarkets based on the housing data's underlying structures. However, although popular in existing literature, current statistical-clustering methods, when handling high-dimensionality housing dataset, has been found to tend to loss low-variance information of the dataset and be deficient in deriving the globally optimal number of submarkets. Accordingly, with the intention of achieving more rigorous housing-market segmentation in the case of high-dimensionality housing dataset, a swarm-inspired projection (SIP) algorithm is introduced by this study. A case study is then conducted using housing dataset of Taipei city to evaluate the predictive accuracy of submarkets' housing prices obtained using hedonic price models, and which are based on the segmentations resulted from both the proposed SIP and a statistical-clustering method using the combination of principal component analysis (PCA) and K-means clustering. The results show that, as compared to the use of PCA and K-means, our proposed SIP algorithm can obtain more optimal number of submarkets for segmentation, and the resulted submarkets are more homogenous and distinctive. This finding highlights the advantages of our proposed SIP algorithm in housing segmentation, and thus it can better help inform the further studies of market segmentation-related problems.

Keywords: Artificial intelligence; Clustering approach; Data mining; Hedonic price model; Housing submarkets.

1 Introduction

Housing market segmentation is an essential research topic in real estate and housing studies. Beginning with a study of Schnare and Struyk (1976) that demonstrated the existence of submarkets within a larger urban housing market, a consensus has been reached in following studies that housing market should be treated as a composition of a number of submarkets rather than a uniform entity (Bourassa et al., 1999; Islam & Asami, 2009). Such consensus on the existence of housing submarkets is due to three elementary characteristics that distinguish housing from other common economic commodities: spatial immobility, durability, and heterogeneity (Adair et al., 1996). Accordingly, the concept of submarkets has been expanded into a broad array of the real estate research, such as predication of housing prices, evaluation of revitalization policy, or understanding of neighborhood effects (Adair et al. 1996; Watkins 2001; Wilson et al. 2011; Hui et al., 2016; Wu et al., 2018).

The key to the identification of submarkets is the segmentation of a total market. In the early stage, the criteria for housing market segmentation are based on theoretical justification, which has two major streams: the geography-based segmentation and the quality-based segmentation (Adair et al., 1996; Islam & Asami, 2009). The geography-based segmentation assumes that the housing market can be stratified based on environmental features such as geographical boundary, while the quality-based segmentation assumes that the housing market can be stratified based on physical features of housing such as dwelling type. However, both these segmentation approaches have become partially obsolete, insofar their deficiencies in subjectivity and requirement for prior domain knowledge (Bourassa et al., 1999). Accordingly, data-driven segmentation approaches have recently been given prominence for their relatively superior capability for delineating the factual number of housing submarkets more objectively and accurately on a basis of the use of data's underlying structure itself (Wu & Sharma, 2012; Helbich et al., 2013).

Although various traditional statistical data-driven approaches have been used to identify housing submarkets, including factor analysis (Dale-Johnson, 1982), discrete choice models (Tu, 1997) and neural networks (Kauko et al., 2002), due to the rapid development of information technology, the clustering analysis has recently been more popular in housing market segmentation research due to such analysis' remarkable computational (Wu et al., 2018). Several clustering methods that have been applied in housing market segmentation are derived from data-mining science, including partitioning methods (Wu & Sharma, 2012), hierarchical methods (Bates, 2006), and density-based methods (Wu et al., 2018). However, slowness of convergence to solutions has been often observed when using these clustering methods to handle high-dimensionality data (Su et al., 2009). Accordingly, principal component analysis (PCA) has proposed as a fundamental preprocessing step for these methods as such analysis is able to effectively reduce data dimension (Han et al., 2012; Helbich et al., 2013) and this combination of clustering analysis and PCA can be termed as statistical-clustering method (Wu et al., 2018).

Nonetheless, this statistical-clustering method is still subjective to two major deficiencies. First, the use of PCA itself may lead to lose a certain degree of crucial low-variance information to correctly distinguishing housing submarkets (Reif 2018), and thus producing poor performance on the segmentation for succeeding clustering analysis. Second, most of current clustering approaches, such as K-means clustering, are susceptible to result in a locally optimal number of clusters rather than a global optimum one (Bourassa et al., 1999; Su et al., 2009). However, more homogenous and distinctive housing submarkets can only be identified when a globally optimal number of clusters (i.e., submarkets) are determined, where the clusters are deemed to have the highest level of intra-cluster similarity and the lowest level of inter-cluster dissimilarity (Han et al., 2012). Consequently, the aforementioned two major deficiencies in the statistical-clustering method may lead policymakers to have little confidence in using this method.

With the intention of remedying these two deficiencies of the statistical-clustering method, the present study innovatively introduces a swarm-intelligence-based clustering algorithm: the swarm-inspired projection (SIP) algorithm, for the segmentation of housing submarkets, involving a high-dimensionality housing dataset. Due to the self-organizing and data-projecting features, the SIP algorithm is capable of effectively determining the globally optimal number of clusters by directly projecting high-dimensionality data into a low-dimensionality space while avoiding the loss of essential low-variance information (Su et al., 2009). The present study will examine the effectiveness of the proposed SIP algorithm in housing-market segmentation using a high-dimensionality housing dataset from the Taipei city, which contains 5,012 samples of residential properties with the transaction prices and their associated 38 attributes collected from 2008 to 2010. The submarket-segmentation performance resulted from the proposed SIP algorithm, in terms of the predictive accuracy of hedonic price models for each segmented submarket, will then be compared with a statistical clustering method, i.e., the combination of PCA and K-means clustering. It is hoped that the proposed SIP-based segmentation method can improve the computational efficiency and accuracy of current housing-market segmentation approaches.

2 Literature Review

2.1 Housing Market Segmentation

Housing market segmentation is a process of delineating several small-scale homogeneous housing submarkets from a large-scale heterogeneous housing market (Adair et al., 1996). Within a housing submarket, dwellings are close substitutes (i.e., the dwellings have similar marginal hedonic prices) for each other but weak substitutes for the dwellings in other submarkets (Bourassa et al., 1999; Wu et al., 2018). The concept of housing market segmentation is original from the study of Schnare and Struyk (1976), which argued that a set of compartmentalized and unique housing submarkets generally occur within a larger urban housing market due to the highly inelastic demands of households for particular structural and locational attributes of dwellings, coupled with these inelastic supplies. Since then, housing market segmentation has been widely applied in the real estate research for its capability for better understanding the neighborhood effects, predicting housing prices, and evaluating revitalization policies (Wu & Sharma 2012; Manganelli et al. 2014; Wu et al., 2018).

Although a general agreement has been reached on the existence of housing submarkets in the current literature, a less consensus has been achieved on the universally accepted empirical criteria and methodologies for housing market segmentation (Wu et al., 2018). In the early stage, housing submarkets are commonly defined in an ad hoc manner, i.e., housing submarkets are segmented on the basis of empirical criteria predefined by expert judgment (Bourassa et al., 1999). This way of segmentation is also called a priori segmentation. A priori segmentation can be further classified into two major streams: the geography-based segmentation and the quality-based segmentation (Islam & Asami, 2009). The geography-based segmentation focuses on the identification of housing submarkets using geographically contiguous boundaries, including planning subareas such as inner city and outer city and administrative areas such as different districts. For example, by analyzing 1,080 samples of the residential-property transactions using hedonic price models, Adair et al. (1996) discovered the existence of housing submarkets in the inner city, the middle city, and the outer city of the Belfast, UK. On the other hand, the quality-based segmentation delineates housing submarkets based on the quality attributes of dwellings, including dwelling types such as detached and semi-detached and dwelling heights such as low-rise and high-rise. For example, by examining 544 dwellings sampled from the market sales in the Glasgow city, Scotland, Watkins (2001) found that dwelling-type submarkets, i.e., flat, detached, semi-detached, and terraced submarkets, actually exist in the city's housing market. Although convenient, intuitive, and straightforward, considering the over-reliance on the experts' subjective knowledge on the formation mechanism of housing submarkets, a priori segmentation has been considered to be deficient in objectively achieving the factual number of submarkets that possess the highest levels of internal homogeneity and external heterogeneity (Bourassa et al., 1999; Watkins, 2001).

2.2 Data-driven Approach in Housing Market Segmentation

Considering the deficiency of a priori segmentation approach in deriving the factual number of submarkets, the data-driven segmentation has been given much prominence in housing market segmentation research for its capability for delineating the factual number of housing submarkets more objectively and accurately. Unlike a priori segmentation that requires experts' intuition or experience, the data-driven segmentation aims to uncover the housing market's underlying patterns and then decompose the single market into several distinctive submarkets on a basis of data structure (Bourassa et al., 1999). Compared with the priori segmentation approaches, the data-driven segmentation approaches are considered to be more objective and accurate if the data are analyzed using robust statistical tools (Wu et al., 2018). Several statistical tools have been developed for the identification of housing submarkets, including factor analysis (Dale-Johnson, 1982), discrete choice models (Tu, 1997) and neural networks (Kauko et al., 2002). Due to the rapid development of information technology, clustering analysis has become one of the most popular and widely adopted data-driven methods for its superior computational efficiency (Wu et al., 2018).

