REMV
REAL ESTATE MANAGEMENT AND VALUATION

Journal homepage:
www.remv-journal.com

REVIEW OF CLUSTERING METHODS USED IN DATA-DRIVEN HOUSING MARKET SEGMENTATION

Štěpán Skovajsa[1]*

[1]*Institute of Forensic Engineering, Brno University of Technology, Purkyňova 464/118, 612 00 Brno, Czech Republic, e-mail: Stepan.Skovajsa@vut.cz*

* *Correspondingauthor*

ARTICLE INFO

Keywords:

clustering algorithms, housing market analysis, housing market segmentation, data-driven segmentation

JEL Classification:

R31

ABSTRACT

A huge effort has already been made to prove the existence of housing market segments, as well as how to utilize them to improve valuation accuracy and gain knowledge about the inner structure of the entire superior housing market. Accordingly, many different methods on the topic have been explored, but no universal framework is yet known. The aim of this article is to review some previous studies on data-driven housing market segmentation methods with a focus on clustering methods and their ability to capture market segments with respect to the shape of clusters, fuzziness and hierarchical structure.

Citation: Skovajsa, Š. (2023). Review of clustering methods used in data-driven housing market segmentation. *Real Estate Management and Valuation, 31*(3), 67-74. https://doi.org/10.2478/remav-2023-0022

1. Introduction

The segmentation of the housing market can be roughly defined as the classification of all real estates on the market into smaller groups, known as submarkets or segments, in such a way that all estates in the given segment have the same or similar values of housing characteristics, but different values in comparison with estates in another segment. This means the low entropy in housing characteristics is present within estates of one particular segment, but high entropy is present among estates in two different segments.

The motivation behind the housing market segmentation lies in the heterogeneity of the real estate market. According to the theory of revealed preference, the preferences of each consumer are projected to his internal valuation of each specific good, and therefore the utility function can somehow be determined by the observation of purchasing habits. Hedonic pricing extends this theory to heterogeneous goods. Similarly, as the utility function can be determined by observation of purchasing habits, the value of each characteristic of heterogeneous goods can be determined by observing the willingness to pay for that characteristic through the prices (Rosen, 1974), as every consumer, whether consciously or subconsciously, compares the utility for each housing characteristic with given price and available substitutes. The hedonic valuation function can be defined as $\pi(z, \rho)$, where ρ is a vector of hedonic regression coefficients, and z is a characteristics vector representing the quantity of each characteristic. Assume U be an (infinite) set of all possible locations in the particular market. The spatial heterogeneity is the possibility to have different regression coefficients ρ_u and ρ_v for different locations $u, v \in U$. If we keep ρ_u constant for all $u \in U$, we may encounter aggregation bias (references to related studies can be found in Usman et al.,2020), i.e., invalid valuation assumption about constant values for all housing characteristics in each location u, in other words, assuming one spatial equilibrium for the whole market. This may apply also for the structural heterogeneity (e.g., two characteristically very distinct apartments in the same building).

In housing market segmentation, we distinguish between two basic approaches: a priori segmentation (also known as ad hoc or subjective segmentation or experience-oriented segmentation), and data-driven segmentation (also known as objective segmentation) (Usman et al., 2020). A priori segmentation is based on

subjectively predefined criteria (e.g., geographically or administratively defined boundaries, postal codes, boundaries defined by local experts). On the contrary, the data-driven approach utilizes data-science to find such segments by observing latent structures in the input data. The a priori approach is not very scalable as there is a need for a human expert to manually delineate segments and update each particular information when it is going to deprecate. However, most of the critique is directed to the arbitrariness and non-scientificness of the a priori approach (Usman et al., 2020), which brings a data-driven approach to the foreground of current housing segmentation research.

The segmentation itself is, in most cases, an application of clustering (Shi et al., 2015), the statistical method of determining groups of similar objects. There are many clustering algorithms with various properties and different approaches to how the clusters are formed. From the high-level perspective, the algorithms can be divided into categories by the way how they form clusters (e.g., by distance from the mean, by the density of neighbors, utilizing probability theory). Each approach has its pros and cons, which should be taken into account when the particular algorithm is being chosen. This article reviews some popular algorithms, highlights their pros and cons, and references some applications in housing market segmentation for each of them. The aim is to compare the categories of algorithms concerning the questions about possible shapes of clusters, fuzziness possibilities, overlapping of segments, and their hierarchical structure to analyze the way the submarkets nest. In the end, the discussion tries to summarize the findings.

