Applied Spatial Analysis and Policy (2020) 13:223–256
https://doi.org/10.1007/s12061-019-09301-x

Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets

Yangyi Wu[1] · Yehua Dennis Wei[1] · Han Li[2]

Received: 1 February 2018 / Accepted: 30 April 2019/
Published online: 21 May 2019

Abstract

As an obstacle to the hedonic model's reliability, housing submarkets have drawn plenty of scholarly attention because they lack an integrated and standardized classification framework and validation methods. By incorporating multiple spatial statistics and data mining techniques into a hybrid spatial data mining method, this study develops an innovative classification methodology that replaces spatial continuity with spatial connectivity. Employing Salt Lake County as the case, we identify 43 housing submarkets based on differentiation among structural differences, the complexity of urban space, and neighborhood characteristics. With the introduction of urban amenities into the validation framework, the comparison between the submarket-based model and non-submarket regression shows our classification not only enhances prediction accuracy but also achieves better theoretical comprehension of local housing markets. Besides contributing to an understanding of urban spatial heterogeneity, our study also provides a feasible spatial modeling method which is capable of processing a large dataset with more than 200,000 observations.

Keywords Housing submarkets · Spatial connectivity · Hedonic model · Big data

✉ Yehua Dennis Wei
wei@geog.utah.edu

Yangyi Wu
yangyi.wu@geog.utah.edu

Han Li
han.li@miami.edu

[1] Department of Geography, University of Utah, Salt Lake City, UT 84112-9155, USA

[2] Department of Geography and Regional Studies, University of Miami, Coral Gables, FL 33124, USA

Introduction

The hedonic model has been widely employed as a standard approach in understanding the determinants of housing value, assessing housing values, and calculating property taxes. It is an integrated model that assumes a uniform market when evaluating housing structures, neighborhood conditions, and other determinants of housing values. Since the late 1990s, concepts of submarkets, as well as their spatially heterogeneous effects, have been brought into the discussion and thus cast doubt on the accuracy of the housing value predictions of the traditional hedonic model. These effects may lead to the spatial autocorrelation in both housing values and explanatory variables, which affects their spatial independence, resulting in biased estimations (Adair et al. 1996; Goodman and Thibodeau 2003).

To control the effects of spatial heterogeneity, the application of spatial econometric methods in the hedonic model has been broadly accepted as a solution because they usually account for spatial autocorrelation among the observations (Yu et al. 2007). However, in the United States, many counties in metropolitan areas consist of more than 200,000 housing parcels (Jia and Gaughan 2016; Lee and Moudon 2006), which are far beyond the operational capability of traditional spatial econometrics (Anselin 2013; Fotheringham and Rogerson 2008) and makes it nearly impossible for researchers and tax accessor offices to process them at an affordable cost. Current studies still rely on sampling to reduce the computing cost, which is not as informative as using all of the observations (Li et al. 2016). Thus, to improve the accuracy and efficiency of housing value estimation, a feasible approach coping with large datasets to classify submarkets is urgently needed.

Other studies try to focus on the division of housing submarkets, and these studies could be grouped into two categories. One is based on geographical and administrative boundaries like census tracts, postal zones (Goodman and Thibodeau 2003), and municipalities (Brasington 1999; Tian et al. 2017), or the concentric rings to represent urban spatial structure (Li et al. 2018). The other category is based on the data-driven methodology that has thrived since 2005, thanks to its high objectivity and ability to account for various effects (Wu and Sharma 2012). K-means clustering (Bourassa et al. 2003), Ward's clustering (Bates 2006), and the CART decision tree (Clapp and Wang 2006) are the common algorithms for housing submarket classification. Nonetheless, these approaches have also been questioned because of algorithm shortcomings.

This paper seeks to develop an efficient and accurate method for data-driven housing-submarket classification, a method that is capable of processing large datasets without cloud computing or cluster computing support. We develop a hybrid spatial clustering method to integrate the strengths of various methods such as Getis and Ord's spatial statistics, k-means clustering, hierarchical clustering, density-based spatial clustering of applications with noise (DBSCAN), and decision tree, to retain substitutability, similarity, and spatial connectivity simultaneously. To examine the soundness of the classification, three goals—improvement of prediction accuracy, consideration of local effects, and better theoretical rationality, are established. The comparison between non-submarket and submarket-based regressions shows that our method not only improves general model prediction accuracy, which is reflected by adjusted r-squared, but also local prediction ability, which is reflected by the spatial autocorrelation in residuals. More importantly, it reflects the spatial variation in the effects of the housing price determinants with the hedonic model prediction, and proves its usefulness in understanding housing prices.

Literature Review

Housing Submarkets: Definition, Principles, and Identification

The housing submarket is critical in the analysis of urban housing markets, because the structure of an urban housing market is too complex for an equilibrium model (Whitehead 1999). As a housing market varies in housing characteristics and prices by location, a combination of supply and demand-related factors results in the segmentation of housing markets (Goodman and Thibodeau 2003). With the widespread use of the hedonic model, such a problem has become more important not only in the academy, but also in the assessment of housing values for property taxes. Housing submarkets form from the interplay of four characteristics: heterogeneity, durability, location fixity, and cost of supply (Rothenberg 1991; Bourassa et al. 1999; Hwang and Thill 2009; Watkins 2001). In the current literature, three main principles of housing submarkets are widely accepted: substitutability, similarity, and spatial continuity (Wu and Sharma 2012). Substitutability is the core principle in defining and validating housing submarkets. High substitutability means that attributes of houses in the same housing submarket should have similar contribution in estimating the overall housing prices with the hedonic model. Thus, the best way to evaluate substitutability is to compare the model performances at the local submarkets (Watkins 2001). Similarity refers to the similarity of housing structural attributes, such as room counts, house size, and housing condition, as well as local neighborhood socioeconomic conditions (Bourassa et al. 2003; Watkins 2001; Bates 2006). Spatial continuity means that a housing submarket must occupy a continuous spatial space and submarkets should also have clear geographic boundaries, which can be actual barriers such as administrative boundaries and highways, or invisible segments (Wu and Sharma 2012).

Based on the above three principles of housing submarkets, scholars have developed a series of submarket simulation approaches in practice. Spatial regression models, such as the spatial lag model, spatial error model, and spatial filtering regression, are the most frequently-applied approaches to detect and deal with the spatial autocorrelation of underlying determinants. They are superb at improving the prediction accuracy of the hedonic model, and preventing geographical bias (Griffith 2002; Li et al. 2016; Li et al. 2018). However, most spatial regression models only give a global regression result with respect to spatial autocorrelation, and thus the detailed local effects are often ignored. Moreover, they are difficult to use with large datasets. Housing submarket classification is an alternative way to analyze the actual structure of a housing market.

At the initial stage, boundaries of prior definitions such as zip code and census tract in the United States were commonly accepted as housing submarkets. However, later on studies found that a prior definition fails to address the principles of substitutability and similarity (Wu and Sharma 2012). Consequently, data-driven methods emerged that account for substitutability and similarity. Typically, such studies employ several kinds of distance variables, for example, distance to the CBD or distance to the city's outer limit, to capture the spatial organization of housing submarkets (Bates 2006;

Bailey 1999; Bourassa et al. 2003; Clapp and Wang 2006). These studies successfully capture the principle of similarity, but neglect the role of spatial continuity, because such distance variables are more likely to represent locational characteristics rather than spatial patterns.

To include more aspects of spatial continuity while maintaining substitutability and similarity, scholars have focused on the methods of spatial clustering to classify housing markets (Hwang and Thill 2009). Spatial clustering combines data similarity and spatial pattern, so it reflects similarity and spatial continuity at the same time. The algorithms of spatial clustering can be divided into two broad groups.

The first group integrates spatial constraints into the traditional data clustering methods such as hierarchical and non-hierarchical clustering in three ways. First, spatial data, such as geographical coordinates, are treated as variables in the clustering process, and a proper weight for the spatial variables is the key to maintaining an organized spatial pattern of the classification result (Wise et al. 1997). The second method is to define clusters of structural and neighborhood variables alone without any spatial constraints, then merge the "isolated" observations and small clusters into the most similar and spatially adjacent clusters (Fovell and Fovell 1993). The third way is to locate the potential cluster for each observation by evaluating the similarity of its adjacent neighbors, which is usually called "regionalization" (Guo 2008; Mennis and Guo 2009).

The other broad group includes methods like spatial autocorrelation statistics and spatial scan, which use spatial patterns to classify clusters (She et al. 2015). They cluster observations due to the attributes of the observations themselves, their neighborhoods, and the global condition. For example, spatial scan relies on a "geographical window," and clusters the observations by judging the levels of spatial density and data similarity in the geographical windows (Kulldorff 1997).

However, these methods cannot comprehensively fulfill the three principles of the housing submarket simultaneously. First, some of the methods are imprecise because of their algorithms. For example, k-means clustering is not good at non-vertex dataset, and spatial autocorrelation statistics are unable to precisely judge the boundaries of clusters as they rely on statistical significance to divide the clusters but usually produce insignificant results around the boundary area (Grubesic et al. 2014). Second, some of them are defective when considering spatial continuity. For example, if geographical coordinates are considered as variables, a proper weight is important (Wu and Sharma 2012). Moreover, house densities vary across urban space, which means the impossibility of finding an appropriate and universalized spatial weight to maintain spatial continuity. For instance, since city centers usually have denser clusters of dwellings than the suburbs, the coordinate differences between nearby houses in city centers are smaller than suburbs. Assuming the variations of other characteristics are similar between these two areas, city centers are less likely to be divided into smaller clusters because of the smaller differences in coordinates.

Third, some methods lack objectivity. Many algorithms need pre-given parameters which affect the clustering result. However, as the setting standard for these parameters varies due to local conditions, they are hard to precisely define, such as the desired cluster number of k-means clustering.

Last, some methods are not practical in the context of big data and social data revolution. Space and time complexity of the algorithms, which requires the processing memory capacity and processing time, make many well-developed methods unsuitable for processing large datasets at a low cost (Guo 2008; Helbich et al. 2013; Li et al. 2017).

Recent studies have demonstrated that substitutability, similarity and spatial continuity might be insufficient to classify urban housing markets (Torrens 2008). First, as a core principle of the housing submarket, substitutability does not have a stable and accurate indicator in many studies. Second, similarity is problematic because the levels of similarity within a single submarket and degrees of dissimilarity between submarkets are arbitrarily defined. Third, spatial continuity, as a concept of spatial constraint, theoretically underestimates the complexity of urban spaces, especially in sprawling areas, where leapfrog development is common.

Therefore, in this study, to better reflect the complexity of urban space, we propose the notion of spatial connectivity to replace spatial continuity. First, concentrating on spatial continuity would result in small and numerous clusters in those "complex areas" which impact the model's accuracy (Batty and Xie 1996; Ewing et al. 2014; Hamidi and Ewing 2014; Xie et al. 2007; Wei 2016; Wei and Ewing 2018; Wilhelmsson 2004). Second, spatial continuity is hard to define when houses are treated individually since they are recorded as discrete observations in the dataset. Third, aggregating houses into continuous units, an alternative way to achieve spatial continuity, omits a wealth of specific characteristics (Helbich et al. 2013). Thus, we regard small clusters with similar characteristics in the same neighborhood as one submarket and name the spatial relationship spatial connectivity.

Validation

The assessment of the classification result should comply with the principles of substitutability, similarity, and spatial connectivity. Moreover, a proper validation framework should improve prediction accuracy and theoretical explanation, as well as consider local effects. While spatial connectivity can be examined by geospatial information system (GIS) visualization, and data clustering naturally indicates within-group similarity, the measurement of substitutability is usually the most concerned part in the validation process for housing submarket classification in current studies (Helbich et al. 2013; Wu and Sharma 2012). Various model accuracy indicators, such as weighted mean squared error (WMSE), root mean squared error (RMSE), and adjusted R-squared are commonly used to compare the goodness-of-fit of different submarket classifications (Goodman and Thibodeau 2003; Helbich et al. 2013; Manganelli et al. 2014). However, these measurements can only assess general model performance and not the complexity of substitutability.

The optimal housing submarket classification should also be able to capture spatial heterogeneity, which means that with a proper model, spatial autocorrelation in residuals should be insignificant. Furthermore, previous validation processes usually use identical variables in classification and validation, and add housing submarkets as dummy variables, which overestimate the performance of housing submarkets because these variables are already found related within groups and differentiated between groups. Therefore, some new variables need to be included for validation to test the ability of our models to explain these new variables. The results should also be examined under a well-studied framework in order to validate the classification not only in terms of model prediction but also theoretical explanation.