Clustering analysis is a data-mining technique for dividing a set of data observations into multiple groups or clusters according to their similarity so that the observations within a cluster are similar to each other but dissimilar to the observations in other clusters (Han et al., 2012).

Considering that traditional clustering methods, such as partitioning methods, hierarchical methods, and density-based methods, are deficient in dealing with a high-dimensionality dataset due to the slowness of the convergence (Su et al., 2009), PCA has generally been utilized as a preprocessing step of the traditional clustering methods for the reduction of data dimension (Han et al., 2012; Helbich et al., 2013). Specifically, PCA is capable of reducing the dimensions of the original data into several orthogonal principal components (PCs), and then the PCs with highest variations are selected to serve as the data input of clustering analysis. The combination of PCA and clustering analysis for housing market segmentation was firstly proposed by the study of Bourassa et al. (1999), which integrated PCA into K-means clustering and Wald's clustering for identifying housing submarkets from 4,600 individual dwellings in the Sydney and Melbourne metropolitan areas. Following this work, a great number of studies have adopted different clustering methods to combine with PCA for housing market segmentation. For example, Wu and Sharma (2012) adopted a spatially constrained K-means clustering coupled with the PCA to identify housing submarkets from 86,000 single-family housing units in the city of Milwaukee, Wisconsin. Helbich et al. (2013) developed a data-driven framework, combining the Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) algorithm with the PCA, to segment the Austrian housing market with 3,800 geocoded homes into a set of spatially contiguous submarkets. Most recently, Wu et al. (2018) proposed a data-driven framework, combining a density-based spatial clustering algorithm with a geographically weighted PCA, to segment the housing market in Shenzhen, China. Therefore, the data-driven framework that combines the PCA and traditional clustering methods have been widely applied in existing housing market segmentation research, which is also called the statistical clustering method (Wu et al., 2018).

However, the statistical clustering method for housing market segmentation is subjective to two major limitations. The first is the use of PCA within the statistical clustering method tends to lose some important low-variance information that can distinguish different housing submarkets. Statistical clustering method assumes that the PCs that contain the highest variance of the original dataset are the most useful information for the use of clustering analysis to segment housing submarkets. However, this assumption is not always true when the separation of housing submarkets is more pronounced in the direction of lower variance (Reif 2018). In this case, the selected PCs fail to reflect the underlying data structure of the housing dataset, which can lead to the poor segmentation performance of further clustering analysis. The other deficiency is observed in the identification of an optimal number of clusters (i.e., housing submarkets) by using traditional clustering methods (Bourassa et al., 1999). During the traditional clustering process, the optimal number of clusters is generally determined by optimizing a pre-defined cluster-validation function (e.g., minimizing the total within-cluster variation) over a range of possible value of clusters using the built-in gradient descent algorithm. However, such a built-in gradient descent algorithm of traditional clustering methods is easy to fall into a local optimum rather than a global optimum (Su et al., 2009). Considering the aforementioned two deficiencies of the statistical clustering method, it is essential to adopt a novel clustering approach that is capable of dealing with a high-dimensionality housing dataset without losing the key low-variance information and meanwhile, derives a globally optimal number of housing submarkets.

Swarm-intelligence-based clustering is one of the most advanced and novel clustering methods in the computer science field (Thrun, 2018). Swarm intelligence is a branch of artificial intelligence that is inspired by the collective behaviors of living things (Rana et al., 2011). For example, particle swarm optimization (POS) algorithm is one typical example of the swarm-intelligence-based algorithm. POS algorithm is a population-based globalized search algorithm that mimics the flocking behaviors of birds (Kennedy and Eberhart (1995). Besides the POS algorithm, other typical swarm-intelligence-based algorithms include the ant-based algorithm that mimics the social behaviors of ants (Labroche et al., 2003) and the information-flocking-based algorithm that mimics the behaviors of fishes (Picarougne et al., 2004). The swarm-intelligence-based algorithm has been seen as an effective tool for clustering problems due to its flexibility, robustness, decentralization, and self-organization (Su et al., 2009; Thrun, 2018), which can provide us with new insight in housing market segmentation research.

For dealing with the high-dimensionality dataset, Su et al. (2009) developed the swarm-inspired projection (SIP) algorithm, which is inspired by the collective behaviors of doves. The SIP algorithm is capable of directly projecting high-dimensionality data into a low-dimensionality space for visually identifying the inherent clusters within the dataset. The SIP algorithm is expected to overcome the two major deficiencies of the statistical clustering method for housing market segmentation for the following two reasons: First, the self-organizing feature of the SIP algorithm makes it capable of reducing the data dimension while preserving the topological properties of the input space, which avoids the loss of essential low-variance information for distinguishing different clusters. Second, the data-projecting feature of the SIP algorithm makes it capable of determining a globally optimal number of clusters, which avoids the local minimization problem incurred from the use of the gradient descent algorithms built-in within the traditional clustering methods. Therefore, compared with the statistical clustering method, the SIP algorithm is expected to have a better performance in identifying a factual number of housing submarkets from a high-dimensionality housing dataset.

2.3 Comments on Previous Work

Identifying the factual number of submarkets from single heterogeneous housing market can contribute to the real estate research, for example, improving the predictive accuracy of advanced real-estate price modelling like hedonic price modelling. Data-driven housing market segmentation has been given much prominence in recent years for its capability for objectively and accurately delineating housing submarkets based on the underlying structure of housing dataset. However, most existing data-driven housing market segmentation studies adopt the statistical clustering method that combines PCA and traditional clustering methods such as K-mean clustering, which is deficient in deriving a globally optimal number of housing submarkets due to the tendency of losing the key low-variance information and the susceptibility of converging on a locally optimal number of clusters. On the other hand, in the computer science field, the application of swarm intelligence in clustering problems provide us with an opportunity to overcome the aforementioned weaknesses of the statistical clustering method. Considering that the self-organizing and data-projecting features of the SIP algorithm make it possible to determine a globally optimal number of clusters while avoiding the loss of essential low-variance information, the present study aims to examine the effectiveness of the SIP algorithm in housing market segmentation using a high-dimensionality housing dataset.

3 Methodology

As shown in Fig. 1, to investigate the capability of the swarm-inspired projection (SIP) algorithm for segmenting housing submarkets from a high-dimensionality housing dataset, the present study compares the segmentation performance of the SIP algorithm with a statistical clustering method – the combination of PCA and K-means clustering – in delineating the factual number of housing submarkets from the Taipei city's housing market. It is expected that the factual number of submarkets possess the highest level of internal homogeneity and external heterogeneity, which can derive the housing price models with high predictive accuracy. Therefore, hedonic price modelling is used to examine the segmentation performance of the SIP algorithm and the combination of PCA and K-means, and three performance-evaluation measures are adopted to measure the predictive accuracy of the hedonic models established for each segmented housing submarkets: the adjusted R-squared (R^2), the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE). The following subsections discuss more details about the study area and housing dataset, the SIP-based segmentation, the PCA and K-means-based segmentation, hedonic price modelling, and the performance-evaluation measures.

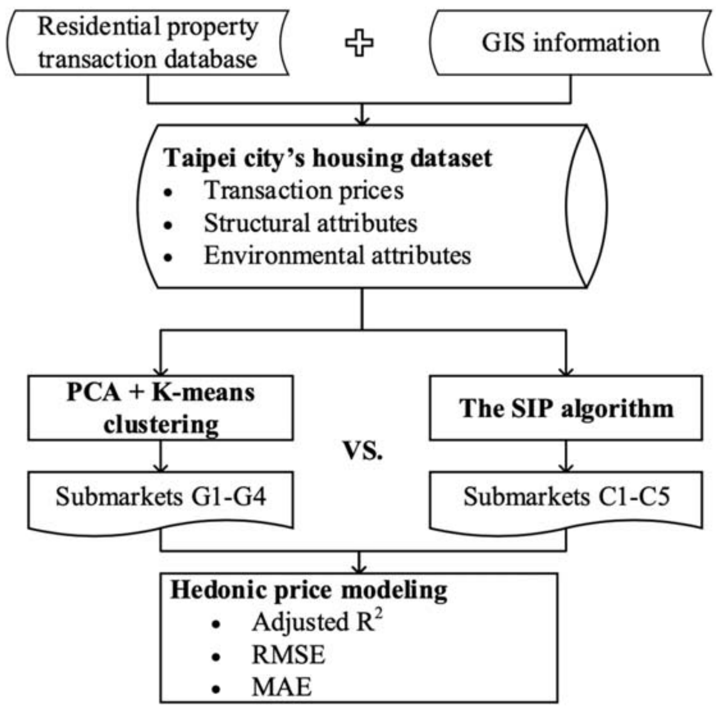

Fig. 1. Methodological diagram

3.1 Study Area and Housing dataset

Taipei city, the capital city of Taiwan is chosen as the study area as the city is one of the most densely populated and well-developed modern metropolises in Eastern Asian, covering an area of 271.80 km^2 with the population of 2.69 million (Taipei City Government, 2017). Taipei city presents itself as an interesting study area for analysing the structures of the housing market as a large number of residential property transactions occur within the city. Although there exist several studies on housing-price prediction in Taipei city (e.g., Chen et al. (2011) and Chen et al. (2017)), the underlying assumption of these studies is that Taipei city's housing market is a uniform entity without inherent distinctive housing submarkets. Therefore, it is worthy of investigating the existence of housing submarkets and its influence on the housing-price prediction through our proposed SIP algorithm.