2. Fundaments of clustering

Clustering can be defined as classifying data points $\boldsymbol{x_n} \in \boldsymbol{X}$, where $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N})$ is an input dataset, into $\boldsymbol{K}$ clusters $\boldsymbol{C_1}, \boldsymbol{C_2}, \dots, \boldsymbol{C_K}$such that for each cluster $\boldsymbol{C_k}$ and data point $\boldsymbol{x_n}$, we have a membership function $\boldsymbol{\mu_k}(\boldsymbol{x_n}) \in [\boldsymbol{0}, \boldsymbol{1}]$ indicating strongness of inclusion $\boldsymbol{x_n}$ into cluster $\boldsymbol{C_k}$. The clustering of a sample $\boldsymbol{x_n}$ can be also expressed as a vector of sample inclusion into each cluster:

$$\boldsymbol{w}(\boldsymbol{x_n}) = \big(\mu_1(\boldsymbol{x_n}), \mu_2(\boldsymbol{x_n}), \dots, \mu_K(\boldsymbol{x_n})\big) = (w_{n1}, w_{n2}, \dots, w_{nK}) \quad (1)$$

And similarly, the inclusion of each point to a particular cluster can be represented as a matrix

$$W(X) = \begin{bmatrix} \boldsymbol{w}(\boldsymbol{x_1}) \\ \boldsymbol{w}(\boldsymbol{x_2}) \\ \vdots \\ \boldsymbol{w}(\boldsymbol{x_N}) \end{bmatrix} = \begin{bmatrix} \mu_1(\boldsymbol{x_1}) & \mu_2(\boldsymbol{x_1}) & \cdots & \mu_K(\boldsymbol{x_1}) \\ \mu_1(\boldsymbol{x_2}) & \mu_2(\boldsymbol{x_2}) & \cdots & \mu_K(\boldsymbol{x_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1(\boldsymbol{x_N}) & \mu_2(\boldsymbol{x_N}) & \cdots & \mu_K(\boldsymbol{x_N}) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NK} \end{bmatrix} \quad (2)$$

For all data points$x_n \in X$, the sum of membership into each cluster should be equal to 1. Formally:

$$\sum_{k=1}^{K} \mu_k(\boldsymbol{x_n}) = \sum_{k=1}^{K} w_{nk} = 1 \text{ for all } n \in [1, N] \quad (3)$$

From the perspective of the co-domain of membership functionμ, the clustering methods can be split into hard clustering (or crips clustering) and soft clustering methods. Soft clustering is based on the fuzzy-set theory, such that each data point $\boldsymbol{x_n}$can belong to more than one cluster C_k, where the strongness of membership is indicated by $\mu_k(\boldsymbol{x_n})$ or w_{nk}.Hard clustering is a special case of soft clustering. It is based on crisp-set theory ($x_n$either belongs to the cluster or not, i.e., the codomain of μ is restricted to $\{0, 1\}$), and the membership of data point to a cluster is exclusive, i.e., the equation (3) still holds. Formally, the membership function μ for the hard clustering can be defined as:

$$\mu_k(\boldsymbol{x_n}) = w_{nk} = \begin{cases} 1 & \text{iff} \boldsymbol{x_n} \text{belongs to cluster} C_k \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

This means w_{nk} is 1 if n-th example is contained in the k-th cluster, 0 otherwise.

3. Base clustering algorithm types used in housing market segmentation

3.1. Centroid-based clustering

This type of clustering is based on the measuring distance from $\boldsymbol{x_n}$ to $c_k$via δ_k, where c_k is the centroid of the cluster C_k, and δ_k is a distance function to that centroid. The centroid can be seen as some representative data point for the particular cluster. Usually, δ_k measures distance as the Euclidean, Minkowski, or Manhattan distance.

3.1.1. K-means algorithm

From the view of the membership function μ, the algorithm performs hard clustering, as it always assigns just one particular cluster C_k to the data point x_n, so that the centroid of the cluster is the nearest centroid of all clusters to the point x_n.This basically

forms clusters as Voronoi cells, which is a big advantage in simplicity and interpretability, but also a big disadvantage - we cannot fit non-convex shapes of clusters and/or detect outliers. In K-means, the centroid of a cluster is computed as the mean of all data points belonging to that cluster.

