Sosyoekonomi

2021, Vol. 29(49), 11-32

RESEARCH ARTICLE

ISSN: 1305-5577
DOI: 10.17233/sosyoekonomi.2021.03.01
Date Submitted: 14.01.2021
Date Revised: 26.04.2021
Date Accepted: 02.05.2021

Cluster Analysis for Housing Market Segmentation[1]

Shihomi ARA-AKSOY *(https://orcid.org/0000-0003-3424-2561), Department of Economics, Hacettepe University, Turkey; e-mail: sara@hacettepe.edu.tr*

Elena IRWIN *(https://orcid.org/0000-0001-7908-5964), Department of Agricultural, Environmental and Development Economics, The Ohio State University, USA; e-mail: irwin.78@osu.edu*

Konut Piyasası Bölümlendirmesinde Kümelenme Analizi[2]

Abstract

Cluster analysis is often used to determine housing submarkets. However, commonly used methods cannot handle mixed-mode data when variables of different types and units are combined. We propose new similarity measures that handle both continuous and categorical variables using normalization and discretization steps and partial match criteria. These measures are used in agglomerative hierarchical clustering with a formulation where the optimal number of clusters is automatically determined without a priori information regarding the number of submarkets. The experiments using housing sales data show that the proposed measures perform better than the commonly used standardized Euclidean distance in identifying submarkets.

Keywords : Housing Market Segmentation, Hierarchical Clustering, Mixed-Mode Data, Hedonic Price Model.

JEL Classification Codes : R21, R31.

Öz

Kümeleme analizi, konutların bir dizi değişkene dayalı olarak benzerliklerine göre gruplandırıldığı alt pazarları belirlemek için kullanılan popüler bir yöntemdir. Ancak, yaygın olarak kullanılan yöntemler, farklı tür ve birimlerdeki değişkenlerin bir arada kullanıldığı verileri doğrudan işleyemez. Bu çalışmada, düzgeleme ve ayrıklaştırma adımlarını ve kısmî eşleşme kriterlerini kullanarak hem sürekli hem de kategorik değişkenleri aynı çerçevede ele alabilen yeni benzerlik ölçümleri öneriyoruz. Bu ölçümler, alt pazarların sayısına ilişkin ön bilgi olmadan optimum küme sayısının otomatik olarak belirlendiği bir formülasyon ile aglomeratif hiyerarşik kümelemede kullanılmaktadır. Konut satış verilerini kullanan deneylerde, önerilen benzerlik ölçümleri, alt pazarların belirlenmesinde yaygın olarak kullanılan standartlaştırılmış Öklid mesafesinden daha iyi performans göstermektedir.

Anahtar Sözcükler : Konut Piyasası Bölümlendirmesi, Hiyerarşik Kümeleme, Karışık Veri, Hedonic Fiyat Modeli.

---

[1] *This article was produced based on the Ph.D. Dissertation titled "The Influence of Water Quality on the Demand for Residential Development around Lake Erie", submitted to the Department of Agricultural, Environmental and Development Economics, The Ohio State University. This paper is based upon work supported by the National Science Foundation Grant DEB-0410336 and the National Atmospheric and Oceanic Administration Ohio Sea Grant Program.*

[2] *Bu makale Ohio Eyalet Üniversitesi, Tarım, Çevre ve Kalkınma Ekonomisi Bölümü'ne sunulan "Su Kalitesinin Erie Gölü Çevresinde Konut Geliştirme Talebi Üzerindeki Etkisi" başlıklı doktora tezine dayanılarak hazırlanmıştır.*

1. Introduction

The determination of housing submarkets is critical for statistical inference, prediction and structural estimation of second-stage hedonic models that rely on variation across submarkets to identify a household's demand for a particular locational good or service. While traditional approaches are largely adhoc, relying on pre-existing geographic boundaries or structural features to define boundaries, recent data-driven approaches seek to use the underlying structure of the raw data to uncover groups of homogeneous observations. Two primary methods have been advanced in the literature: spatial statistical modelling, in which the spatial autocorrelation among hedonic residuals is used to group houses (e.g. Gillen et al., 2001; Tu et al., 2007), and classification methods that group observations based on measures of similarity (e.g. Goetzmann & Wachter, 1995; Maclennan & Tu, 1996; Hoesli et al., 1997; Bourassa et al.,1999; Day, 2003; Kim & Park, 2005; Clapp & Wang, 2006; Bates, 2006). These data-driven methods are a significant improvement over traditional adhoc approaches but are also subject to their own limitations. For example, methods relying on spatial autocorrelations use residuals from a hedonic model in which price is regressed on a set of structural variables to reveal similarity in neighbourhood characteristics. The maintained assumption is that unobserved spatial autocorrelations are not spurious and instead reflect systematic variations in neighbourhood structure. On the other hand, classification methods rely on key assumptions about the input variables used to cluster observations and the specification of similarity measures used to gauge homogeneous grouping. This approach relies on observed features to group observations and thus unmeasured neighbourhood-specific effects are omitted.

Because there is not a proven "best" data-driven method for identifying housing submarkets, the preferred method depends on the purpose of submarket determination as well as the characteristics of the study areas. For example, in regions in which local public goods are highly spatially heterogeneous and permit more extensive household sorting across local areas, it is reasonable to expect that housing submarkets can be estimated using observed variables that capture the primary variations in local public goods and services across the region. In such cases, a classification method that uses a parsimonious set of observed variables may be preferred as it avoids assumptions about the structure of the unobserved variation. Alternatively, in regions in which public goods do not vary widely across neighbourhoods or household sorting is limited for other reasons, identification of housing submarkets using observed variation may not be feasible and thus, an approach that backs out this information using spatially autocorrelated hedonic residuals may be more appropriate. In many settings in the U.S., public goods vary widely within a region due to the many forms of local government (e.g., school districts, townships, municipalities) that are present within a single region. Thus we argue that in such cases, a classification method that employs a set of variables to identify variation in local public goods and services and other major location features, such as location relative to the central city, is the preferred method for estimated housing submarkets in this case.

Cluster analysis is perhaps the most common of the classification methods and has become a popular method for determining submarkets (e.g., Goetzmann & Wachter, 1995; Hoesli et al., 1997; Bourassa et al., 1999; Day, 2003; Kim & Park, 2005; Bates, 2006). While a variety of clustering techniques exist, all are built on a similarity measure that is used to quantify the strength of similarity between observations and cluster observations into homogeneous groups. The choice of a similarity measure depends on the types (continuous, binary, or categorical) of variables and characteristics (if categorical, ordered or not; if continuous, units) of variables. The quantification of similarity is made more complicated if the set of variables contains both categorical and continuous variables (so-called mixed-mode variables) and variables with vastly different units, since ordinal differences in the values of these variables are not comparable across variables. Binary and categorical variables are common in housing submarket studies, e.g., political jurisdiction, construction type, existence of a fireplace or air-conditioning are all commonly used variables, and thus the treatment of mixed-mode variables is an important methodological consideration in these studies.