A well-established housing dataset is essential for further data analysis. It is expected that each data record represents a residential property with its transaction price and associated attributes, including structural attributes and environmental attributes. Structural attributes, also called physical attributes, are unique internal characteristics of each residential property, including the age and floor area of the property. Environmental attributes are the external environmental characteristics attached to each property, which can be further classified into two categories: transportation-related attributes (e.g., the distance to airport or railway station) and facility-related attributes (e.g., the number of libraries or art centres nearby) (Chen et al., 2017). Environmental attributes can be measured using two types of variables, i.e., distance-based variable (e.g., the distance from the nearest mass transit system) and quantity-based variable (e.g., the number of shopping malls within 800 meters). The 800-meter threshold is adopted as the equivalence of a commuter's maximum walking distance (i.e., 0.5 miles) (O'Sullivan & Morrall, 1996).

The data were primarily drawn from the Gigahouse Taiwan's Real Estate Portal database, containing 5,012 transactions of residential properties in the Taipei city's housing market from 2008 to 2010. Each transaction record contains one transaction price and 10 structural attributes of the property. 28 environmental attributes of each property were extracted through GIS analysis, including 10 transportation-related attributes and 18 facility-related attributes, all of which were further combined with each transaction record as a new data sample. A housing dataset containing 5,012 data samples with 39 variables (or dimensions) were initially established. To guarantee data quality for further analysis, data cleaning was conducted for the dataset. After eliminating both missing values and extreme values, 4,136 valid samples remain within the dataset. To improve the accuracy and efficiency of the further clustering algorithm, data normalization was applied for the cleaned dataset, where the data were scaled to fall within a range between 0 to 1. Table A.1. shows the descriptive details of the variables, their names, descriptions, and some basic statistics (i.e., mean and standard deviation).

3.2 The SIP-based Segmentation

As one form of swarm intelligence, the SIP algorithm is a novel data projection algorithm inspired by the foraging behaviors of doves. Given that most of the previous swarm-intelligence-based algorithms were deficient in clustering high-dimensionality data, Su et al. (2009) invented the SIP algorithm by integrating the double self-organizing feature map (DSOM) algorithm (Su and Chang, 2001) into the basic concept of the swarm-intelligence-based algorithm. In the Su et al. (2009)'s SIP algorithm, each data pattern, i.e., one record of data with multiple dimensions, is regarded as one artificial crumb. All the data records are tossed as artificial crumbs for feeding a flock of doves that are initially set by the user. All the doves have their artificial sense organ (measured by a multi-dimensional sense organ vector) to perceive the existence of data patterns and their initial positions (measured by a two-dimensional position vector). The flock of doves then adjust their positions to seek for the crumbs based on their degrees of satiety (measured by a satiety parameter). The individual doves know each other's foraging status and mimic the behaviors of the doves with the best performance in foraging crumbs. Each dove has its foraging strategy, which is adjusted according to their degree of satiety. For example, the individual dove with a lower degree of satiety is more likely to imitate other successful doves. In contrast, the individual dove with a higher degree of satiety is more conservative and less likely to change its foraging strategy. Consequently, the flock of doves gradually form into different groups of doves based on the distribution of artificial crumbs. The number of inherent clusters from the initial dataset can be observed by viewing the distribution of doves. The details of the SIP algorithm can refer to the study of Su et al. (2009).

3.3 The PCA and K-means-based Segmentation

The combination of PCA and K-means clustering is then applied to segment the same housing dataset. This method has been widely used in previous housing market segmentation studies (e.g., Bourassa et al. (1999), Bates (2006), and Wu and Sharma (2012)). PCA is first applied to the housing dataset for deriving a set of orthogonal PCs that are sorted in descending order by the explained proportion of variance. To retrieve an appropriate number of PCs that can explain the maximum proportion of variance, the criteria that the chosen PCs should at least explain 70% of the total variance (Jolliffe and Cadima, 2016) is adopted in the present study. Hence, 15 PCs are finally selected (Table A.2.). The scores of these selected PCs serve as the data input for K-means clustering analysis. The number of optimal clusters (i.e., the number of K) from K-means clustering is determined by the Elbow method, which aims to look for the number of K that can achieve minimum intra-cluster variation (Han et al., 2012).

3.4 Hedonic Price Modelling and Performance-evaluation Measures

In the real estate studies, hedonic price modelling has been seen as a promising method for understanding housing market structure, estimating the value of properties, and predicting the housing price (Islam & Asami, 2009). By regressing the housing price on a vector of structural and environmental attributes, the implicit or shadow prices of these attributes, also called hedonic prices, can be revealed from the estimated coefficients of the regression models (Rosen, 1974; Adair et al., 1996). Considering that the hedonic prices for housing attributes vary among distinctive housing submarkets, it is expected that the accurate identification of the factual number of housing submarkets can contribute to a high level of predictive accuracy of hedonic price models (Bourassa et al., 1999).

Therefore, hedonic price modelling is conducted for each SIP-segmented and PCA and K-means-segmented housing submarket. Multiple linear regression model is chosen as the functional form to fit the explanatory variables, i.e., the attributes of the property, to the response variable, i.e., the unit price of the property. The formula is shown as Eq. (1).

$$P_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (1)$$

where P_i refers to the unit price of the ith sample; X_{ki} refers to the value of the kth attribute of the ith sample; β_0 and β_k refer to the intercept and the estimated coefficients for the kth attribute, respectively; ε_i refers to the error term.

The parameters of the linear model, β_0 and β_k, are estimated by the Ordinary Least Squares (OLS) method for minimizing the sum of squared residuals. Backward elimination based on Akaike information criterion (AIC) is adopted to iteratively remove the least important explanatory variables for deriving a best-performing model. Model diagnostics are conducted to check whether the assumptions of the linear regression model are met or not. The assumptions, in terms of the linearity of the data, the normality of residuals, and the homogeneity of residuals variance, are examined using residual plots. Cook's distance is used to examine the presence of influential values (i.e., high-leverage points). Multicollinearity of the explanatory variables is assessed by the variance inflation factor (VIF), which should be less than five as suggested by James et al. (2013).

To examine the predictive accuracy of the hedonic price models for each submarket, three performance-evaluation measures are adopted: 1) Adjusted R-squared (R^2), an adjusted version of R^2 to measure the proportion of variation in the outcome that can be explained by the predictors with a penalty. 2) The Root Mean Squared Error (RMSE), the square root of the average squared difference between the observed and predicted outcome. 3) The Mean Absolute Error (MAE), the average absolute difference between the observed and predicted outcome, which is less sensitive to the outliers compared to RMSE. Higher Adjusted R^2 means the model has a higher level of statistical explanation, and lower RMSE and MAE mean the model has a lower level of prediction error (Kassambara, 2018).

4 Analysis Results

Fig. 2. shows the clustering process of the swarm-inspired projection (SIP) algorithm for the Taipei city's housing dataset. Initially, 49 doves are uniformly deployed on a two-dimensional artificial ground, as shown in Fig. 2(a). After five iterations (Fig. 2(b) - (f)), these doves gradually move to different clusters of artificial crumbs (i.e., data records) according to the underlying structure of the dataset (i.e., the unique features of 39 data attributes). Five clusters can be visually identified, shown as five distinct dense regions in the scatter plot (Fig. 2(f)). Therefore, the SIP algorithm identifies five housing submarkets within the Taipei city.

