www.degruyter.com/view/j/remav

IDENTIFYING HOUSING SUBMARKETS IN JOHOR BAHRU AND KULAI, MALAYSIA: A DATA-DRIVEN METHOD

Nur Asyikin Mohd Sairi
Faculty of Technology Management and Business
UniversitiTun Hussein Onn Malaysia, Malaysia
e-mail: asyikinsairi@gmail.com

Burhaida Burhan
Faculty of Technology Management and Business
UniversitiTun Hussein Onn Malaysia, Malaysia
e-mail:burhaida@uthm.edu.my

Edie Ezwan Mohd Safian
Faculty of Technology Management and Business
UniversitiTun Hussein Onn Malaysia, Malaysia
e-mail: edie@uthm.edu.my

Abstract

Generally, housing within the same area tends to share similar structural and locational characteristics as well as being characterized by the same homebuyers' preferences. This information provides insight on the existence of housing submarkets. Thus, this study aims to identify the existence of housing submarkets in Johor Bahru and Kulai, Malaysia through principal component analysis and cluster analysis. A total of 29 housing attributes were included in the study. The results from the principal component analysis have successfully identified four principal components that represent the basic facility, main infrastructure, less common facility and housing quality characteristics. On the other hand, cluster analysis managed to identify the existence of four housing submarkets in Johor Bahru and Kulai, where three submarkets showed the characteristics of housing quality, whereas one submarket appeared to exhibit the characteristic of main infrastructure. The findings of this study provide valuable information on the existence of four housing submarkets in Johor Bahru and Kulai areas, which will benefit the city governments, real estate developers, mortgage lenders, non-profit organizations, as well as individual homeowners in making critical choices and developing market-appropriate strategies.

Key words: *housing submarkets, home buyers' preferences, principal component analysis, cluster analysis*.

JEL Classification: *D49, R30*.

Citation: Mohd Sairi, N. A., Burhan, B., & Mohd Safian, E.E.(2022). Identifying housing submarkets in Johor Bahru and Kulai, Malaysia: a data-driven method. *Real Estate Management and Valuation, 30*(2), 01-11.

DOI: https://doi.org/10.2478/remav-2022-0009

1. Introduction

Residential housing is a multi-attribute commodity with numerous and diverse characteristics (Ligus & Peternek, 2017). Housing attributes, in general, include both structural attributes, such as the

number of bedrooms, and locational/neighborhood attributes, such as proximity to certain public amenities (Yu et al., 2007). Basically, the locational and structural characteristics of a property aid in determining the dimensions of the housing submarket as buyer subgroup preferences are based on their view of the locational and structural characteristics of the available housing units (Watkins, 2001). In fact, submarkets can be defined specifically based on residential characteristics (Song & Quercia, 2008).The previous arguments imply that structural and locational housing characteristics can be used to identify housing submarkets. To attain the research goal, the study makes use of the housing characteristics of housing units in Johor Bahru and Kulai.

The real estate market appears to be divided into submarkets, each with its own set of spatial and non-spatial characteristics (Kopczewska & Cwiakowski, 2021). Submarkets are described as groups that are homogeneous within themselves but heterogeneous in comparison to other groups (Wu & Sharma, 2012). Houses within the same submarket are substitutable (Gavu & Owusu-Ansah, 2019).In addition, housing submarkets are diverse types of housing in local areas of a metropolitan area that are comparable substitutes for potential buyers (Hwang,2015). The conventional housing submarket assumption is that properties in a given area are identical in several aspects and that residents in that location belong to the same social and economic community and have similar locational preferences (Bangura & Lee, 2020; Xiao et al., 2016).

Moreover, submarkets are operationalized in most applications by dividing the entire market into separate, different and distinct areas (Helbich et al., 2013). However, inside a submarket, there is a distinct market process because housing units from the same submarket are comparable, whereas housing units from different housing submarkets are not (Hwang, Thill 2009). Furthermore, residents value the specific attributes within a housing submarket, and developers accommodate to meet the demand (Hwang, 2015).