Overall, the classification of urban housing submarkets deserves further research. While concentrating on spatial continuity, current studies have not considered the complexity of urban space and the application of large datasets. Such lack of consideration may lead to submarket fragments. A suitable and feasible solution for housing submarket classification in the context of big data, which can retain the three principles of substitutability, similarity, and spatial connectivity at the same time, is also needed. Moreover, while similarity and spatial connectivity are ensured by their algorithms, a proper and objective validation is needed to test substitutability by evaluating model improvement in terms of prediction accuracy, ability to consider local effects, and theoretical rationality.

Methodology

Data and Study Area

Our study area is Salt Lake County (Fig. 1).It constitutes most of the Salt Lake City metropolitan area, which is the largest metropolitan area in Utah and one of the fastest growing areas in the U.S (Li et al. 2016). The county had 1.03 million population in 2010, which represented an increase of 14.6% from 2000 (U.S. Census Bureau 2001, 2011). We selected Salt Lake County as the study area for three main reasons. First, the Salt Lake County Assessor's Office provides comprehensive information for over 24,000 single-family houses, which is an appropriate size for testing our methodologies in both classification and validation. Second, there have been plenty of empirical studies focusing on the underlying determinants of housing values in Salt Lake County, which provide enough theoretical support and comparative examples for our validation (Dong et al. 2016; Li et al. 2016; Liao et al. 2015; Lowry and Lowry 2014; Wei et al. 2018; Wei et al. 2016). Third, because of the region's rapid growth, studies have raised the concern of residential segregation in Salt Lake County, which would mean a high level of spatial heterogeneity (Iceland and Sharp 2013; Korinek and Maloney 2010).

The property attributes were derived from the Salt Lake County tax assessor's dataset for the year 2011, which includes basic information about each building such as building structure, land value, and final housing value. These values have been found to be a little bit underestimated because of the delay in updating (Jarosz 2008), but that shortcoming is compensated for by the fact that it is a comprehensive and integrated dataset. Other sources are too dependent on real estate information, which cannot reflect the whole housing market distribution and result in biased estimation. The other advantage of this dataset is that it includes not only the structures and values, but also the building use types. From this dataset, we can also easily extract non-residential sites such as hospitals, schools, and supermarkets.

The socioeconomic data were collected from the American Community Survey (ACS) dataset on census tract level. The Utah Automated Geographic Reference Center (AGRC) offered the GIS data of part of the amenities, such as transportation infrastructures, parks, churches, rivers, and streams. The vegetation information was derived from Landsat TM 5 satellite images from U.S. Geological Survey by calculating the Normalized Difference Vegetation Index (NDVI).

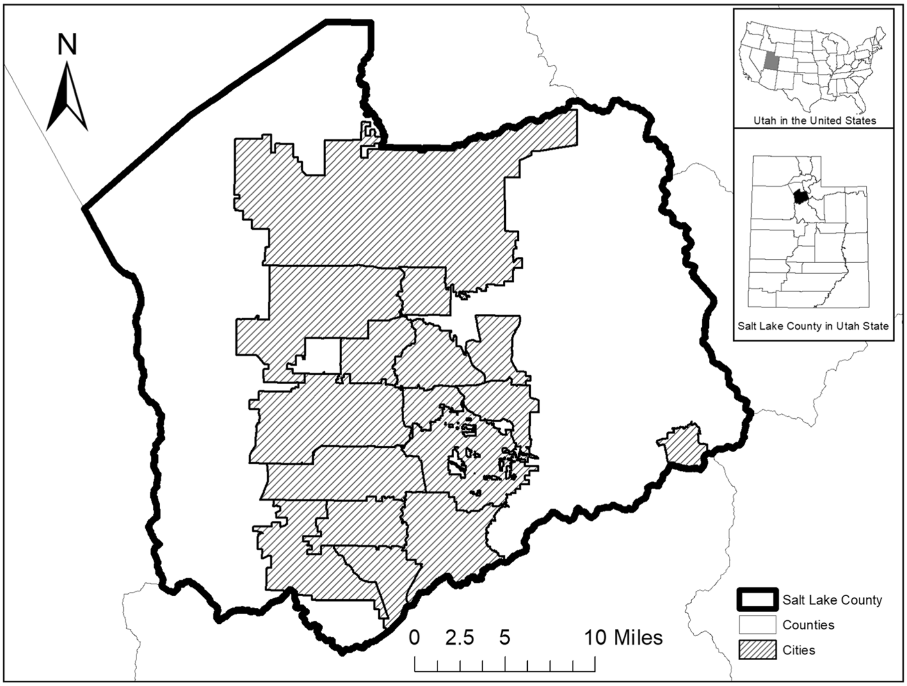

Fig. 1 Salt Lake County

Hybrid Spatial Clustering of Housing Submarkets

There were more than 240,000 single-family houses in Salt Lake County in 2011. This large dataset requires an enormous computing capacity for running spatial clustering and spatial regression models. We found that to merely construct a spatial weight matrix of about 240,000 observations in the spatial regression models would require over 500 GB memory in R, which exceed the computing power of a personal computer. Currently, there is no classification framework that can simultaneously deal with a large dataset, maintain a reasonable spatial organization, and remain objective within pre-given parameters. Therefore, we developed a hybrid spatial clustering method that combines substitutability, similarity, spatial connectivity and computing complexity. To measure substitutability, we included the residual of a simple hedonic model as a dimension in our data clustering, since spatial clusters of residuals reflect the similarity of price mechanisms (Tu et al. 2007). Since there is always a trade-off between similarity and spatial connectivity, we combined several different clustering and processing methods to maintain them simultaneously. For spatial connectivity, our first goal was to add spatial constraints into the observations. As adding spatial variables into tranditional data clustering distorts the results, and regionalization algorithms are too costy in computing time when applied to large datasets, we chose local spatial autocorrelation statistics, which are widely accepted in spatial clustering and more stable with large datasets (Assunção et al. 2006). They are also compatible with other algorithms in multivariate spatial clustering (Peeters et al. 2015). To ensure the best similarity result, we applied a two-step clustering combining k-means and hierahical clustering, which is capitable to handle large dataset processing and is more precise than either of the single algorithm (Day 2003). Finally, the results of the two-setp clustering may contain huge clusters which consist of several remote housing submarkets, so we used density-based spatial clustering of applications with noise (DBSCAN) to break them down, and then use the decision tree algorithm to reclassify the outliers defined the DBSCAN (Fig. 2).

Variables for Clustering

As spatial connectivity could be ensured by the hybrid spatial clustering method using spatial information, the selection of variables for spatial clustering should consider the dimensions of similarity and substitutability at the same time, which includes three parts: housing structure, neighborhood social-economic conditions, and locational indicators. These variables are widely accepted as determinants of housing value and the basic components of the hedonic model, and they are also widely regarded as the characteristics of housing submarket (Li et al. 2016; Li et al. 2018). Similarity was obtained directly by the variable values, and substitutability was evaluated by using the residuals of a linear model to reflect the similarity of housing value mechanism. The

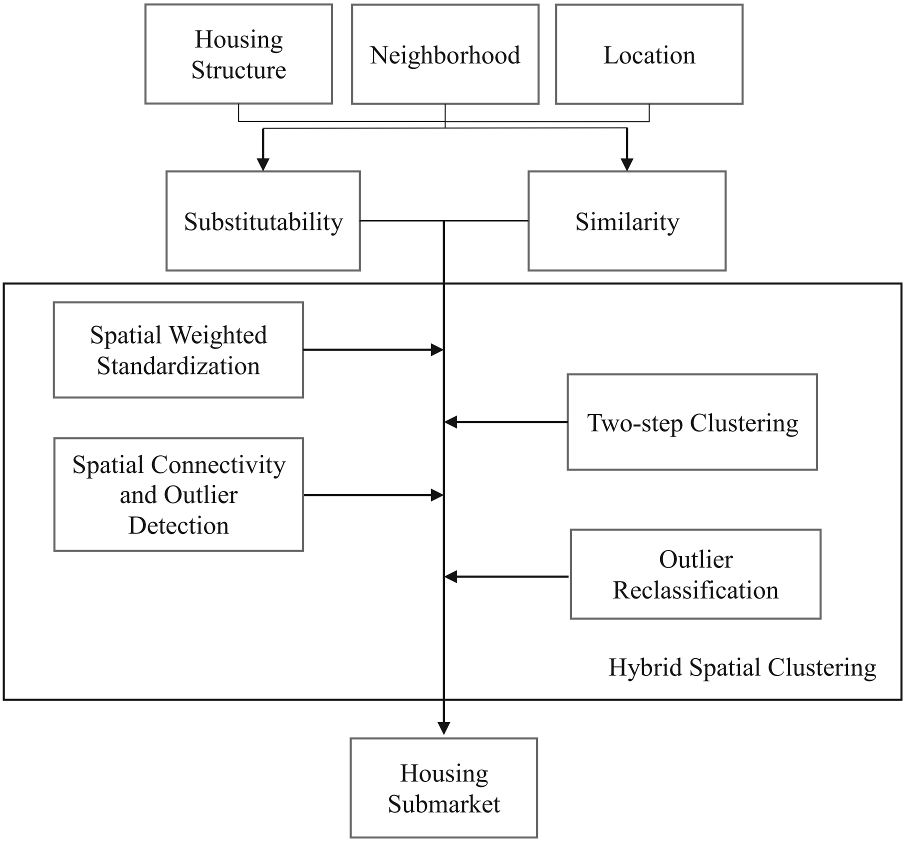

Fig. 2 Theoretical and analytical framework for housing submarkets

definitions and descriptions of the variables which were involved in the clustering process are given in Table 1.

The structural variables represent the basic structures and quality of houses. We included the house occupation status (number of families indicate if the house is for rent/share or single family) (Wu 1996), inner house structure (main floor area, number of rooms, number of full baths), and house general condition (year built, housing value, building value which does not consider land value, and house condition rating).

Neighborhood socioeconomic conditions are also of great importance in defining housing submarkets, since these conditions are highly associated with family welfare, such as tax rate and educational zoning, and neighborhoods are an important component in social network (Adair et al. 1996). We used neighborhood economic conditions, which include tax rate and local income level, and demographic conditions, which include local median age, population density, proportions of the population accounted for by non-Hispanic Whites, African-Americans, and Asian-Americans. Hispanics are not included because of strong multicollinearity.

Table 1 Variables for housing submarket classification

			Housing structure	Variable name
Similarity	Housing Structure Variables	Occupation	Number of Families	NUM_FAMILI
		Inner Structure	Main Floor Area	MAIN_FL_AR
			Total Rooms	TOTAL_ROOM
			Full Bath Numbers	FULL_BATHS
		Condition	Year of Built	YEAR_BUILT
			Housing Value	FINL_VALUE
			Building Value	BLD_VALUE
			House Condition	OverallCon
	Neighborhood Social-Economic Condition	Economic Condition	Local Housing Tax Rate	Tax
			Neighborhood Median Household Income	Median_age
		Demographic Condition	Percentage White Population	WhiteR
			Percentage Black Population	BlackR
			Percentage Asian Population	AsianR
			Neighborhood Median Age	Median_inc
			Neighborhood Population Density	PopDen
	Location Indicator		Distance to Nearest Agricultural Land	AgriDis
			Distance to Nearest Industrial Land	IndustryDi
Substitutability			Residuals of the OLS regression model	Residual

Location is the most signficant factor in housing choices, which is highly related to accessibility, commuting, and environment, contributing significantly to the formation of housing submarket (Cho et al. 2006; Dubin 1992). We used two distances to represents the basic zoning status of each house: its distance from the nearest agricultural land, which is regarded as the urban edge of undeveloped area, and the distance from the nearest industrial land, which is regarded as the traditional employment region.

As substitutability should reflect people's reference to the factors mentioned above when buying houses, this paper used an OLS regression and its residuals to simulate substitutability. Independent variables are the variables mentioned above except housing value, which are usually regarded as basic components of the hedonic model. The dependent variable is the housing value. The residuals indicate the underlying substitutability as the observations with similar residuals are likely to be explained by a same mechanism (Tu et al. 2007).

Spatial Weighted Standardization (Getis and Ord's Gi* Statistics)

Local spatial autocorrelation statistics, such as local Moran's I and Getis and Ord's Gi*, have been proven effective as spatial clustering methods. The z-score of local spatial autocorrelation statistics reflects not only the degree but also the significance of local spatial autocorrelation of the pre-defined neighborhood. Thus, local spatial autocorrelation statistics have the potential to provide spatially weighted standardization for our input data to remove outlier observations of the neighborhood and reflect area differences.