Despite the increasing application of cluster methods to identifying housing submarkets, many existing studies do not provide a clear rationale for the specification of the similarity measure. In addition, the literature has failed to adequately consider the complications that arise in combining disparate types of data that, if not done carefully, can introduce substantial bias into the cluster estimation. For example, existing studies treat both categorical and continuous variables in the same manner by using a standardized Euclidean distance measure for both types of variables. However, neither the rationale for doing so nor the potential bias that may be introduced if this approach is inappropriate has been clearly articulated in the literature.

This paper considers the methodological issues involved in developing and applying a similarity measure for mixed-mode variables used in a cluster analysis to identify housing submarkets. The study region is comprised of 10,655 observations on housing sales between 1985 and 1996 from four adjacent counties arrayed along the Lake Erie coast in Ohio, USA. These data include both continuous and categorical structural variables that are of importance to the estimation. We propose similarity measures that can treat both continuous and categorical variables together while minimizing the loss of the relative similarity information. These measures are applied to the data and used with a clustering technique, the agglomerative hierarchical method, in which the optimal number of clusters is automatically determined without a priori information regarding the number of submarkets. A comparison of the goodness of fit of the clusters generated using different similarity measures shows that the proposed similarity measures generate clusters that are significantly better in fit and substantially different in structure in comparison to the commonly used measure that is based on standardized Euclidean distance.

2. Cluster Analysis Methods and Applications to Housing Submarkets

A popular cluster analysis algorithm used to iteratively group and regroup observations is the k-means algorithm. This method is initialized by assigning k cluster centroids and then proceeds to partition all observations with respect to their similarities to these centroids. The similarities are computed using the Euclidean distance between the attributes of the observations and the cluster centroids. The centroids are iteratively updated using the mean values of the observations assigned to each cluster, and the observations are regrouped according to their distances to the new cluster means. This procedure minimizes the sum-of-squared errors between the observations and the clusters that are represented using their centroids. The main drawbacks of the k-means algorithm are the requirement of a priori knowledge of the number of clusters (i.e., the number of submarkets) and the random initialization of the cluster centroids that may give different results in each run. An alternative approach is the agglomerative hierarchical clustering method, which produces multi-level groupings by initializing a separate cluster corresponding to each observation and then, at each successive iteration, merges two clusters that are determined to be the most similar.

Existing studies typically include a single variable for the clustering analysis. For example, Goetzmann and Wachter (1995) use effective rents and vacancy rates of the U.S. office market in two separate clustering studies with the k-means algorithm and find the clusters characterized with bicoastal association and oil-related cities. Kim and Park (2005) conduct two cluster analyses, one with housing sales price and the other with housing price changes in Seoul and its five satellite towns. They find that the clusters generated from two variables were very different regardless of the absolute value of housing prices. Hoesli et al. (1997) include property returns in their clustering of UK commercial properties. Using two similarity measures (squared correlation coefficient and Euclidean distance) and both the hierarchical and k-means algorithms, they find that the dominant factor in their clustering is property type.

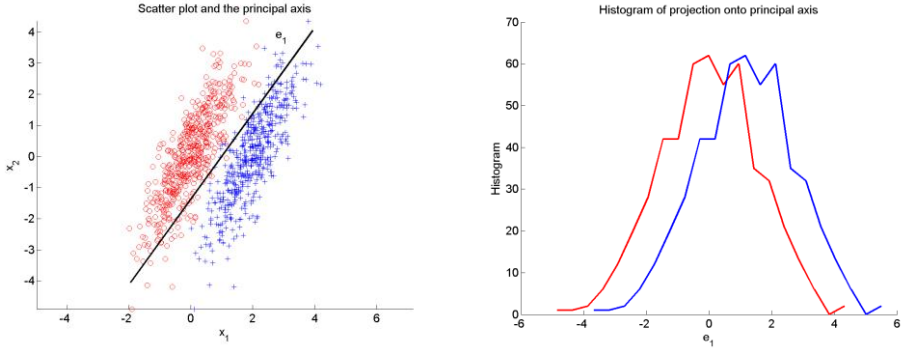
2.1. Variable Aggregation: Principal Component Analysis

When multiple variables are used in the cluster analysis, some means of aggregating the variables for comparison purposes is needed. Many studies in the housing market segmentation literature use principal components analysis (PCA) (Bishop, 2006) to reduce the original set of variables to a smaller set of factors (Maclennan & Tu, 1996; Bourassa et al., 1999; Day, 2003; Bates, 2006). PCA uses the eigenvectors of the covariance matrix of the data to obtain a new set of bases. The projection of the data onto a subset of the eigenvectors corresponding to the largest eigenvalues provides a data representation and summarization that are optimal in the least-squares sense. This projection corresponds to a linear transformation where the new representation (variables) is obtained as a linear combination of the existing variables. A potential problem with the use of PCA with housing data is the ambiguity of a linear combination of mixed-mode (continuous and categorical) variables. For example, how does one reasonably compute a weighted combination of

proximity to city centre (continuous), number of floors (categorical) and presence of a fireplace (binary)? Another important problem is that PCA ignores any information regarding the separability of the clusters (submarkets) as it only tries to find the linear projection that maximizes the variance of the projected data (Bishop, 2006). Therefore, the transformation to the new subspace may result in wrong or imprecise separation of the submarkets, as illustrated in Figure 1. If the principal components and the corresponding new variables are defined in such a potentially misleading manner, the cluster analysis performed using these variables may produce clusters that do not reveal the underlying group structure of observations.

Unfortunately, many studies that use cluster analysis with multiple variables do not clearly discuss their choice of the clustering algorithm and the similarity measure. In cases in which a discussion of how the similarity between observations is evaluated, it is most likely that either the Euclidean distance or the standardized Euclidean distance - the default settings of most clustering software packages - are used. However, as we demonstrate in the next section, even standardized Euclidean distance is not appropriate in the case of mixed-mode variables.

Figure: 1
Illustration of PCA for Two-Dimensional Synthetic Data with Two Known Clusters



The principal component resulting from the application of PCA on the whole data set (collection of blue + and red points ○) is shown as the black line (left figure). When all points are projected onto this line, the separability of the points from two clusters actually decreases as shown in the histograms (right figure) even though the clusters are quite separable in the original two-dimensional space.