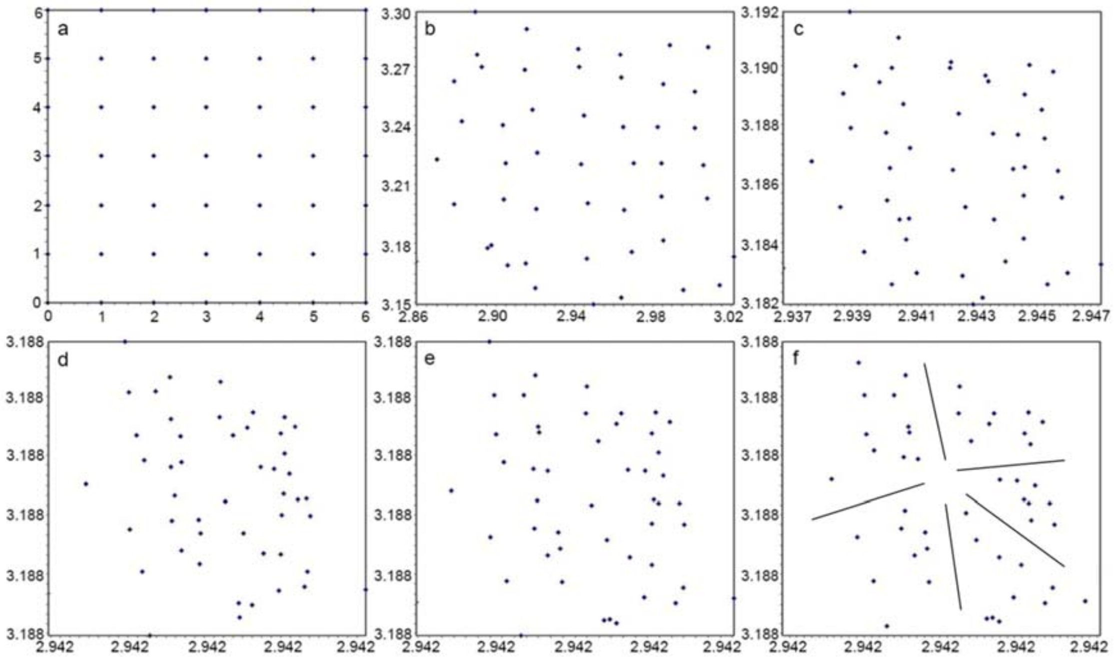

Fig. 2. The SIP clustering process for the Taipei city's housing dataset

Table 1. shows the housing market segmentation results derived from the SIP algorithm and the combination of PCA and K-means. The number of housing submarkets derived from these two approaches is inconsistent. Five submarkets (C1 to C5) are found using the SIP algorithm, while four submarkets (G1 to G4) are found using the PCA and K-means. To compare the segmentation results of these two approaches, hedonic price modelling is conducted for each submarket. The results show that the linear-regression assumptions for hedonic price models all hold (Fig. A.1 to A.10) and no multicollinearity is observed for the predictors of each model (Table A.3), which indicates that the statistical measures derived from the hedonic price models are convincing.

Table 1. The housing market segmentation results

	The SIP algorithm					The PCA and K-means			
Submarket	C1	C2	C3	C4	C5	G1	G2	G3	G4
Quantity	1065	1019	521	591	940	609	1587	559	1380
Ave. unit price[a]	0.441	0.439	0.537	0.454	0.456	0.492	0.521	0.402	0.393

Note: Ave. unit price[a] refers to the average housing price per unit area.

Table 2. shows the comparison of the hedonic price modelling results from the single market (i.e., the original dataset) and the five SIP-segmented submarkets. As suggested by the study of Adair et al., (1996), well-defined housing submarkets can be observed from the following two aspects. First, the level of statistical explanation of the hedonic price models for submarkets should be higher than the one for the original single market, given that the submarkets should be more homogenous than the single market. Such phenomenon can be observed from the hedonic price models for the SIP-segmented submarkets: the adjusted R^2 for each submarket is relatively high ranging from 0.36 to 0.58; especially for the submarket C3, the adjusted R^2 (0.58) increases by nearly 66% compared with the single market (0.35). Second, different combinations of significant variables and their diverse hedonic prices can be observed for each submarket, given that the properties are close substitutes within a submarket but weak substitutes for other submarkets' properties. As shown in column three and four of Table 2, the significant positive and negative attributes with a p-value less than 0.001 are sorted in descending order by their hedonic prices (i.e., the model coefficients). It can be found that the significant attributes differ to a certain extent between the hedonic models due to the disparate compositions of the housing submarkets. For example, the attribute with the highest hedonic price varies among submarkets. To be specific, the *holding_ratio* has the highest hedonic price within the submarket C1, while the *quan_ArtCenter* has the highest hedonic price within the submarket C3; The *quan_lib* is the most influential attribute affecting the housing price within the submarket C4 but turns to be insignificant within the submarket C5. Therefore, the results suggest that the SIP algorithm is capable of delineating more homogenous and distinctive housing submarkets from a high-dimensionality housing dataset.

Table 2. The hedonic price modelling results for the single market and SIP-segmented submarkets

Housing Market	Adjusted R^2	Significant positive attributes[a]	Significant negative attributes[a]	Number of Attributes
Single market S	0.35	quan_hospital, Holding_ratio, quan_ArtCenter, quan_university, Hall, quan_MRT, quan_supermarket, dis_airport, dis_expressway,	quan_police, dis_university, dis_CBD, quan_expressway,	20

		quan_DepartStore, Height, dis_police, quan_nightmarket, dis_hospital	dis_interchange, dis_MRT	
Submarket C1	0.44	Holding_ratio, quan_DepartStore, quan_university, quan_supermarket, dis_expressway, dis_RailwayStation, quan_MRT, quan_nightmarket, dis_police	quan_police, dis_university, dis_CBD, quan_expressway	13
Submarket C2	0.36	quan_hospital, Holding_ratio, quan_ArtCenter, Bedroom, quan_MRT, Height, quan_DepartStore, dis_nightmarket,	quan_police, dis_university, dis_MRT	11
Submarket C3	0.58	quan_ArtCenter, dis_expressway, quan_DepartStore	quan_police, dis_CBD, dis_RailwayStation, quan_expressway	7
Submarket C4	0.54	quan_lib, quan_hospital, Hall, quan_MRT, dis_RailwayStation, dis_police	quan_police, dis_CBD, dis_fire	9
Submarket C5	0.45	quan_DepartStore, dis_expressway, quan_university, quan_LargeRetail, dis_airport	dis_CBD, quan_expressway	7

Note: [a]The attributes are sorted in descending order by their coefficients.

Table 3 shows the predictive accuracy of the hedonic price models for the SIP-segmented submarkets and the PCA and K-means-segmented submarkets, making the single market as a benchmark. In terms of the level of statistical explanation, the adjusted R^2 for each SIP-segmented submarket increases on an average of 35.9%, ranging from 3.1% to 66.4%. Especially for the submarket C3 and C4, the Adjusted R^2 exceeds over 50% compared to the single market. Although the combination of PCA and K-means delineates the submarket G1 to G3 with high Adjusted R^2, this method also derives the submarket G4 with poor statistical explanation (Adjusted R^2=0.263). On the other hand, the hedonic price models for the SIP-segmented submarkets show good performance in terms of the reduced level of prediction error. Specifically, the RMSE and MAE for each SIP-segmented submarket decrease on an average of 14.4% and 13.9%, respectively. However, the reduced RMSE and MAE are unstable across the PCA and K-means-segmented submarkets, representing in a low level of the prediction error of the hedonic price models for the submarket G1, G3, and G4 but a high level of prediction error for the submarket G2. Even worse, the MAE of the hedonic price model for the submarket G2 is higher than the one for the original single market. Therefore, the overall predictive accuracy of the hedonic price models for the SIP-segmented submarkets is more satisfying compared to the ones for the PCA and K-means-segmented submarkets. The results imply that the SIP algorithm generally outperforms the PCA and K-means in segmenting an optimal number of housing submarkets for which the hedonic price models can achieve a higher level of predictive accuracy.

Table 3. The comparison of the predictive accuracy of the hedonic price models

Housing	Adjusted R^2		RMSE		MAE	
market	Value	Improvement	Value	Improvement	Value	Improvement
Single market						
S	0.347	-	0.133	-	0.105	-
The SIP-segmented submarkets						
C1	0.443	27.6%	0.116	-12.4%	0.092	-12.0%
C2	0.358	3.1%	0.112	-15.6%	0.089	-15.3%
C3	0.578	66.4%	0.113	-15.2%	0.088	-16.0%
C4	0.535	54.1%	0.111	-16.7%	0.089	-14.5%

C5	0.445	28.1%	0.117	-12.1%	0.093	-11.6%
The PCA and K-means-segmented submarkets						
G1	0.380	9.4%	0.115	-13.4%	0.092	-12.5%
G2	0.466	34.2%	0.132	-1.0%	0.105	**0.1%**
G3	0.466	34.2%	0.090	-32.5%	0.071	-32.3%
G4	0.263	**-24.3%**	0.104	-21.7%	0.082	-21.5%

5 Discussion and Conclusions

With the rapid development of information technology and geographic information systems, it is easier to assemble a high-dimensionality housing dataset with numerous structural and environmental attributes from online resources. For both researchers and practitioners in the real estate field, more objective and accurate delineation of housing submarkets from a high-dimensionality housing dataset is challenging, especially in the identification of a globally optimal number of housing submarkets without losing essential low-variance information which the statistical clustering method (e.g., the combination of PCA and K-means clustering) is deficient in. Therefore, the present study introduces the swarm-inspired projection (SIP) algorithm for identifying housing submarkets from a high-dimensionality housing dataset. The usefulness of the SIP algorithm for housing market segmentation is illustrated by segmenting the Taipei city's housing dataset containing the transaction prices of residential properties and associated 38 attributes. The segmentation performance of the SIP algorithm is competed with the combination of PCA and K-means clustering through evaluating the levels of statistical explanation and prediction error of the hedonic price models constructed for each submarket. Two major research findings can be found from the analysis results:

First, the SIP algorithm is effective in identifying more homogenous and distinctive housing submarkets from a high-dimensionality housing dataset. By analysing the Taipei city's housing dataset using the SIP algorithm, five housing submarkets can be visually identified from a two-dimensional plot. The statistical measures derived from the hedonic price models established for each SIP-segmented submarket further demonstrate the effectiveness of the SIP algorithm in housing market segmentation, representing in 1) the hedonic price model for each SIP-segmented submarket shows a higher level of statistical explanation than the one for the original single market, and 2) different combinations of significant attributes and their diverse hedonic prices can be observed for each SIP-segmented submarket. Second, the SIP algorithm outperforms the use of PCA and K-means in deriving a globally optimal number of housing submarkets for which the hedonic price models can achieve a higher level of predictive accuracy. Specifically, the improved statistical explanation and the reduced prediction error of the hedonic price models for the SIP-segmented submarkets are more stable than the ones for the PCA and K-means-segmented submarkets, representing in 1) all the performance-evaluation measures (i.e., Adjusted R^2, RMSE, and MAE) of the hedonic models for each SIP-segmented submarket (C1 to C5) are better than the ones for the original single market (S), however, 2) the poor explanation power and the high prediction error can be observed in the PCA and K-means-segmented submarket G4 and G2, respectively. These two research findings add support to the previous argument that the SIP algorithm can overcome the weaknesses of the statistical clustering method, namely, the tendency of losing some essential low-variance information that can distinguish housing submarkets and the susceptibility of converging on a locally optimal number of clusters rather than a global optimum. Therefore, the SIP algorithm can serve as a powerful data-driven approach for identifying a factual number of homogenous and distinctive housing submarkets from a high-dimensionality housing dataset for which hedonic price models can achieve a higher level of predictive accuracy.