Local Moran's *I* index compares an observation to its neighborhood, which is effective in detecting outliers and clusters (Anselin 1995). In this case, the effect of outliers is strengthened, not weakened, which makes it difficult to classify these outliers. On the other hand, Getis and Ord's Gi* statistic compares the whole neighborhood to all the observations to determine if the neighborhood is significantly different. Thus, in this study, we used Getis and Ord's Gi* statistics to spatially standardize our dataset, whose spatial weight is based on the nearest neighborhoods to avoid the influence of house densities. Getis and Ord's Gi* can be formulated as follows (Getis and Ord, 1992):

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \overline{X} \sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}} \qquad (1)$$

Where x_j is the attribute value for observation j, $w_{i,\,j}$ is the spatial weight between observation i and observation j, n is the observation count, $\overline{X}$ is the mean value of all the observations and S is the standard deviation of all the observations.

Local spatial autocorrelation is superb at finding clusters of selected attributes. However, the results are not enough to define housing submarkets. First, spatial clustering solely relying on spatial autocorrelation highly depends on the statistical signficance. However, the z-scores of spatial autocorrelation statistics are not always signficant, resulting in difficulties in classifying the insignificant observations. Second, it is difficult to integrate the results of multiple variables, as their spatial patterns may be different. To solve these problems, previous studies usually combine it with traditional data clustering methods (Peeters et al. 2015). For example, k-means clustering is frequently combined with local spatial autocorrelation statistics and is also able to process large datasets. However, k-means clustering is sensitive to a pre-given parameter, the number of clusters k (Day 2003; Poudyal et al. 2009). To fix this, Day (2003) proposed a two-step hybrid clustering method integrating k-means clustering and hierarchical clustering by first doing k-means clustering with a large number of pre-clusters and then organizing these pre-clusters into final clusters using hierarchical clustering. In this way, an optimized number is defined relying on the hierarchical tree of the hierarchical clustering. Moreover, it avoids the inability of k-means to process non-convex datasets and the large memory requirments of hierachical clustering. Thus, we applied the two-step hybrid clustering to identify housing submarket, which consists of small cluster detection by k-means clustering, and small cluster aggregation by hierarchical clustering.

Two-Step Clustering: K-Means Clustering and Hierarchical Clustering

Since k-means clustering is efficient with large datasets but relies on pre-given cluster counts, it was applied to organize the whole field into 2500 small clusters (Hartigan and Wong 1979). Each of the output clusters contained a small group of houses with similar attributes standardized by the spatial weighted standardization. With such a large number of clusters, the influence of the cluster number is minimized.

Ward’s hierarchical agglomerative clustering was then applied to the small clusters derived from the k-means using their centroids to merge them based on similarity (Murtagh and Legendre 2014). The number of the final clusters was determined by R package “Nbclust,” which uses over 20 indices to select the best solution by different evaluations of within-group difference (Charrad et al. 2014). Nbclust indicated that 7 and 27 are the optimal cluster numbers. As previous studies have demonstrated most metropolitan counties in the United States have more than 20 housing submarkets (Bourassa et al. 2003; Wu and Sharma 2012; Royuela and Duque 2013), we selected 27 as the optimal cluster number (Fig. 3).

Although the two-step method is efficient in the clustering process, the spatial connectivity has not been examined. Similar housing submarkets in different urban sectors are regarded as the same submarket if they have similar local spatial autocorrelation z-scores as shown in Fig. 3. Therefore, we needed to break the large indentified clusters into smaller housing submarkets in accordance with their spatial patterns. The algorithm should be able to ascertain if a part of a cluster is close to other parts of the cluster. To achieve this, we applied a density-based clustering method with the coordinate information of each house to detect the spatial connectivity within each cluster indentified in the previous steps.

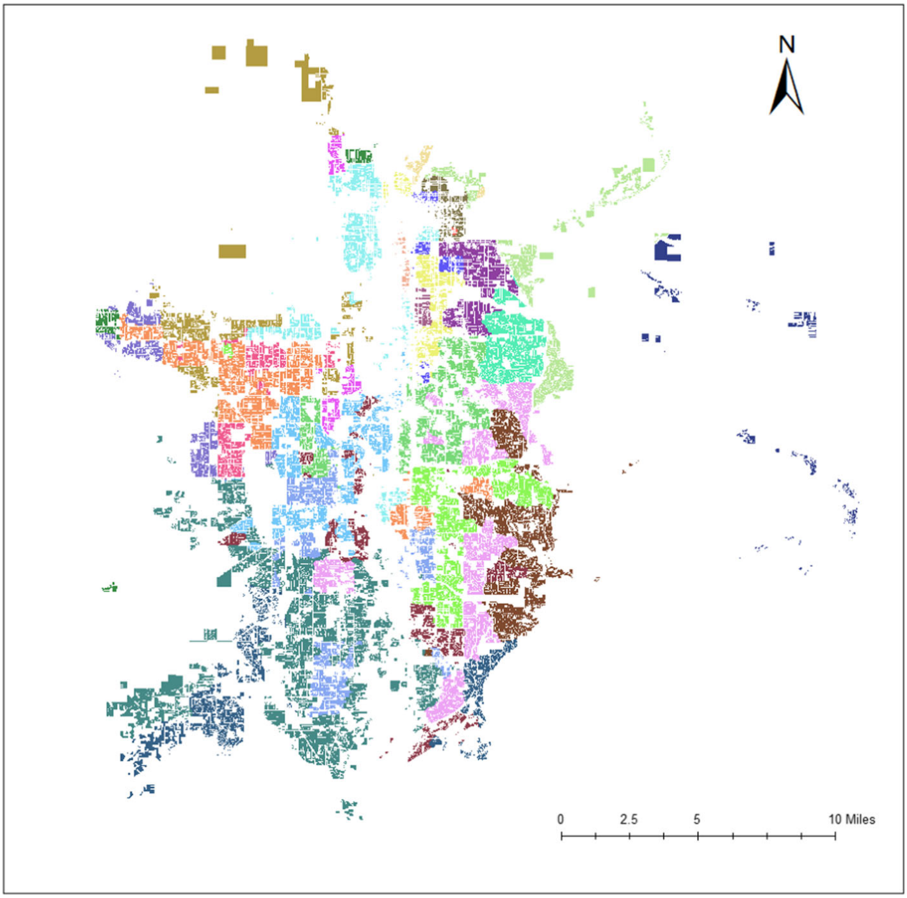

Fig. 3 Housing submarket classification results of hierarchical clustering

Spatial Connectivity Test and Outlier Detection: DBSCAN

As we can see in Fig. 3, there were several clusters having several separated housing sectors. To distinguish these spatially separated clusters, we needed to assess the spatial connectivity within a cluster and separate it if there is not enough spatial connectivity between sectors. An algorithm which captures the distribution of the houses and breaks each cluster due to a standard rule of spatial connectivity (distance, density, neighbors of Thiessen polygons, etc.), was thus needed. At the same time, urban sprawl should also be accounted in the classification process, which requires the tolerance of leapfrog housing submarket. Therefore, the combination of distance and density would be a good rule for spatial connectivity for Salt Lake County: a density threshold which defines that the leapfrog housing cluster as a part of the submarket or just outlier, and a distance which take the qualified leapfrog submarket into account. Considering these expectations, we applied a density-based clustering algorithm, which is called Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al. 1996). Because spatial information should be exclusively considered in this step, only the locational information (the x coordinates and y coordinates of the centroid) of each house were treated as clustering variables. With the DBSCAN algorithm, a spatially connected cluster forms when there is enough density of observations within a pre-given radius, and other observations are tagged as outliers. In this way observations in a spatially connected cluster may not fully maintain spatial continuity, but they obtain a high spatial connectivity for which a minimum density is required over a given distance. These arguments of connectivity are flexible due to the reality requirements. We set the search radius as 1.75 miles, which is equivalent to the average block group size in Salt Lake County. The density requirement was set as 30. If the threshold was too low, many isolated outliers would be wrongly classified, and if it was too high, the houses in the sparsely-settled east edge in Salt Lake County would be clustered fragmentally. After all the spatial clusters are properly processed, the outliers were picked out and further reclassified because they were imprecisely clustered.

Two kinds of observations were considered as outliers. The first one consisted of the outlier observations identified by the DBSCAN model, who had no similar nearby neighborhood. The second one consisted of the small clusters that lacked enough observation counts to be regarded as a complete and functional housing submarket. The threshold value of defining an outlier cluster depends on the optimal definition of housing submarket. We selected 400 as the threshold for the minimum acceptable submarket size because the observation count of each cluster started to increase steadily when it reaches 424. The output of DBSCAN is represented in Table 2. To maintain the general size of submarkets, we further classified these observations into the most similar nearby submartkets. As most houses are tagged with proper submarkets, a classification method is suitable at this stage, which relies on a tagged dataset to determine the categories of untagged observations.

Outlier Reclassification: Decision Tree

In the outlier detection process, unconnected observations and clusters that are too small are tagged as outliers. Even if they were imprecisely or randomly clustered in previous processes, they were supposed to belong to a nearby housing market and should be reclassified by similarity. To ensure spatial connectivity, we used a decision tree to reclassify these outliers with their coordinates considered. Conditional inference trees algorithm was applied, which estimates a regression relationship by binary recursive partitioning in a conditional inference framework. Compared with traditional decision tree algorithms, the conditional inference trees algorithm is more likely to be unbiased (Hothorn et al. 2006a; Hothorn et al. 2006b; Strasser and Weber 1999; Strobl et al. 2009).

We used 70% of the non-outlier dataset as the training data to build the decision tree, and the remaining 30% as the test data to test the precision of the decision tree. 70% and 30% are widely-applied ratios in model assessment, as most data are involved in model training and rest small part plays as the assessment dataset. After training the model with the test dataset, the correctly classified observation count was 72,778 out of 74,283, indicating a high precision of 97.97%. Then the outliers were reclassified by the model.

In sum, five specific steps and methods were employed to identify the housing submarkets: selection of clustering variables, spatial weighted standardization (Getis

Table 2 Clustering results of DBSCAN

Hierarchical clustering results	DBSCAN results	Counts	Hierarchical clustering results	DBSCAN results	Counts	Hierarchical clustering results	DBSCAN results	Counts
1	1	1540	10	1	13,415	18	1	17,073
2	Outlier	6*		2	1033	19	1	1495
	1	297*	11	1	9598		2	2302
	2	6578	12	1	4403	20	1	4209
	3	2242		2	2418	21	1	12,454
	4	64*		3	3243	22	Outlier	5*
3	1	228*		4	181*		1	6416
	2	16,366	13	1	2958	23	Outlier	12*
4	Outlier	2*		2	255*		1	10,624
	2	24,951	14	1	8567		2	40*
5	1	3573	15	1	11,051	24	1	790
	2	3873		2	3168	25	1	1197
6	1	19,661		3	53*		2	1543
	2	3172	16	1	10,873		3	275*
7	1	1030		2	2430	26	Outlier	6*
	2	6944	17	Outlier	17*		1	93*
8	1	424		1	3142		2	48*
	2	71*		2	90*		3	57*
9	1	5646		3	606		4	113*
	2	7944		4	1517		5	607
	3	1546		5	3126		6	85*
				6	2109	27	1	116*

*: Outlier clusters or observations

and Ord's Gi* statistics), two-step clustering consisting of (k-means clustering and hierarchical clustering), spatial connectivity test and outlier detection (DBSCAN), and outlier reclassification (decision tree).

Validation

In this study, using a comparison between a non-submarket model and a submarket-based model, we assessed the housing submarket classification regarding the following aspects: the accuracy of prediction and the abilities of controlling local effects and providing theoretical explanation.

Model Design

In the validation process, we built and compared two hedonic models, a non-submarket regression model and a submarket-based regression model in terms of the following aspects: housing structure, community, and amenities. The amenities category was divided into four subgroups: service and consumer goods, aesthetics and physical settings, public services and accessibility (Fig. 4). The non-submarket regression model represents the traditional application of the hedonic model, and the equation of the non-submarket regression is as follows:

$$HV_i = \beta_0 + \beta_i X_i + \varepsilon_i \tag{2}$$

Where β_0 is the intercept, β_i is the coefficient of X_i, X_i is the vector of attributes of property i, ε_i is the vector of residuals.

As to the submarket-based model, it should be able to capture local effects for further validation. Thus, we used a linear regression hedonic model to which the classification results of the housing submarket are added as dummy variables in the submarket-based regression model, and a matrix instead of a single linear equation is built to make both slope coefficients and intercept coefficients unique and comparable for each housing submarket (Gujarati 1970).

The equation of the submarket-based regression model is as follows.

$$HV_i = \beta_j + G_{ij}\beta_i X_i + \varepsilon_i \tag{3}$$

Where β_j is the intercept for submarket j, G_{ij} is the vectors of dummy variables of the relationship between submarkets j and property i. β_i is the coefficient of X_i, X_i is the vector of attributes of property i, ε_i is the vector of residuals.