2.2. Similarity Measures: Euclidean Distance

Similarity between two observations can be measured using distance metrics in the attribute space in which an increasing order of distance is assumed to correspond to a decreasing order of similarity. In such an approach, the Euclidean distance implicitly assigns more weighting to variables with large ranges than those with small ranges. Furthermore, adding the difference of one variable to the one from another variable is ambiguous and unintuitive when the variables have different units. To overcome the latter problem,

standardized Euclidean distance $\tilde{x}$ is calculated either by using linear scaling to unit variance,

$$\tilde{x} = \frac{x-\mu}{\sigma} \tag{1}$$

where μ and σ are the sample mean and the sample standard deviation of x, respectively, and $\tilde{x}$ is the normalized value, or by using linear scaling to unit range as

$$\tilde{x} = \frac{x-l}{u-l} \tag{2}$$

where l and u are the lower and upper bounds (minimum and maximum values) of x respectively.

This normalization procedure transforms the variables so that they are unitless and approximately equalizes the ranges of the variables and make them have approximately the same effect in the computation of similarity (Aksoy, 2001). However, standardization does not solve the ambiguity that arises in combining continuous and categorical variables when the Euclidean distance is used because normalization of categorical or binary variables using these methods can produce undefined (or impossible) values. In addition, standardization by linear scaling to unit variance implicitly assumes a normal distribution, an assumption that may or may not be appropriate depending on the data.

To our knowledge, there is no existing housing market segmentation study which treats multiple variables with different characteristics by taking into account the limitation of the use of (standardized) Euclidean distance as discussed above.

2.3. Clustering Algorithms: Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is one of the most widely used algorithms in the area of market segmentation. Hierarchical clustering tries to capture multi-level groupings using hierarchical representations rather than flat partitions as used by the k-means clustering algorithm. One of the main advantages of hierarchical clustering over the k-means algorithm is that it does not require a priori knowledge of the number of clusters. Furthermore, its input consists of the pairwise similarity values of all observations, so it can use any similarity measure, including the ones proposed in this paper, unlike the fixed Euclidean distance in k-means.

The agglomerative hierarchical clustering algorithm produces a dendrogram where each level corresponds to grouping into a different number of clusters. Boberg and Salakoski (1993) propose a method to automatically determine the optimal number of clusters. This method defines a self-similarity measure

$$h(C) = \max\{1 - s_{ij} \mid i, j \in C\} \tag{3}$$

that quantifies the dissimilarity between the observations of the same cluster C (Everitt et al., 2001). Then, the optimal number of clusters is found at level t when the following condition is met:

$$\exists C_n \in R_{t+1} : h(C_n) > \theta \tag{4}$$

where C_n is the n'th cluster in the R_{t+1} level with $t+1$ clusters and θ is a threshold. The threshold θ can be defined as

$$\theta = \mu + \alpha\sigma \tag{5}$$

where μ is the average and σ is the standard deviation of dissimilarity between any pairs among all observations and α is a user-defined parameter. The threshold can be selected by utilizing the distribution of h where a significant h value is defined as the one that lies in the upper tail of this distribution. The α parameter in equation (5) is selected to define a 5% significance level for h using a Gaussian distribution assumption in the experiments in Section 5. Given this parameter, when the condition in equation (4) is met at the R_{t+1} level, the t clusters at level R_t are used as the defined clusters.

3. Developing Similarity Measures for Mixed-mode Variables

We seek a method for constructing a similarity measure for mixed-mode data that permits aggregation of different types, scales and distributions of data such that the resulting measure of similarity is unbiased. One possibility is to dichotomize all variables and use a similarity measure for binary data (Everitt et al., 2001). However, dichotomization causes a significant loss of information. Instead, we adopt Gower's (1971) definition of similarity that provides a more general measure as

$$s_{ij} = \frac{1}{K}\sum_{k=1}^{K} s_{ijk}^{m} \tag{6}$$

where s_{ijk}^{m} is the similarity between the i'th and j'th observations according to the k'th variable and K is the number of variables. The superscript m represents the similarity measure used. The individual similarities s_{ijk}^{m} can be defined differently for different types of variables as long as the resulting value is between zero and one so that the combined similarity measure s_{ij} is also in the [0,1] range.

We adapt the so-called Hamming distance for categorical variables. This similarity measure is defined over [0,1] as:

$$s_{ijk}^{1} = \begin{cases} 1, \text{ if exact match} \\ 0, \text{ otherwise} \end{cases} \tag{7}$$

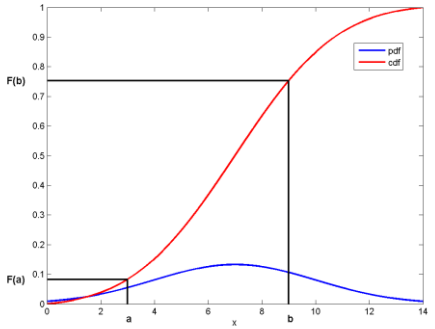
where s_{ijk}^{1} is set to one or zero depending on whether the two observations i and j are the same or different, respectively, on the k'th variable.

In handling continuous variables, we first normalize the values using an approach that does not rely on any distributional assumptions and then discretize the data by assigning a certain number of bins using uniform quantization. The non-parametric normalization step described below, and the following quantization of the normalized values correspond to equal-frequency discretization that is shown to yield good results compared to methods that make parametric density assumptions about the attribute values (Witten & Frank, 2005). Furthermore, the resulting ordinal data allow the use of well-defined similarity measures for categorical variables and a straightforward combination of the similarities from all variables in equation (6).

Our first step in handling the continuous data is to normalize these variables using a method based on the cumulative distribution function (CDF), which can be estimated from the data in a non-parametric way with no density assumption requirement. Given a random variable x with a cumulative distribution function $F_x(x)$, the random variable $\tilde{x}$ resulting from the transformation $\tilde{x} = F_x(x)$ is uniformly distributed in the [0,1] range. The concept of this transformation can be visualized in Figure 2. Even though this transformation modifies the distribution of the original values, the motivation behind transforming the variable to have a uniform distribution in the [0,1] range is to make the values spread as much as possible in that range so that the discrimination ability of that attribute is increased.

The choice for the uniform distribution as a target for the transformed range comes from the fact that the uniform distribution on an interval is the maximum entropy distribution among all continuous distributions which are supported in that interval. Entropy is the amount of information contained in a random variable. An ideal attribute for identifying the similarity between observations is the one that has different values for different observations and similar values for similar observations. If there is no prior information about the distribution of the similarity, it is important to select attributes with lots of variation among items in order to distinguish different items better. For example, in order to define dissimilarity among multiple people, the attribute gender gives very little information about distinguishing one from the other. This kind of variable with very similar values for the items has very low entropy. On the other hand, the attributes such as height, weight, and age have higher entropy. Having maximum entropy is important because it ensures to describe the differences between observations as much as possible. If the range of the value $[a, b]$ is the only information given, the uniform distribution is the one that has the maximum entropy (Theodoridis & Koutroumbas, 2006). Furthermore, we observed that most of our variables do not have a normal distribution. Given the fact that the CDF transformation does not assume any distributional form, this transformation is preferred over the normalization in equation (1) used in the standardized Euclidean distance.