The principle of housing submarkets describes residential sorting in terms of different preference structures that determine whether the observed units are close substitutes or otherwise (Hwang, 2015). This indicates that housing preferences are similar within the same housing submarket because houses in the same submarket are more similar than houses in other submarkets. Thus, by identifying housing submarkets in Johor Bahru and Kulai, information on home owners' housing preferences can also be obtained. Furthermore, submarkets are groups of similar entities that are different from other groups (Wu & Sharma, 2012). Subdividing urban housing markets into submarkets is said to have three distinct advantages (Keskin & Watkins, 2017). First, it has been demonstrated that assigning housing units to submarkets, as the first step in the estimation procedure, improves the predictive accuracy of statistical models. Second, submarkets provide an effective framework for planners and policymakers to investigate dynamic change in the housing system. Lastly, understanding the submarket structures can assist a number of real estate market players in making better decisions (Keskin & Watkins, 2017; Leishman et al., 2013).

In addition, research on housing submarkets will aid in the study of urban issues such as residential segregation and the accumulation of wealth, as well as the evaluation of policies such as community revitalization and equal housing opportunities (Wu & Sharma, 2012). Similarly, Hwang and Thill (2009) claim that the study on housing submarkets contributes to a better understanding of the problem within the specific submarket areas, which is beneficial to tax assessment, community development, market area analysis, and formulating well-informed urban policies. Nevertheless, housing submarkets have been proposed as a method for assessing urban housing difficulties and modelling house prices (Hwang, 2015). In addition, understanding the submarket structure can facilitate a variety of housing-sector stakeholders in making more informed decisions (Leishman et al., 2013).

Therefore, the purpose of this research is to determine the existence of housing submarkets in the area of Johor Bahru and Kulai, Malaysia. The paper begins by providing an overview of the identification of housing submarkets. The next section clarifies the data and methodology used. Afterwards, a comprehensive description of the empirical results of the analysis is provided. The last section focuses on the discussion and conclusions drawn from the study.

2. Literature review – housing submarkets identification

In the current literature, the three key aspects of housing submarkets that are commonly accepted are substitutability, similarity, and spatial continuity (Wu & Sharma, 2012). High substitutability implies that the attributes of houses in the same housing submarket should contribute similarly to estimating

overall housing prices (Wu et al., 2019; Watkins, 2001). On the other hand, similarity refers to the similarity of housing structural attributes as well as the socioeconomic conditions of the local neighborhood (Watkins, 2001). Next, spatial continuity requires that a housing submarket occupy a continuous space and that submarkets have distinct geographic boundaries, which can be physical barriers such as administrative boundaries or invisible segments (Wu & Sharma, 2012; Wu et al., 2020).

Despite substantial research, there seems to be no strong agreement on how to better identify spatial submarkets for housing studies (Xiao et al., 2016; Leishman et al., 2013; Watkins, 2001). In addition to this, Keskin and Watkins (2017) claim that there is no strong conceptual justification for the existence of submarkets, despite the universal acceptance that submarkets exist and are analytically practical. Seemingly, the scarcity of well-regarded methods for identifying housing submarkets contributes to the lack of consensus on defining and delineating housing submarkets (Hwang &Thill, 2009).

According to Xiao et al. (2016), there are three common techniques for identifying submarkets which are based on fixed geographical areas, data-driven and lastly, through an *ad hoc* specification of spatial boundaries by experts. The first technique is based on the idea that the characteristics of buildings determine and reflect people's preferences on the supply side, with a willingness to pay for individual housing characteristics. On the other hand, the second technique allows for systematic statistical methods such as principal component analysis and cluster analysis to delineate housing submarkets. The third technique relies on the specifications of submarkets' boundaries based on the evaluations of experts such as valuers, market analysts and sales agents (Xiao et al., 2016).

Leishman et al. (2013), on the other hand, emphasized two basic techniques, the first of which is to use statistical methods such as principal components analysis and cluster analysis. The second approach is through market experts, such as valuers and estate agents, to define the segments (Leishman et al., 2013).These techniques are identical to those mentioned by Alas (2020) and Goodman and Thibodeau (2007). In addition, Goodman and Thibodeau (2007) mentioned that some researchers used principal component analysis and statistical clustering techniques to divide small geographic areas into housing submarkets, while others generated procedures that explicitly model submarket boundaries.

Besides, Goodman and Thibodeau (2007) added that some submarket construction techniques focus on supply-side determinants of house prices and construct submarkets using housing stock characteristics and neighborhood characteristics, such as the dwelling type and the quality of neighborhood schools. In contrast, other submarket construction techniques concentrate on demand-side determinants of house prices, forming housing submarkets based on household income or other socioeconomic and demographic characteristics (Goodman & Thibodeau, 2007).