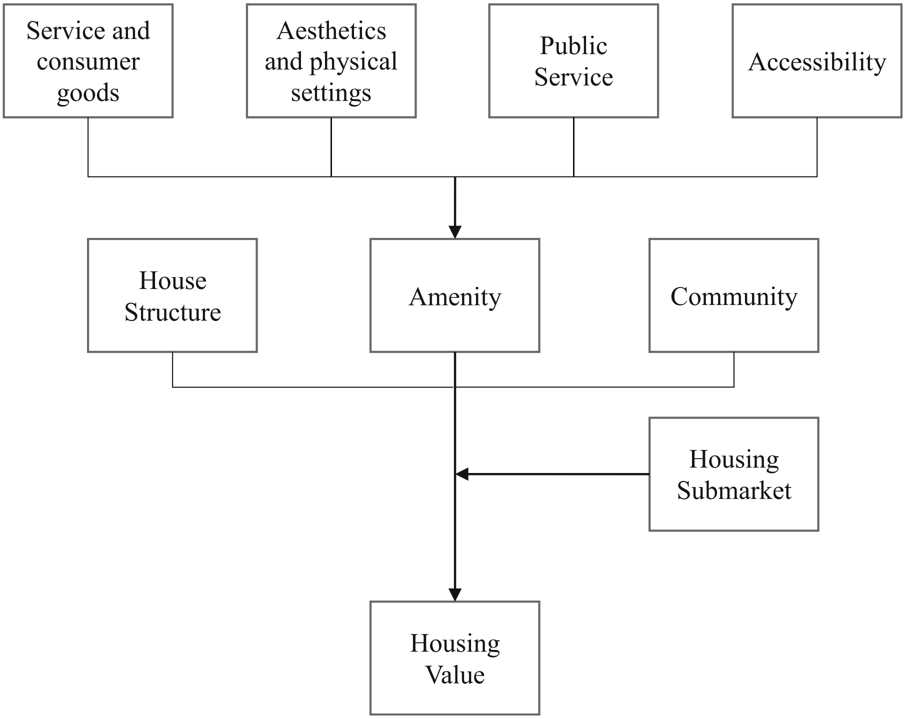

Fig. 4 Theoretical framework for housing value

This model only reports the significance of each variable for a single submarket, as the significances for the dummy variables describe the differences but not the coefficients of other submarkets directly. Thus, an iteration for each submarket were done and we extracted the coefficients and significances for each of the submarkets as the result.

Variables

The independent variables for validation are shown in Table 3, and the dependent variable is the estimated housing value.

Among these variables, housing structure, neighborhood social-economic conditions, and locational variables are also involved in the classification process. Under the framework of the hedonic model, neighborhood social-economic conditions and locational variables were merged as community characteristics to distinguish them from the physical environment variables which are combined with amenity variables.

We selected urban amenities as the "extra component" of the validation process to test the theoretical explanation ability of the submarkets since urban amenities are well-studied and widely used housing value determinants (Li et al. 2016; Hui et al. 2018; Jim and Chen 2007; Tian et al. 2017; Waltert and Schläpfer 2010). Housing prices respond to urban amenities variously in different regions, as the consequences of various preferences and the presence of associated negative factors such as noises or safety issues (Boustan 2013; Chi and Marcouiller 2013; Geoghegan 2002; Li et al. 2016; Osland and Thorsen 2013). For example, both the distance to CBD and forest coverage have been found to have spatially varying effects. (Gao and Asami 2007; Kong et al. 2007; Li et al. 2016; Sander and Polasky 2009; Wu et al. 2017). Geographically sensitive effects further result in strong spatial autocorrelations in residuals when using a simple linear model to estimate housing values, in which case considering housing submarkets should mitigate spatial autocorrelation. Thus, urban amenities are suitable to test the rationality of the classification of housing submarkets not only by enhancing validation robustness by introducing new determinants, but also by providing theoretical cross-validation from existing studies.

Following Glaeser (2000), we divided amenities into four general categories: consumer goods and services, public services, physical environment, and accessibility. The group of consumer goods and services includes theaters, sports and exercise services, laundries, banks, restaurants and auto services. The attractiveness of these amenities is widely recognized, but they also have offsetting negative effects on their neighborhoods (Gao and Asami 2007; Sander and Polasky 2009). Public services include hospitals and clinics, libraries, schools, and religious facilities, which have long been regarded as important factors in housing studies (Diao and Ferreira Jr 2010; Huh and Kwak 1997; Park and Lah 2006). The physical environment consists of rivers, streams, parks, lakes, and surrounding greenery, whose effects are highly related to their locations (Kong et al. 2007; Li et al. 2016; Wu et al. 2017). Transportation facilities are represented by commuter railway stations, bus stops, main roads, and light rail stations. They are similar to consumer services in that they have offsetting negative effects (Diao and Ferreira Jr 2010; Duncan 2011).

Most of the variables were calculated by the distance to the nearest facility, and the distances and values were applied by a logarithm transformation to ensure approximate

Table 3 Variables for validation

General category	Detailed category	Description	Variable name
Housing Structure Variables		Number of Families	NUM_FAMILI
		Main Floor Area	MAIN_FL_AR
		Total Rooms	TOTAL_ROOM
		Full Bath Numbers	FULL_BATHS
		Year of Built	YEAR_BUILT
		Building Value	Bldv
		House Condition	OverallCon
Community and locational condition Location Indicator	Neighborhood Social-Economic Condition	Local Housing Tax Rate	Tax
		Median Age	Median_age
		Percentage White Population	WhiteR
		Percentage Black Population	BlackR
		Percentage Asian Population	AsianR
		Median Household Income	Median_inc
		Population Density	PopDen
	Location Indicator	Distance to Nearest Agricutural Land	AgriDis
		Distance to Nearest Industrial Land	IndustryDi
Amenity	Service and Consumer goods	Theaters	DisThea
		Sport & Exercise Services	DisExec
		Laundries	DisLaun
		Banks	DisBank
		Retail Services	DisReta
		Restaurants	DisRest
		Auto Services	DisAuto
	Public Service	Hospitals & Clinics	DisHeal
		Libraries	Libr
		Schools	DisEduc
		Religion Places	Worship
	Aesthetics and physical settings	River	River
		Streams	Streams
		Parks	Parks
		Lakes	Lakes
		Neighborhood Vegetation Index	NDVI
	Mobility, Accessibility, and Commuting	Commuter Rail Stations	CommuterRa
		Bus Stops	Bus
		Main Roads	DisRoad
		Light Rail Stations	LightRail

normal distributions. The vegetation information was derived from Landsat TM 5 satellite image from U.S. Geological Survey. An NDVI index was calculated, and negative values were set to zero indicating that no vegetation exists. The neighborhood vegetation index of a single house was the average NDVI index in 0.25-mile range.

Classification Results

The final clustering result can be seen in Fig. 5, including 43 submarkets. The average size of our submarket is 5813 houses, and to our surprise, it nearly matches the average size of housing submarkets in the study of Wu and Sharma (2012), which is 5748. Figure 6 better portraits the general location of specific submarkets.

We can find several general patterns of our results. First, we observe a blank zone in the horizontal middle of the county, which is the central business district (CBD) separating the whole county into west and east parts. Interestingly, the submarkets are also divided by the main business area, as all the submarkets have their majorities on either side and none of them are evenly divided by the blank zone. The widest part

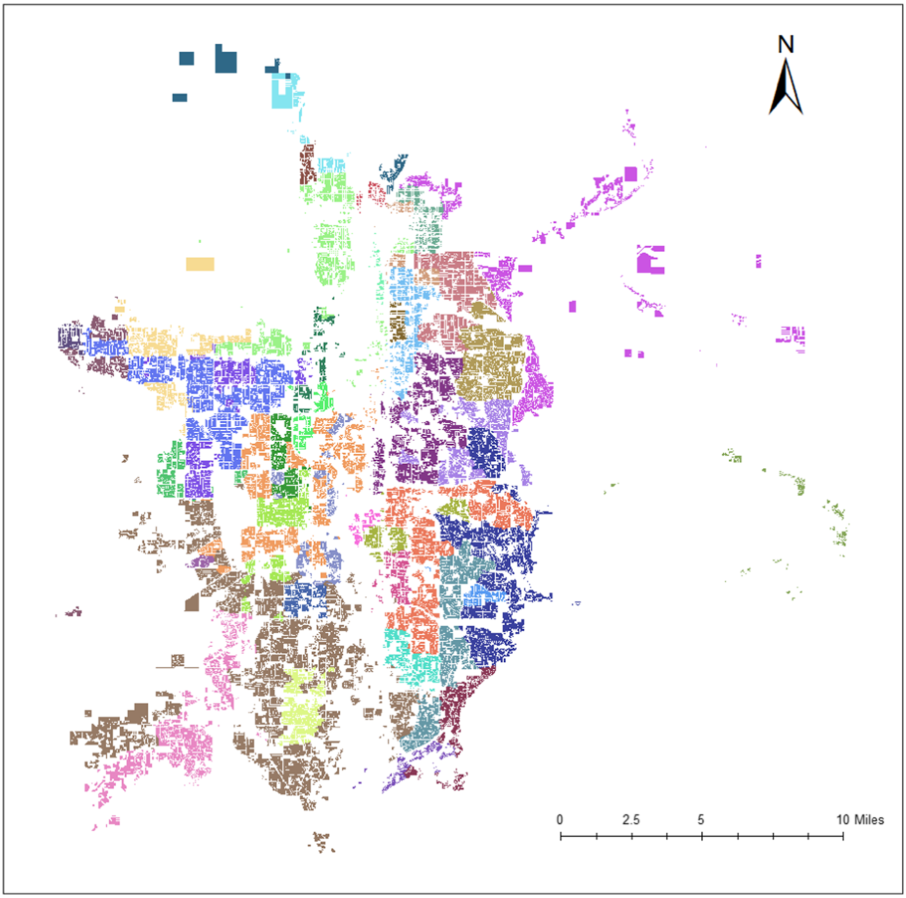

Fig. 5 Final housing submarket classification results

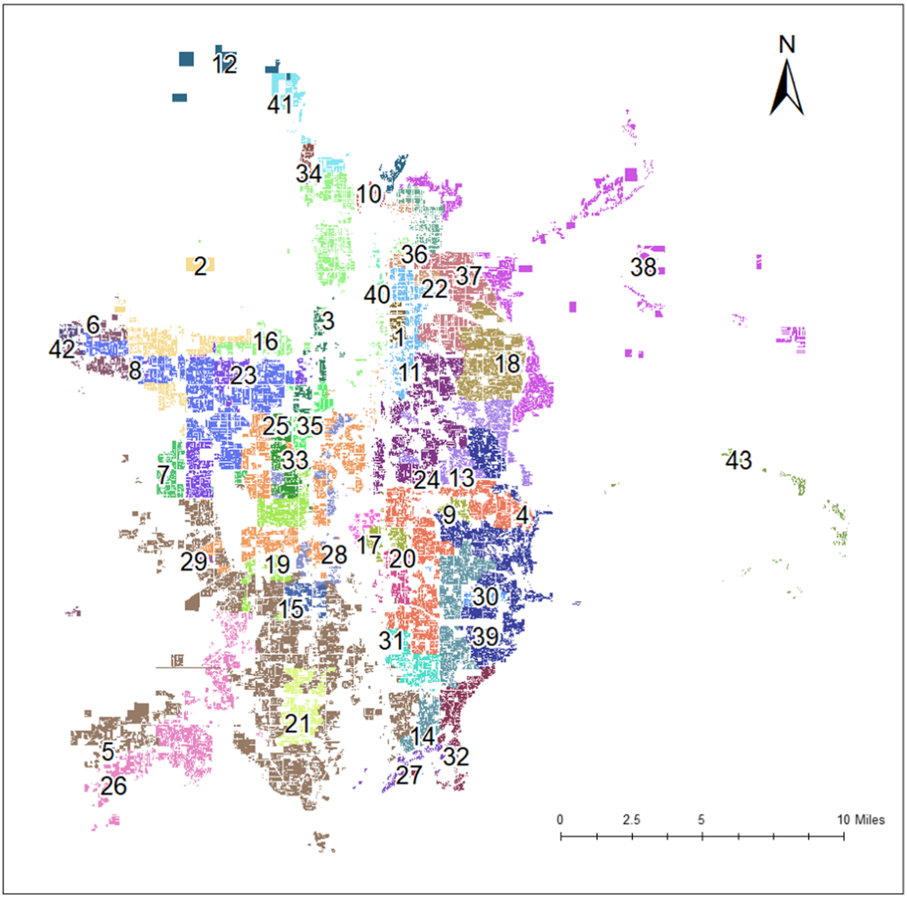

Fig. 6 Final housing submarket classification results with submarket index

of the main business area is lower than the break threshold of spatial connectivity in the DBSCAN process(1.75 miles), thus the east-west division is not caused by the classification algorithm. The spatial pattern of the submarkets is concordant with the current residential pattern in Salt Lake County, which could generally be divided into three parts. As the CBD divides the county into the east and the west part, there is also a north-south pattern in the west part. The east part consists of mid-aged, middle-class, and white majority communities. The northwest part has more low-income and Hispanic families, while the southwest part, as a newly developed area, is a common choice for recent arrivals who are young and middle-class.

