Using Fuzzy Clustering Methods for Delineating Urban Housing Submarkets

Sungsoon Hwang
Department of Geography
DePaul University
990 W Fullerton Ave, Suite 4300
Chicago, IL 60614
+1-773-325-8668
shwang9@depaul.edu

Jean-Claude Thill
Department of Geography & Earth Sciences
University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223
+1-704-687-5909
jfthill@uncc.edu

ABSTRACT
This study investigates whether a fuzzy clustering method is of any practical value in delineating urban housing submarkets relative to clustering methods based on classic (or crisp) set theory. A fuzzy *c*-means algorithm is applied to obtain fuzzy set membership degree of census tracts to housing submarkets defined within a metropolitan area. Issues of choosing algorithm parameters are discussed on the basis of applying fuzzy clustering to 85 metropolitan areas in the U.S. The comparison between results of fuzzy clustering and those of crisp set counterpart shows that fuzzy clustering yields statistically more desirable clusters.

Categories and Subject Descriptors
I.5.3 [**Pattern Recognition**]

General Terms
Algorithms, Performance

Keywords
Fuzzy clustering, Data mining, Housing submarket, GIS

1. INTRODUCTION
This study attempts to examine the potential of fuzzy clustering in enriching methods for identifying housing submarkets. Identifying housing submarkets have many pragmatic values. It concisely reveals socio-demographic structure of city, and helps understand geographic process surrounding neighborhood formation. The most commonly used method for identifying housing submarkets (see [19] for the survey) is based on the hedonic analysis followed by cluster analysis. Hedonic analysis [11] [13] is used to derive dimensionality of housing market by estimating what array of attributes would be significant factors influencing housing price; those predictors of housing price provide input vector for cluster analysis (or clustering), defining the dimension of attribute space in which clustering is conducted.

Clustering methods can be roughly classified into hierarchical method and partitioning method [20]. Partitioning clustering can be divided into exclusive and overlapping methods depending on which set theory the algorithm is built on. Exclusive clustering (e.g., *k*-means) is built on classic set theory where an element is an exclusive member of a set. Overlapping clustering is based on fuzzy set theory where an element can be a member of one or more sets. For instance, some aggregate unit of housing (mainly delineated by census unit) is composed of a mix of different housing types and diverse demographics. In such cases, it is logical to consider that the housing unit belongs to more than one housing submarket. This study examines whether fuzzification offers any advantage to methodology for (housing) market segmentation over counterpart based on classic set theory.

The validity of clustering results varies by algorithm parameter values. The effect of parameter values (such as the number of clusters c and fuzziness exponent m) on clustering results has not been entirely understood by any token [14]. Recent years have observed a growth in researches on cluster validity index; body of literature under this line of research aims at developing an index by which the optimality of parameter values is determined [1] [2] [21] [18]; see [10] for the survey of cluster validity index. In practice, choosing the optimal number of housing submarkets or degree of fuzzification for classification is not a straightforward task. It is unquestionable that clustering algorithms should be equipped with a mechanism for self-validation. This study explores the route in which way fuzzy clustering can be improved with regard to validating results in response to parameter values.

The objectives of this study are threefold. First, this paper will demonstrate how fuzzy clustering is applied to identifying housing submarkets. The methodology described in this paper can be generalized as market segmentation techniques. We focus on classifying aggregate census units to housing submarkets drawing upon researches from housing studies and pattern recognition field. It is notable that this study can fill the gap in the application of fuzzy set theory to social science. Second, the

ACMGIS'07, November 7-9, 2007, Seattle, WA Copyright 2007 ACM
ISBN 978-1-59593-914-2/07/11...$5.00.

paper presents findings on empirical behaviors of fuzzy *c*-means algorithm in the particular application over a range of parameter values. Focus will be placed on validating clustering results on the basis of cluster validity index, as a means to make the choice of parameter values non-arbitrary. Results further allow us to make comparative assessment of fuzzy clustering in different application areas (e.g., housing market segmentation using census data vs. land use classification using satellite image). Third, the study examines whether fuzzy clustering outperforms traditional clustering methods based on classic set theory. 85 empirical cases will help evaluate the performance of fuzzy clustering. Such way, the utility of fuzzy clustering in (housing) market segmentation techniques can be better informed.

