Journal of International Technology and Information Management

Volume 24 | Issue 1 Article 2

2015

An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction

Donghui Shi
Anhui Jianzhu University

Jian Guan
University of Louisville

Jozef Zurada
University of Louisville

Alan S. Levitan
University of Lousiville

An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction

Donghui Shi
Department of Computer Engineering
School of Electronics and Information Engineering
Anhui Jianzhu University
CHINA
Universidad Técnica Particular de Loja
ECUADOR

Jian Guan
Department of Computer Information Systems
College of Business
University of Louisville
USA

Jozef Zurada
Department of Computer Information Systems
College of Business
University of Louisville
USA

Alan S. Levitan
School of Accountancy
College of Business
University of Louisville
USA

ABSTRACT

A main obstacle to accurate prediction is often the heterogeneous nature of data. Existing studies have pointed to data clustering as a potential solution to reduce heterogeneity, and therefore increase prediction accuracy. This paper describes an innovative clustering approach based on a novel adaptation of the Fuzzy C-Means algorithm and its application to market segmentation in real estate. Over 15,000 actual home sales transactions were used to evaluate our approach. The test results demonstrate that the accuracy in price prediction shows notable improvement for some clustered market segments. In comparison with existing methods our approach is simple to implement. It does not require additional collection of data or costly development of models to incorporate social-economic factors on segmentation. Finally our approach is not market specific and can be easily applied across different housing markets.

Keywords: Fuzzy c-means clustering, mass assessment, k-means clustering, real estate submarkets, cluster homogeneity, ANFIS

  ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

INTRODUCTION

There is general agreement that a housing market consists of a set of submarkets. Various methods have been proposed for reliable detection of submarkets. Common methods in determining submarkets are mostly geographic, administratively determined, or statistics-based. Administratively determined boundaries, such as those used by local government assessment offices, can be ineffective (Zurada, Levitan, & Guan, 2011). Though geographic boundaries can be more effective (Fik, Ling, & Mulligan, 2003), such boundaries also restrict segmentation to only spatial considerations. As Goodman and Thibodeau (2007) point out, consumers do not necessarily limit their housing search to spatially contiguous areas when house hunting. Statistical methods often use specially developed neighborhood characteristics such as quality of schools and level of public safety. However, the development of such neighborhood characteristics can be challenging and costly (Goodman & Thibodeau, 2007). This paper explores the feasibility of an untested approach to construct submarkets with the objective to better disaggregate the properties in a given market into more homogeneous clusters. The approach is based on an adaptation of the fuzzy C-means (FCM) algorithm. Our approach is based on a proven and common practice by humans, uses data features readily available to local governments or appraisal offices, and is simple to implement without requiring complicated model development. As is common in such studies, we evaluated the validity of this newly introduced housing segmentation approach by measuring the accuracy of price prediction in the resulting submarkets. A new input variable, based on the prices of comparable properties in their cluster, is added to the typical set of housing characteristics within the newly formed submarkets. Two different methods were used to predict the prices of properties in the clusters, adaptive neural fuzzy inference system (ANFIS) and the traditional multiple regression analysis (MRA). The test results, based on actual sales transactions in Louisville, KY, show that the price prediction accuracy noticeably improves for some of the resulting clusters.

The paper is organized as follows. The next section provides a brief literature review, which is followed in section 3 by a presentation of the adapted FCM algorithm and its application to create homogeneous clusters for properties in a given market. Section 4 provides a description of the data set used in this study. Section 5 presents and discusses the results of testing using both MRA and ANFIS. Finally Section 6 provides a summary and conclusions.

RELEVANT LITERATURE

A main obstacle to efforts to obtain an accurate valuation of real estate properties is the heterogeneous nature of real estate data (Goodman & Thibodeau, 2007; Mark & Goldberg, 1988). An increasingly common approach to reduce data variability and improve accuracy is to cluster the data set into submarkets that are more homogeneous (Bourassa, Hamelink, Hoesli, & MacGregor, 1999; Fletcher, Gallimore, & Mangan, 2000; Goodman & Thibodeau, 1998, 2003, 2007; Wilhelmsson, 2004). It is well recognized in the real estate literature that housing markets are typically segmented into clusters/submarkets and these clusters/submarkets should be included in the price prediction process (Wilhelmsson, 2004). A cluster/submarket can be defined as a set of properties within which implicit pricing and/or property features differ from those of another area/set (Goodman & Thibodeau, 1998). It is argued that each of these different clusters/submarkets should have its own price equation/model (Straszheim, 1974). The expectation is that these separate models should provide better estimates for pricing than an aggregate model

  ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

because an aggregate model will have biased coefficients (Michaels & Smith, 1990). In addition to improving price prediction accuracy, identification of submarkets also allows researchers to better model spatial and temporal variation in prices, helps lenders assess risk in home ownership, and reduces home buyers' search costs (Goodman & Thibodeau, 2007).