Figure: 2
Illustration of the CDF Transformation



The blue curve is the probability density function. The red curve is the corresponding cumulative distribution. The points a and b are mapped to F(a) and F(b), respectively, using this transformation.

The second step in the handling of the continuous data is to discretize the variables into a certain number of bins using uniform quantization. We experiment with three definitions of the similarity measure that differ in how the discretization (or binning as it is also called) is performed. The first is the application of the strict measure of Hamming distance, in which s_{ijk}^1 is set to one or zero for an exact match or no match, respectively. However, this clearly ignores the fact that those observations that have values in neighbouring bins are more similar to each other than the ones with values in more distant bins and thus the strict application of Hamming distance will result in a substantial loss of information. By giving a partial score to the "not an exact match, but close" case, we attempt to include more information regarding the similarity between two observations from a continuous variable which otherwise could have been lost in the simple match-mismatch setting. We define two alternative measures that incorporate partial matches as

$$s_{ijk}^2 = \begin{cases} 1, \text{ if exact match} \\ 0.5, \text{ if one-mismatch} \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

in which the one-mismatch is applied to the cases where the attribute of two observations fall into neighbouring bins, or

$$s_{ijk}^3 = \begin{cases} 1, \text{ if exact match} \\ 0.6, \text{ if one-mismatch} \\ 0.3, \text{ if two-mismatch} \\ 0, \text{ otherwise} \end{cases} \tag{9}$$

in which the two-mismatch is counted when observation i is found to be two bins away from where observation j is. The similarity measures s_{ijk}^2 and s_{ijk}^3 defined in equations (8) and (9) are referred to as modified Hamming 1 and modified Hamming 2, respectively, in the rest

of the paper. An example case with five bins using modified Hamming 2 is illustrated in Table 1. These measures are designed to reduce the sensitivity to the exact bin locations used in the quantization. Together with the non-parametric normalization step, the resulting equal-frequency discretization process is proposed as an alternative to the Euclidean or standardized Euclidean distance-based similarity computation that implicitly uses a Gaussian density assumption for the attribute values.

Finally, because the exact geographic location is a consideration in defining housing submarkets, we define a similarity measure for these variables as

$$s_{ijk}^{4} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k} \tag{10}$$

where r_k is the range of observations for the k'th variable, defined for continuous variables in (Everitt et al., 2001). This measure uses the normalization by linear scaling to unit range as in equation (2). The CDF normalization is not used for these two variables because the absolute difference in geographical location between two observations matters.

Table: 1
Match Scores for the Modified Hamming 2 Measure

		Observation *i*				
	Bin	I	II	III	IV	V
Observation *j*	I	1	0.6	0.3	0	0
	II	0.6	1	0.6	0.3	0
	III	0.3	0.6	1	0.6	0.3
	IV	0	0.3	0.6	1	0.6
	V	0	0	0.3	0.6	1

For example, for the attribute ***k****, if observation* ***i****'s value is found in bin III, and observation* ***j*** *is in bin V, the score for two-mismatch (0.3) is applied.*

4. Comparison and Evaluation of Methods

Given the similarity measures for categorical and continuous data defined by $s_{ijk}^{1},\ldots, s_{ijk}^{4}$, we compute the overall similarity measure defined in (6) using s^1 for categorical variables, using one of s^1, s^2 or s^3 for continuous variables, and using s^4 for geographical locations. This yields three different overall similarity measures that differ by the binning procedure used to compute the similarity of the continuous variables (s^1, s^2 or s^3). We then use the agglomerative hierarchical clustering algorithm to perform the cluster analysis for each of these three similarity measures. The average linkage criterion is used as the inter-cluster similarity measure in which the distance between two clusters is measured as the average of the distance between all pairs of observations that are made up of one observation from each cluster (Everitt et al., 2001). This criterion is chosen because it is relatively more robust than other criteria such as the Ward's method, that is commonly used in the housing market segmentation literature, which tends to find same size, spherically shaped clusters, uses the squared error, and is sensitive to outliers (Everitt et al., 2001). The number of clusters is automatically determined using the procedure described in Section 2.3.

To determine which similarity measure fits the data best, we adopt the method used by Bourassa et al. (1999) and compute the weighted mean squared error (WMSE) based on the estimated hedonic pricing equation to compare each clustering outcome. Given the hedonic model estimates, the WMSE for a particular clustering outcome is calculated as follows:

$$\text{WMSE} = \frac{\sum_{i=1}^{t}(n_i - m - 1)\text{SE}_i^2}{\sum_{i=1}^{t}(n_i - m - 1)} \tag{11}$$

where n_i is the number of observations in the i'th cluster, t is the number of clusters, m is the number of independent variables in the hedonic equation, and SE_i^2 is the variance based on the estimation of hedonic equation for the i'th cluster (Bourassa et al., 1999). The smaller the WMSE value, the better the clustering method. However, as stated in Bourassa et al. (1999), the WMSE value decreases as the number of clusters increases. Therefore, we use this value to compare clustering outcomes for the same number of clusters.

We also compare the results from different methods using the adjusted Rand index (Everitt et al., 2001). This index measures the agreement between two cluster structures according to the proportion of pairs of observations that agree in the cluster labels. Agreement occurs if two observations that belong to the same cluster in one method are put into the same cluster by the other method, or two observations that belong to different clusters in one method are also put into different clusters by the other method. The index has the advantage of allowing the comparison of two sets of clustering outcomes with different numbers of clusters. It is computed as

$$\text{ARI} = \frac{\sum_{u=1}^{U}\sum_{v=1}^{V}\binom{n_{uv}}{2} - \left[\sum_{u=1}^{U}\binom{n_{u\cdot}}{2}\sum_{v=1}^{V}\binom{n_{\cdot v}}{2}\right]/\binom{N}{2}}{\left[\sum_{u=1}^{U}\binom{n_{u\cdot}}{2} + \sum_{v=1}^{V}\binom{n_{\cdot v}}{2}\right]/2 - \left[\sum_{u=1}^{U}\binom{n_{u\cdot}}{2}\sum_{v=1}^{V}\binom{n_{\cdot v}}{2}\right]/\binom{N}{2}} \tag{12}$$

where n_{uv} denote the number of observations in cluster u of the first method and cluster v of the second method, $n_{u\cdot} = \sum_{v=1}^{V} n_{uv}$ denote the number of observations in cluster u of the first method, $n_{\cdot v} = \sum_{u=1}^{U} n_{uv}$ denote the number of observations in cluster v of the second method, U is the number of clusters obtained by the first method, V is the number of clusters obtained by the second method, and N is the total number of observations. The index has a maximum value of one, meaning perfect agreement between two clustering outputs, and an expected value of zero (Milligan & Cooper, 1986; Everitt et al., 2001). We use the adjusted Rand index to quantify the significance of the difference between the outcomes of different clustering methods with small WMSE values and also use it to compare the results of the proposed similarity measures to those of the commonly used standardized Euclidean distance.