Remainder of this paper is organized as follows. We begin with describing a fuzzy *c*-means algorithm that forms the basis of this study in Section 2. Then the fuzzy clustering-based methodology for classifying housing market is described in Section 3. In Section 4, the methodology is illustrated using the case of Buffalo-Niagara Falls MSA (Metropolitan Statistical Area). We report statistical significance test results to determine whether fuzzy clustering produces better results than traditional methods based on 85 sample metropolitan areas in the U.S. Finally, we conclude this study by summarizing the work, and remarking on future research.

2. FUZZY CLUSTERING

In operational terms, fuzzy clustering can be described as the problem of determining the fuzzy set membership of data point k to cluster i [3]. The total number of data points can be denoted by n, and the a priori specified number of clusters can be denoted by c. Fuzzy partition U of data points is obtained such that the following objective function can be minimized:

$$\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^{m}\left\|x_k - v_i\right\|_A^2 \qquad \text{(Eq. 1)}$$

where u_{ik} is the membership degree of data point k to cluster i, x_k is the vector of data point, v_i is the ith cluster center ($1 \le i \le c$), and m is the fuzziness exponent ($m \in [1, \infty]$). Cluster membership is fuzzier when m is larger. If m is 1, it becomes hard (i.e., crisp set-based) clustering. In a sense, hard clustering (e.g., k-means) can be seen as a special case of fuzzy clustering. Therefore, both methods can be described under the rubric of fuzzy clustering algorithm. Necessary conditions for solutions to Eq. 1 are

$$v_i = \sum_{k=1}^{n}(u_{ik})^{m}\,x_k \Big/ \sum_{k=1}^{n}(u_{ik})^{m} \qquad \text{(Eq. 2)}$$

$$u_{ik} = \left(\sum_{j=1}^{c}\left(\frac{\left\|x_k - v_i\right\|}{\left\|x_k - v_j\right\|}\right)^{2/(m-1)}\right)^{-1} \qquad \text{(Eq. 3)}$$

where u_{ik} and v_i are updated at each iteration until the difference between the current u_{ik} and the previous u_{ik} is less than a predefined cut-off threshold ε_L (i.e., convergence criteria). The fuzzy *c*-means algorithm can be stated as follows:

Algorithm 1 Fuzzy c-means algorithm from [3]

Step 1: Fix c, $2 \le c < n$; and fix m, $1 \le m < \infty$, Initialize $U^{(0)}$ so that it becomes the fuzzy matrix. Then at step l, l = 0, 1, 2, ….;

Step 2: Calculate the c fuzzy cluster centers $\{v_i^{(l)}\}$ with (Eq. 2) and $U^{(l)}$

Step 3: Update $U^{(l+1)}$ using (Eq. 3) and $\{v_i^{(l)}\}$

Step 4: Compare $U^{(l)}$ to $U^{(l+1)}$ in a matrix norm; if $\|U^{(l+1)} - U^{(l)}\| \le \varepsilon_L$ stop. Otherwise, return to Step 2.

Algorithm 1 attempts to find the optimal fuzzy partition given a fixed value of c. The fuzzy *c*-means algorithm stated above does not address what value of c is optimal for the fuzzy partitioning of the instances. The criterion used to evaluate the optimal number of cluster can be formulated as a fuzzy cluster validity index. Cluster validity index has been proposed by many researchers since fuzzy clustering was introduced. Despite recent growth in this line of research, the search for a robust validity index remains open.

The literature on fuzzy clustering algorithms fails to suggest any standard procedures for determining an appropriate value of m. Instead, the rule of thumb of a value in the range [1.5, 2.5] is recommended in the pattern recognition literature [14]. It is not known whether such a range would work for different domains such as the classification of geographically defined housing markets. Therefore, empirical tests are needed to calibrate the range of fuzziness exponent m that is appropriate to the application in hand.