Second, small submarkets are likely to be close to the CBD and the periphery area. This phenomenon suggests that even though we replaced spatial continuity with spatial connectivity to avoid submarket fragments, city centers and city edges still appear of higher complexity as there are more small submarkets in these areas. Figure 7 shows the submarkets around the city center of Salt Lake City, consisting of 11 submarkets, many of which are small.

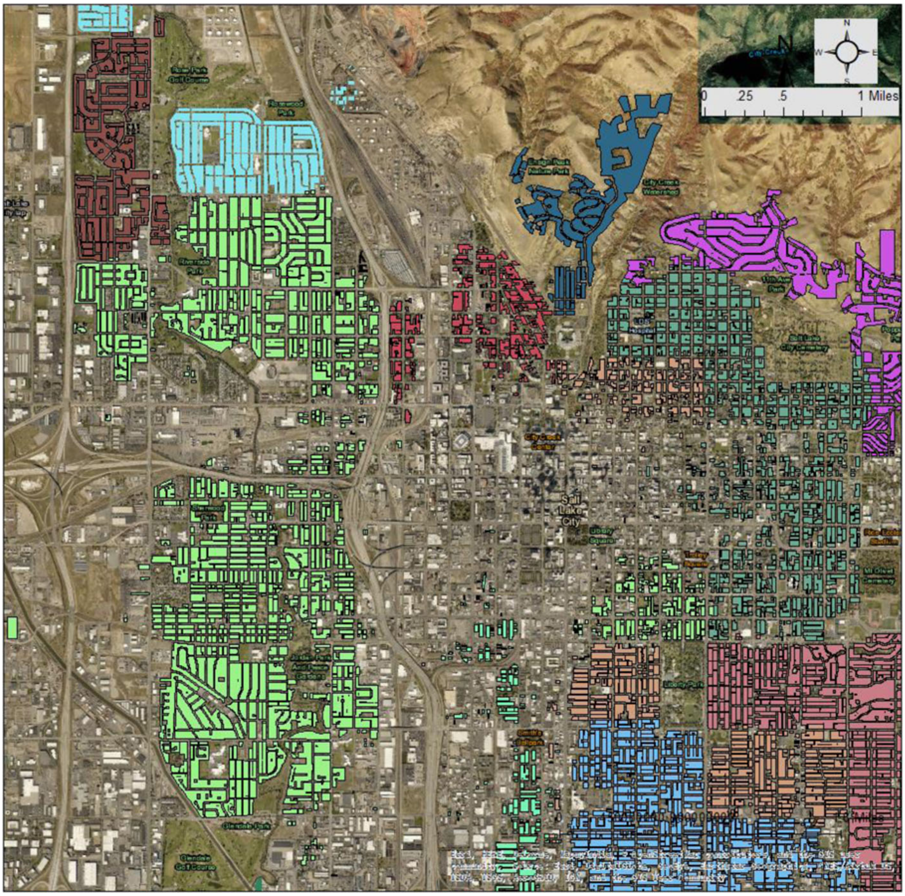

Fig. 7 Submarkets around Salt Lake City center

Third, main roads frequently appear as submarket boundaries. For example, for housing submarkets 18, 37, and 38, the largest submarkets in the northeast area (Fig. 8), Foothill Drive serves as an important boundary. State Route 266 (4500 S) is the south edge of submarket 18. In addition, a small part of 700 E is the west edge of submarket 37. Other local roads, such as 1300 E, 1300 S, Parleys Way, and 2100 S are also important boundaries for submarkets 18 and 37.

Besides the characteristics of boundaries and submarket sizes, according to Table 4 which provides the spatial statistical variations between submarkets using the z-scores of Getis and Ord's spatial statistics, we can also find that adjacent submarkets can be easily distinguished by their attributes of housing structures, neighborhood socioeconomic conditions, and locational indicators. For example, submarkets 18, 37, and 38 have considerable differences in the distance to agricultural land. Submarket 37 is nearer to the downtown area and has the smallest houses when compared to other two submarkets (Fig. 8). Almost all of the housing structural indicators are different. For example, submarket 18 is to the south of submarket 37, and larger houses can be found in this area. Submarket 37 is to the east of the other two submarkets, along or in the Wasatch Mountains, and has larger houses with mountain views. According to the

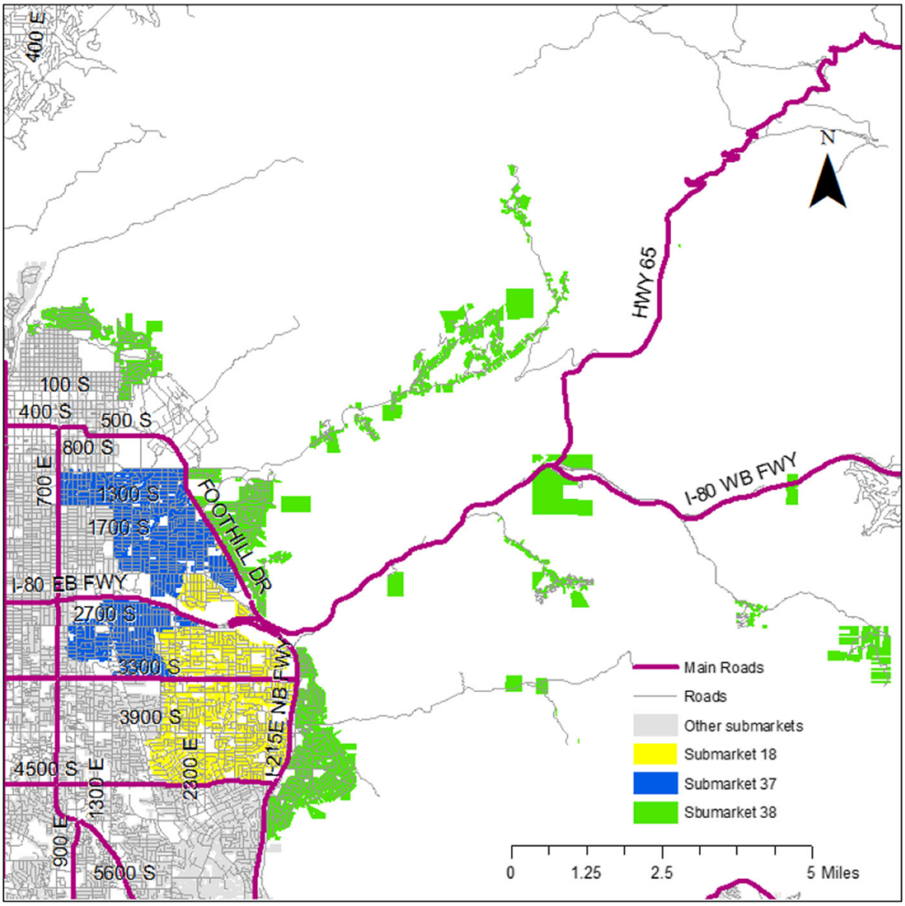

Fig. 8 Submarket 18, 37 and 38

dataset, the average house size is 1180 square feet in submarket 37, 1475 square feet in submarket 18, and 1898 square feet in submarket 38. These differences are significant according to the Mann-Whitney U test. Their communities also differ from others in terms of Asian rate, income, age, and population density.

On the other hand, in some area, neighborhood submarkets are similar in some dimensions, yet in other dimensions, one is significantly "better" than the other one. For example, submarket 5 and submarket 26 in the southeast corner of Salt Lake County are very similar to each other. The houses are newly constructed and in good condition, and community constitutions are dominated by young white people of high incomes. However, even though submarket 26 has smaller main floor areas, the other characteristics of submarket 26 are "better" than submarket 5, since it is away from industries, the houses are newer and of better condition, and the residents tend to be younger and have higher incomes. Moreover, the differences in residuals indicate that submarket 5 and submarket 26 may have different market mechanisms. These differentiations also indicate that the submarkets are able to capture spatial heterogeneities of the attributes.

Table 4 Submarket characteristics

C	AD	ID	NF	AR	TR	FB	Y	BV	FV	T	OC	WR	BR	AR	I	Age	PD	Re
1	24.6	−5.5	2.5	−5.9	−8.5	−4.8	−10.7	−1.1	−2.0	−7.6	−8.2	−13.4	−0.4	36.1	−15.8	−5.2	14.2	1.2
2	−5.8	−7.6	−1.5	−3.7	−3.6	1.2	7.8	−0.6	−1.5	6.0	4.6	−15.5	1.1	28.7	−6.8	−7.3	−9.5	−0.2
3	−5.7	−6.7	−0.7	−4.0	−4.0	−0.9	1.2	−0.9	−1.8	−1.7	−2.2	−14.8	7.4	29.0	−11.3	−1.5	0.1	−1.4
4	−2.6	−0.2	−0.5	−1.4	0.5	−2.5	0.3	−0.4	−0.6	6.8	−1.7	6.6	−5.7	−0.4	0.6	5.4	2.0	−1.2
5	−7.8	−4.2	−1.5	4.9	3.9	7.0	8.9	0.5	0.8	−5.8	10.1	5.6	−4.5	−3.8	4.8	−7.5	−9.3	0.1
6	−4.3	−8.6	−1.1	−5.3	−5.8	−2.2	4.0	−1.1	−2.0	12.7	−0.8	−6.9	−4.1	−9.5	−4.8	−12.3	−4.0	1.1
7	−4.5	−6.9	−1.6	−4.9	−3.6	1.4	8.2	−0.7	−1.4	6.6	4.5	−4.9	−0.7	−1.3	2.3	−10.0	−4.3	0.6
8	−3.2	−3.9	−0.9	−3.6	−4.7	−4.7	−0.9	−1.2	−2.1	6.5	−6.6	−9.1	0.7	−6.4	−7.9	−2.1	9.0	−0.7
9	−4.5	−6.6	0.9	−1.8	−0.2	−2.7	−1.6	−0.3	−0.8	7.1	−3.5	−1.8	8.5	−6.5	−3.5	−2.4	2.5	−0.7
10	20.4	−9.3	9.5	−4.0	−3.9	−0.4	−17.8	−0.5	−1.2	8.7	−5.0	−3.2	11.7	−1.4	−5.2	−2.8	0.4	−0.6
11	26.5	−1.4	3.6	−5.1	−6.4	−4.2	−14.4	−0.6	−1.4	4.2	−6.8	−2.1	7.2	−0.5	−9.9	2.3	12.8	−0.3
12	17.1	−6.0	9.4	12.4	9.9	7.2	−2.4	3.2	6.5	9.0	0.1	1.7	0.4	34.7	22.3	3.2	−17.6	9.5
13	−0.4	2.3	0.2	6.7	6.0	1.5	−0.8	0.8	2.5	−13.8	0.9	8.3	−5.6	−5.3	1.3	18.8	−2.8	1.9
14	−0.1	7.1	−1.4	4.2	6.3	1.5	3.6	−0.1	1.0	0.5	2.3	8.5	−3.9	−6.3	2.1	5.9	−3.2	1.5
15	−8.8	−5.7	−1.4	8.2	4.7	5.5	6.5	0.7	1.2	−7.0	6.6	10.3	−6.8	−6.9	−5.7	16.5	−5.5	−2.3
16	11.7	−7.5	2.0	−5.2	−6.5	−4.4	−8.0	−1.4	−2.6	7.3	−7.8	−22.6	9.9	1.2	−15.4	−6.2	6.1	0.4
17	1.9	−9.8	1.3	−5.4	−6.8	−4.6	−10.8	−0.9	−2.0	4.7	−5.6	−19.1	2.3	−8.8	−12.0	−6.1	−3.3	−0.8
18	6.6	19.3	0.6	3.1	2.8	−2.3	−5.5	0.0	1.1	−5.9	−3.5	8.9	−4.1	−2.4	2.6	7.2	0.5	0.8
19	−6.6	−4.2	−1.5	−1.1	−1.2	−2.7	2.7	−0.6	−1.1	−9.2	−1.5	4.8	−4.1	−9.4	−1.5	−4.4	−1.6	−1.4
20	−7.0	−2.8	−0.8	−2.7	−3.4	−4.3	−3.2	−0.6	−1.4	2.6	−5.4	5.2	−2.9	−4.0	−4.2	−2.2	−4.7	−2.6
21	−6.2	−0.2	−1.5	−0.4	−0.8	−0.6	3.8	−0.7	−0.7	−2.4	0.8	9.5	−6.6	−9.9	−2.1	−4.2	−3.5	0.3
22	18.7	2.9	6.2	−3.8	−5.3	−2.9	−19.9	−0.2	−0.6	8.9	−7.0	−2.8	37.2	−2.0	−5.0	0.5	17.4	0.2
23	−4.7	−3.8	−1.5	−5.3	−4.2	−2.6	3.9	−1.0	−1.8	5.3	−2.1	−10.2	−2.8	6.2	−3.5	−6.3	12.2	0.8