Various methods have been proposed to determine boundaries for clusters and they include geographical, administrative, and statistically determined boundaries. Dale-Johnson (Dale-Johnson, 1982) uses factor analysis of property features and pre-defined administrative boundaries (zip codes) in segmenting his data. Dale-Johnson finds improvement in model predictability but the study does not use out-of-sample tests. Fletcher et al. (Fletcher et al., 2000) add property type and age to zip codes in their approach to submarket construction. Though their disaggregated models yield statistically better price prediction results, the results are not good enough to offer practical value. Statistical methods, such as principal component analysis and cluster analysis, can also be used to define submarkets and the resulting submarkets may or may not correspond to spatially defined submarkets (Bourassa, Cantoni, & Hoesli, 2010). Bourassa et al. (1999) use a statistical technique to derive clusters/submarkets in which they combine principal component analysis and cluster analysis. The factors extracted from the principal component analysis are used in two different clustering methods, k-means and Ward. Their results show an improvement of about 25% when compared with those of the aggregated model. More recently Bourassa et al. (2010) confirm the benefit of submarkets in price prediction through a comparative study of various submarket construction techniques, which include ordinary least squares and geostatistical methods. Their results confirm that the inclusion of submarket variables improve price prediction accuracy. Goodman and Thibodeau (2003) use hierarchical linear models to define clusters/submarkets for the Dallas metropolitan area. They use two different submarket construction methods, one using zip codes and the other using census tracts. Both methods lead to improved estimates of property values using regression models. In their more recent study Goodman and Thibodeau (2007) find that their model based on school quality and public safety significantly improves price prediction accuracy. However, they also point out that development of such models can be costly and challenging. More recently Belasco et al. use a finite mixture model to identify latent submarkets from household survey data (Belasco, Farmer, & Lipscomb, 2012). Their results suggest homogenous preferences of residents within identified submarkets.

In general these submarket construction/clustering studies show that clusters/submarkets in real estate data exist and property valuation models that incorporate such clusters/submarkets can potentially result in better price prediction accuracy. Some of these studies use spatial boundaries in determining clusters/submarkets and these boundaries include zip codes, census tracts, and local government defined boundaries. If one follows a more general definition of a cluster/submarket as a set of properties that are relatively close substitutes of one another and poor substitutes for properties outside the submarket, then spatial boundaries are not necessarily the only way to segregate real estate data (Bourassa et al., 1999). Zurada et al. (2011) show that administrative boundaries as used in mass appraisal by local governments may not be effective in segmenting real estate data. In fact, as Goodman and Thibodeau (2007) point out, consumers do not necessarily limit their housing search to spatially contiguous areas when house hunting. Consumers are more driven by price and features as well as location. Development and incorporation of neighborhood characteristics in cluster/submarket construction provide an alternative approach to geographic and administrative boundaries but such development can be challenging (Goodman & Thibodeau, 2007).

This paper introduces a unique clustering method often applied in other fields, the fuzzy c-means clustering method. The objective is to offer an alternative method to housing market segmentation and to better understand why certain properties cluster well and others do not. In addition we incorporated a common and proven practice of using similar properties ("comparables") by expert human appraisers in property appraisal. After segmentation by FCM, a comparable properties-based index is introduced for each property as an additional input in price prediction. Our approach relies on real estate transactions data readily available in various appraisal contexts and does not require the development of neighborhood or submarket characteristics. Neither does our approach require the collection of additional data. Lastly our FCM-based approach is simple to implement.

Often the measure of the effectiveness of a clustering/submarket construction approach is the improvement in price prediction accuracy (Bourassa et al., 2010; Goodman & Thibodeau, 2007). The standard and dominant approach to price prediction is MRA-based (McCluskey, Cornia, & Walters, 2012). The use of MRA in property price prediction has been discussed extensively because MRA is thought to be a poor model for handling the inherent complexity and high level of data variability that exist in real estate data (Mark & Goldberg, 1988; Stevenson, 2004). Notable issues with the application of MRA include non-linearity, multi-collinearity, and heteroscedasticity (Kilpatrick, 2011; Mark & Goldberg, 1988). Various data mining models have been proposed and tested (Antipov & Pokryshevskaya, 2012; Do & Grudnitski, 1992; Gonzalez & Laureano-Ortiz, 1992; Guan, Zurada, & Levitan, 2008; Peterson & Flanagan, 2009; Selim, 2009; Zurada et al., 2011), but the results are rather mixed. Some studies find data mining methods to be superior (Antipov & Pokryshevskaya, 2012; Do & Grudnitski, 1992; Peterson & Flanagan, 2009) but others find little or no improvement when compared to MRA (Guan et al., 2008; Zurada et al., 2011). A major criticism of the use of data mining methods such as artificial neural networks in mass appraisal is their lack of interpretability (McCluskey et al., 2012). Guan et al. (2008) propose the use of ANFIS as ANFIS combines the benefits of neural networks and interpretable results. The resulting fuzzy rules offer a mechanism to capture the inherent imprecision in mass appraisal (Bagnoli & Smith, 1998; Byrne, 1995). Therefore in this study we used both MRA and the adaptive neuro-fuzzy inference system (ANFIS) to test our new approach.

METHODOLOGY

This section describes an alternative approach to group properties into more homogeneous clusters. FCM is a well-known clustering method popularized by Bezdek (1984). Like most clustering methods FCM segments data elements (properties in our case) into clusters so that data elements in the same cluster are as similar as possible and data elements in different clusters are as dissimilar as possible. FCM belongs to the class of soft clustering methods where a data element can be assigned to different clusters at the same time. For each cluster to which a data element is assigned, a value called membership level indicates the degree to which the data element belongs to that cluster. In the context of housing segmentation a property can belong to more than one segment through FCM clustering. For each resulting segment FCM finds a level of association of a property with that segment. Market segments can then be determined by assigning each property to a segment with which the property has the highest level of association. The level of association can be determined by any meaningful criterion such as proximity, prices, etc. A more formal description of FCM for market segmentation follows.