3. DELINEATING HOUSING SUBMARKETS

In this section we present the methodology for delineating housing submarkets defined within a metropolitan area. The problem is to partition a metropolitan area (as a single housing market) into the most homogeneous housing submarkets. The methodology is divided into two steps. The first step is to identify what constitutes dimensionality of housing market. Hedonic analysis is conducted to extract predictors of housing price. The predictors form the vector of data point x_k required for the next step, that is clustering. We extend fuzzy *c*-means algorithm so that the optimality of parameters c and m can be tested. Therefore, fuzzy clustering algorithm to be described below maps the fuzzy set membership degree of each data point (census tract in this study) to housing submarkets given a range of parameter values c and m.

3.1 Hedonic Analysis

The considerable literature on hedonic analysis indicates that housing price is associated with socioeconomic characteristics of residents, structural characteristics of housing units, and locational characteristics of neighborhoods [5] [7]. An array of variables considered to determine house prices are compiled at a census tract level in this study. Candidate explanatory variables for hedonic analysis are listed in Table 1. They encompass economic indicator, educational attainment, occupation, life cycle, characteristics of housing units, ethnicity, length of residence, school quality, crime, and job accessibility.

Data sources for this analysis include the U.S. Population and Housing Census, school district surveys, crime reports, and Census Transportation Planning Package. Variables unavailable

at the level of Census Tract (e.g., crime, school quality) are aggregated to Census Tract level by spatial overlay in Geographic Information Systems (GIS). Job accessibility (jobacm) is calculated following the gravity model [9] where travel time is used as a measure of spatial separation. Spatial impedance parameters are calibrated using maximum likelihood estimation [8].

Table 1 Candidate Predictors of Housing Price

VarNM	Variable Definition	Data	Year
pcincome	per capita income	Census	2000
college	% college degree	Census	2000
managep	% management workers	Census	2000
prodp	% production workers	Census	2000
famcpchl	% family with children	Census	2000
nfmalone	% nonfamily living alone	Census	2000
black_p	% black	Census	2000
nhwht_p	% non-hispanic white	Census	2000
nativebr	% native born	Census	2000
medroom	median number of room	Census	2000
hudetp	% detached housing unit	Census	2000
yrhublt	median year structure built	Census	2000
ptratio	pupil to teacher ratio	NCES*	2002
schexp	school expenditure per student	NCES	2002
vrlcrime	violent crime rate	FBI**	2003
prpcrime	property crime rate	FBI	2003
jobacm	job accessibility	CTPP***	2000

* National Center for Education Statistics, Common Core of Data

** FBI's annual report, "Crime in the United States 2003"

*** Census Transportation Planning Package

Stepwise regression is employed as a method of hedonic analysis to control for multicollinearity. Stepwise regression is chosen over factor analysis mainly because results of statistical analysis are more easily interpreted [15]. The dependent variable of the census tract-level hedonic analysis is the median price of owner-occupied housing units. U.S. homeownership rate was 68.6 percent in the fourth quarter of 2003 [12]. When the value of dependent variable is missing, the value for the corresponding census tract is either excluded (e.g., park) or derived for the analysis using the simple regression model based on the relationship with median rent (e.g., downtown). This treatment of missing values turned out to make cluster analysis results more reliable. Clustering results were highly sensitive to untreated outliers.

3.2 Clustering

The result of hedonic analysis (i.e., a set of significant price predictors selected from Table 1) is fed into fuzzy clustering. The vector comprised of selected predictors of house price is converted into *z*-score for normalization. The dimensions of clustering algorithms will vary since the number of selected predictors will differ by a metropolitan area. For instance, it is likely that large metropolitan areas will be divided into housing submarkets in attribute space of higher dimensionality than small metropolitan areas. Likewise, the optimal number of housing submarkets (denoted by c^*) in a particular metropolitan area is likely to vary. It is quite conceivable that the quality of clustering will vary as the number of cluster c changes. This leads us to examine the procedure for the statistically superior choice of c.