The major contribution of the present study is to propose a novel swarm-inspired data-driven segmentation approach for complementing current housing market segmentation research. In most existing studies, the data-driven framework for identifying housing submarkets is the statistical clustering method that combines the use of PCA and traditional clustering methods (e.g., K-means clustering), which tends to loss essential low-variance information and converges on a locally optimal number of clusters when handling a high-dimensionality dataset. The SIP algorithm, due to its self-organizing feature and data-projecting feature, overcomes the aforementioned weaknesses of the statistical clustering method by directly projecting high-dimensionality data into a low-dimensionality space for visually identifying the inherent clusters within the dataset while preserving the topological properties of the input space. Compared with the combination of PCA and K-means approach, the SIP-algorithm can better reveal a globally optimal number of homogenous and distinctive housing submarkets from the heterogeneous market structure. The SIP-segmented submarkets can derive the hedonic price models with a higher level of predictive accuracy, which can help inform the decision making of the stakeholders involved in the real state field. For example, as for property appraisers, a more accurate estimation of hedonic prices for the property's attributes and more accurate the prediction of housing price can be achieved within the homogenous and distinctive housing submarkets. Accordingly, the property-valuation strategies can target for each housing submarket for attracting potential property buyers. Not limited in the case of Taipei city, the data-independent feature of the SIP-based segmentation method makes it applicable to different urban areas.

The proposed SIP-algorithm approach has some limitations that should be acknowledged. For example, the establishment of the high-dimensionality housing dataset does not consider the characteristics of inhabitants, e.g., the socio-demographic background of inhabitants. Previous studies pointed out that housing submarkets also emerge from the diverse economic and ethnic background of inhabitants (Adair et al., 1996; Islam & Asami, 2009). To derive more accurate data analysis results, future studies are expected to establish a more comprehensive housing dataset containing not only the structural and environmental features of properties but also the socio-demographic characteristics of inhabitants. The proposed SIP-based segmentation method can also be extended to other fields that require market segmentation for supporting more accurate and effective marketing or investment strategies.

Appendixes

Table A.1. The descriptive details of the normalized variables in the housing dataset

Variable name	Description	Mean	SD
Unit_price	Selling price per unit area	0.458	0.166
Structural attributes			
Floor_area	Floor area of the property (pin, about 36 square feet)	0.203	0.107
Land_area	Land area of the property (pin, about 36 square feet)	0.179	0.097
Holding_ratio	The percentage of the property held by the buyer	0.006	0.016
Age	Age of the property	0.453	0.216
Bedroom	The number of bedrooms	0.196	0.098

Hall	The number of living halls	0.125	0.059
Bathroom	The number of bathrooms	0.084	0.089
Type	The type of property	0.220	0.141
Height	The height of the property	0.254	0.133
Sales_duration	The number of days before sale	0.617	0.227
Environmental (transportation-related) attributes			
Dis_MRT	Distance to the nearest MRT station	0.357	0.320
Quan_MRT	Quantity of nearby MRT stations	0.236	0.196
Dis_railwaystation	Distance to the nearest Railway Station	0.074	0.225
Dis_airport	Distance to the nearest airport	0.034	0.151
Dis_interchange	Distance to the nearest interchange	0.066	0.214
Dis_expressway	Distance to the nearest expressway	0.150	0.174
Quan_expressway	Quantity of nearby expressways	0.269	0.270
Environmental (facility-related) attributes			
Dis_university	Distance to the nearest university	0.054	0.059
Quan_university	Quantity of nearby universities	0.158	0.185
Dis_lib	Distance to the nearest library	0.446	0.288
Quan_lib	Quantity of nearby libraries	0.003	0.032
Dis_artcenter	Distance to the nearest libraries	0.306	0.325
Quan_artcenter	Quantity of nearby art centers	0.066	0.084
Dis_largeretail	Distance to the nearest large-scale retail stores	0.146	0.286
Quan_largeretail	Quantity of nearby large-scale retail stores	0.066	0.124
Dis_departstore	Distance to the nearest department store	0.189	0.304
Quan_departstore	Quantity of nearby department stores	0.092	0.178
Dis_supermarket	Distance to the nearest supermarket	0.310	0.191
Quan_supermarket	Quantity of nearby supermarkets	0.311	0.149
Dis_nightmarket	Distance to the nearest night market	0.149	0.276

Quan_nightmarket	Quantity of nearby night markets	0.157	0.261
Dis_hospital	Distance to the nearest hospital	0.370	0.327
Quan_hospital	Quantity of nearby hospitals	0.003	0.016
Dis_police	Distance to the nearest police station	0.446	0.268
Quan_police	Quantity of nearby police stations	0.206	0.172
Dis_fire	Distance to the nearest fire department	0.378	0.343
Quan_fire	Quantity of nearby fire department	0.165	0.152
Dis_CBD	Distance to Central Business District	0.383	0.227

Table A.2. Summary of the selected principle components (PCs)

Principle component	Eigenvalue	PTV (%)	Cumulative PTV (%)
PC1	5.50	14.10	14.10
PC2	2.73	7.01	21.10
PC3	2.49	6.40	27.50
PC4	2.25	5.78	33.28
PC5	1.92	4.93	38.21
PC6	1.59	4.08	42.29
PC7	1.58	4.04	46.33
PC8	1.49	3.82	50.15
PC9	1.37	3.51	53.67
PC10	1.23	3.16	56.83
PC11	1.10	2.83	59.66
PC12	1.04	2.66	62.32
PC13	1.00	2.56	64.88
PC14	0.98	2.51	67.38
PC15	0.96	2.46	69.84

Note: PTV refers to proportion of total variance.

Table A.3. Summary of the variance inflation factor (VIF) test for each hedonic price model

	Min	Max	Mean
Single market			
S	1.064	3.196	1.562
The SIP-segmented submarkets			
C1	1.195	3.939	2.090
C2	1.038	4.469	1.944
C3	1.195	4.944	2.215
C4	1.142	3.434	1.812
C5	1.099	3.263	1.648
The PCA and K-means-segmented submarkets			
G1	1.144	2.374	1.656
G2	1.093	4.681	1.991
G3	1.147	2.800	1.787
G4	1.073	4.142	1.984