On the other hand, Kopczewska and Cwiakowski (2021) and Wu et al. (2020) explain that the main approach to defining the housing submarkets is through division or administration boundaries where a housing unit is assumed to be nearly substitutable. Next, the housing submarkets are identified through a data-driven model that takes into consideration many characteristics using a quantitative model (Kopczewska & Cwiakowski, 2021; Wu et al., 2020).

A study conducted by Keskin and Watkins (2017) discussed three main steps in identifying housing submarkets. First, the data is divided to identify possible submarkets. Second, house price modeling techniques (primarily hedonic) are used to assess the price of standardized houses. Third, statistical techniques are applied to determine the existence of significant differences between the submarket-specific standardized price estimates (Keskin & Watkins, 2017).

Since housing submarkets are not observable theoretical constructs, data-driven methods that do not depend on the assessors' intuition and professional expertise seem to be a reasonable choice worth considering (Helbich et al., 2013). Moreover, the data-driven approach accounts for both substitutability and similarity (Wu et al. 2020). In fact, some of the most accepted data-driven approaches applied in delineating housing submarkets are principal component analysis and cluster analysis (Keskin & Watkins, 2017; Hwang,2015; Leishmanet et al., 2013; Wu & Sharma, 2012). Typically, housing submarkets are identified by either geographic locations or the physical characteristics of the houses (Bourassa et al., 2003). Furthermore, housing submarkets are defined not only by the similarity of their dwellings but also by their housing characteristics (Cox & Hurtubia, 2020). These housing characteristics are, in fact, used in data-driven methods.

Based on the facts stated above by previous scholars, it is clear that data-driven methods have

many advantages over other known methods. Thus, the data-driven methods of principal component analysis and cluster analysis were used in this research to identify housing submarkets in Johor Bahru and Kulai districts based on the generated housing characteristics.

3. Data and Methods

The analysis involves a combination of both principal component analysis and cluster analysis using the Statistical Package for the Social Sciences (SPSS) software. The principal component analysis was first conducted to minimize the number of variables in a data set while retaining as much information as possible. This technique also aids in the extraction of significant components that account for the majority of the variance in the data. Next, cluster analysis facilitates grouping the components into clusters that are homogeneous within themselves and heterogeneous between each other. Cluster analysis uses the component scores obtained from the principal component analysis.

Cluster analysis can be done in a number of methods, where the hierarchical clustering (agglomerative) and non-hierarchical clustering (k-means) methods are primarily used (Ebelinget et al., 2013; Antonenko et al., 2012). According to Saenz et al. (2011), before the results can be analyzed, the cluster approach requires two steps, i.e., first, deciding the best number of groupings inherent in the data and next, conducting the cluster analysis itself to assign each observation to its best-fit group. Consequently, this research utilized both hierarchical and non-hierarchical methods. In addition,to select the required number of clusters for the hierarchical approaches, agglomerative clustering using Ward's method was applied. After determining the number of clusters, k-means clustering was used to form the clusters in a non-hierarchical manner.

Expressly, this study used both methodologies to determine the optimal number of housing submarkets by taking into account both structural and locational characteristics of housing. A total of 43,610 latest housing data in Johor Bahru and Kulai, Malaysia were included in this research. Furthermore, this study examined 29 variables as mentioned in Table 1, including 8 structural characteristics and 21 locational characteristics of housing units in Johor Bahru and Kulai. The variables in this study were chosen based on a literature review and the availability of data as Ligus and Peternak (2017) emphasized that housing attribute selection is strongly limited by the availability of the data in practical application. The structural characteristics were derived from housing data provided by the Valuation and Property Services Department (JPPH). On the other hand, the locational characteristics were generated using a distance matrix analysis via a software called the Quantum Geographical Information System (QGIS) using the data obtained from the Town and Country Planning Department, Johor (JPBDJ). Both data were received from each department in April 2019 and December 2019, respectively.

Table 1

Configuration of variables

Variables name		
Structural characteristics	Locational characteristics	
Number of bedrooms	Distance to CBD	Distance to college
Floor level	Distance to shopping complex	Distance to university
Age of building	Distance to hypermarket	Distance to hospital
Property type	Distance recreational park	Distance to clinic
Land area	Distance to sports complex	Distance to police station
Building size	Distance to stadium	Distance to fire station
Condition of the property	Distance to golf course	Distance to city road
Types of construction	Distance to lake	Distance to highway
	Distance to kindergarten	Distance to airport
	Distance to primary school	Distance to bus terminal
	Distance to secondary school	

Source: own study.