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

Let $X = \{x_1, \dots, x_n\}$ be a set of n properties where each x_i represents a property. The FCM clustering method returns k clusters $C = \{c_1, \dots, c_k\}$ and a membership matrix

$$U = \left\{u_{i,j} \middle| u_{i,j} \in [0,1], i = 1, \dots, n, j = 1, \dots, k, \sum_{j=1}^{k} u_{i,j} = 1\right\} \quad (1)$$

where each u_{ij} is the degree to which property x_i belongs to cluster c_j. In FCM the data set is partitioned into k clusters by minimizing the following objective function

$$J(X; U, C) = \sum_{j=1}^{k} \sum_{i=1}^{n} (u_{ij})^m \left\|x_j - c_j\right\|^2, 1 < m < \infty \quad (2)$$

where m is a weighting exponent, $\|\cdot\|$ is the Euclidean norm, and

$$\sum_{j=1}^{k} u_{ij} = 1, 1 \le i \le n \quad (3)$$

Minimization of the objective function is an iterative process and consists of the following steps:

1. Select a value for k as the number of clusters, m as the weighting exponent, and ϵ as the termination threshold.
2. Initialize the membership matrix U^0 with random values between 0 and 1
3. Initialize iteration counter t with 0
4. $t = t + 1$
5. Calculate the cluster centers as follows
$$c_i^{(t)} = \frac{\sum_{j=1}^{n} (u_{ij}{}^{(t-1)})^m x_j}{\sum_{j=1}^{n} (u_{ij}{}^{(t-1)})^m}, 1 \le i \le k$$
6. Update the membership matrix as follows
$$u_{ij}^{(t)} = \frac{1}{\sum_{l=1}^{k} \left(\frac{\left\|x_i - c_j^{(t)}\right\|}{\|x_i - c_l\|}\right)^{\frac{2}{m-1}}}, 1 \le j \le k, 1 \le i \le n$$
7. If $\left\|U^{(t)} - U^{(t-1)}\right\| < \epsilon$, stop; otherwise go to step 4.

Although FCM allows each property to belong to one or more clusters, in our study a property belongs to only one cluster, where its membership in the cluster u_{ij} > threshold[1].

Once the clusters are created, a new input variable, the mean price of comparables, or MPC, is defined for each property in its cluster to improve prediction accuracy. The MPC index is based on a common process used by human experts in determining comparables for a given property for either sales or appraisal purposes. For any given property *x*, comparables are those properties whose features and sale prices are most likely to reflect the features and sale price of property *x*. The objective of our approach is to group properties into homogeneous clusters so that the most similar properties are more likely to be used as comparables.

[1] The threshold value is 0.8. For details please see the results section.

Since the use of FCM allows the resulting clusters to become more homogeneous, the resulting MPC of each property calculated using the *k* nearest neighbors within a cluster is more likely to be close to the actual price of the property. The process for calculating MPCs is defined as follows. For each property in each cluster the *k* nearest neighbors for the property are first determined by location and the distance between the property *p1* and its neighbor *p2* as follows:

$$d = \sqrt{(p1_{Longitude} - p2_{Longitude})^2 + (p1_{Latitude} - p2_{Latitude})^2} \qquad (4)$$

Once the *k* nearest neighbors are found, the MPC of property $x_i$using the *k* nearest neighbors is calculated as follows:

$$MPC_i = \frac{1}{k}\sum_{j=1}^{k} SalePrice_j \qquad (5)$$

where *k* is the number of neighbors for x_i.

As noted above, the larger the threshold value is, the more homogeneous the resulting clusters created by FCM will be. However, a larger threshold value also causes a greater number of properties to remain unassigned to any cluster. As can be seen in Figure 2 the number of properties assigned to clusters is inversely related to the threshold value. For example if the threshold value is 0.9, the total number of properties assigned to clusters is 6,422 (or about 38.7% of the total number properties) but becomes 15,904 (or about 95.8%) if the threshold value is reduced to 0.5.

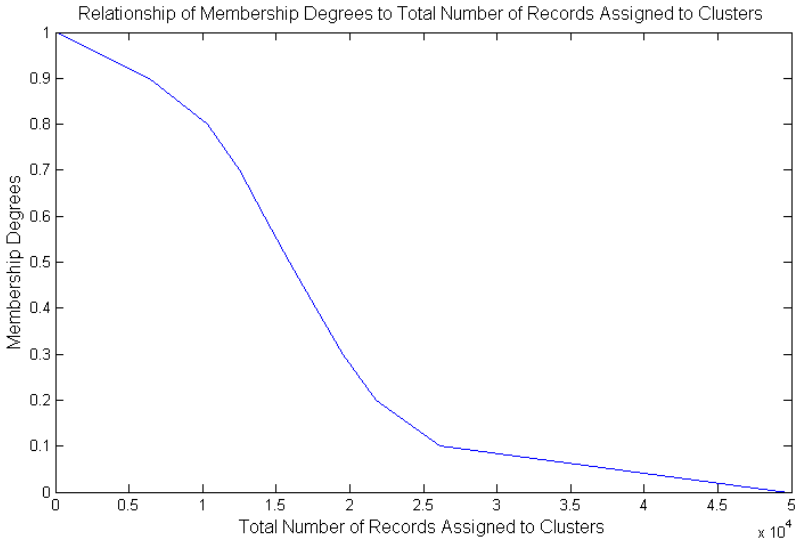

Figure 1. Relationship between membership degrees and total records in clusters

Therefore, those properties whose membership degree is not greater than or equal to the threshold are not assigned to any cluster. This paper introduces a modification of the FCM algorithm so that these unassigned properties are further grouped into additional clusters.

Next we describe the algorithm.