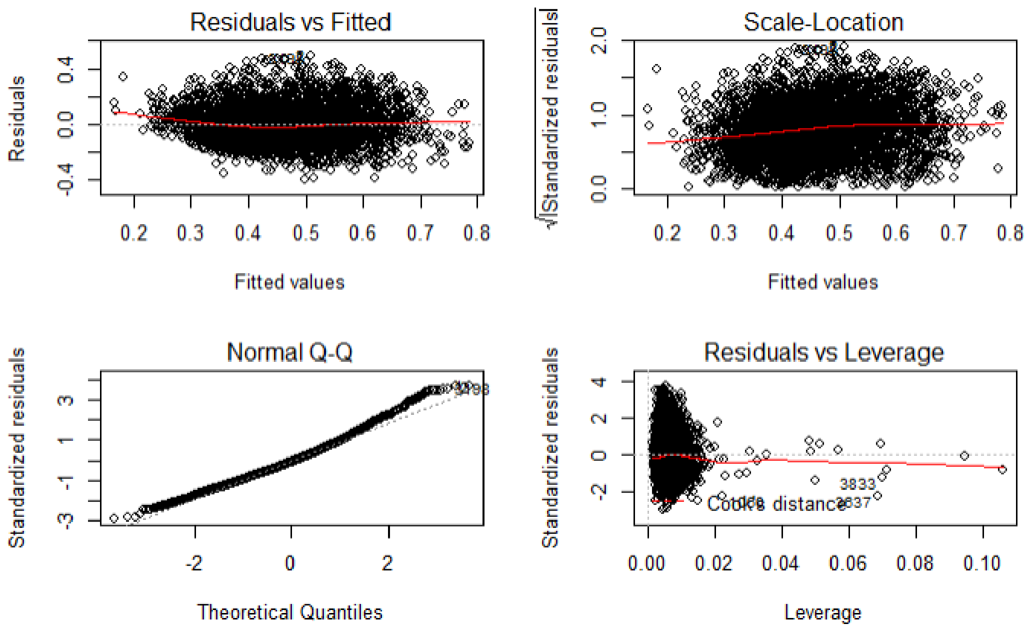

Fig. A.1. Residual plots for the single market's hedonic price model

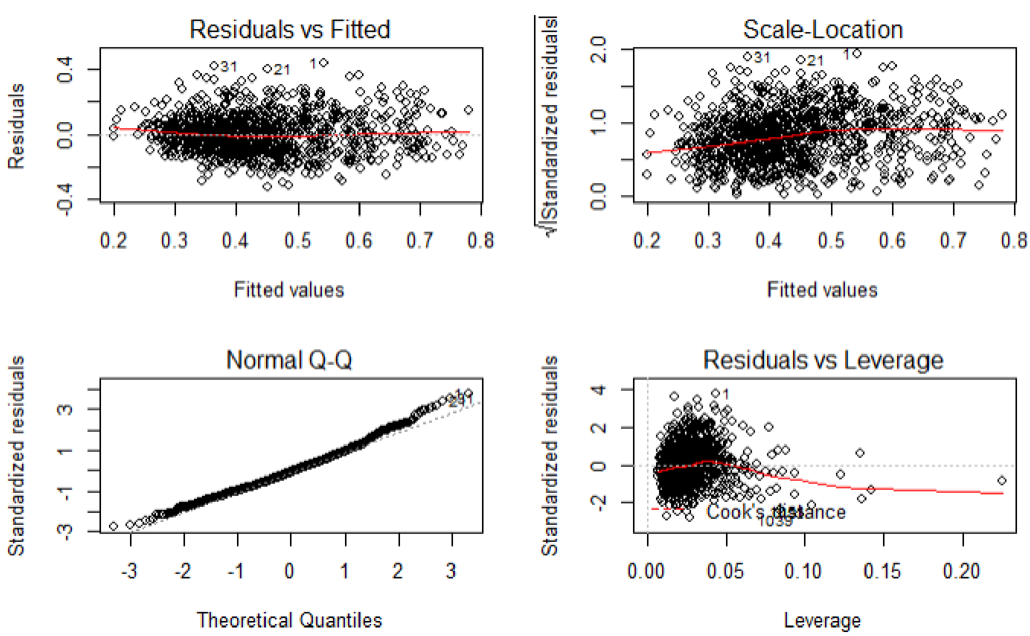

Fig. A.2. Residual plots for the submarket C1's hedonic price model

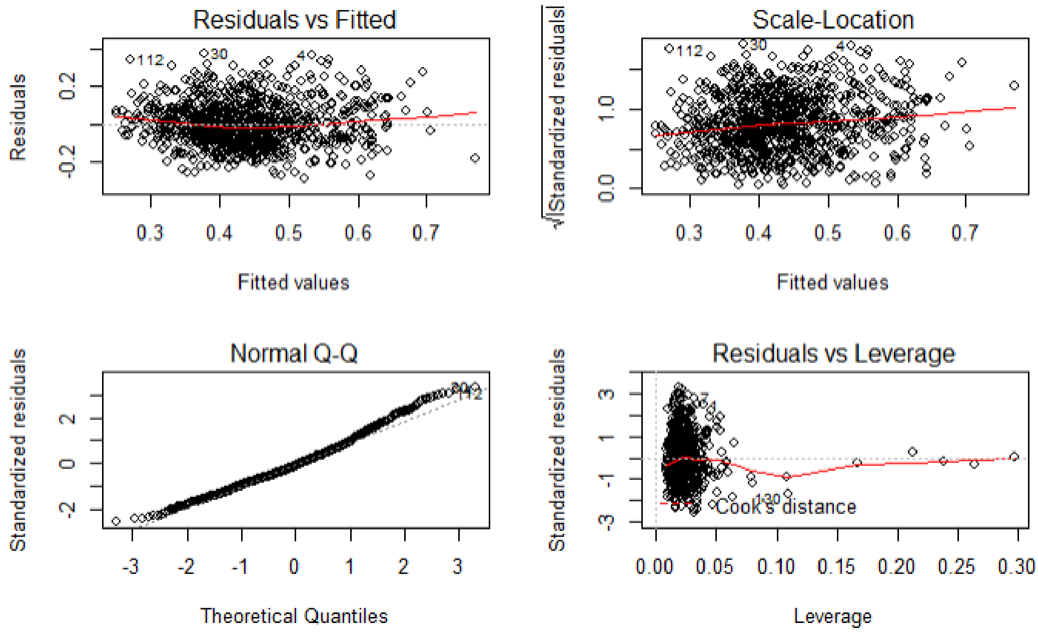

Fig. A.3. Residual plots for the submarket C2's hedonic price model

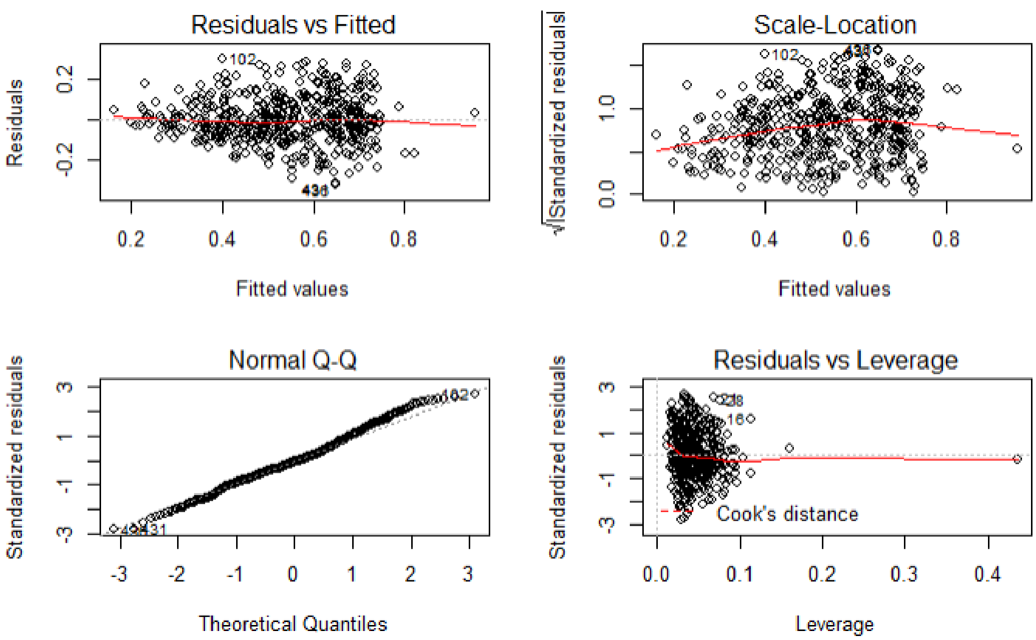

Fig. A.4. Residual plots for the submarket C3's hedonic price model

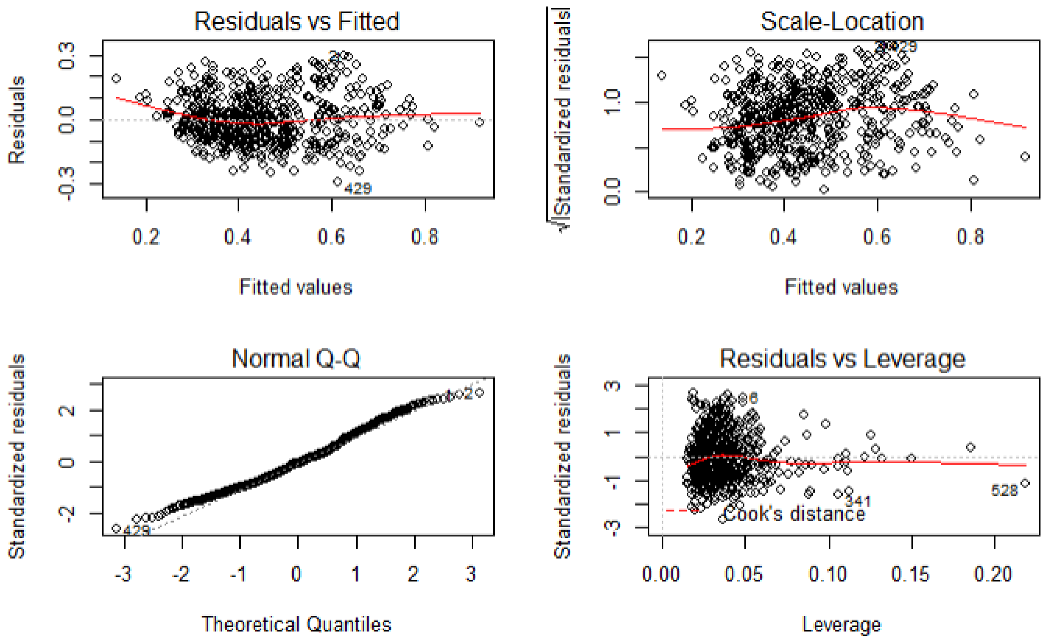

Fig. A.5. Residual plots for the submarket C4's hedonic price model

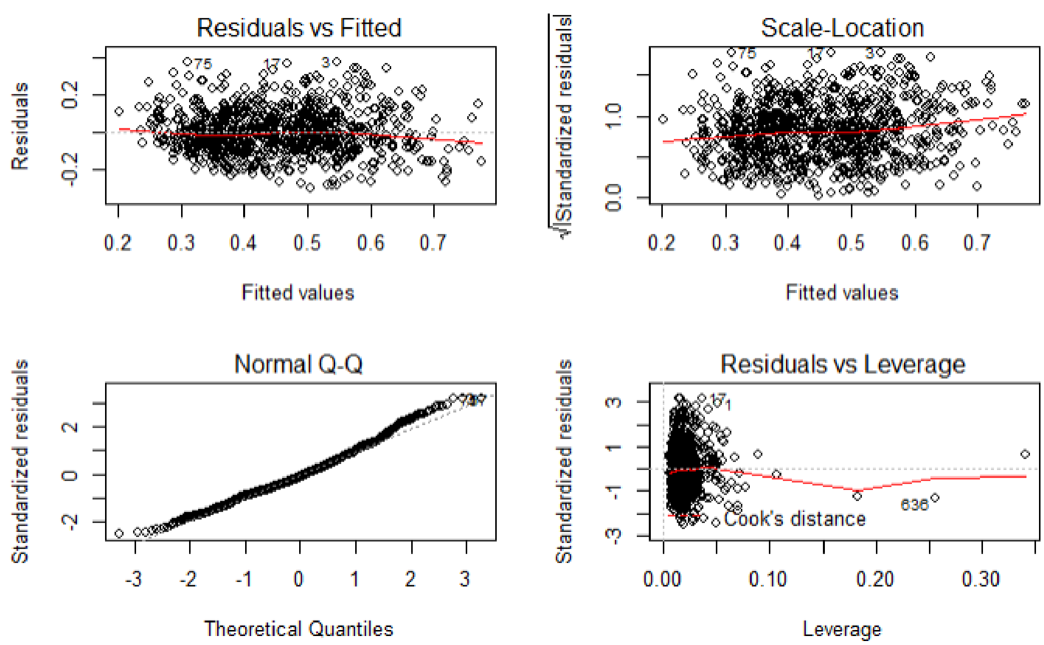

Fig. A.6. Residual plots for the submarket C5's hedonic price model

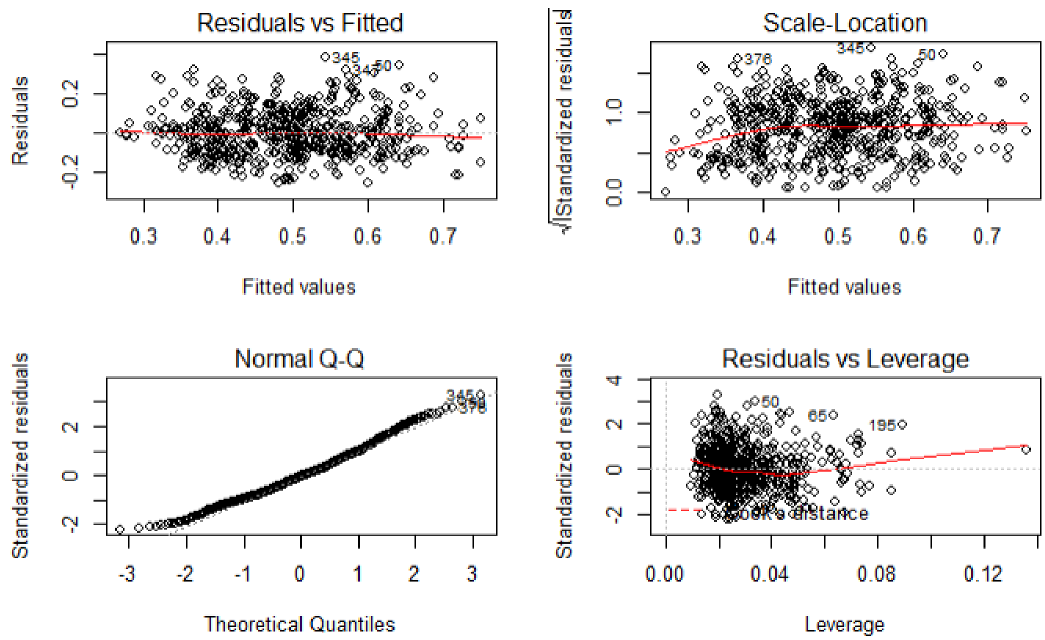

Fig. A.7. Residual plots for the submarket G1's hedonic price model

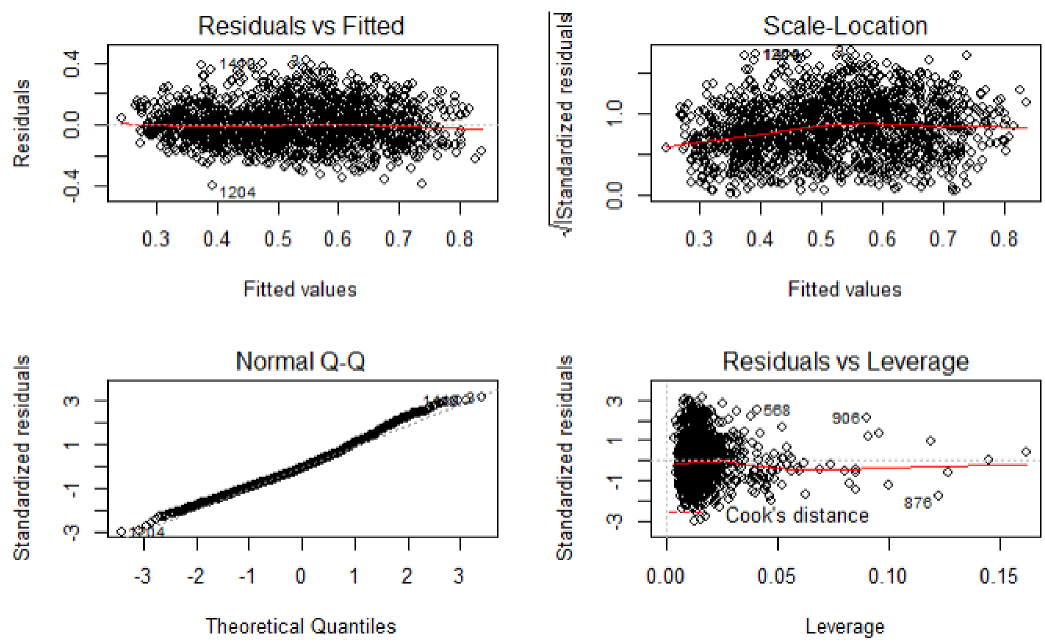

Fig. A.8. Residual plots for the submarket G2's hedonic price model

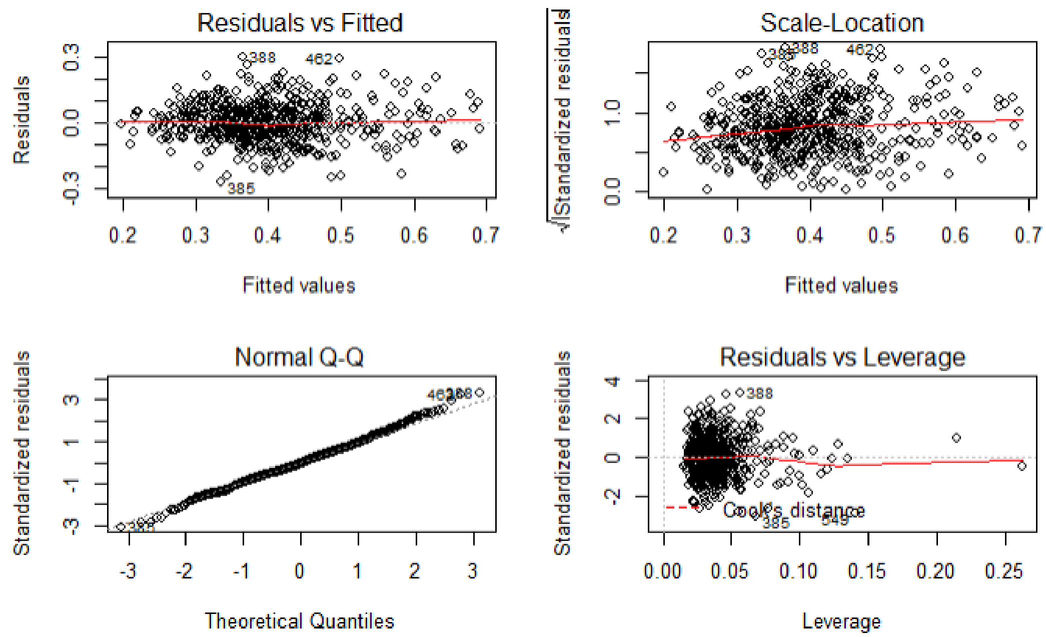

Fig. A.9. Residual plots for the submarket G3's hedonic price model

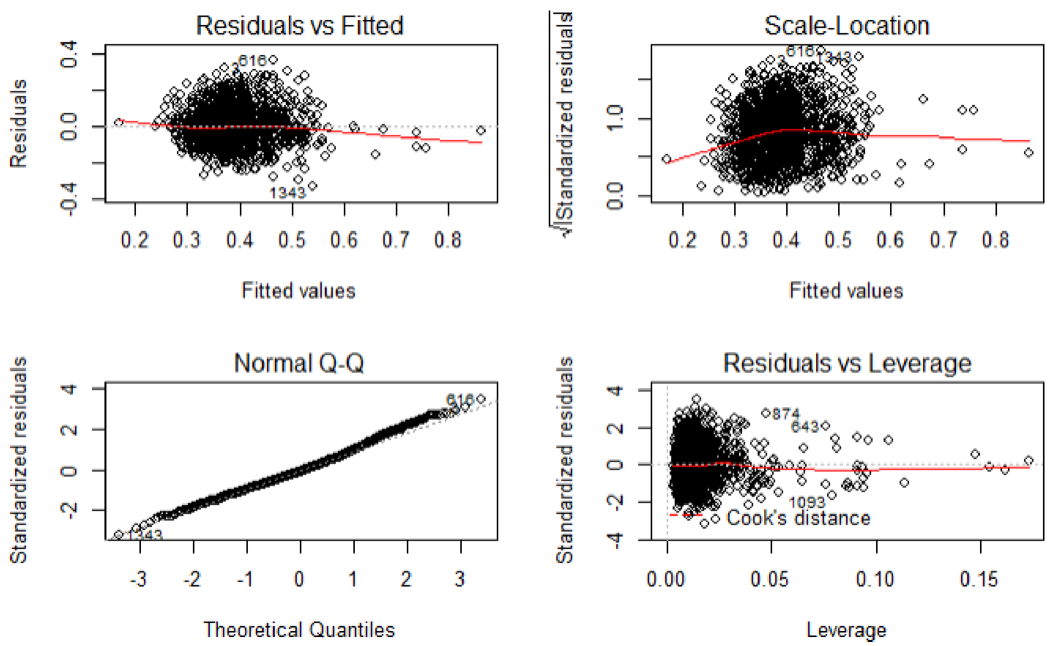

Fig. A.10. Residual plots for the submarket G4's hedonic price model

References

Adair, A. S., Berry, J. N., & McGreal, W. S. 1996. "Hedonic modelling, housing submarkets and residential valuation." *Journal of Property Research*, 13(1), 67-83. https://doi.org/10.1080/095999196368899.

Bates, L. K. 2006. "Does neighborhood really matter? Comparing historically defined neighborhood boundaries with housing submarkets." *Journal of Planning Education and Research*, 26(1), 5-17. https://doi.org/10.1177/0739456x05283254.

Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. 1999. "Defining housing submarkets." *Journal of Housing Economics*, 8(2), 160-183. https://doi.org/10.1006/jhec.1999.0246.

Chen, J.-H., Ong, C. F., Zheng, L., & Hsu, S.-C. 2017. "Forecasting spatial dynamics of the housing market using Support Vector Machine." *International Journal of Strategic Property Management*, 21(3), 273-283. https://doi.org/10.3846/1648715X.2016.1259190.

Chen, P.-F., Chien, M.-S., & Lee, C.-C. 2011. "Dynamic modeling of regional house price diffusion in Taiwan." *Journal of Housing Economics*, 20(4), 315-332. https://doi.org/10.1016/j.jhe.2011.09.002.

Dale-Johnson, D. 1982. "An alternative approach to housing market segmentation using hedonic price data." *Journal of Urban Economics*, 11(3), 311-332. https://doi.org/10.1016/0094-1190(82)90078-X.

Han, J., Kamber, M., & Pei, J. 2012. *Data mining concepts and techniques* (3rd ed.). Waltham, Mass: Elsevier.

Helbich, M., Brunauer, W., Hagenauer, J., & Leitner, M. 2013. "Data-driven regionalization of housing markets." *Annals of the Association of American Geographers*, 103(4), 871-889. https://doi.org/10.1080/00045608.2012.707587.