4. Empirical results

4.1. Principal component analysis (PCA)

The Principal Component Analysis (PCA) was repeated three times until every variable (housing characteristics) included in the analysis fulfilled the following requirements:

1) The Kaiser-Meyer-Olkin Measure of Sampling Adequacy(KMO) value is 0.6 or above and the Bartlett's Test of Sphericity value is significant ($p < 0.05$).
2) The existence of some correlations greater than 0.30 between the variables.
3) The communality value of the variables exceeds 0.50.

The total number of variables was reduced from 29 to 17 because it did not meet the PCA requirements. During the first PCA, the variables of distance to a city road, distance to an airport, distance to a lake, distance to a kindergarten, number of bedrooms, property types, building condition and types of construction were excluded. Afterward, during the second PCA, more variables were excluded, including the distance to the police station, distance to the CBD, land area and building age. Each variable in the third PCA fulfilled all the requirements, hence no variables needed to be excluded.

In this analysis, the KMO value is 0.764 and Bartlett's test is significant ($p= 0.000$), therefore PCA is appropriate. Table 2 demonstrates the total variance explained by each component as well as the eigenvalues for each component. Only the first four components have eigenvalues greater than one (7.184, 2.748, 1.385, 1.349).

Table 2

Total variance explained

Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	7.184	42.256	42.256	7.184	42.256	42.256
2	2.748	16.164	58.420	2.748	16.164	58.420
3	1.385	8.150	66.570	1.385	8.150	66.570
4	1.349	7.937	74.507	1.349	7.937	74.507
5	0.843	4.961	79.467			
6	0.635	3.735	83.203			
7	0.591	3.476	86.679			
8	0.528	3.107	89.786			
9	0.485	2.854	92.640			
10	0.341	2.003	94.644			
11	0.245	1.443	96.087			
12	0.203	1.193	97.280			
13	0.185	1.088	98.368			
14	0.158	0.931	99.300			
15	0.076	0.447	99.747			
16	0.033	0.193	99.940			
17	0.010	0.060	100.00			

Source: own study.

Table 2 shows that the first component, with 42.25 percent of variation explained, accounted for the highest variance in the data set. Following that, the second component explains 16.16 percent of variance. The third component explains 8.15 percent of its variance, while the fourth component explains the lowest proportion, at 7.93 percent. These four components explained a total of 74.50 percent of the variance in the data. The result thus indicates that there are four components to be extracted for these variables.

After deciding on the number of factors to extract, the next step was to interpret the factor loadings of each of the variables. Factor loadings indicate how much a variable contributes to a specific component and how similar one variable is to others (LU et al. 2011). Table 3 displays the information on the factor loadings and is provided under the rotated component matrix table.

The rotated component matrix aids in determining what the components represent, which is the

key output of principal component analysis. The component matrix for the variables distance to a recreational park (0.831), distance to a hypermarket (0.812), distance to a shopping complex (0.781), distance to a sport complex (0.756), distance to a bus terminal (0.753), distance to a fire station (0.746) and distance to a secondary school (0.718) were all discovered to have a high loading for the first component.

Table 3

Rotated component matrix

	Component			
	1	2	3	4
Distance to Recreational Park (km)	0.831	004	0.068	-0.067
Distance to Hypermarket (km)	0.812	0.314	0.138	0.032
Distance to Shopping Complex (km)	0.781	-0.236	0.316	0.035
Distance to Sport Complex (km)	0.756	0.160	0.236	0.013
Distance to Bus Terminal (km)	0.753	0.302	0.244	-0.001
Distance to Fire Station (km)	0.746	0.355	0.247	-0.038
Distance to Secondary School (km)	0.718	0.300	0.447	-0.031
Distance to Stadium (km)	0.100	0.964	0.150	-0.052
Distance to University (km)	0.071	0.943	0.102	-0.059
Distance to College (km)	0.123	0.934	0.097	-0.056
Distance to Hospital (km)	0.491	0.669	0.134	-0.039
Distance to Highway (km)	0.527	0.618	-0.203	0.007
Distance to Golf Course (km)	0.140	0.101	0.832	-0.052
Distance to Primary School (km)	0.282	-0.039	0.738	0.029
Distance to Clinic (km)	0.287	0.214	0.649	0.045
Building Size	0.057	-0.018	-0.008	0.845
Floor Level	-0.083	-0.095	0.017	0.827

Source: own study.