Table 4 (continued)

C	AD	ID	NF	AR	TR	FB	Y	BV	FV	T	OC	WR	BR	AR	I	Age	PD	Re
24	2.1	0.9	0.6	0.7	−0.3	−2.5	−4.0	−0.8	−0.4	−12.4	−4.0	3.8	3.0	−0.8	−2.1	6.2	1.4	0.2
25	−3.0	−4.7	−1.5	−1.3	−1.0	−3.6	2.1	−0.7	−1.1	−10.6	−3.5	2.4	−4.4	−1.0	2.1	8.4	3.4	−2.5
26	−5.4	3.1	−1.5	0.1	2.1	8.5	11.8	0.5	0.0	−5.7	14.7	6.5	−4.6	−5.7	14.0	−13.7	−9.2	−2.4
27	−5.9	11.9	−1.4	12.0	13.4	13.1	8.9	3.2	4.6	5.8	10.3	7.5	4.2	−0.8	15.5	−1.8	−13.9	0.3
28	−6.1	−5.1	−1.3	2.5	1.6	3.9	7.2	0.2	0.0	−9.9	6.1	1.3	−3.0	10.9	1.2	−1.0	−2.7	−1.5
29	−6.5	−8.2	−1.4	2.1	3.1	7.8	9.3	−0.3	−0.4	−9.0	10.2	3.3	−6.7	−1.7	−1.3	−10.1	5.7	1.0
30	−2.4	−1.8	−1.4	1.0	4.9	0.9	5.0	−0.3	0.5	0.2	5.5	5.2	−0.6	11.6	1.8	7.2	−0.4	1.8
31	−5.6	1.8	−1.5	−0.7	3.0	3.3	6.4	−0.2	0.2	0.5	5.1	5.9	−0.9	7.4	2.1	−1.5	−1.1	1.8
32	−2.7	−2.4	−1.5	0.2	5.6	10.4	10.5	1.4	1.0	2.6	9.8	3.2	8.2	12.5	18.0	−1.5	−12.5	−3.5
33	−3.4	−5.6	−1.3	−3.7	−2.8	−2.2	3.4	−0.6	−1.4	−11.6	−1.2	−2.6	1.7	0.6	−2.3	−3.4	1.7	−1.1
34	−2.5	−8.4	−1.0	−6.0	−3.8	−1.0	5.4	−0.7	−1.7	9.0	−0.7	−29.0	37.5	3.5	−4.0	−16.5	15.8	2.8
35	−4.8	−3.3	−0.7	−2.8	−2.2	−4.7	−0.8	−0.9	−1.7	−10.6	−5.4	−10.6	26.8	0.7	−7.7	−1.5	5.2	−1.6
36	1.5	4.6	10.0	−2.2	−1.4	−0.8	−21.2	6.6	6.1	8.8	−5.6	5.3	−3.5	7.9	−6.2	1.4	12.3	−1.6
37	17.7	14.3	2.9	−2.7	−1.1	−3.0	−12.6	0.5	0.8	7.8	−5.2	7.5	−2.7	−1.6	6.7	3.8	8.9	0.0
38	0.2	19.6	−0.5	10.9	8.0	1.3	−2.9	2.3	4.8	1.2	−1.4	8.7	−4.3	4.0	25.8	24.8	−14.1	−0.1
39	−4.9	0.7	−1.0	9.6	10.3	4.4	3.2	1.7	3.7	2.9	3.6	8.1	−4.6	3.0	10.0	12.5	−6.8	2.9
40	19.4	−7.7	2.6	−6.7	−9.5	−5.0	−18.3	−1.4	−2.6	9.0	1.2	−13.5	52.5	4.1	−22.3	−2.4	−8.2	−1.1
41	0.7	−8.0	−1.5	−2.4	−4.6	0.7	−3.7	−1.2	−1.9	8.7	−4.0	−14.1	−5.7	5.4	−9.5	5.4	−14.3	−2.4
42	−4.5	−7.8	−0.2	−6.3	−8.6	−5.4	−11.8	−1.6	−2.8	12.7	−9.8	−3.0	−0.4	−6.2	−11.7	−4.1	−18.5	−2.8
43	−10.7	75.1	−0.5	−6.3	−8.4	−9.8	−1.2	4.0	7.1	1.6	−4.5	10.3	−6.2	−3.8	16.2	13.5	−18.7	5.6

Variable names are abbreviated for better visualization. *C* Cluster/Submarket index, *AD* AgriDis, *ID* IndustryDi, *NF* NUM_FAMILI, *AR* MAIN_FL_AR, *TR* TOTAL_ROOM, *FB* FULL_BATHS, *Y* YEAR_BUILT, *BV* BLD_VALUE, *FV* FINL_VALUE, *T* Tax, *OC* OverallCon, *WR* WhiteR, *BR* BlackR, *AR* AsianR, *I* Median_inc, *Age* Median_age, *PD* PopDen, *Re* Residual

Furthermore, submarket segmentation is also sensitive to accessibility to various facilities, despite the fact that these facilities are not included as the variables in the classification process. For example, submarket 2 and submarket 16, as shown in Fig. 9, are two adjacent submarkets located near the industrial area of Salt Lake County. While submarket 2 is located around open spaces, natural landscapes and golf clubs, submarket 16 is not directly adjacent to open spaces and natural landscapes, as it is surrounded by, or even mixed, with other submarkets or non-residential facilities as shown in Fig. 9. The west side of submarket 16 is a large area of large stores and companies, and the east side of it is the city center of Salt Lake City. Thus, whereas submarket 2 has great accessibility to natural landscapes, submarket 16 is located near commercial facilities and the urban center. This indicates that our submarket classification is able to capture the different underlying preferences for urban amenities. These spatial patterns indicate that our classification fits the local conditions. The validation process further assesses the classification results quantitatively and comprehensively.

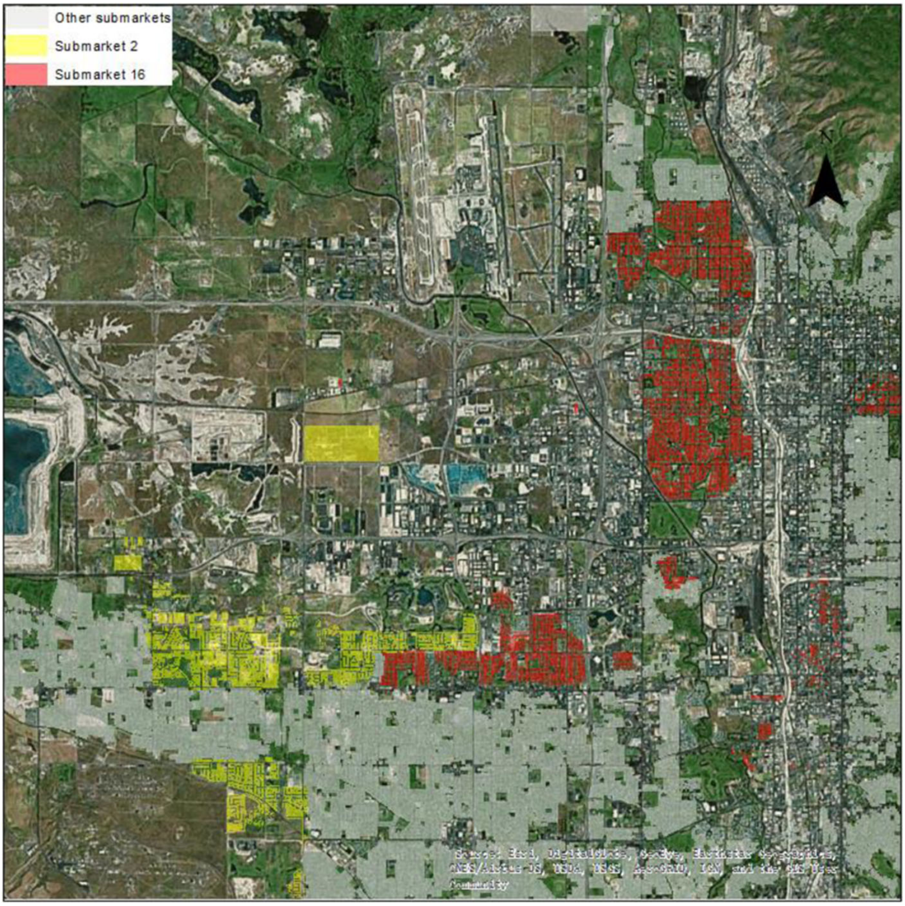
Other submarkets
Submarket 2
Submarket 16
N
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Fig. 9 Submarket 2 and 16

Validation Results

The regression results of the non-submarket model can be found in Table 5. The basic summary of the submarket-based model can be found in Table 6. All the coefficients are multiplied by 1000 to ensure better visualization. We need to point out that the relationship between effects and distances are inverted because higher value with a nearer distance will have a positive effect. So, with all the distances, the coefficient of a positive value represents a negative effect on the neighborhood house values.

Even though the adjusted r-squared value of the non-submarket model is relatively high, we can still observe a higher adjusted r-squared value of 0.922 with the submarket-based model. The higher adjusted r-square value of the non-submarket indicates that the amenity is a suitable addition to the traditional hedonic model and a proper validation framework for housing submarkets and our housing submarket classification explains more when removing the influence of the degree of freedom. Besides the improvements in general prediction accuracy, a better ability to explain local effects can also be found in the submarket-based model. Measuring the global Moran's *I* index based on the spatial relationship of the nearest 500 observations, we can observe a significant and high spatial autocorrelation of the non-submarket model, indicating that there is a strong spatial heterogeneity among the effects of various amenities, which the model is unable to capture. On the other hand, the same measurement shows a very low global Moran's I index (0.02) for the submarket-based model. The spatial autocorrelation is still significant because perfect substitutability only exists in a perfect theoretical submarket, while the effects of most amenities vary with distance in the real world (Li et al. 2016). The improvements could be found directly in the value distribution of the coefficients, as some of which are different compared with the non-submarket model results, which represents that the submarket-based model can account for spatial heterogeneity. For example, light rail stations show a strong areal variation of their effects and some of them even show inconsistent effects with the non-submarket model (Fig. 10), which illustrates that the submarket-based model is more precise in capturing local effects and that sometimes the non-submarket model may be misleading.

These improvements demonstrate that the submarket-based model is more efficient in various kinds of predication accuracies. However, they cannot prove that the coefficients and housing submarkets are meaningful in the real world. Both our classification and current validation comparisons come from data science technologies, which may lead to a common fallacy: the algorithms are good in finding underlying relationships, but they cannot guarantee that the relationships they find make sense. Thus, we still need to find out if our submarket-based model could help us better interpret the effects according to existing theories.