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

Input:

- Original data set, *X*;
- The initial number of clusters, *c;* here *c=3*
- The threshold record number of clusters Θ_0 =3000.

Output

- *C* : the set of resulting clusters;

1. Find *c* clusters using the dataset *X* using FCM
2. For each of these *c* clusters find the MPC for each property within each cluster
3. Set Θ = the number of *R*
4. If Θ < threshold Θ_0 retain *R* as a cluster and Exit
5. Otherwise set *X* = *R*
6. Repeat steps 1-5 until the number of properties in *R* is less than the threshold Θ_0

The modified FCM process will create an initial *c* number of fuzzy clusters (see step1). Then for each property in each of these newly created clusters the MPC is calculated in step 2. Next the number of properties in the remaining, unclustered set *R*, Θ is compared to a threshold size, Θ_0, to determine if *R* needs to be further segmented into smaller subclusters (steps 3, 4). If the number of properties in *R* is less than the threshold Θ_0, it is kept as a new cluster and the clustering process terminates. Otherwise the properties in *R* will go through another round of clustering. This process continues until in the number of properties in *R* is less than the threshold Θ_0.

DATA DESCRIPTION

The original data set used in this study contains 20,192 sales transactions of residential properties in Louisville, Kentucky, US from 2003-2007. The Jefferson County Tax Assessor in Louisville, Kentucky provided access to the entire database of over 300K properties and 143 variables. About 220K and 80K properties were identified as residential and commercial properties, respectively. After the removal of commercial properties, the data set was reduced to approximately 200,000 records and included only residential properties with actual sales dates and prices for years 2003–2007. Vacant lots, properties sold for less than $10,000, properties having less than 500 square feet on the floors, and several of those built before the year 1800 were excluded. Next the records were cleansed to eliminate repeated records representing group sales, inconsistent coding, missing values, and other obvious errors common in any database that large. After the data were preprocessed to eliminate invalid and incomplete records, 16,592 records remained. Table 2 shows the fields used in this study. Table 3 shows their descriptive statistics. Each sales record contains fields describing the typical features of a property, the sale date, and the sale price. The fields shown in Table 2 are known to have influence on the price of a property and are normally included in any model of real estate price prediction (Zurada et al., 2011). Since the data set contains sales records spanning several years the price of each property is CPI-index adjusted by the formula in Table 1.

YEAR	BASE YEAR 1983	BASE YEAR 2007

2003	184.000	1.127
2004	188.900	1.098
2005	195.300	1.062
2006	201.600	1.028
2007	207.342	1.000

Table 1. CPI-U (Consumer Price Index – All urban consumers)

Variable Name	Sample Record	Explanation
Sale price [$] (Dependent variable)	390000	Actual sale price
Year Built	1968	Year in which the property was built
Age	37	Age in years when the property was sold
Square footage in the basement [Feet]	900	Square feet in basement
Square footage on the floors [Feet]	2931	Total square feet above basement
Number of fireplaces	1	One fireplace (range 0-3)
Garage size (number of cars)	2	Two-car garage (range 0-2)
Number of baths	4	0=substandard bath; 1=1 bath; 2=1 ½ baths; 3=2 baths;4=2 ½ baths, etc. up to 6=more than 3 baths
Presence of central air	1	0=no central air; 1=central air is present
Lot type	1	1=up to one-fourth acre; 2=one-fourth to one-half acre; 3=one-half to 1 acre; 4=over 1 acre
Construction type	3	1=1 story; 2=1 ½ story; 3=2 story; 4=2 ½ story; 5=split-level; 6=bi-level; 7=condominium
Wall type	2	1=frame; 2=brick; 3=other
Basement type	1	0=none; 1=partial; 2=full
Basement code	1	0=none; 1=standard; 2=half standard; 3=walk-out
Garage type	3	0=none; 1=carport; 2=detached; 3=attached; 4=garage in basement; 5=built-in garage
Longitude	-85.573	Degrees west
Latitude	38.216	Degrees north

Table 2. Fields of a property used in the study

Variable	Mean	Std Dev	Minimum	Maximum	Median
Sale price	166562	105585	10700	889220	139674
Year Built	1967	32	1864	2006	1966
Age	38	31.8	-1	141	39
Square footage in the basement [Feet]	187.8	400.2	0	2952	0
Square footage on the floors [Feet]	1577.5	649.2	525	7459	1399
Number of fireplaces	0.51	0.50	0.00	1.00	1.00

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

Garage size (number of cars)	1.13	0.87	0.00	2.00	1.00
Number of baths	2.65	1.49	0.00	6.00	3.00
Presence of central air	2.65	1.49	0.00	6.00	3.00
Lot type	1.18	0.52	1.00	4.00	1.00
Construction Type	1.63	0.82	1.00	3.00	1.00
Wall type	1.59	0.52	1.00	3.00	2.00
Basement type	1.13	0.94	0.00	2.00	2.00
Basement code	0.62	0.49	0.00	1.00	1.00
Garage type	1.82	1.33	0.00	5.00	2.00
Longitude	-85.90	2.86	-157.91	-71.05	-85.65
Latitude	38.15	0.73	21.37	47.12	38.20

Table 3. Descriptive statistics of data set

A careful examination of our data shows that the data are very heterogeneous. As can be seen in the descriptive statistics several data fields exhibit large variation such as Age (as shown through Year Built), Sale Price, and Floor Size. This type of large variation in data values may affect prediction accuracy.