Each iteration of the algorithm proceeds in two steps. First, it assigns all data points to clusters given by current centroids $c_1, c_2, \ldots, c_K$. In the second step, the position of centroids is reevaluated, and the particular update for each of them is calculated accordingly. The algorithm ends when the required number of steps is completed or when some convergence criteria is fulfilled (e.g., negligibly small movement of centroids).The learning process lies in the minimization of the following objective function:

$$\sum_k^K \sum_n^N w_{nk} \delta_k(x_n) \quad (5)$$

wherew_{nk}can be seen as an indicator function.

This algorithm was also used in the first housing segmentation framework considered by Bourassa et al. (1999), where the clustering was performed over PCA identified factors. Although the K-means algorithm is very simple and has strong assumptions, it is still used nowadays. For example, Calka (2019) proposes the two-stage model, where K-means is used to cluster by housing characteristics in the first stage, and then the geostatistical approach to incorporate spatial dependence in the second stage. Moreover, in most cases, the K-means algorithm serves as a basis clustering algorithm for comparison when any new segmentation algorithm arises (e.g., Hwang & Thill, (2009); Wu et al., (2018); Liu et al., (2021); Chen et al., (2021)).

3.1.2. Fuzzy C-means algorithm (FCM)

This algorithm proceeds in a similar way as K-means, except for the fact that the distance to all centroids is assumed for each data point, not just the nearest one. The objective function is defined as:

$$\sum_k^K \sum_n^N (w_{nk})^m \delta_k(x_n) \quad (6)$$

where the hyper parameter $m \in [1, \infty]$is a fuzziness exponent (or fuzzifying factor).

The higher the value of m the more fuzziness is allowed. Note that when $m = 1$, we obtain the same objective function as in the case of K-means. Malinowski et al. (2018)compare several clustering algorithms, including both K-means and FCM. In that study, K-means slightly outperformed the FCM algorithm. On the other hand, Hwangand Thill (2009) conclude that the fuzzy C-means algorithm provides better results than K-means.

Shi et al. (2015) uses an adapted version of FCM with a homogeneity index, which acts as some sort of filter to delineate outliers and make clusters more homogeneous. The approach is very simple and general.

3.2. Density-based clustering (model-based clustering)

In this type of clustering, the distance of the data points $x \in X$ is also assumed, but in another manner, where there are no such representants as centroids, and rather the distance between individual points is assumed. The idea is that the cluster should be a more contagious area with a higher density of points. This means that the close data points are more likely to be similar, hence more likely to form a cluster, and vice versa. This allows density-based clustering methods to fit non-linear shapes of clusters and identify outliers more easily as standalone distant points from the others. Moreover, the hyperparameter K is no longer needed as the algorithms can figure the corresponding number of clusters itself, i.e., K is no longer an input but an output.

3.2.1. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

The algorithm proceeds by identifying the core points in the first step. The data point x_n is the core point if there are at least minPts other points in the ϵ distance from x_n.The non-core points are then connected to the core points if there is at least one core point in the ϵ distance, i.e., the core point is directly reachable (Ester et al., (1996)).The remaining points are classified as noise (outliers).

The problem with this algorithm is that each cluster might have a different density, therefore it is hard to find proper ϵ and minPts hyperparameters.This is bypassed in successive algorithms like OPTICS, which is similar in nature but addresses this problem via sorting by spatial closeness.

For example, Guoet al. (2012) employed DBSCAN to analyze time series and found specific clusters of growth associated with policy changes in real estate markets in China. This is not a housing market segmentation problem as presented here, but still worth mentioning as a higher-level segmentation method useful for identifying similar cities according to their real estate market growth in time.

3.2.2. Density-Based Spatial Clustering (DBSC)

This kind of algorithm can distinguish between the

eISSN: 2300-5289 | Received 2023-01-31 | Revised 2023-03-27 | Accepted 2023-03-29

spatial proximity of the data point and its attributes. In the first phase, it utilizes Delaunay triangulation and several metrics to detect long edges, which are later removed. In the second phase, attribute similarity is assumed, which means Euclidean attributes distance is computed between neighboring points to determine borders between clusters (two points might be neighbors, but if they possess very different attributes, they are very unlikely to be in the same cluster). This algorithm was proposed in Liu et al., (2012), while in Wu et al. (2018), it was used for housing market segmentation along with *geographically weighted PCA* (GWPCA). It was shown that DBSC outperforms the basic K-means approach.