Next, the variables of the distance to a stadium (0.964), distance to a university (0.943), distance to a college (0.934), distance to a hospital (0.669) and distance to a highway (0.618) were found to have high component 2 loadings. On the other hand, the distance to a golf course (0.832), the distance to a primary school (0.738) and the distance to a clinic (0.649) have a high loading for component 3.Only two variables, i.e. the building size (0.845) and floor level (0.827) load highly on the fourth component.Technically, a component represents whatever its variables have in common. Thus, an interpretation of the underlying traits measured by each variable is displayed in Table 4.

Table 4

Component description

PCA component	Description of component
1st component	Basic facilities Recreational parks, hypermarkets, shopping complexes, sport complexes, bus terminals, fire stations and secondary schools fall under the basic facilities provided within a neighborhood
2nd component	Main infrastructure Stadiums, universities, colleges, hospitals and highways represent the main infrastructure in a city
3rd component	Less common facility Golf course, clinic and primary school are among the less common facilities in a neighborhood
4th component	Housing quality A bigger building size and higher floor level indicate better housing quality

Source: own study.

Based on the PCA, four components were extracted, where the first component represents a basic facility, the second component represents the main infrastructure, the third component represents a less common facility, and lastly, the fourth component represents housing quality. As a result, after recognizing the representation of all four components, the component scores were saved to be used in the cluster analysis.

4.2. Cluster analysis (CA)

The effective technique for determining the number of clusters is to combine information from the agglomeration schedule and the dendrogram (Yim & Ramdeen, 2015). Both the agglomeration schedule and the dendrogram were obtained from the hierarchical clustering analysis. According to Yim and Ramdeen (2015), the agglomeration schedule helps in determining the point at which two clusters being combined are perceived too dissimilar to form a homogeneous group.

When there is a large difference between the coefficients of two consecutive stages, it indicates that the clusters being merged are becoming more heterogeneous and that it would be perfect to halt the clustering process (Yim & Ramdeen, 2015). The agglomeration schedule generated from this research is presented in Table 5.

Table 5

Agglomeration schedule

Stage	Cluster combined		Coefficients	Stage cluster first appears		Next stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	16	17	57.085	0	0	15
2	4	13	172.376	0	0	3
3	4	15	386.766	2	0	16
4	1	14	640.425	0	0	7
5	3	10	1005.315	0	0	8
6	2	8	1387.621	0	0	12
7	1	6	1780.097	4	0	9
8	3	11	2217.927	5	0	11
9	1	7	2666.673	7	0	12
10	5	9	3144.074	0	0	13
11	3	12	3825.696	8	0	14
12	1	2	4779.584	9	6	13
13	1	5	5817.358	12	10	14
14	1	3	7088.998	13	11	15
15	1	16	9393.792	14	1	16
16	1	4	15277.386	15	3	0

Source: own study.

As can be seen from Table 5, there is a sudden jump in the coefficient values from stage 13 to stage 14. The first noticeable increase is from stage 13 to stage 14, with a difference of 1,271.64 (7088.998-5817.358) in the coefficient value. This indicates that the optimum stopping point for merging the clusters is after stage 13. Thus, counting the remaining stages from the sudden jump (cluster 13 to cluster 16) leads to a 4 cluster solution.

Next, a straight vertical line needs to be drawn on the dendrogram to support the findings from the agglomeration schedule. The number of intersections between the cut-off line and the horizontal lines stipulates the number of clusters. The dendrogram generated from the hierarchical clustering analysis is presented in Figure 1.

According to the agglomeration schedule, there is a sudden jump from stage 13 to stage 14, which also specifies that it is best to stop the clustering after stage 13 and eliminate the last three stages (Stage 14, Stage 15 & Stage 16). The last three vertical lines in the dendrogram represent the last three stages in the agglomeration schedule. Hence, a straight vertical line needed to be drawn after the last three vertical lines in the dendrogram. The cut-off line crosses four horizontal lines (refer to Fig. 1). This would identify four clusters, one for each point where a branch intersects the line drawn.

The dendrogram, as mentioned above, helps to identify the number of clusters which were later

used for the k-means clustering. Therefore, based on the results of both the agglomeration schedule and the dendrogram, a total of four clusters were used for the k-means clustering. The results from the k-means clustering are displayed in Table 6.