Finally, we use a Chow test to derive housing market submarkets from the neighbourhoods produced by the clustering analysis. As Galster (2003) and Tu et al. (2007) point out, a neighbourhood is defined as a cluster of housing units which share the similar set of housing attributes, including housing structures, neighbourhoods, political and environmental characteristics. A neighbourhood acts as a building block of a submarket and

thus a submarket may be composed of one or more neighbourhoods. Thus, a method of aggregating similar neighbourhoods into one submarket is necessary. Following others in the literature, (e.g. Bourassa et al., 1999; Day, 2003; Goodman & Thibodeau, 2003) we use a Chow test. However, unlike these previous studies, we apply a spatial Chow test that allows for the presence of spatial error autocorrelation in the hedonic model, which, if unaccounted, will invalidate the test results (Anselin, 1988, 1990). The spatial Chow test statistics for the model specification with spatial error autocorrelation is expressed as

$$C_{SP} = [\boldsymbol{e}_R'(\boldsymbol{I} - \lambda \boldsymbol{W})'(\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{e}_R - \boldsymbol{e}_U'(\boldsymbol{I} - \lambda \boldsymbol{W})'(\boldsymbol{I} - \lambda \boldsymbol{W})\boldsymbol{e}_U]/\sigma^2 \sim \chi_K^2 \quad (13)$$

where $\boldsymbol{e}_R$ and $\boldsymbol{e}_U$ are the consistent estimates of the restricted and unrestricted residuals, **I** is an identity matrix of dimension K by K where K is the number of restrictions, λ is the coefficient of spatial error autoregression, **W** is the spatial weight matrix and σ^2 is the estimate for the error variance for the restricted model (Lagrange Multiplier (LM) test), the unrestricted model (Wald test), or both (Likelihood Ratio (LR) test).

5. Experiments and Results

5.1. Data

Data on arms-length transactions of single-family occupancy houses from 1985 to 1996 from four Ohio counties arrayed along the Lake Erie coastline (Erie, Lorain, Ottawa and Sandusky counties) yield a total of 10,655 observations[3]. These houses are characterized with the average sales price of 111,503 dollars (1996$) and the average age (the year of sales - the year built) of 30 years. The observations are located both in urban and rural areas, and the major cities within the study area include Lorain (71,245 (population from the 1990 census)), Sandusky (29,764) and Elyria city (56,746). A distinguishing feature of this study region is that these counties are adjacent to Lake Erie, one of the five Great Lakes and whose presence has a substantial influence on the regional housing market.

5.2. Implementation of Cluster Analysis

Cluster analysis is implemented using the agglomerative hierarchical clustering algorithm with the average linkage criterion. MATLAB is used for all experimentation. The six attributes included in cluster analysis are median household income, distance to the closest city, distance to the closest lake coastline, municipality, and the geographical coordinates (latitude and longitude). Five different specifications are considered for the similarity measurement. The first specification uses the standardized Euclidean distance as the baseline similarity method. The second specification uses the CDF-based normalization as a straightforward alternative to the normalization in the standardized Euclidean. These two specifications do not involve any discretization of the continuous variables.

[3] *The data were obtained from the Center for Urban and Regional Analysis at the Ohio State University.*

The remaining three specifications use the proposed similarity measures (equations 7- 9). These specifications differ according to how the continuous variables are handled. The continuous variables of median household income, distance to the closest city and distance to the coast line are normalized using the CDF transformation and discretized using six different cases for the number of bins (5, 10, 20, 30, 40, 50) to study the effects of discretization. The similarity for these variables is computed using the Hamming, modified Hamming 1, and modified Hamming 2 measures defined in equations (7), (8) and (9), respectively. In all three specifications, the categorical municipality variable has values from 1 through 211 and the similarity is coded using the Hamming distance in which the code is set to one if the observations are in the same municipality and zero otherwise (equation 7). The geographic coordinates are normalized by scaling to unit range without changing the distribution of the original (raw) values as in equation (2) and similarity is measured using equation (10).

Variables used in clustering should reflect households' decision-making process as well as the formation of submarkets. Although one may think that including as many variables as possible for clustering may help determine more realistic submarkets, it is not necessarily the case for two reasons. First, because as the number of attributes increases in clustering, the more "noise" is introduced in the clustering process. Furthermore, the distance computed between objects starts becoming more ambiguous because many dissimilar objects could have very similar distances in a high-dimensional attribute space as adding large differences in a few variables and adding small differences in many variables may produce the same results. This is a commonly known problem, called "the curse of dimensionality" in the pattern recognition and machine learning literature (Theodoridis & Koutroumbas, 2006). Second, variables that are used in both the cluster analysis and hedonic model estimation are subject to reduced variance within each cluster and potential bias. For example, if we use school district ranking as our clustering variable, the variation of this variable within each cluster is smaller than the case of not including it in the clustering. Therefore, if we include school district ranking in both clustering and the estimation of the hedonic price function, both the magnitude and the variance of the coefficient estimated are affected. This can result in insignificant estimates or structural change in the estimates across submarkets. However, it is also true that the variables which determine the submarkets are attributes considered in the housing purchase decisions as well. Thus, it is important to choose the variables that best represent the market segmentation. If the market is truly segmented by a particular variable, then it is reasonable that this variable may affect housing prices differently across submarkets and thus any structural change in the estimated coefficient reflects true structural change. However, if the market is not segmented by a variable, but the variable is inadvertently included in the cluster analysis, then its inclusion in the hedonic model is likely to result in insignificant or biased estimates due to the spurious reduction in variance and grouping of similar values within each cluster.

We address these issues by using a parsimonious set of variables in the cluster analysis that is mutually exclusive from the set used in the hedonic model. As discussed previously, this approach is sensible in our case because of the substantial spatial

heterogeneity in neighbourhood characteristics that exist at a local scale, which enables greater household sorting across local areas and thus makes it feasible to reasonably estimate neighbourhood boundaries based on observed features. In Ohio, the provision of local public goods and services varies by township, a highly localized scale. There are 211 distinguishable municipalities (townships, cities, villages) in our dataset and each observation is assigned to one of 211 municipalities. To capture potential distinctions within municipalities, we also include the following variables: median household income (calculated at the census block group level), distance to the closest city and distance to the Lake Erie coastline. Distance to the closest city is computed by using the major roads network from individual houses. Distance to the lake coastline is measured as the straight-line distance to the closest coastline. Lastly, the latitude and longitude of the house is included to account for highly localized similarities in geographic location.