3.3. Distribution-based clustering

Distribution based clustering is a probabilistic type of clustering, i.e., the clusters are represented by probabilistic distributions and the learning proceeds by fitting parameters of those distributions so that the likelihood of the generative model is maximized (the probability that data point $x \in X$ comes from the specific cluster C_k).This kind of clustering allows for a higher level of flexibility because of the flexibility of probabilistic modeling. The probability theory itself does not specify how to solve the problems, but rather provides the theory for model composition. If the model is analytically intractable, various algorithms may be used. They are divided into two distinct groups: deterministic such as EM (Expectation-Maximization) or Variational Inference, and non-deterministic/probabilistic – like Monte-Carlo simulation methods.

Liu et al., (2018) uses the Bayesian approach, where the presence in a particular cluster is present as the prior belief, and the clusters themselves are represented by multivariate Gaussian distributions. There is also a comparison with the K-means approach, which is outperformed.

3.4. Connectivity-based clustering (hierarchical clustering)

This type of clustering extends the basic representation of membership (let us recall the matrix W) with a hierarchical structure between all clusters. This structure can be represented as a tree from the data modeling perspective (among hierarchical clustering methods, the term *dendrogram* is used for the plot of that tree). The connectivity-based algorithms can be divided according to which side of the tree the hierarchical structure is constructed. The divisive approach (top-down, i.e., from the root to the leaves) begins with all points under one big cluster which is gradually divided into smaller subclusters, and the agglomerative approach (bottom-up, i.e., from the leaves to the root), where each point represents its own cluster (singleton) in the beginning, with gradual merging to form higher-level clusters. Various criteria can be applied to compare specific clusters to form a tree. For example, variance (e.g., Ward's minimum variance method used in Bourassaet al., (1999); Calka (2019), entropy and the probability that the clusters were obtained by the same or very similar random process in Liuet al., 2021), etc.

This kind of clustering can also serve as an ancillary method for other clustering methods, so that own clusters are constructed with different clustering methods, and next the hierarchical clustering is performed to obtain a hierarchical structure of previously constructed segments. For example, Liu et al., (2021) use distribution-based clustering to construct segments, and then another algorithm, based on agglomerative technique with Bayesian hypothesis testing, is utilized to obtain a hierarchical structure for previously obtained clusters.

3.5. Population-based clustering

Generally, population algorithms are founded in managing a population of individuals, where each individual acts as a solution to a given problem. This approach of a whole population of solutions should overcome the problem of getting stuck in local minima (or maxima) upon classical algorithms, where usually only one solution is initialized, and then modified through several iterations (Talbi, 2009).

3.5.1.Evolutionary clustering methods

Evolutionary algorithms (EA) are inspired by the evolution process, where mostly the strongest individuals (solutions, in the case of EA) survive. At the beginning, a population with solutions is initialized (see p. 193 in Talbi, 2009). Then, in consecutive iterations, the population is modified by operators to produce a new generation. The main operators used within EA are *selection*, *crossover*, and *mutation*.

The same principles apply for the genetic algorithms (GA) as a subclass of EA. Manganelli et al., (2015) uses geographically weighted regression for preliminary segmentation and genetic algorithm for identifying marginal contribution of submarkets, but they did not find a huge improvement over multiple linear regression.

3.5.2.Swarm-intelligence based clustering methods

The swarm-intelligence based algorithms are mostly inspired by collective intelligence occurring in nature (e.g., ants, bees, birds). The algorithms usually operate by simulating the swarm of individuals living in a given environment, possessing their internal states, allowing to change these states by observing surroundings, action triggering, etc.

Chen et al., (2021) introduces a housing segmentation method based on the swarm-inspired projection (SIP) algorithm. The SIP algorithm is based on the way doves seek crumbs. The results provide better performance for Taipei's market compared to the K-means approach.