Cluster validity index is used to determine the optimal number of clusters c^* and m^*. In general, cluster validity indices measure statistical properties of clustering results; they are largely formulated as a function of compactness within a cluster and/or separation between clusters. In the subsequent sections, we review four cluster validity indices that are considered for calibrations. We rationalize the choice of validity index through calibration test, which allows us to modify Algorithm 1. The extended algorithm (Algorithm 2) determines the fuzzy set membership degree of data point x_k to submarket v_i given statistically better value of c and m, rather than given any arbitrary c and m.

3.2.1 Cluster Validity Index

After an in-depth review of the literature and pilot tests, four validity indices are chosen for calibrations. The goal is to assess which validity index would provide the most robust method for validating clustering results. The four indices are the partition coefficient [1], the partition entropy [2], the Xie-Beni Index [21], and the SVi index [18]. Their mathematical formulations are given in Eq. 4 to 7.

$$PC(U) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^2}{n} \qquad \text{(Eq. 4: Partition coefficient)}$$

$$PE(U) = -\frac{\sum_{i=1}^{c}\sum_{k=1}^{c}[u_{ik}\log_2(u_{ik})]}{n} \qquad \text{(Eq. 5: Partition entropy)}$$

$$U_{XB} = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^2\left\|x_k - v_i\right\|_A^2}{n\min_{i,j}\left\|v_i - v_j\right\|^2} \qquad \text{(Eq. 6: Xie-Beni index)}$$

$$S_{VI} = \frac{\sum_{i=1}^{c}\dfrac{\sum_{k=1}^{n}(u_{ik})^m\left\|x_k - v_i\right\|_A^2}{\sum_{k=1}^{n}u_{ik}}}{\sum_{i=1}^{c+1}\sum_{j=1}^{c+1}(\mu_{ij})^{(2+w)/w}\left\|z_j - z_i\right\|_A^2} \qquad \text{(Eq. 7: Svi index)}$$

where w is set to 2 in this study

where

$$[z_1, z_2, ..., z_c, z_{c+1}]^T = [v_1, v_2, ..., v_c, \bar{x}]^T$$

$$1 \le i \le c+1, 1 \le j \le c+1, j \ne i$$

$$\mu_{ij} = \frac{1}{\sum_{\substack{l=1\\ l\ne j}}^{c+1}\left(\dfrac{\left\|z_j - z_i\right\|_A}{\left\|z_j - z_l\right\|_A}\right)^w}$$

3.2.2 Calibration Test Results: How Many Clusters?

A good validity index should meet the following requirements: (1) the index should not exhibit a monotonic tendency with the increasing value of *c*. A monotonic tendency would always favor either the smallest or the largest value of *c*; (2) changes in the index value should be relatively stable within the reasonable range of *c*. It is not expected that the index value would exhibit an erratic pattern (i.e., pulse-like) for small ranges of *c*; (3) fuzzy

partitioning results should be somewhat in accordance with critical element of x_k. For example, spatial arrangement of housing submarkets is likely to resemble spatial distribution of income to some degree.

Figure 1 shows typical calibration test results where m is fixed to 1.5. Figure 1 plots how the values of validity indices change with the increasing value of c as a line graph. The value of the Xie-Beni index (shaded in blue with diamond marker) does not change either monotonically nor exponentially. Xie-Beni index predicts c^* at 6 or 7. This pattern is consistent across study set (that is 85 metropolitan areas). Empirical tests suggest that the Xie-Beni index is most robust.