5.3. Cluster Analysis Results

In the rest of the discussion, the five settings for similarity computation are abbreviated as SED, CDF, H, MH1, and MH2 for the standardized Euclidean, CDF normalization, Hamming, modified Hamming 1 and modified Hamming 2, respectively, and a specific case is expressed as [method, number of bins, number of clusters]. For example, the case with the modified Hamming 2 method with five bins and 11 clusters is written as [MH2, 5, 11]. Figure 3 shows an example dendrogram generated for the case of modified Hamming 2 with five bins.

The method described in Section 2.3 is used to find the optimal number of clusters for each setting from the corresponding dendrogram. The range for the number of clusters evaluated is varied from two to 20 clusters. The results are given in Table 2. Except SED, the optimal number of clusters for all settings are less than or equal to 11.

Figure: 3
An Example Dendrogram for the Case of Modified Hamming 2 with Five Bins

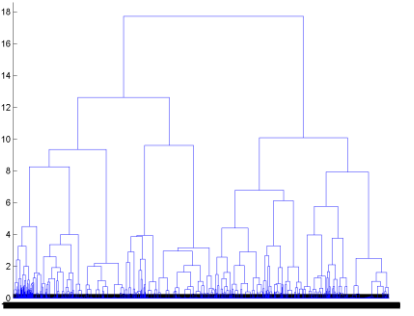


Table: 2
Optimal Number of Clusters for Each Method

Method	Bins	Number of Clusters
SED		14
CDF		11
Hamming	5	5
	10	5
	20	2
	30	10
	40	11
	50	3
MH1	5	8
	10	4
	20	4
	30	11
	40	4
	50	5
MH2	5	11
	10	8
	20	8
	30	3
	40	8
	50	4

Table 3 shows the settings that result in the minimum WMSE for a given number of clusters. We observe that modified Hamming 2 with 5 bins gives the minimum WMSE for most of the cases with the number of clusters varying from four to 13. When the number of clusters is increased further, WMSE cannot be computed for certain cases due either to the generation of a cluster with a very small number of observations or because of all observations in a cluster having the same value for at least one variable.

Table: 3
The Settings that Result in the Minimum WMSE for a Given Number of Clusters

Number of Clusters	Method	Number of Bins	Min(WMSE)
2	Ham	50	0.0579
3	Ham	40	0.0573
4	MH2	5	0.0543
5	MH2	5	0.0536
6	MH2	30	0.0510
7	MH1	30	0.0509
8	MH2	5	0.0499
9	MH2	5	0.0496
10	MH2	5	0.0489
11	MH2	5	0.0488
12	MH2	5	0.0487
13	MH2	5	0.0487
14	Ham	10	0.0485
15	Ham	10	0.0484

Table 4 shows different similarity measures and the corresponding optimal number of clusters sorted in increasing order of WMSE values. Although it is not possible to directly compare WMSE values for different number of clusters, modified Hamming 2 with five bins and 11 clusters has the smallest WMSE compared to other methods with 11 clusters as the optimal number of clusters, such as CDF, [MH1,30], and [H,40]. Furthermore, given that WMSE decreases as the number of clusters increases, it is obvious that the [MH2,5] setting with 11 clusters produces a better clustering than the commonly used standardized Euclidean

distance even when the latter uses 14 clusters as its optimal setting. Together with the fact that the [MH2,5] method results in the smallest WMSE value for a large range of clusters (Table 3) and its WMSE value for 11 clusters having the smallest WMSE among others (Table 4), we select the best similarity measure as the modified Hamming 2 method with five bins and the optimal number of clusters as 11 for our housing data.

Table: 4
WMSE for Different Similarity Measures and the Corresponding Optimal Number of Clusters

Method	Optimal Number of Clusters	WMSE
[MH2,5]	11	0.0488
CDF	11	0.0493
SED	14	0.0495
[MH1,30]	11	0.0502
[H,30]	10	0.0504
[H,40]	11	0.0508
[MH1,5]	8	0.0512
[MH2,40]	8	0.0525
[MH2,20]	8	0.0528
[MH2,10]	8	0.0533
[MH1,10]	4	0.0546
[H,5]	5	0.0550
[H,10]	5	0.0553
[MH1,50]	5	0.0569
[MH1,20]	4	0.0570
[MH2,30]	3	0.0576
[MH2,50]	4	0.0576
[MH1,40]	4	0.0576
[H,50]	3	0.0577
[H,20]	2	0.0580

The settings are sorted in increasing order of WMSE.

Table: 5
Adjusted Rand Index Values Computed Pairwise for Settings That Give the Lowest WMSE Values

	[SED,14]	[MH2,5,11]	[CDF,11]	[MH1,30,11]	[H,30,10]
[SED,14]*		0.254	0.408	0.248	0.199
[MH2,5,11]			0.751	0.972	0.629
[CDF,11]				0.742	0.535
[MH1,30,11]					0.629
[H,30,10]					

Upper bound for the index is one (perfect agreement) and the lower bound is less than but close to zero (perfect disagreement).
** Adjusted Rand index (ARI) value does not change significantly for the case of [SED,11]. e.g., for the comparison between [SED,11] and [MH2,5,11], ARI = 0.2531.*

The adjusted Rand index is also computed to measure the significance of the differences in the cluster structure produced by the commonly used standardized Euclidean distance and the cluster structures produced by the similarity measures proposed in this paper. The adjusted Rand index values for [SED,14], [MH2,5,11] and several other methods with low WMSE are listed in Table 5. We can conclude that the cluster structure obtained with the standardized Euclidean distance differs significantly from the structure from other methods, especially from the discretized similarity measures. We can also observe that the cluster structures from [MH1,30] and [MH2,5] have a very high agreement rate. The

experiments show that, although the standardized Euclidean distance is commonly used in the cluster analysis and housing market segmentation studies, other similarity measures that rely on a more intuitive and effective combination of continuous and categorical variables produce results that are significantly better in terms of fit and markedly different in terms of their cluster structures.

The produced clusters and the houses located in each cluster are shown in Figure 4 for the selected [MH2,5,11] case. We can identify five close-to-coastline clusters [Cluster 3, 4, 7, 8, 10]. Clusters 2, 4, 5, and 7 are formed with one distinguishable geographical area, clusters 1, 3, 6, 8 and 10 are grouped into two or more distinct areas, while cluster 9 and 11 are spread over a larger spatial extent. The clusters generated by the [SED,14] case are shown in Figure 5 for comparison purposes. Although some clusters look similar to the selected case, many clusters, especially the ones located on the east side of the map are formed quite differently.

Figure 6 illustrates the descriptive statistics for selected variables for each cluster produced by the [MH2,5,11] method. The most distinguishable cluster is cluster 7 which is located very close to the coastline and is characterized with the highest median household income, the highest housing price and the newest housings.