3.6. Affinity propagation clustering

This is also quite a novel approach introduced by Frey and Dueck (2007), based on "passing messages" between data points to measure mutual attractivity. Similarly to density-based clustering methods, this method does not require the number of clusters to be specified. When it comes to its application in housing market segmentation, we found only Hu et al., (2022), who used a combination of geographically and temporally weighted regression (GTWR) and affinity propagation clustering.

3.7. Constrained clustering (semi-supervised learning clustering)

According to the input data, the methods of data-driven clustering can be divided into two distinct categories. First are *unsupervised learning methods*, where the input data does not contain any prior information or domain knowledge; these have all been introduced earlier. Sometimes, if we have some intuition regarding the problem, we may want to explicitly specify some prior knowledge, constraint, or domain to help an algorithm proceed faster by limiting the searched state space. This is the case of *constraint-based* or *semi-supervised* clustering algorithms, which are, in most cases, derived from unsupervised clustering methods. In that sense, constraint-based clustering is a mix of data-driven and a priori segmentation approaches, as constraints are specified manually by an expert.

For example, Zhang et al., (2022) uses a constrained-based clustering approach, where constraints are represented as a set of functions with conditioned values. These constraints can incorporate economic rationality resulting from the economic theory, like transportation advantage in the above-mentioned example.

For housing market segmentation, it is typical to use some spatial constraints (as submarkets are assumed to be spatially contiguous). The previously mentioned DBSC algorithm is an example of such a case, as it uses Delaunay triangulation for spatial coordinates (Wu et al., 2018), which can be understood as spatial constraining. The next example of spatial constraining can be found in the approach of Royuela and Duque (2013). Similarly, Kryvobokov (2013) uses Voronoi cells (also known as Thiessen polygons) around each determined centroid based on spatial location.

4. Discussion

The advantages, disadvantages and examples of algorithms for each clustering approach are summarized in Table 1. The properties of the desired algorithm should be considered when using any clustering method for housing market segmentation. This depends on the structure of a particular market and the requirements for such a segmentation. For this reason, we should discuss the algorithms concerning what shapes of clusters they can fit; whether or not it allows fuzziness, and if so, how the overlapping can be measured. Next, we might also be interested in the hierarchical structure between segments to make a deeper analysis of the market structure.

4.1. Shape of the clusters

According to the shape of the clusters, the algorithms can be divided into those that can fit arbitrary shapes of clusters and those that cannot. This is summarized in Table 1 in Liu et al. (2012) for more algorithms than presented in this paper. For centroid-based algorithms, it can be clearly stated that they cannot fit arbitrary shapes based on the principle of how they work; moreover, the number of clusters should usually be explicitly given (at least in the case of K-means and FCM).This is an advantage of density-based algorithms, such as DBSCAN and OPTICS, or more advanced DBSC, which can fit arbitrary shapes and implicitly find out the number of clusters. However, in Malinowski et al. (2018) the OPTICS was outperformed both by simpler K-means and FCM, which had been mentioned above, and therefore the overall advantage of density-based over centroid-based algorithms cannot be stated. Distribution-based algorithms can lie somewhere in between because, in theory, the arbitrary shape can be represented by a specific probability distribution or a mixture of distributions,

eISSN: 2300-5289 | Received 2023-01-31 | Revised 2023-03-27 | Accepted 2023-03-29

though this is quite impractical, as finding proper distribution(s) manually may be exhaustive. In many applications, only Gaussian or other symmetric distribution is assumed (e.g., Liu et al., 2021).From the data point of view, the choice of the algorithm should depend on what subset of data we want to use for clustering. For internal housing characteristics, it can be sufficient to pick K-means or FCM. An example of such usage is Calka (2019). However, if we want to cluster on spatial location, we have to make a strong assumption that clusters are convex (because of the nature of Voronoi cells) and possess rather circular shapes than elongated ones. Novel approaches, such as swarm-intelligence-based clustering (such as Chen et al.,2021) and constraint-based clustering (such as Zhang et al., 2022), should be more deeply explored and compared with "classic" approaches. For these reasons, an "ideal" clustering algorithm cannot be easily chosen according to cluster shape criteria.