Hui, E. C. M., Zhong, J., & Yu, K. 2016. "Heterogeneity in spatial correlation and influential factors on property prices of submarkets categorized by urban dwelling spaces." *Journal of Urban Planning and Development*, 142(1), 04014047. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000270.

Islam, K. S., & Asami, Y. 2009. "Housing market segmentation: A review." *Review of Urban & Regional Development Studies*, 21(2-3), 93-109. https://doi.org/10.1111/j.1467-940X.2009.00161.x.

James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. *An introduction to statistical learning* (Vol. 112). New York, USA: Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kassambara, A. 2018. *Machine learning essentials: Practical guide in R*: STHDA.

Kauko, T. O. M., Hooimeijer, P., & Hakfoort, J. 2002. "Capturing housing market segmentation: An alternative approach based on neural network modelling." *Housing Studies*, 17(6), 875-894. https://doi.org/10.1080/02673030215999.

Kennedy, J., & Eberhart, R. 1995. "Particle swarm optimization." In *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia: IEEE.

Kumar, N. 2019. "Advantages and disadvantages of principal component analysis in machine learning." *The Professionals Point*. http://theprofessionalspoint.blogspot.com/2019/03/advantages-and-disadvantages-of_4.html. Accessed December 12, 2019.

Labroche, N., Monmarché, N., & Venturini, G. 2003. "Visual clustering based on chemical recognition system of ants." In *Genetic and Evolutionary Computation Conference*, Chicago, Ilinois, France: HAL.

Manganelli, B., Pontrandolfi, P., Azzato, A., & Murgante, B. 2014. "Using geographically weighted regression for housing market segmentation." *International Journal of Business Intelligence and Data Mining*, 9(2), 161-177.

Mei, Y., Zhao, X., Lin, L., & Gao, L. 2018. "Capitalization of urban green vegetation in a housing market with poor environmental quality: Evidence from beijing." *Journal of Urban Planning and Development*, 144(3), 05018011. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000458.

Mok, H. M. K., Chan, P. P. K., & Cho, Y.-S. 1995. "A hedonic price model for private properties in Hong Kong." *The Journal of Real Estate Finance and Economics*, 10(1), 37-48. https://doi.org/10.1007/bf01099610.

O'Sullivan, S., & Morrall, J. 1996. "Walking distances to and from light-rail transit stations." *Transportation Research Record*, 1538(1), 19-26. https://doi.org/10.1177/0361198196153800103.