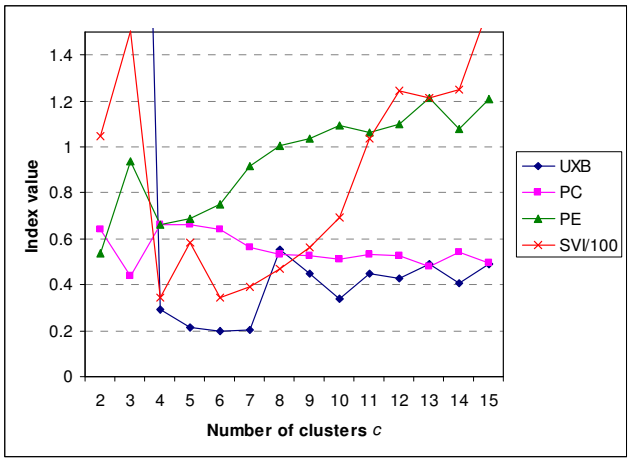


Figure 1 Cluster validity index with *c*

3.2.3 Calibration Test Results: How Much to Fuzzify?

Pilot tests for determining m^* across a range of m [1, 5] given validity indices suggest that the partitioning with m larger than 2 does not lend itself to an intuitive interpretation of fuzzy set membership degree[1]. Thus, m values greater than 2 are not considered in this study. In addition, the partition coefficient (PC) and partition entropy (PE) are not considered for calibration tests because the relation of m with PC and PE is predictable[2].

Table 2 tabulates the changing values of two validity indices in rows across a range of m in a range [1, 1.9] incremented by 0.1 in columns. Index value shown in each cell is the minimum value among those computed in a range of c [2, c_{max}] given the m in columns. m^* is determined when the index values reach the optimum (minimum in the case of two indices shown) over a range of m, as underlined in Table 2. Table 3 shows how predicted c^* values change over a range of m. According to the Xie-Beni index, the optimal fuzziness exponent m^* is 1.5 and the optimal number of cluster c^* is 6. The Svi index predicts that the partitioning may be optimal when m = 1.1 and the optimal number of clusters is 3. However, the Svi index shows the tendency to monotonically increase as m increases.

Table 2 The values of cluster validity indices over a range of *m*

	1	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**
UXB	0.354	0.305	0.266	0.239	0.21	**0.196**	0.222	0.233	0.487	0.675
SVI/100	0.179	**0.178**	0.191	0.213	0.247	0.344	0.433	0.429	0.897	0.81

Table 3 provides auxiliary information on the validity of index. Stability of c^* values over a range of m is considered desirable because it is quite unlikely that the optimal number of clusters would change abruptly over a range of m values. 6, 7, and 10 are candidates for c^* according to the Xie-Beni index. On the other hand, 3, 6, 7, 10, and 12 are candidates for c^* given the Svi index which lacks stability of c^*.

Table 3 Optimal number of clusters *c given *m***

	1	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**
UXB	6	7	7	7	7	**6**	7	7	10	10
SVI/100	3	**3**	3	3	3	6	6	7	10	12

It can be noted from Table 2 that the Xie-Beni index values exhibit U-curve pattern within the range of m [1, 1.9]. Most interestingly, this U-curve pattern was duplicated in other cases. It indicates that optimal clustering results can be obtained between the range of m values 1 and 1.9. However, readers should be cautioned against generalizing it given the number of validity index calibrated, and constrained range of m values tested. It remains to be seen whether this pattern persists beyond the range [1, 1.9] in future research.

3.2.4 Extended Fuzzy Clustering Algorithm

Algorithm 1 can be extended to address cluster validity issues discussed above. The extended algorithm (Algorithm 2) has two more outer loops in comparison with Algorithm 1; one for checking the validity index in a range of m, and the other for checking the validity index in a range of c. Thus, Algorithm 2 has three nested loops: m by c by Algorithm 1. The total number of iterations will be x^*y^*z where x is $(m_{max}\text{-}1)/m_{inc}$, y is c_{max}-2, and z is the number of iteration l determined at step 5. It can be noted that the choice of validity index v is central to internalizing a validation mechanism of fuzzy c-means clustering algorithm.

Algorithm 2 Extended fuzzy *c*-means algorithm

Step 1: Initialize the parameters related to fuzzy partitioning: $c = 2$ ($2 \le c < c_{max}$), $m = 1$ ($1 \le m < m_{max}$), where c is an

[1] For instance if m becomes larger than 2, test results show that membership degree of the most extreme case (let's say census tract k with highest income in a metropolitan area) to a particular cluster becomes fairly small (even less than 0.5), to the point that census tract k cannot be interpreted as a distinct submarket any more.