5.4. Spatial Chow Test Results

In order to group clusters into submarkets, the spatial Chow test (equation 13) is used. The test statistics for all possible combinations of clusters are computed by using generalized method of moments (GMM) estimates (Anselin, 1988; Kelejian & Prucha, 1998). All three test statistics [LM, Wald, LR] are computed by using error variance for restricted model, unrestricted model and both, respectively from estimated GLS procedure. The relative magnitude of the test statistics for all cases are found to be LM > Wald > LR. The most conservative statistics in our case, LR test statistics, are reported in Table 6. The test result suggests that clusters 1 and 2, clusters 4 and 5 and clusters 3, 8, 10 and 11 can be combined to form separate submarkets[4]. Therefore, we reduce eleven clusters into six submarkets.

The result of the regular Chow test suggests combining only clusters 3 and 11, and for all other combinations the structural instabilities are found at least at 1 percent statistical significance level. Thus, we find that the use of the spatial Chow test is less likely to return the result of rejecting the null hypothesis (structural stability between two clusters) compared to the case of the uncorrected Chow test.

[4] *The result of the LM test shows that the structural instability could not be observed for clusters 3-8 and 3-11 at the 1% significant level. For the Wald test, the structural instability was not observed for clusters 1-2, 1-3, 1-8, 2-3, 3-8, 3-11 and 4-5 at the 1% level.*

Table: 6
Spatial Chow Test Statistics [LR test] for Individual and Merged Clusters

	1	2	3	4	5	6	7	8	9	10	11	1 & 2	4 & 5	3&8& 10&11
1		**8.0**	**9.1**	**17.3**	*24.0*	38.9	39.9	**8.7**	**18.1**	*19.7*	**17.3**		*23.7*	**12.7**
2			**8.5**	*23.9*	42.6	74.9	72.2	**17.7**	35.9	31.6	*21.9*		43.9	**18.7**
3				*20.5*	29.1	45.2	47.5	**5.2**	**17.5**	**15.2**	**4.8**	**9.4**	38.9	
4					**8.3**	26.9	118.8	32.0	31.1	40.3	43.5	26.7		38.5
5						*21.2*	112.5	40.6	**13.8**	25.1	48.9	48.7		46.6
6							120.6	61.3	*21.8*	*22.9*	67.5	82.7	31.2	71.9
7								50.4	111.3	53.5	102.2	77.7	172.3	119.3
8									34.3	*20.5*	**16.1**	**18.3**	69.6	
9										25.8	*23.0*	40.7	27.0	42.7
10											*20.2*	35.6	41.8	
11												28.5	57.8	
1 & 2													52.0	*24.6*
4 & 5														73.5
3&8& 10&11														

**Cut-off values of Chi-squared distribution with 11 restrictions are 19.68 for p = 0.05 and 24.73 for p = 0.01.*
*** The values below 5% and 1% cut-off points are in bold and italic, respectively.*

6. Conclusions

This paper presents new methods for computing similarity measures, a critical component of cluster analysis, to determine housing submarkets using individual housing sales data. These methods focus on handling mixed-mode variables, in which some variables are continuous and others categorical. Commonly used similarity measures such as the standardized Euclidean and attribute combination and reduction methods such as the principal components analysis cannot directly handle mixed-mode data with variables of different types and units. Our proposed measures are innovative because they use both normalization and discretization techniques that permit all the variables to be handled in the same framework.

In applying these methods to our data, we find that similarity measures that rely on a more intuitive and effective combination of continuous and categorical variables produce markedly different cluster structures than the commonly used methods for housing market segmentation. To compare the various similarity measures, we calculate a goodness of fit measure for the cluster structure produced by each measure and use the adjusted Rand index to quantify significant differences in the results. The experiments show that the proposed similarity measure based on a moderated form of discretization (match, partial match, and no match scores) produces the best results. In comparing these results with those produced by standardized Euclidean distance, which is the default settings of most clustering software packages, we find that the proposed measure produces results that are significantly better in fit and substantially different in terms of the resulting cluster structure.