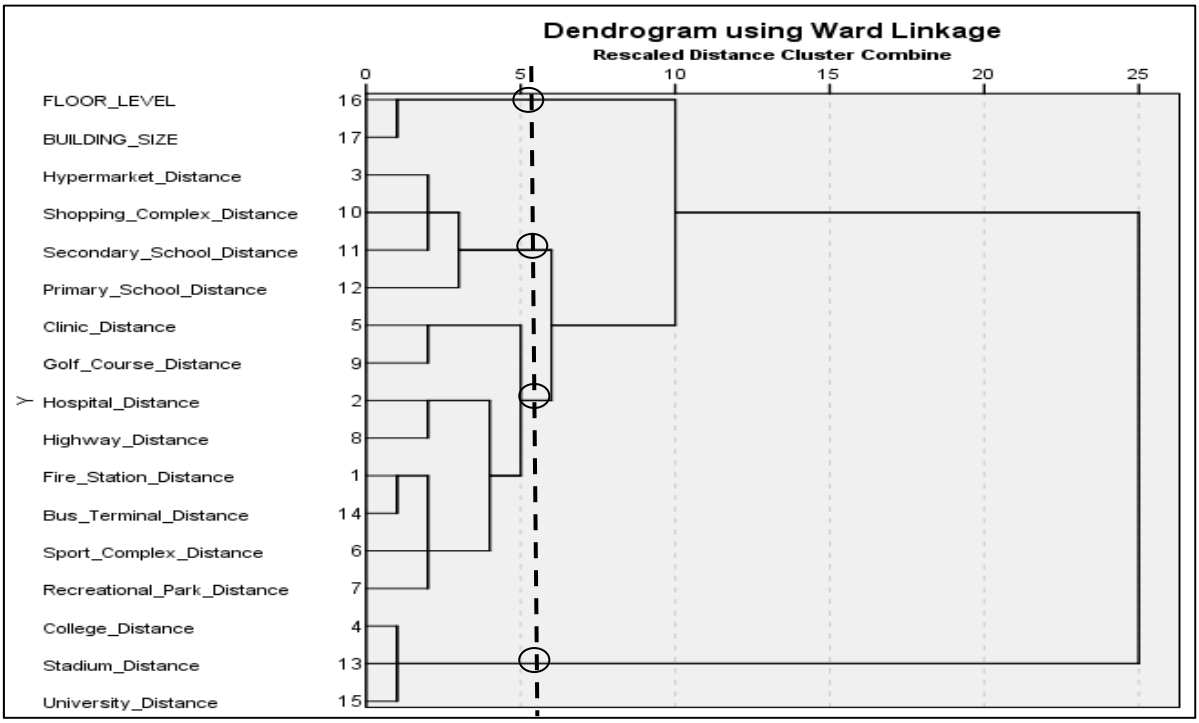


Fig. 1. Dendrogram of hierarchical cluster analysis. *Source*: own study.

Table 6

Final cluster centers

	Cluster			
	1	2	3	4
REGR factor score 1 for analysis 1	0.00126	-0.00144	-0.04381	1.11537
REGR factor score 2 for analysis 1	-0.21029	0.22878	0.06865	4.11659
REGR factor score 3 for analysis 1	-0.02253	0.02483	-0.11932	-1.74993
REGR factor score 4 for analysis 1	0.74235	-0.82246	11.43597	35.42280

Source: own study.

The final cluster centers, which contain information on the mean value for each component, reflect the characteristics of the typical case for each cluster. As presented in Table 6, Cluster 1 is very far from factor score 2 (-0.21029) and more similar to factor score 4 (0.74235). On the other hand, cluster 2 is similar to factor score 2 (0.22878) and very far from factor score 4 (-0.82246). Next, cluster 3 is very similar to factor score 4 (11.43597) and very far from factor score 3 (-0.11932). Lastly, cluster 4 is extremely similar to factor score 4 (35.42280) and very far from factor score 3 (-1.74993). Thus, Table 7 displays the components that are dominant for each cluster based on information gained from Table 6.

Table 7

Clusters' dominant component

Cluster	Dominant component
Cluster 1	4th component
Cluster 2	2nd component
Cluster 3	4th component
Cluster 4	4th component

Source: own study.

It is evident from Table 7 that the fourth component, which represents the housing quality, appeared to be dominant for three clusters, i.e. Cluster 1, Cluster 3 and Cluster 4. On the other hand, the second component, which indicates the main infrastructure, is dominant for Cluster 2.