Location is considered a critical data element in real estate property price prediction and is used in creating clusters of properties, i.e., submarkets, as described in the literature review. Location is represented by longitudes and latitudes in this study. Another common practice in representing location is to use administrative boundaries. In the city where the sales data were collected, administrative boundaries are used to describe each property. The sales records represent properties belonging to about 20 Tax Assessor (TA) districts. The TA districts are then further divided into more than 400 TA neighborhoods, which are in turn divided into about 8,000 TA blocks. One of the ways to assess (or reassess) the value of a property for tax purposes in the city is to sum the sale prices of all similar properties sold recently in the immediate neighborhood of the house, divide the sum by the total square footage of these properties. The resulting value is the price per square foot of similar properties. This price per square foot value is then multiplied by the square footage of the property to be assessed. This practice is actually very similar to the use of the so-called comparables. In real estate appraisal the comparables of a property are the nearby properties with similar features. The mean price of the comparables for a property is often considered a good predictor of the sale price of the property. However, these comparables are often manually determined property by property by expert human agents or appraisal experts. Zurada et al. (2011) show that these administratively determined boundaries may not be very effective in grouping similar properties and propose the creation of different comparables as a feature to improve prediction of property values. As discussed in the literature review administrative or geographical boundaries in general may not be effective in grouping similar properties or submarkets (Goodman & Thibodeau, 2007).

The use of comparables has proven to be effective when human experts (such as real estate agents) determine the comparables of a property as the human experts tend to have very intimate knowledge of the neighborhood of the property. For example there may be several neighborhood properties that are physically close but vary greatly in price because one such neighborhood

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

property has a relatively low price due to aging appliances and/or conditions of the exterior of the property. Apparently the accuracy of a property's price prediction is critically dependent on the correct determination of the comparables for the property. In this study we introduce a new input variable, MPC, as described in the previous section, to better emulate this human appraisal process. MPC is computed for each property after clustering, the objective being to use comparables that are more similar.

TEST RESULTS

This section describes the test results. As consistent with many existing studies on clustering/segmentation prediction results both before and after segmentation are shown and compared. The tests were performed with the Matlab software from Math Works. In all the tests we randomly partitioned the data set into three subsets: training set (40%), validation set (30%), and test set (30%) (Witten & Frank, 2005). Three performance measures: Mean Absolute Percentage Error (MAPE), Root Mean-squared Error (RMSE) and Mean Absolute Error (MAE) were used to measure the performance of the methods on the test sets. The performance measures are as defined in Table 4. In each of the test scenarios 50 random generations of the training, validation and test subsets were created and tested and the results are the averages over the 50 runs. In addition in each of the test scenarios both ANFIS and MRA were used for prediction and their results are compared in the rest of this section. The results across the different test scenarios are also compared.

Error Measure	Formula
Root Mean-squared Error (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n}(actual\ price_i - predicted\ price_i)^2}{n}}$
Mean Absolute Error (MAE)	$\frac{\sum_{i=1}^{n}\lvert actual\ price_i - predicted\ price_i\rvert}{n}$
Mean Absolute Percentage Error (MAPE)	$\frac{\sum_{i=1}^{n}\left\lvert\frac{actual\ price_i - predicted\ price_i}{actual\ price_i}\right\rvert}{n}$

Table 4. Error measures

Tables 5 and 6 show the results of price prediction without clustering. Both ANFIS and MRA were used in the prediction and results of two different test scenarios are shown in the tables. The first set of results was obtained without MPC as an input variable and the second set of results with MPC as an input. One can make two observations regarding the results. First, the addition of MPC as an input variable improved the prediction accuracy for both ANFIS and MRA. For example the percentage of properties with cumulative MAPE less than or equal to 10% are 38.6% (20.1+18.5) of all the properties for ANFIS and 34% (17.6+16.4) for MRA. After the MPC was added as an

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

input, the prediction results improved to 41.5% for ANFIS and 38.5% for MRA. Second, in both cases ANFIS outperformed MRA. For example, with MPC added to the input, the percentage of properties with cumulative MAPE less than or equal to 10% improved from 38.6% to 41.5% for ANFIS and from 34% to 38.5% for MRA. Similarly for RMSE and MAE the addition of MPC improves the prediction results and again ANFIS outperformed MRA. See Table 6 for the results. These results demonstrate that the use of MPC as an input is likely to improve the prediction accuracy.

%	Without MPC as Input		With MPC as Input	
	ANFIS	MRA	ANFIS	MRA
≤ 5 MAPE	20.1	17.6	22.3	20.7
(5,10] MAPE	18.5	16.4	19.2	17.6
(10,15] MAPE	14.5	13.8	14.6	13.5
(15,20] MAPE	10.6	10.6	10.4	10.3
(20,25] MAPE	7.7	8.3	7.6	7.7
>25 MAPE	28.7	33.4	26	30.1
Total	100.0%	100.0%	100.0%	100.0%
Average MAPE	28.6	30.1	26.8	28.2

Table 5. MAPE prediction results without clustering

$	Without MPC as Input				With MPC as Input			
	RMSE		MAE		RMSE		MAE	
	ANFIS	MRA	ANFIS	MRA	ANFIS	MRA	ANFIS	MRA
Average	37742	41345	27805	31147	35289	37763	25747	27957
Max	39146	42094	28551	31687	36427	38740	26202	28687
Min	36904	40121	27107	30185	34399	36931	25103	27342
Std. Dev	519	452	336	368	444	437	272	358

Table 6. RMSE and MAE prediction results without clustering

Next we describe the prediction results after clustering. The modified FCM approach described in the Methodology section was applied to our dataset and as a result 10 clusters were created. Table 7 shows the descriptive statistics of the 10 clusters. Because of space constraints only the statistics for a select number of variables are shown. Three variables were used in clustering: Floor Size, Year-Built, and Basement Size. Other combinations of variables were tried but these three yielded the best clustering results. One can see that Cluster 1 includes relatively moderately priced houses with a mean price about $99,355 and with a mean year built of 1959. Cluster 4 contains the most

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

expensive houses and more recently built houses. Their mean price is around $347,533 and the average year built is 1999.