4.2. Fuzziness

In the real fuzzy world, there are usually no obvious borders separating the segments, but rather smoother transitions. For example, it is very difficult to say where the urban part ends and the suburban or rural part begins. Although there can be a strict administrative boundary, it might not hold true for the market segments (as mentioned in: Wu et al., (2018); Usman et al., (2020), and Shi et al., (2015)), and there can rather be a mixture of overlapping submarkets in some sense. For this purpose, the soft/fuzzy clustering techniques provide a solution to classify each data point to a cluster, with some weight of membership. As the fuzzy clustering allows multiple memberships for each data point, the overlapping of segments should be somehow evaluated (for example, overlapping measures introduced in Hwang and Thill (2009), or studies referenced in Usman et al. (2020)). Fuzzy clustering might have more sense for real-world applications, but on the other hand, the problem of fuzzy clustering might be validation, as we are not usually familiar with the true membership distribution (some approaches are discussed in Hwang and Thill (2009). The next issue might be in the availability of fuzzy clustering algorithms, as most algorithms (including the ones presented here, except FCM) only perform hard clustering.

4.3. Hierarchy

The hierarchy between clusters allows for displaying submarkets in the different levels of granularity, thus allowing the hidden structure of the market to be explored. Goodman and Thibodeau (1998) point out that there might be housing characteristics that involve the nested structure of the market segments (e.g., particular segments around some school districts might be considered as a single superior segment in the higher-level point of view).There can be also a problem with validation as the ground truth for the hierarchy is not usually available (for example, Liu et al., (2021) compares the obtained structure with an urban plan).

Table 1

Summary of clustering approaches

Method	Advantages	Disadvantages	Examples of algorithms
Centroid-based	• Simple, fast, and straightforward. • Can incorporate fuzziness.	• There is no natural way to detect outliers. • Restricted to data where centers (centroids) are easily distinguishable. • Cannot fit arbitrary (non-convex) cluster shapes.	• K-means • Fuzzy C-means
Density-based clustering	• Outlier detection as a side effect of clustering. • Number of clusters is determined by the algorithm. • Arbitrary cluster shapes.	• The determined number of clusters might not correspond to our expectations (e.g., too many clusters are determined).	• DBSCAN • OPTICS
Distribution-based	• Higher flexibility – model can be explicitly defined to incorporate problem specifics. • We can determine the	• Computationally demanding. • Mispecified model might lead to unexpected results.	• Expectation-Maximization • Variational Inference • Monte-Carlo simulation

	strongness of cluster membership via probability.		
Connectivity-based	• Detects hierarchical structure.	• Outlier detection is quite problematic.	• Ward's method
Population-based	• Promising for the future – simulation might somehow mimic the behavior of real estate markets more precisely.	• Currently rather experimental – no standardized and well-proven libraries available. • Usually, a population-based algorithm is tailored to solve a specific problem, i.e., it requires deeper knowledge to implement.	• Genetic algorithms (or Evolutionary in general) • Swarm-intelligence based algorithms (Swarm inspired projection, Ant/Bee Colony Optimization, …)
Affinity propagation	• Can detect the number of clusters on its own. • Arbitrary cluster shapes. • Prone to algorithm initialization.	• As in the case of density-based algorithms, the number of determined clusters might be higher than expected.	• AP is an algorithm itself
Constrained clustering	• Similarly, as distribution-based clustering, there can be some domain knowledge incorporated.	• The constraints might be market-specific; in that case, it requires constraints to be specified and validated for each market.	• Basically any unsupervised clustering algorithm + defined constraints

Source: own study.

5. Conclusion

In this article, several algorithms for clustering in housing market segmentation have been described. Although the lists of algorithms and referenced literature are not very exhaustive, the main clustering approaches have been presented. Unfortunately, there is still currently no universal method proven to work indifferent market conditions, but a gradual development of the methods can be observed, which is probably the effect of three things, i.e.: first, knowledge of the problem has deepened over time; second, the improvement in computational methods; and last, the rise of data availability. What is very common for modern segmentation approaches is that they try to distinguish between spatial proximity and attribute similarity right in the clustering algorithm (e.g. Liu et al., (2021), Liu et al., (2012) extended by Wu et al., (2018)), or at least process the two separately as Calka (2019). This distinction is understandable seeing as how there is no guarantee that both can be weighted equally, so this is probably the right way to go about designing a housing market segmentation algorithm nowadays.