According to the results, we can find certain types of spatial patterns for various amenities. For instance, as shown in Fig. 10, submarkets that have light rail stations or light railways in them do not have strong positive reactions to the light rail system. Most of the light rail stations show negative effects on local housing values. However, we can see that many housing submarkets which are not directly adjacent to but near the light rail system show positive reactions. It is mainly because the light rail transit

Table 5 Result of the non-submarket model

	Variable	Coefficient	
	Intercept	4193.720	***
Housing structure	NUM_FAMILI	−21.532	***
	MAIN_FL_AR	0.072	***
	TOTAL_ROOM	6.768	***
	FULL_BATHS	5.353	***
	YEAR_BUILT	−0.655	***
	Bldv	405.719	***
	OverallCon	16.318	***
Community and location	Tax	−3068.394	***
	Age_csv_Ag	1.709	***
	WhiteR	1.584	***
	BlackR	2.440	***
	AsianR	2.717	***
	Income_csv	0.002	***
	PopDensity	−0.002	***
	AgriDis	−14.774	***
	IndustryDi	3.909	***
Consumer and service	DisThea	13.480	***
	DisExec	−31.191	***
	DisLaun	−19.709	***
	DisBank	0.005	
	DisReta	0.444	
	DisRest	7.996	***
	DisAuto	−5.456	***
Public service	DisHeal	−0.159	
	Libr	4.393	***
	DisEduc	13.044	***
	Worship	−7.516	***
Physical environment	River	34.698	***
	Streams	−12.865	***
	Parks	4.226	***
	Lakes	−14.031	***
	NDVI	337.911	***
Accessibility	CommuterRa	−38.345	***
	Bus	−0.084	
	DisRoad	1.477	***
	LightRail	25.370	***

Adjusted R-Squared: 0.8511

Spatial Autocorrelation in Residuals: 0.23***

0.05 ~ 0.1* 0.01 ~ 0.05** <0.05***

All the coefficients are multiplied by 1000 to ensure better visualization

Table 6 Basic statistics of submarket-based model

	Positive submarket		Negative submarket	
	Counts	Median	Counts	Median
(Intercept)	33	5522.977	4	−42,074.248
DisThea	19	179.381	13	−604.087
DisReta	13	16.164	14	−16.815
DisRest	12	17.310	11	−60.162
AgriDis	4	26.994	28	−26.123
IndustryDi	10	57.221	20	−39.817
DisRoad	26	16.836	14	−30.967
DisEduc	16	36.728	10	−313.537
DisLaun	16	122.585	13	−68.826
DisBank	19	44.643	15	−73.814
DisHeal	16	23.839	18	−31.500
DisAuto	9	28.266	15	−15.560
DisExec	12	253.543	18	−42.916
NUM_FAMILI	3	13.699	23	−23.127
MAIN_FL_AR	38	0.049	0	0.000
TOTAL_ROOM	33	4.092	2	−1.565
FULL_BATHS	26	7.758	2	−5.946
YEAR_BUILT	7	0.156	29	−1.019
Bldv	43	469.793	0	0.000
Tax	7	3303.340	15	−3975.319
Age_csv_Ag	17	7.940	15	−9.393
WhiteR	23	1.938	8	−13.191
BlackR	9	10.179	23	−11.388
AsianR	26	9.158	8	−18.917
Income_csv	12	0.009	28	−0.003
PopDensity	8	0.029	33	−0.030
OverallCon	33	15.759	3	−7.203
Lakes	5	22.187	26	−39.847
CommuterRa	9	1017.806	29	−536.237
Bus	17	14.302	13	−14.080
Libr	13	57.471	16	−130.098
LightRail	19	76.548	12	−143.762
Parks	12	9.292	15	−18.520
Worship	6	38.286	23	−22.984
River	20	428.914	11	−390.391
Streams	4	44.043	32	−16.237
NDVI	18	289.989	11	−319.924

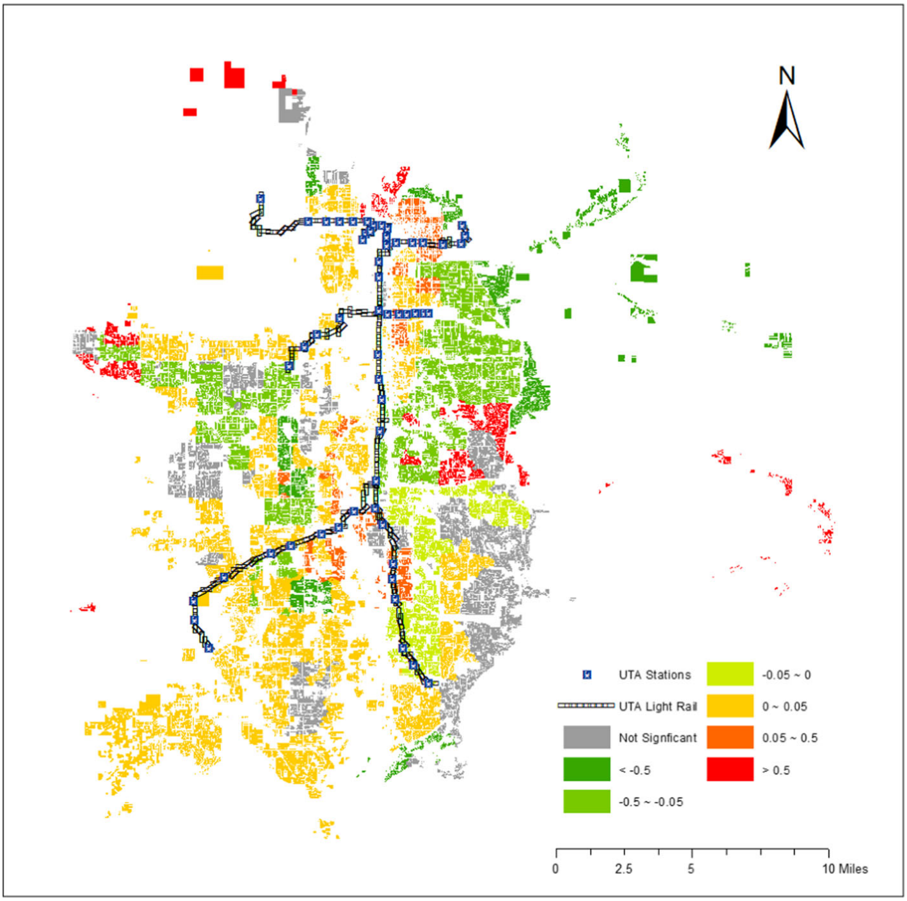

Fig. 10 Effects of light rails stations on housing values in Salt Lake County

system has a strong negative effect on the local level because of safety issues and noise pollution (Li et al. 2016; Liao et al. 2015).

The effects of other amenities may be more related to sectors and locations. Take major roads as an example. Figure 11 illustrates the effects of main roads, from which it is evident that most of the positively influenced housing submarkets concentrate on the county edge area. These areas disproportionately rely on the highway system in daily life. On the other hand, those areas that respond negatively to main roads are found mostly around downtown and the industrial area, where there is less reliance on the highway system for commuting and essential activities. Thus, the positive effect of the main road appears when the local community lives away from employment centers and other amenities.

As these spatial patterns concur with other current studies (Li et al. 2016), our housing submarket classification can explain the underlying effects of various factors, which implies that it is suitable for theoretical explanations. Compared to the non-submarket model, we find improvements in substitutability, which is reflected not only in general prediction accuracy and spatial explanation accuracy, but also in theoretical

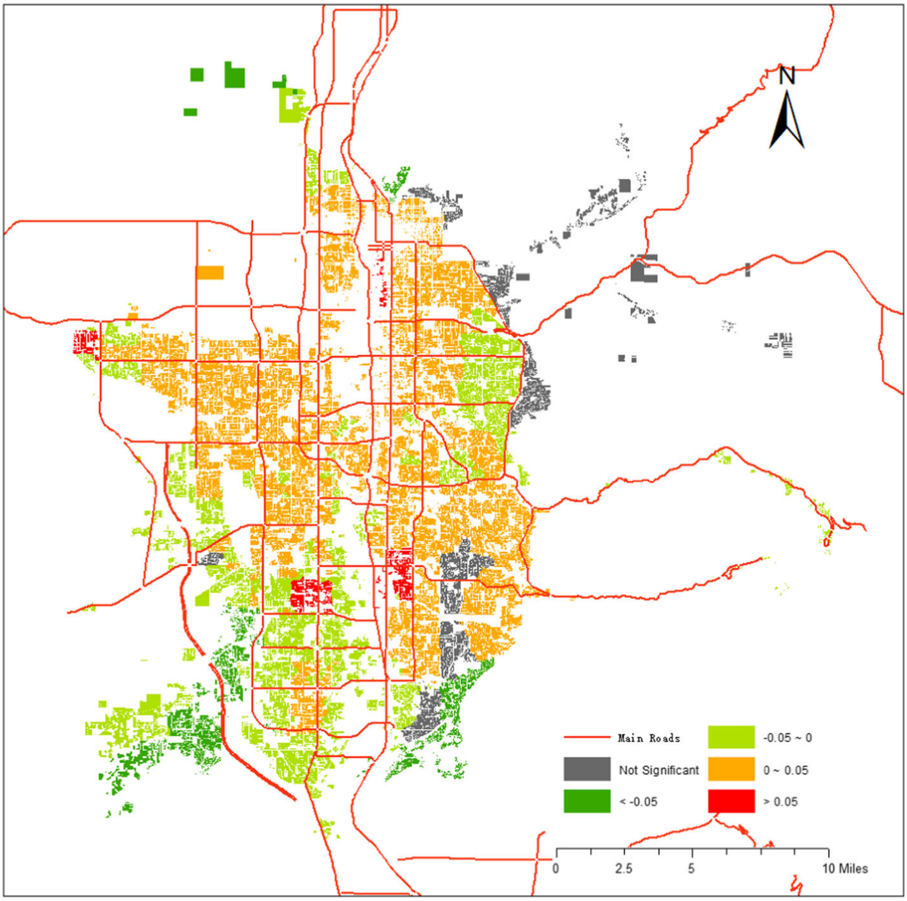

Fig. 11 Effects of main roads on housing values in Salt Lake County

rationality and spatial explanation ability. These improvements show that our classification can maintain substitutability, similarity, and spatial connectivity.

Discussion and Conclusion

The housing submarket is one of the central topics of housing and urban studies, in which regard it is closely associated with urban inequality, sustainable development, and residential segregation (Baker et al. 2016; Wei 2015). Based on the experience of housing studies in economics, sociology, and geography, several principles of the housing submarket have been recognized, and a many data-driven methods have been developed. However, spatial continuity, which is one of the basic principles, may lead to imprecise classification, because it ignores the complexity of urban space. Moreover, most current methods are incompatible with large datasets, and most of the county-level tax assessor's datasets are too large for them. Furthermore, while spatial factors and similarity are easy to assess, the validation process of substitutability in current studies is too centered on model predication accuracy improvement rather than comprehensive assessment. A classification that is unbiased and suitable to process large datasets and validation that account for substitutability are urgently needed.

Relying on various data clustering techniques and emphasizing the importance of spatial connectivity of housing submarkets, this study first puts forward an innovative hybrid classification method for housing submarkets. It is based on various variables to capture similarity, residuals of a simple linear model between housing value and these factors to represent substitutability, and the notion of spatial connectivity instead of spatial integrity in order to reflect the complexity of urban space. This study integrates several clustering algorithms to handle big datasets as well as classify housing submarkets while maintaining an acceptable substitutability, similarity, and spatial connectivity. Using the dataset of Salt Lake County, our results show that, while ensuring similarity based on the local level, our submarket classification aggregates submarket fragments together to avoid two pitfalls: (1) an extremely large number of clusters, which would increase the difficulty of understanding the study area, and (2) small submarkets, which could result in inaccurate local prediction results because of small sample size. We further find that our submarket-based model is able to capture residential differences, the complexity of urban space, and housing structure differences. Moreover, the model analyzes the role of accessibility to various amenities and shows that roads are important boundaries for submarkets. In addition, this method is computationally fast and economical, enabling it to handle county-level data with a personal computer.

We also have developed a comprehensive validation process for substitutability, which complements well-known existing assessments of similarity and spatial connectivity. To avoid a biased assessment, a comprehensive framework is introduced into the validation process, which is based on the three aspects of substitutability: prediction accuracy, the ability to consider local effects, and theoretical rationality. Compared to non-submarket models, our submarket-based model provides improvements in all three of these aspects and, in particular, it excels in discovering local effects. Our validation not only shows that the classification method is suitable for classifying housing submarkets in terms of the three aspects but also provides a potential assessment which is more integrated and comprehensive for other housing submarket classifications.

Our classification method is suitable for planners, estate agencies, tax assessors and decision-makers for modeling housing submarkets and estimate housing values efficiently and economically. The results show that the classification is able to model residential differences, as well as capture the variation of houses and preferences in the neighborhood. The validation framework is also meaningful as it directly tests if a new classification method could be reasonable and useful in reality. The submarket-based model also shows that it is possible to combine the concept of the housing submarket with the hedonic model to assess comparable potential influences of various factors at the local level. The whole process of classification and validation would be beneficial for local governments as it not only provides a feasible way to assess housing values more precisely, but also contributes a better methodology for understanding local urban space, residential segregation, and housing inequality.

Acknowledgments We would like to thank the constructive comments of Robert Argenbright.

Funding The study was funded by National Institute for Transportation & Communities (69A3551747112) and the Ford Foundation (0155–0883).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

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.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93–115.

Anselin, L. (2013). Spatial econometrics: Methods and models. In: Springer Science & Business Media.

Assunção, R. M., Neves, M. C., Câmara, G., & da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science, 20*(7), 797–811.

Bailey, T. J. (1999). Modelling the residential sub-market: Breaking the monocentric mould. *Urban Studies, 36*(7), 1119–1135.