Cluster	n	Mean Sale Price($)	Mean MPC($)	Floor Size	Year-Built	Basement Size	Baths	Garage Size
1	3672	99355	100701	1135	1959	8	1.4	0.8
		25240	36667	225	6	51	0.8	0.9
2	4187	215815	217132	1869	2002	12	3.6	1.5
		77079	64190	469	4	71	0.7	0.7
3	2209	89539	88014	1322	1913	8	1.6	0.6
		61058	45555	386	11	52	0.9	0.8
4	767	347533	352924	2426	1999	1205	5.1	1.9
		84533	61462	479	6	239	1.0	0.3
5	812	135317	136655	1372	1980	21	2.5	1.0
		39853	26773	301	6	81	0.9	0.9
6	1318	106043	106434	1124	1942	34	1.5	0.8
		58943	46082	348	5	108	0.9	0.8
7	233	202665	204627	2166	1962	5	3.4	1.5
		63398	37381	250	7	37	0.9	0.8
8	538	138854	140421	1224	1958	703	2.2	1.1
		36400	25278	193	5	139	1.0	0.8
9	235	283115	282030	2184	1998	722	4.6	1.9
		58193	29115	337	6	83	1.0	0.3
10	2612	224894	223417	1959	1969	567	3.5	1.3
		138839	114806	957	28	548	1.6	0.9

Table 7. Descriptive statistics (the means and standard deviations) for clustered data

Table 8 and 9 show the MAPE results of prediction for all 10 clusters. Though only 3 variables are used in clustering, all 18 variables were used in testing, or price prediction. The tables show the prediction results for all the 10 clusters with and without the use of MPC as an input variable. Please note MPC was calculated after clustering using neighboring properties within the cluster.

Table 8 shows the MAPE results for ANFIS and Table 9 shows the MAPE results for MRA. Freddie Mac's (the Federal Home Loan Mortgage Corp.) criterion states that on the test data, at least half of the predicted sale prices should be within 10% of the actual prices (Fik, Ling, and Mulligan, 2003). The ANFIS MAPE results show that 2 of the 10 clusters meet the Freddie Mac criterion. These clusters (2 and 4) account for about 30% of the records. Results with MPC are slightly better than those without MPC as an additional input. Clusters 2 and 4 with MPC as an additional input have 66.7% and 55.3% of the records in the respective clusters with cumulative MAPE less than or equal to 10% and the same clusters yield 61.5% and 53.1% without the use of MPC as an additional input.

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

The cumulative MAPE results for MRA produce much better results with 5 clusters meeting the Freddie Mac criterion when MPC is used as an input. These are clusters 2, 4, 5, 8, and 9, accounting for about 40% of all the records. The results obtained without MPC as an input are not as good, with only 3 clusters meeting the Freddie Mac criterion.

%	MAPE Results with MPC as Input (k=10, u_{ij}=0.8)									
	1	2	3	4	5	6	7	8	9	10
Average	18.4	10.9	58.2	17.4	82.9	48.7	84.1	21.0	29.1	22.6
<= 5	22.8	38.3	9.4	30.4	15.6	12.7	12.3	21.3	17.2	20.5
(5,10]	20.8	28.4	9.3	24.9	14.6	12.6	13.2	20.3	15	18.5
(10,15]	16.4	15.7	9	17.7	12	11.6	11.4	17	12.4	16
(15,20]	12.5	8	8.3	10.3	9.6	10.2	10.3	11.9	11.2	12.4
(20,25]	7.8	3.7	7.9	5.7	7.8	8.8	7.8	8.7	8.4	8
>25	19.6	5.9	56	10.9	40.3	44.1	45	20.8	35.7	24.7
Without MPC as Input										
Average	21.2	11.8	75.0	15.2	34.0	52.6	80.3	24.1	24.9	23.5
<= 5	20.6	34.1	6.5	28.8	21	10.4	12.6	19.7	24.6	19.1
(5,10]	18.4	27.4	6.8	24.3	18.5	10.8	11.5	18.2	22.7	17.7
(10,15]	14.8	16.7	6.8	19.2	15.7	9.8	12	15	16.2	15.3
(15,20]	11.8	9.5	6.7	11.5	11.3	9.4	9.5	13.4	11.4	12.6
(20,25]	9	4.7	6.9	6	8	8.5	8.4	9.4	6.8	8.6
>25	25.3	7.6	66.2	10.2	25.4	51.1	46	24.3	18.4	26.7