Cluster analysis (CA) aids in the formation of homogeneous submarkets. Therefore, in the context of this study, the results from CA showed the existence of four housing submarkets in Johor Bahru and Kulai, where a total of three housing submarkets have the housing quality characteristics, whereas only one submarket has the main infrastructure characteristics. In other words, three submarkets in Johor Bahru and Kulai areas are characterized by housing quality, whereas one submarket is characterized by the main infrastructure.

As previously stated, within the same submarkets, home purchasers' preferences are comparable. The conclusion drawn from the study results is that homeowners in the three submarkets prefer to live in higher quality houses. They appear to favor the structural characteristics of the house more than the location. On the other hand, one submarket group expressed a preference for the house's locational qualities, notably owning a home that is surrounded by or adjacent to main infrastructure. Furthermore, the analyses revealed that structural characteristics are prominent in three submarkets, whereas locational characteristics are prominent in one submarket.

5. Discussion and conclusions

Generally, submarkets are groups of similar entities that are different from other groups. In addition, housings that belong to the same submarket tends to share similar housing characteristics as well as sharing the same buyers' housing preferences. This study has successfully identified the existence of four submarkets in Johor Bahru and Kulai, Malaysia through incorporating the PCA and CA. Through the PCA, four principal components that represent the basic facility, main infrastructure, less common facility and housing quality were identified. The significant factor scores are then used in the CA, which later generated four homogenous housing submarkets.

The fourth component (housing quality) appears to be dominant in three submarkets (Cluster 1, Cluster 3 and Cluster 4). This implies that the three submarkets have the characteristics of the fourth component (housing quality). On the other hand, one submarket (Cluster 2) appears to have the characteristics of the second component (main infrastructure). These findings signify that home owners in three submarkets prefer higher housing quality, whereas home owners in one remaining submarket prefer a residence that is closer to the main infrastructures.

Furthermore, the findings of this study suggest that the demand for a high-quality home and to reside in a neighborhood surrounded by main infrastructure typically formed the submarket structure in Johor Bahru and Kulai. Given that the identified submarkets are based on housing attributes, it is possible that the rising construction costs and technological advancements in the construction process will result in a long-term dynamic shift in the housing market. In fact, housing prices are expected to escalate over time. In addition, the housing preferences of homebuyers may change in the future, depending on the prevailing trend and their needs, which may also affect future housing market dynamics.

According to Cox and Hurtubia(2020), determining and characterizing housing submarkets is important because it provides information about the development of built space and the presence of spatial and structural housing characteristic segmentations in city expansion areas, which can be one of the factors causing fragmented urban sprawl and residential segregation. Therefore, the findings of this research provide insight into understanding the submarket structure in Johor Bahru and Kulai areas, which will benefit the city governments, real estate developers, mortgage lenders, non-profit organizations, as well as individual homeowners in making critical choices and developing market-appropriate strategies.

The identified housing submarkets also provide information on the preferences of home buyers. As a result, the aforementioned organizations as well as individual home buyers may benefit from these findings. Developers, for instance, can plan future residential areas by taking into account the characteristics of each submarket. Home buyers, on the other hand, could live in their dream home based on the identified submarkets that best suit their needs.

Despite the fact that the data-driven method employed in this research takes into account the locational attributes, the resulting submarkets remain spatially unstructured. According to Wu and Sharma (2012), data-driven approaches do not enforce spatial contiguity on submarkets, hence market

segments formed from them might not be localized. Moreover, the PCA and CA can be utilized to group similar houses into submarkets that may or may not constitute geographical boundaries (Bourassa et al., 2007). Hence, this research recommends further studies using the spatial approach to identify the housing submarkets in the Johor Bahru and Kulai area.

6. Acknowledgements

This research was supported by the Ministry of Higher Education (MOHE) through the Fundamental Research Grant Scheme (FRGS/1/2019/SS08/UTHM/02/1). In addition, the authors would like to thank the Department of Town and Country Planning Johor (JPBDJ) and the Valuation and Property Services Department (JPPH) for providing the data used in this research.