Funding Sources

This work is supported by the grant Segmentation of residential real estate market for property valuation purposes - empirical evidence and statistical modeling (ÚSI-K-21-6920), which is realized within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948, and financed from the OP RDE.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

References

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

Calka, B. (2019). Estimating residential property values on the basis of clustering and geostatistics. *Geosciences, 9*(3), 143. https://doi.org/10.3390/geosciences9030143

Chen, J. H., Ji, T., Su, M. C., Wei, H. H., Azzizi, V. T., & Hsu, S. C. (2021). Swarm-inspired data-driven approach for housing market segmentation: A case study of Taipei city. *Journal of Housing and the Built Environment, 36*(4), 1787–1811. https://doi.org/10.1007/s10901-021-09824-1

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In*: Knowledge Discovery and Data Mining -96 Proceedings, 96*, 34, 226-231.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972-976.

Goodman, A. C., & Thibodeau, T. G. (1998). Housing market segmentation. *Journal of Housing Economics, 7*(2), 121–143. https://doi.org/10.1006/jhec.1998.0229

Guo, K., Wang, J., Shi, G., & Cao, X. (2012). Cluster analysis on city real estate market of China: Based on a new integrated method for time series clustering. *Procedia Computer Science, 9*, 1299–1305. https://doi.org/10.1016/j.procs.2012.04.142

Hu, L., He, S., & Su, S. (2022). A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. *Computers, Environment and Urban Systems, 94*, 101775. https://doi.org/10.1016/j.compenvurbsys.2022.101775

Hwang, S., & Thill, J. C. (2009). Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning. B, Planning & Design, 36*(5), 865–882. https://doi.org/10.1068/b34111t

Kryvobokov, M. (2013). Hedonic price model: defining neighbourhoods with Thiessen polygons. *International Journal of Housing Markets and Analysis, 6*(1), 79-97. https://doi.org 10.1108/17538271311306039

Liu, Q., Deng, M., Shi, Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences, 46*, 296–309. https://doi.org/10.1016/j.cageo.2011.12.017

Liu, Z., Cao, J., Xie, R., Yang, J., & Wang, Q. (2021). Modeling submarket effect for real estate hedonic valuation: A probabilistic approach. *IEEE Transactions on Knowledge and Data Engineering, 33*(7), 2943–2955. https://doi.org/10.1109/TKDE.2020.3010548

Liu, Z., Yan, S., Cao, J., Jin, T., Tang, J., Yang, J., & Wang, Q. (2018). A Bayesian approach to residential property valuation based on built environment and house characteristics. *In IEEE international conference on big data (big data)*. IEEE.

Malinowski, A., Piwowarczyk, M., Telec, Z., Trawiński, B., Kempa, O., & Lasota, T. (2018). An approach to property valuation based on market segmentation with crisp and fuzzy clustering. In *International Conference on Computational Collective Intelligence*, 534-548.Springer, Cham. https://doi.org/10.1007/978-3-319-98443-8_49

Manganelli, B., De Mare, G., & Nesticò, A. (2015). Using genetic algorithms in the housing market analysis. In Computational Science and Its Applications—ICCSA 2015: 15th International Conference, Banff, AB, Canada, June 22-25, 2015 [Springer International Publishing.]. *Proceedings, 15*(Part III 15), 36–45.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169

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. https://doi.org/10.1016/j.compenvurbsys.2012.04.005

Shi, D., Guan, J., Zurada, J., & Levitan, A. S. (2015). An innovative clustering approach to market segmentation for improved price prediction. *Journal of International Technology and Information Management, 24*(1), 2. https://doi.org/10.58729/1941-6679.1033

Talbi, E. G. (2009). Metaheuristics: from design to implementation. John Wiley & Sons. https://doi.org/10.1002/9780470496916

Usman, H., Lizam, M., & Adekunle, M. U. (2020). Property price modelling, market segmentation and submarket classifications: A review. *Real Estate Management and Valuation, 28*(3), 24–35. https://doi.org/10.1515/remav-2020-0021

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

Zhang, X., Zheng, Y., Ye, X., Peng, Q., Wang, W., & Li, S. (2022). Clustering with implicit constraints: A novel approach to housing market segmentation. *Transactions in GIS, 26*(2), 585–608. https://doi.org/10.1111/tgis.12878