Figure: 4
11 Clusters Produced by the Modified Hamming 2 Method

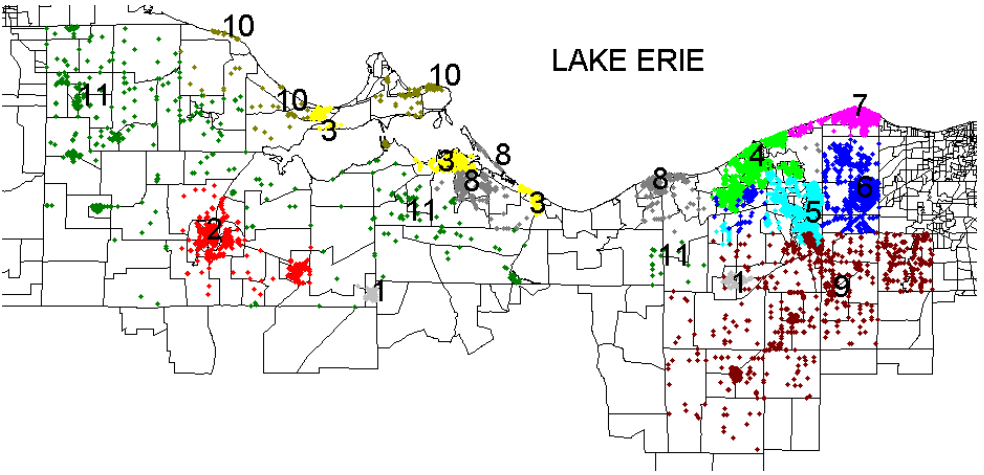


Figure: 5
14 Clusters Produced by the Standardized Euclidean Distance Method

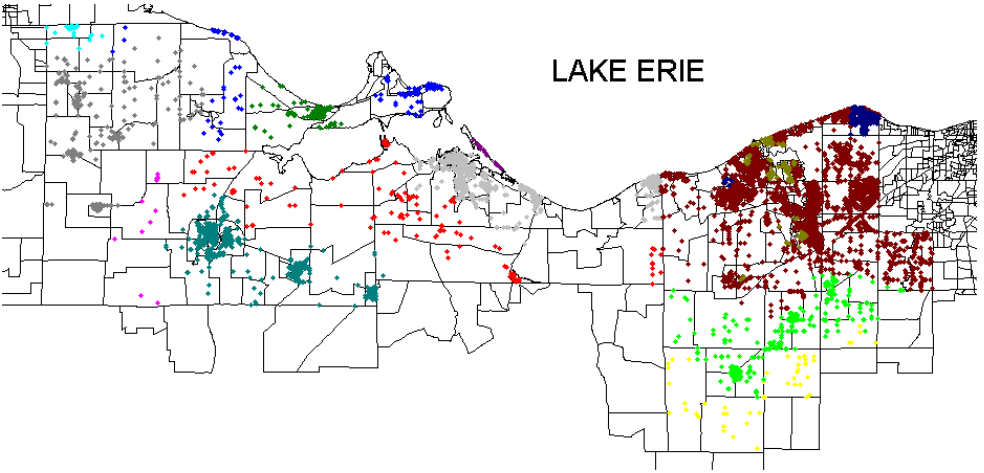
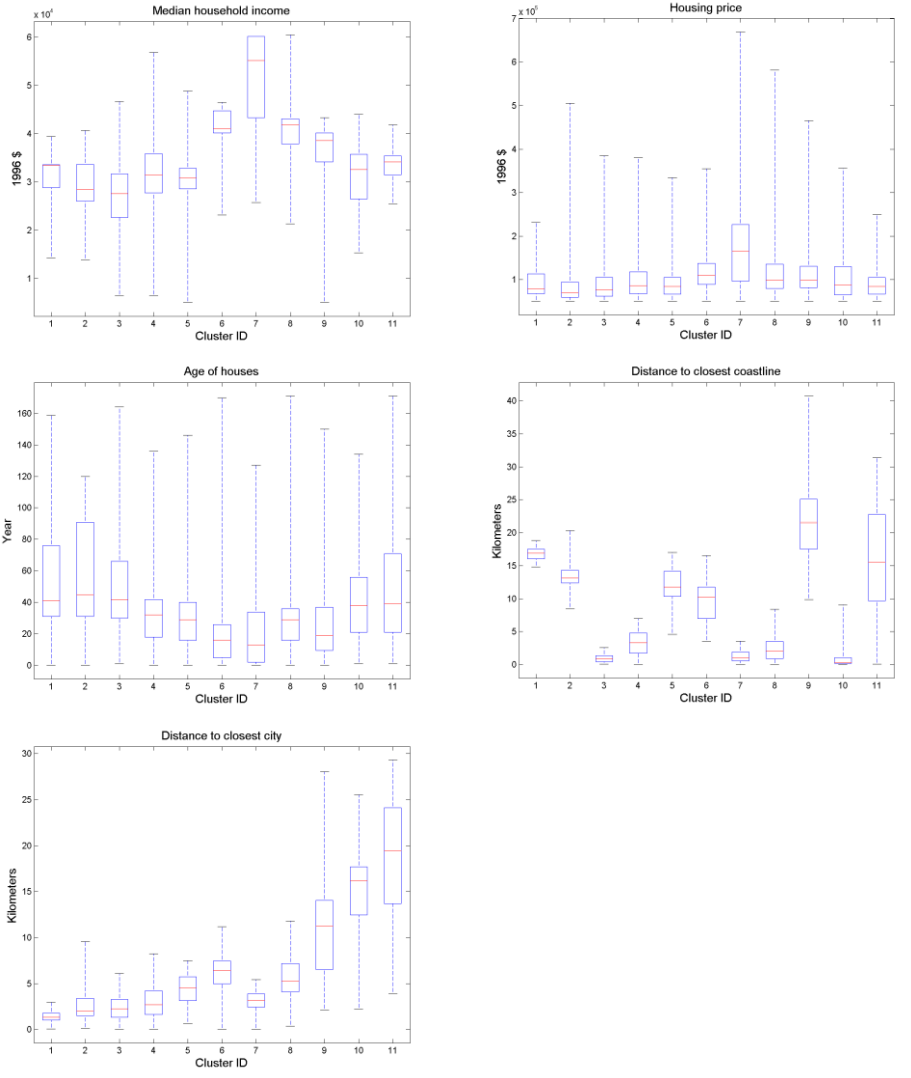


Figure: 6
Descriptive Statistics (Box and Whisker Plots) of Selected Variables for Each Cluster Produced by the [MH2,5,11] Case



The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data.

References

Aksoy, S. & R.M. Haralick (2001), "Feature normalization and likelihood-based similarity measures for image retrieval", *Pattern Recognition Letters*, 22(5), 563-582.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1990), "Spatial dependence and spatial structural instability in applied regression analysis", *Journal of Regional Science*, 30(2), 185-209.

Bates, L.K. (2006), "Does Neighborhood Really Matter? Comparing historically Defined neighborhood boundaries with housing submarkets", *Journal of Planning Education and Research*, 26, 5-17.

Bishop, C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York, USA.

Boberg, J. & T. Salakoski (1993), "General formulation and evaluation of agglomerative clustering methods with metric and non-metric distances", *Pattern Recognition*, 26(9), 1395-1406.

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

Clapp, J.M. & Y. Wang (2006), "Defining neighborhood boundaries: Are census tracts obsolete?", *Journal of Urban Economics*, 59, 259-284.

Day, B. (2003), "Submarket identification in property markets: A hedonic housing price model for Glasgow", *Technical Report*, The Centre for Social and Economic Research on the Global Environment, School of Environmental Sciences, University of East Anglia, Norwich, UK.

Everitt, B.S. & S. Landau & M. Leese (2001), *Cluster Analysis*, Fourth Edition. Arnold, London, UK.

Galster, G.G. (2003), "Neighborhood dynamics and housing markets", in: T. O'Sullivan & K. Gibb (eds), *Housing Economics and Public Policy*, Oxford: Blackwell.

Gillen, K. & T.G. Thibodeau & S. Wachter (2001), "Anisotropic autocorrelation in house prices", *Journal of Real Estate Finance and Economics*, 23(1), 5-30.

Goetzmann, W.N. & S.M. Wachter (1995), "Clustering methods for real estate portfolios", *Real Estate Economics*, 23(3), 271-310.

Goodman, A.C. & T.G. Thibodeau (2003), "Housing market segmentation and hedonic prediction accuracy", *Journal of Housing Economics*, 12, 181-201.

Gower, J.C. (1971), "A general coefficient of similarity and some of its properties", *Biometrics*, 27, 857-872.

Hoesli, M. & C. Lizieri & B. MacGregor (1997), "The spatial dimensions of the investment performance of UK commercial property", *Urban Studies*, 34(9), 1475-1494.

Kelejian, H. & I.R. Prucha (1998), "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances", *Journal of Real Estate Finance and Economics*, 17, 99-121.

Kim, K. & J. Park (2005), "Segmentation of the housing market and its determinants: Seoul and its neighbouring new towns in Korea", *Australian Geographer*, 36(2), 221-232.

Maclennan, D. & Y. Tu (1996), "Economic perspectives on the structure of local housing systems", *Housing Studies*, 11(3), 387-406.

Milligan, G.W. & M.C. Cooper (1985), "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, 50, 159-179.

Theodoridis, S. & K. Koutroumbas (2006), *Pattern Recognition*, 3rd edition. Elsevier, USA.

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

Witten, I.H. & E. Frank (2005), *Data Mining*, 2nd edition. Elsevier, USA.