References

Alas, B. (2020). A multilevel analysis of housing submarkets defined by the municipal boundaries and by the street connections in the metropolitan area: Istanbul. *Journal of Housing and the Built Environment, 35*, 1201–1217. https://doi.org/10.1007/s10901-020-09735-7

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60*(3), 383–398. https://doi.org/10.1007/s11423-012-9235-8

Bangura, M., & Lee, C. L. (2020). House price diffusion of housing submarkets in Greater Sydney. *Housing Studies, 35*(6), 1110–1141. https://doi.org/10.1080/02673037.2019.1648772

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics, 35*, 143–160. https://doi.org/10.1007/s11146-007-9036-8

Bourassa, S. C., Hoesli, M., & Peng, V. S. (2003). Do housing submarkets really matters? *Journal of Housing Economics, 12*(1), 12–28. https://doi.org/10.1016/S1051-1377(03)00003-2

Cox, T., & Hurtubia, R. (2020). Subdividing the sprawl: Endogenous segmentation of housing submarkets in expansion areas of Santiago, Chile. *Environment and Planning. B, Urban Analytics and City Science, 29*, 355.

Ebeling, B., Vargas, C., & Hubo, S. (2013). Combined cluster analysis and principal component analysis to reduce data complexity for exhaust air purification. *The Open Food Science Journal, 7*(1), 8–22. https://doi.org/10.2174/1874256401307010008

Gavu, E. K., & Owusu-Ansah, A. (2019). Empirical analysis of residential submarket conceptualisation in Ghana. *International Journal of Housing Markets and Analysis, 12*(4), 763–787. https://doi.org/10.1108/IJHMA-10-2018-0080

Goodman, A. C., & Thibodeau, T. G. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics, 35*(2), 209–232. https://doi.org/10.1111/j.1540-6229.2007.00188.x

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

Hwang, S. (2015). Residential segregation, housing submarkets and spatial analysis: St. Louis and Cincinnati as a case study. *Housing Policy Debate, 25*(1), 91–115. https://doi.org/10.1080/10511482.2014.934703

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

Keskin, B., & Watkins, C. (2017). Defining spatial housing submarkets: Exploring the case for expert delineated boundaries. *Urban Studies (Edinburgh, Scotland), 54*(6), 1446–1462. https://doi.org/10.1177/0042098015620351

Kopczewska, K., Cwiakowski P. 2021. Spatio-temporal stability of housing submarkets. Tracking spatial location of clusters of geographically weighted regression estimates of price determinants. *Land Use Policy, vol. 103*(2), pp. 1-18.

Leishman, C., Costello, G., Rowley, S., & Watkins, C. (2013). The predictive performance of multilevel models of housing sub-markets: A comparative analysis. *Urban Studies (Edinburgh, Scotland), 50*(6), 1201–1220. https://doi.org/10.1177/0042098012466603

Ligus, M., Peternek, P. 2017. Impacts of urban environmental attributes on residential housing prices in Warsaw (Poland): Spatial hedonic analysis of city districts. *Procedia - Social and Behavioral Sciences I, 220*(251), 155-164.

Lu, W. Z., He, H. D., & Dong, L. (2011). Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. *Building and Environment, 46*(3), 577–583. https://doi.org/10.1016/j.buildenv.2010.09.004

Saenz, V. B., Hatch, D., Bukoski, B. E., Kim, S., Lee, K., & Valdez, P. (2011). Community college student engagement patterns. *Community College Review, 39*(3), 235–267. https://doi.org/10.1177/0091552111416643

Song, Y., & Quercia, R. G. (2008). How are neighbourhood design features valued across different neighbourhood types? *Journal of Housing and the Built Environment, 23*(4), 297–316. https://doi.org/10.1007/s10901-008-9122-0

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

Wu, C., & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. *Applied Geography (Sevenoaks, England), 32*(2), 746–756. https://doi.org/10.1016/j.apgeog.2011.08.011

Wu, Y., Wei, Y. D., & Li, H. (2020). Analyzing spatial heterogeneity of housing prices using large datasets. *Applied Spatial Analysis and Policy, 13*(1), 223–256. https://doi.org/10.1007/s12061-019-09301-x

Xiao, Y., Webster, C., & Orford, S. (2016). Can street segments indexed for accessibility form the basis for housing submarket delineation? *Housing Studies, 31*(7), 829–851. https://doi.org/10.1080/02673037.2016.1150433

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology, 11*(1), 8–21. https://doi.org/10.20982/tqmp.11.1.p008

Yu, D., Wei, Y. D., & Wu, C. (2007). Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning. B, Planning & Design, 34*(6), 1085–1102. https://doi.org/10.1068/b32119