Table 8. MAPE results for clustered data using ANFIS

%	Results with MPC as Input (k=10, u_{ij}=0.8)									
	1	2	3	4	5	6	7	8	9	10
Average	18.2	11.9	56.5	8.9	15.3	34.0	14.2	14.2	7.8	24.1
<= 5	23.1	34	9.7	37.2	27.9	14.2	24.7	26.7	41.5	18.4
(5,10]	21	27.2	10	28.9	22.1	13.9	20.9	23.4	29.2	17.7
(10,15]	16.2	17	9.8	18.3	16.4	12.6	17.3	19.8	15.6	15.3
(15,20]	12.4	10.5	8.8	8.2	12.1	10.8	13.2	11.5	8.2	12.5
(20,25]	8.1	4.4	8.3	3.2	7.7	9.6	9	6.5	3.8	8.5
>25	19.2	6.9	53.4	4.2	13.8	38.9	15	12	1.7	27.7
Results without MPC as Input										
Average	20.7	13.0	71.4	10.7	16.1	39.3	15.6	16.3	8.3	25.0
<= 5	20.6	29.7	6.9	30.5	26.1	11.3	22	21.6	40.1	17.2
(5,10]	18.7	24.3	7	26.5	22.3	11.1	19.7	20.7	28.1	15.8
(10,15]	15.5	17.8	7.1	20.5	16.3	11.3	16.2	18	15.9	15.5
(15,20]	12.1	12.2	7.2	10.9	12.1	10.6	13.5	13.9	8	12.6
(20,25]	9.1	6.6	7	4.9	7.6	9.5	10.2	9.2	5.2	9

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

>25	24.1	9.3	64.8	6.7	15.5	46.2	18.5	16.5	2.7	30

Table 9. MAPE results for clustered data using MRA

Significantly different patterns can be observed when analyzing the RMSE and MAE results with and without MPC as an additional input to the ANFIS and MRA models for clustered data (Tables 10 and 11). Table 10 shows that the ANFIS models with MPC as input generated the lowest average RMSE=20966 and MAE=15610 for Cluster 1, and the highest average RMSE=373130 and MAE=131273 for Cluster 7. The average RMSE and MAE over all 10 clusters with MPC are 125989 and 51723, respectively. Table 10 also depicts that the ANFIS models without MPC as input generated the lowest average RMSE=24252 and MAE=17986 for Cluster 1, and the highest average RMSE=378183 and MAE=129471 for Cluster 7. The average RMSE and MAE over all 10 clusters without MPC are 107562 and 46504, respectively.

Table 11 shows that for all 10 clusters MRA-based models produce much lower errors than the same models created with ANFIS. For example, the models with MPC yielded the lowest average RMSE=20655 and MAE=15429 for Cluster 1, and the highest average RMSE=47538 and MAE=36584 for Cluster 10; whereas the models without MPC generated the lowest average RMSE=23092 and MAE=17512 for Cluster 1, and the highest average RMSE=50471 and MAE=39306 for Cluster 10. The average RMSE and MAE over all 10 clusters with MPC are 31014 and 23691, respectively. The average RMSE and MAE over all 10 clusters without MPC are 35708 and 27203, respectively.

Thus for clustered data the MRA-based models produce significantly lower minimum, maximum, and average errors than ANFIS-based models. However, adding MPC as input to the ANFIS and MRA models reduces the minimum, maximum, and average errors only in the MRA-based models.

ANFIS Results with MPC (k=10, u_{ij}=0.8)										
	1		2		3		4		5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	20966	15610	27430	19271	39352	29696	177597	53826	341767	96650
Max	21762	16261	29419	20185	42634	31982	1121857	138641	998127	259150
Min	20099	14998	25305	17877	36674	27948	41584	31881	80434	37433
Std. Dev	427	303	762	470	1483	1014	192772	20332	189901	47211
	6		7		8		9		10	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	70400	31606	373130	131273	42061	24902	120484	79628	46707	34771
Max	265738	47729	1968765	527001	245094	59434	637260	298607	50263	37421
Min	34558	25403	49669	39435	24595	18879	53022	38054	44419	32986
Std. Dev	48703	4941	398401	109298	32450	6182	108661	52797	1396	1022
ANFIS Results without MPC as Input										
	1		2		3		4		5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE

Avg	24542	17986	31278	21655	52158	38723	116439	48160	132191	38746
Max	51014	20556	41023	23474	76836	44722	318738	76245	362078	69516
Min	22664	16998	28694	20671	47357	36332	46322	36081	24998	19379
Std. Dev	4058	597	2534	639	6053	1586	72435	8820	93514	13531
	6		7		8		9		10	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	74604	36460	378183	129471	52675	28887	163649	67773	49900	37181
Max	365356	68075	2267246	445391	147917	53655	1194930	368236	55765	40005
Min	39578	29977	53709	38548	27426	20000	31652	25525	46615	34707
Std. Dev	61183	6657	428252	103042	29659	6797	235130	66197	1825	1157

Table 10. RMSE and MAE results for clustered data using ANFIS

MRA with MPC as Input (k=10, u_{ij}=0.8)										
	1		2		3		4		5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	20655	15429	28623	20793	37295	28434	38053	29524	22656	17043
Max	21606	16065	30582	21718	39535	29940	43547	33590	26474	19237
Min	19578	14737	26810	19418	34888	26465	34122	26723	19802	15472
Std. Dev	473	329	799	505	1091	848	1889	1497	1429	880
	6		7		8		9		10	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	31309	23763	33935	26560	22658	17285	27416	21497	47538	36584
Max	34440	26275	41069	31796	25805	19672	34241	27039	49773	38321
Min	28527	21848	24542	19261	18687	14823	21130	15669	45506	34836
Std. Dev	1400	975	3926	3053	1466	1051	2931	2203	1054	903
MRA Results without MPC as Input										
	1		2		3		4		5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	23092	17512	32141	23846	46405	36593	44022	34870	24030	18121
Max	23934	18318	33733	25201	48861	38811	47771	38160	27646	19874
Min	22091	16852	30647	22663	44060	34560	39973	31777	21308	16352
Std. Dev	410	346	731	533	1132	993	1884	1536	1454	892
	6		7		8		9		10	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Avg	37228	28880	37983	29704	26136	20337	29269	22863	50471	39306
Max	39545	31186	45698	35691	29685	22813	35171	27945	52914	41382
Min	33959	26818	28489	22468	22858	17990	23032	18704	46981	36305
Std. Dev	1131	909	3505	2782	1512	1127	2664	2218	1274	1140

ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy

Table 11. RMSE and MAE results for clustered data using MRA

CONCLUSIONS

Data heterogeneity can reduce the predictive accuracy of data mining methods. Segmentation of a real estate market has been recognized as a viable method for addressing the issue of data heterogeneity. This paper proposes an innovative method for data clustering to improve predictive performance. The existing solutions of using spatial and statistics-based methods to segment/cluster the data have demonstrated the importance of clustering in improving price prediction. In this paper an approach based on an adaptation of the FCM algorithm, defined on a newly introduced homogeneity index, was used to create clusters that are more homogeneous. These resulting clusters were in turn used to predict sale prices. Two different classification methods were employed in the testing: the traditional and commonly used MRA approach, and ANFIS. As discussed in the results section, FCM-based clustering and the use of comparable properties yielded improved and interesting results. Though the prediction results are not uniform across all clusters, some clusters yield very good prediction results.

In addition to its promising performance when applied to our dataset this new clustering approach is also simple to implement. The required data are readily available in any real estate sale transaction. For example any assessment office in a local government already collects and tracks the type of property records used in this study. The adaptation of the FCM algorithm and the implementation of the homogeneity index are straightforward and do not require costly collection and modeling of data. And unlike those clustering methods that depend on the modeling of socioeconomic data that can be market specific, this new approach can be easily applied across different housing markets, thus potentially serving as a common tool for property valuation, risk assessment for lenders, and a consistent basis for government policy formulation.

Acknowledgment: This work was partially supported by the Prometeo Project of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador.

REFERENCES

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39*(2), 1772-1778.

Bagnoli, C., & Smith, H. C. (1998). The Theory of Fuzzy Logic and its Application to Real Estate Valuation. *Journal of Real Estate Research, 16*(2), 169-200.

Belasco, Eric, Farmer, Michael C, & Lipscomb, Clifford A. (2012). Using a Finite Mixture Model of Heterogeneous Households to Delineate Housing Submarkets. *Journal of Real Estate Research, 34*(4), 577-594.

Bezdek, J C, Ehrlich, R, & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2), 191-203.

Bourassa, S.C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *Journal of Real Estate Resach, 32*(2), 139-159.

Bourassa, Steven C, Hamelink, Foort, Hoesli, Martin, & MacGregor, Bryan D. (1999). Defining housing submarkets. *Journal of Housing Economics, 8*(2), 160-183.

Byrne, P. (1995). Fuzzy analysis: A vague way of dealing with uncertainty in real estate analysis? *Journal of Property Valuation and Investment, 13*(3), 22-41.

Dale-Johnson, D. (1982). An alternative approach to housing market segmentation using hedonic price data. *Journal of Urban Economics, 11*(3), 311-332.

Do, A. Q., & Grudnitski, G. (1992). A Neural Network Approach to Residential Property Appraisal. *The Real Estate Appraiser, 58*(3), 38–45.

Fik, T. J., Ling, D. C., & Mulligan, G. F. (2003). Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach. *Real Estate Economics, 31*(4), 623-646.

Fletcher, M., Gallimore, P., & Mangan, J. (2000). The modelling of housing submarkets. *Journal of Property Investment & Finance, 18*(4), 473-487.

Gonzalez, AJ, & Laureano-Ortiz, R. (1992). A case-based reasoning approach to real estate property appraisal. *Expert Systems With Applications, 4*(2), 229-246.

Goodman, A.C., & Thibodeau, T.G. (1998). Housing market segmentation. *Journal of Housing Economics, 7*(2), 121-143.

Goodman, A.C., & Thibodeau, T.G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics, 12*(3), 181-201.

Goodman, A.C., & Thibodeau, T.G. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics, 35*(2), 209-232.

Guan, J., Zurada, J., & Levitan, A.S. (2008). An Adaptive Neuro-Fuzzy Inference System Based Approach to Real Estate Property Assessment. *Journal of Real Estate Research, 30*(4), 395-420.

Kilpatrick, J. (2011). Expert systems and mass appraisal. *Journal of Property Investment & Finance, 29*(4/5), 529-550.

Mark, J., & Goldberg, M. (1988). Multiple regression analysis and mass assessment: A review of the issues. *Appraisal Journal, 56*(1), 89-109.

McCluskey, W. J., Cornia, G. C., & Walters, L. C. (2012). *A primer on property tax: Administration and policy*: John Wiley & Sons.

Michaels, R.G., & Smith, V.K. (1990). Market segmentation and valuing amenities with hedonic models: the case of hazardous waste sites. *Journal of Urban Economics, 28*, 223-242.

Peterson, S., & Flanagan, A. B. (2009). Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. *Journal of Real Estate Research, 31*(2), 147-164.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843-2852. doi: DOI 10.1016/j.eswa.2008.01.044

Stevenson, S. (2004). New empirical evidence on heteroscedasticity in hedonic housing models. *Journal of Housing Economics, 13*(2), 136-153.

Straszheim, M. (1974). Hedonic estimation of housing market prices: A further comment. *The Review of Economics and Statistics, 56*(3), 404-406.

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

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

Zurada, J., Levitan, A.S., & Guan, J. (2011). A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research, 33*(3), 349-387.

  ISSN: 1543-5962-Printed Copy ISSN: 1941-6679-On-line Copy