[2] A large m makes the value of u_{ik} less variable (i.e., for instance, u_{ik} = {0.25, 0.25, 0.25, 0.25} instead of u_{ik} = {1, 0, 0, 0}), leading to a small value of the partition coefficient and a large value of the partition entropy. As a consequence, the smallest value of m (i.e., 1) is always favored according these two indices.

integer, m is a real number; Fix m_{inc} where m_{inc} is incremental value of m ($0 < m_{inc} \leq 0.1$); Fix cut-off threshold ε_L; Choose validity index v

Step 2: Given c and m, initialize $U^{(0)}$ so that it becomes the fuzzy matrix. Then at step l, $l = 0, 1, 2, \ldots$;

Step 3: Calculate the c fuzzy cluster centers $\{v_i^{(l)}\}$ with (Eq. 2) and $U^{(l)}$

Step 4: Update $U^{(l+1)}$ using (Eq. 3) and $\{v_i^{(l)}\}$

Step 5: Compare $U^{(l)}$ to $U^{(l+1)}$ in a convenient matrix norm; if $\| U^{(l+1)} - U^{(l)} \| \leq \varepsilon_L$ to go step 6; otherwise return to Step 3.

Step 6: Compute the validity index v_c^m for given c and m

Step 7: If $c < c_{max}$, then increase $c \leftarrow c + 1$ and go to step 3; otherwise go to step 8

Step 8: If $m < m_{max}$, then increase $m \leftarrow m + m_{inc}$ and go to step 3; otherwise go to step 9

Step 9: Obtain the optimal validity index $v_{c^*}^{m^*}$ from v_c^m, optimal number of clusters c^*, and optimal amount of fuzziness exponent m^*; The optimal fuzzy partition U is obtained given c^* and m^*

4. RESULTS

4.1 Descriptive summary

Algorithm 2 was implemented to identify housing submarkets of each of 85 metropolitan areas where Xie-Beni index is set to validity index v. 85 metropolitan areas (Figure 2) are random samples stratified by four Census Regions (i.e., Northeast, Midwest, South, and West) and population (i.e., over a million, 250K-1 million, 100K-250K, and less than 100K). They include 75 MSA and 10 CMSA.

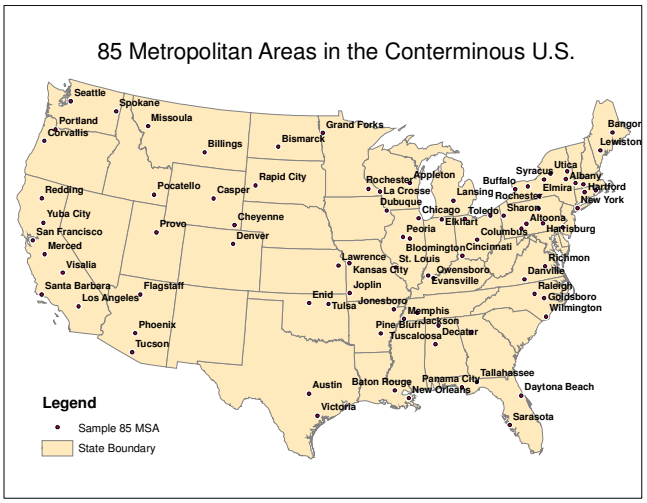


Figure 2 Study Set: 85 Metropolitan Areas

Being space limited, descriptive statistics of clustering results will be presented here. In Figure 3, p denotes the dimension of clustering, $c\_$ denotes the optimal number of clusters, and $m\_$ denotes optimal fuzziness exponent.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
p	85	1	11	4.05	2.464
c_	85	2	10	5.33	2.195
m_	85	1.0	1.9	1.381	.2353
Valid N (listwise)	85				

Figure 3 Descriptive Statistics of Clustering Results

Before looking into descriptive statistics, readers should be warned that the maximum value of c denoted by c_{max}, in Algorithm 2 is set to the minimum value among square root of n and 10, instead of squared root of n. It will necessarily set c_{max}, to 10 when n is sufficiently large. The computation time for fuzzy classification of very large metropolitan areas such as New York or LA, exponentially increases over large values of c. Implementing Algorithm 2 over a range of c values greater than 10 is left to future research. Figure 4 shows frequency distribution of m^* over 85 empirical cases. The average m^* is 1.38, and the median is 1.4; it indicates that optimal fuzziness amount peaks around 1.4 in this particular application. The histogram will provide a reference from which the choice of appropriate range of m can be made in applications alike.