Baker, E., Bentley, R., Lester, L., & Beer, A. (2016). Housing affordability and residential mobility as drivers of locational inequality. *Applied Geography, 72*, 65–75. https://doi.org/10.1016/j.apgeog.2016.05.007

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

Batty, M., & Xie, Y. (1996). Preliminary evidence for a theory of the fractal city. *Environment and Planning A, 28*(10), 1745–1762.

Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics, 8*(2), 160–183.

Bourassa, S. C., Hoesli, M., & Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics, 12*(1), 12–28. https://doi.org/10.1016/s1051-1377(03)00003-2

Boustan, L. P. (2013). Local public goods and the demand for high-income municipalities. *Journal of Urban Economics, 76*, 71–82. https://doi.org/10.1016/j.jue.2013.02.003

Brasington, D. (1999). Which measures of school quality does the housing market value? *Journal of Real Estate Research, 18*(3), 395–413. https://doi.org/10.5555/rees.18.3.g1n4hq8212111j5m

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'NbClust. *Journal of Statistical Software, 61*, 1–36.

Chi, G., & Marcouiller, D. W. (2013). Natural amenities and their effects on migration along the urban–rural continuum. *The Annals of Regional Science, 50*(3), 861–883.

Cho, S. H., Bowker, J. M., & Park, W. M. (2006). Measuring the contribution of water and green space amenities to housing values: An application and comparison of spatially weighted hedonic models. *Journal of Agricultural and Resource Economics*, 485–507.

Clapp, J. M., & Wang, Y. (2006). Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics, 59*(2), 259–284. https://doi.org/10.1016/j.jue.2005.10.003

Day, B. (2003). Submarket identification in property markets: A hedonic housing price model for Glasgow (no. 03-09). *CSERGE Working Paper EDM.*

Diao, M., & Ferreira, J., Jr. (2010). Residential property values and the built environment: Empirical study in the Boston, Massachusetts, metropolitan area. *Transportation Research Record: Journal of the Transportation Research Board, 2174*, 138–147.

Dong, E., Liao, F. H. F., & Kang, H. (2016). *Grocery shopping: Geographic scale matters in analyzing effects of the built environment on choice of travel mode.* Paper presented at the Transportation Research Board 95th annual meeting.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics, 22*(3), 433–452.

Duncan, M. (2011). The synergistic influence of light rail stations and zoning on home prices. *Environment and Planning A, 43*(9), 2125–2142.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the Kdd.

Ewing, R., Meakins, G., Hamidi, S., & Nelson, A. C. (2014). Relationship between urban sprawl and physical activity, obesity, and morbidity - update and refinement. *Health & Place, 26*, 118–126. https://doi.org/10.1016/j.healthplace.2013.12.008

Fotheringham, A. S., & Rogerson, P. A. (Eds.). (2008). *The SAGE handbook of spatial analysis*. Sage.

Fovell, R. G., & Fovell, M.-Y. C. (1993). Climate zones of the conterminous United States defined using cluster analysis. *Journal of Climate, 6*(11), 2103–2135.

Gao, X., & Asami, Y. (2007). Effect of urban landscapes on land prices in two Japanese cities. *Landscape and Urban Planning, 81*(1–2), 155–166. https://doi.org/10.1016/j.landurbplan.2006.11.007

Geoghegan, J. (2002). The value of open spaces in residential land use. *Land Use Policy, 19*(1), 91–98. https://doi.org/10.1016/S0264-8377(01)00040-0

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*(3), 189–206.

Glaeser, E. L. (2000). The new economics of urban and regional growth. *The Oxford Handbook of Economic Geography*, 83–98.

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics, 12*(3), 181–201. https://doi.org/10.1016/s1051-1377(03)00031-7

Griffith, D. A. (2002). A spatial filtering specification for the auto-Poisson model. *Statistics & Probability Letters, 58*(3), 245–251.

Grubesic, T. H., Wei, R., & Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers, 104*(6), 1134–1156. https://doi.org/10.1080/00045608.2014.958389

Gujarati, D. (1970). Use of dummy variables in testing for equality between sets of coefficients in two linear regressions: A note. *The American Statistician, 24*(1), 50–52.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science, 22*(7), 801–823. https://doi.org/10.1080/13658810701674970

Hamidi, S., & Ewing, R. (2014). A longitudinal study of changes in urban sprawl between 2000 and 2010 in the United States. *Landscape and Urban Planning, 128*, 72–82. https://doi.org/10.1016/j.landurbplan.2014.04.021

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C: Applied Statistics, 28*(1), 100–108.

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

Hothorn, T., Hornik, K., Wiel, M. A. V. D., & Zeileis, A. (2006a). A Lego system for conditional inference. *American Statistician, 60*(3), 257–263.

Hothorn, T., Hornik, K., & Zeileis, A. (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674.

Huh, S., & Kwak, S. J. (1997). The choice of functional form and variables in the hedonic price model in Seoul. *Urban Studies, 34*(7), 989–998.

Hui, E. C., Liang, C., & Yip, T. L. (2018). Impact of semi-obnoxious facilities and urban renewal strategy on subdivided units. *Applied Geography, 91*, 144–155.

Hwang, S., & Thill, J.-C. (2009). Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning. B, Planning & Design, 36*(5), 865–882.

Iceland, J., & Sharp, G. (2013). White residential segregation in US metropolitan areas: Conceptual issues, patterns, and trends from the US census, 1980 to 2010. *Population Research and Policy Review, 32*(5), 663–686.

Jarosz, B. (2008). Using Assessor parcel data to maintain housing unit counts for small area population estimates. In S. H. Murdock & D. A. Swanson (Eds.), *Applied demography in the 21st century: Selected papers from the biennial conference on applied demography, San Antonio, Texas, January 7–9, 2007* (pp. 89–101). Dordrecht: Springer Netherlands.

Jia, P., & Gaughan, A. E. (2016). Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. *Applied Geography, 66*, 100–108. https://doi.org/10.1016/j.apgeog.2015.11.006

Jim, C., & Chen, W. Y. (2007). Consumption preferences and environmental externalities: A hedonic analysis of the housing market in Guangzhou. *Geoforum, 38*(2), 414–431.

Kong, F., Yin, H., & Nakagoshi, N. (2007). Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning, 79*(3–4), 240–252. https://doi.org/10.1016/j.landurbplan.2006.02.013

Korinek, K., & Maloney, T. N. (Eds.). (2010). *Migration in the 21st century: Rights, outcomes, and policy.* Routledge.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods, 26*(6), 1481–1496.

Lee, C., & Moudon, A. V. (2006). The 3Ds + R: Quantifying land use and urban form correlates of walking. *Transportation Research Part D: Transport and Environment, 11*(3), 204–215. https://doi.org/10.1016/j.trd.2006.02.003

Li, H., Wei, Y. H. D., Yu, Z., & Tian, G. (2016). Amenity, accessibility and housing values in metropolitan USA: A study of Salt Lake County, Utah. *Cities, 59*, 113–125. https://doi.org/10.1016/j.cities.2016.07.001

Li, H., Wei, Y. H. D., & Korinek, K. (2017). Modelling urban expansion in the transitional greater Mekong region. *Urban Studies, 55*, 1729–1748. https://doi.org/10.1177/0042098017700560

Li, H., Wei, Y. H. D., Wu, Y., & Tian, G. (2018). Analyzing housing prices in Shanghai with open data: Amenity, accessibility and urban structure. *Cities*.

Liao, F. H. F., Farber, S., & Ewing, R. (2015). Compact development and preference heterogeneity in residential location choice behaviour: A latent class analysis. *Urban Studies, 52*(2), 314–337.

Lowry, J. H., & Lowry, M. B. (2014). Comparing spatial metrics that quantify urban form. *Computers, Environment and Urban Systems, 44*, 59–67.

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.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems, 33*(6), 403–408. https://doi.org/10.1016/j.compenvurbsys.2009.11.001

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification, 31*(3), 274–295.

Osland, L., & Thorsen, I. (2013). Spatial impacts, local labour market characteristics and housing prices. *Urban Studies*, 0042098012474699.

Park, S., & Lah, T. J. (2006). The impact of WTE facility on housing value. *International Review of Public Administration, 10*(2), 75–83.

Peeters, A., Zude, M., Käthner, J., Ünlü, M., Kanber, R., Hetzroni, A., Gebbers, R., & Ben-Gal, A. (2015). Getis–Ord's hot-and cold-spot statistics as a basis for multivariate spatial clustering of orchard tree data. *Computers and Electronics in Agriculture, 111*, 140–150.

Poudyal, N. C., Hodges, D. G., & Merrett, C. D. (2009). A hedonic analysis of the demand for and benefits of urban recreation parks. *Land Use Policy, 26*(4), 975–983.

Rothenberg, J. (1991). *The maze of urban housing markets: Theory, evidence, and policy.* University of Chicago Press.

Royuela, V., & Duque, J. C. (2013). HouSI: Heuristic for delimitation of housing submarkets and price homogeneous areas. *Computers, Environment and Urban Systems, 37*, 59–69.

Sander, H. A., & Polasky, S. (2009). The value of views and open space: Estimates from a hedonic pricing model for Ramsey County, Minnesota, USA. *Land Use Policy, 26*(3), 837–845. https://doi.org/10.1016/j.landusepol.2008.10.009

She, B., Zhu, X., Ye, X., Guo, W., Su, K., & Lee, J. (2015). Weighted network Voronoi diagrams for local spatial analysis. *Computers, Environment and Urban Systems, 52*, 70–80.

Strasser, H., & Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics, 8*(2), 220–250.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348.

Tian, G., Wei, Y. H. D., & Li, H. (2017). Combined effects of accessibility and environmental health risk on housing Price: A case of Salt Lake County, UT. *Applied Geography*, *89*, *12–21*.

Torrens, P. M. (2008). A toolkit for measuring sprawl. *Applied Spatial Analysis and Policy, 1*(1), 5–36.

Tu, Y., Sun, H., & Yu, S. M. (2007). Spatial autocorrelations and urban housing market segmentation. *The Journal of Real Estate Finance and Economics, 34*(3), 385–406.

U.S. Census Bureau. (2001). Census 2000. *Available online* https://www.census.gov/prod/cen2000/phc-1-46.pdf. Accessed 10 May 2019.

U.S. Census Bureau. (2011). Census 2010. *Available online* https://www.census.gov/quickfacts/fact/table/saltlakecountyutah,ut/PST045218 Accessed 10 May 2019.

Waltert, F., & Schläpfer, F. (2010). Landscape amenities and local development: A review of migration, regional economic and hedonic pricing studies. *Ecological Economics, 70*(2), 141–152. https://doi.org/10.1016/j.ecolecon.2010.09.031

Watkins, C. A. (2001). The definition and identification of housing submarkets. *Environment and Planning A, 33*(12), 2235–2253.

Wei, Y. H. D. (2015). Spatiality of regional inequality. *Applied Geography, 61*, 1–10. https://doi.org/10.1016/j.apgeog.2015.03.013

Wei, Y. H. D. (2016). Towards equitable and sustainable urban space. *Sustainability, 8*(8), 804.

Wei, Y. H. D., & Ewing, R. (2018). Urban expansion, sprawl and inequality. *Landscape and Urban Planning, 177*, 259–265.

Wei, Y. H. D., Xiao, W., Wen, M., & Wei, R. (2016). Walkability, land use and physical activity. *Sustainability, 8*(1), 65.

Wei, Y. H. D., Xiao, W., Simon, C. A., Liu, B., & Ni, Y. (2018). Neighborhood, race and educational inequality. *Cities, 73*, 1–13. https://doi.org/10.1016/j.cities.2017.09.013

Whitehead, C. M. (1999). Urban housing markets: Theory and policy. *Handbook of Regional and Urban Economics, 3*, 1559–1594.

Wilhelmsson, M. (2004). A method to derive housing sub-markets and reduce spatial dependency. *Property Management, 22*(4), 276–288.

Wise, S., Haining, R., & Ma, J. (1997). Regionalisation tools for the exploratory spatial analysis of health data. In *Recent developments in spatial analysis* (pp. 83–100): Springer.

Wu, F. (1996). Changes in the structure of public housing provision in urban China. *Urban Studies, 33*(9), 1601–1627.

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., Du, Q., & Luo, P. (2017). Spatial effects of accessibility to parks on housing prices in Shenzhen, China. *Habitat International, 63*, 45–54.

Xie, Y., Fang, C., Lin, G., Gong, H., & Qiao, B. (2007). Tempo-spatial patterns of land use changes and urban development in globalizing China: A study of Beijing. *Sensors, 7*(11), 2881–2906.

Yu, D., Wei, Y. H. D., & Wu, C. (2007). Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning. B, Planning & Design, 34*(6), 1085–1102.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.