Olszewski, K., Waszczuk, J., & Widłak, M. 2017. "Spatial and hedonic analysis of house price dynamics in Warsaw, Poland." *Journal of Urban Planning and Development*, 143(3), 04017009. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000394.

Picarougne, F., Azzag, H., Venturini, G., & Guinot, C. 2004. "On data clustering with a flock of artificial agents." In *16th IEEE International Conference on Tools with Artificial Intelligence*, Boca Raton, FL, USA: IEEE.

Rana, S., Jasola, S., & Kumar, R. 2011. "A review on particle swarm optimization algorithms and their applications to data clustering." *Artificial Intelligence Review*, 35(3), 211-222. https://doi.org/10.1007/s10462-010-9191-9.

Rosen, S. 1974. "Hedonic prices and implicit markets: Product differentiation in pure competition." *Journal of Political Economy*, 82(1), 34-55.

Schnare, A. B., & Struyk, R. J. 1976. "Segmentation in urban housing markets." *Journal of Urban Economics*, *3*(2), 146-166.

Su, M.-C., & Chang, H.-T. 2001. "A new model of self-organizing neural networks and its application in data projection." *IEEE Transactions on Neural Networks*, 12(1), 153-158. https://doi.org/10.1109/72.896805.

Su, M.-C., Su, S.-Y., & Zhao, Y.-X. 2009. "A swarm-inspired projection algorithm." *Pattern Recognition*, 42(11), 2764-2786. https://doi.org/10.1016/j.patcog.2009.03.020.

Taipei City Government. 2017. "Demographic structure and composition". *Department of Information Technology, Taipei City Government*. https://english.gov.taipei/cp.aspx?n=C619997124A6D293. Accessed December 10, 2019.

Thrun, M. C. 2018. *Projection-based clustering through self-organization and swarm intelligence: Combining cluster analysis with the visualization of high-dimensional data*. Marburg, Germany: Springer Vieweg.

Tu, Y. 1997. "The local housing sub-market structure and its properties." *Urban Studies*, 34(2), 337-353. https://doi.org/10.1080/0042098976203.

Watkins, C. A. 2001. "The definition and identification of housing submarkets." *Environment and Planning A: Economy and Space*, 33(12), 2235-2253. https://doi.org/10.1068/a34162.

Wilson, P., White, M., Dunse, N., Cheong, C., & Zurbruegg, R. 2011. "Modelling price movements in housing micro markets: Identifying long-term components in local housing market dynamics." *Urban Studies*, 48(9), 1853–1874. https://doi.org/10.1177/0042098010380960.

Wu, C., & Sharma, R. 2012. "Housing submarket classification: The role of spatial contiguity."

Applied Geography, 32(2), 746-756. https://doi.org/10.1016/j.apgeog.2011.08.011.

Wu, C., Ye, X., Ren, F., & Du, Q. 2018. "Modified data-driven framework for housing market segmentation." *Journal of Urban Planning and Development*, 144(4), 04018036. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000473.