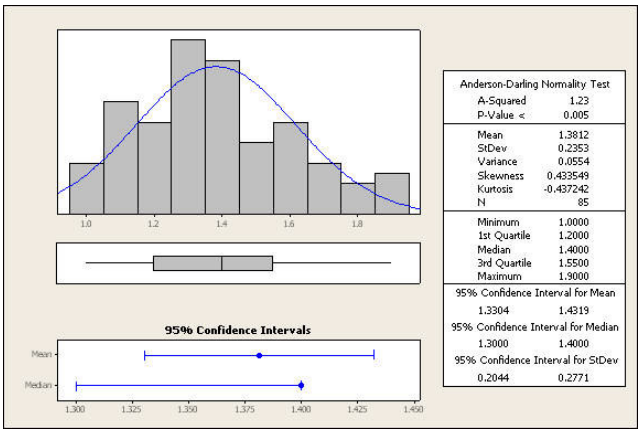


Figure 4 Histogram of m^* derived from 85 empirical cases

To ensure that clustering results do not counter general observation, we conduct the correlation analysis between log of population and parameter values shown in Figure 3. The population of metropolitan areas is positively correlated with the dimension of housing market p (Pearson Correlation .835 with p-value .000) and the optimal number of clusters c^* (Pearson Correlation .560 with p-value .000). It indicates that housing market of large metropolitan areas is more complex (i.e., more factors influencing submarkets formation) and diverse (i.e., divided into more submarkets) than small metropolitan areas.

4.2 Illustrations

This section illustrates the methodology described in Section 3 using the case of Buffalo-Niagara Falls MSA (Buffalo hereafter). Figure 5 shows how home value is spatially distributed over 294 census tracts in Buffalo. Eight (therefore p is 8 in this case) variables (among variables listed in Table 1) are identified as significant predictors of housing price in 85 percent of the variation; those are per capita income (pcincome), the percentage of college degree holder (college), the percentage

of married couples with children (famcpchl), the percentage of detached housing units (hudetp), median year of housing structure built (yrhublt), the percentage of non-hispanic white (nhwht_p), the percentage of native-born residency status (nativebr), and job accessibility (jobacm). The hedonic variables are shown in *x*-axis of Figure 6 in the order listed above.

Z-scores of the 8 explanatory variables in the Buffalo hedonic model form data vector x_k. Following step 9 in Algorithm 2, the optimal validity index $v_{c^*}^{m^*}$ value given Xie-Beni index is 0.3385 as double-underlined in Table 4. Therefore, the optimal number of clusters is 3, and the optimal fuzziness exponent is 1.3. It means that the Buffalo housing market is partitioned into three submarkets with a fuzziness exponent of 1.3. The second best alternative would be five submarkets with fuzziness exponent of 1.3 as underlined in Table 4.

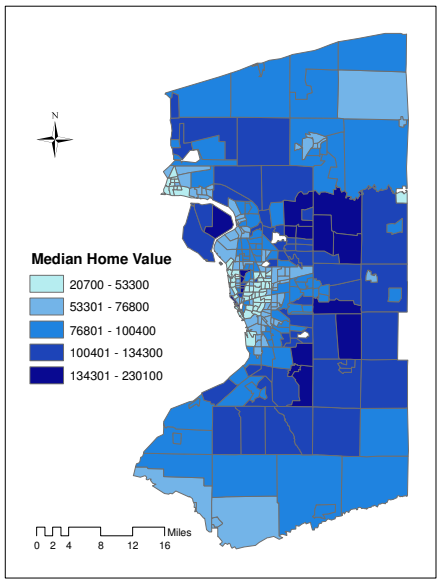


Figure 5 Choropleth map of median home value in the Buffalo-Niagara Falls MSA (Metropolitan Statistical Area)

Table 4 Validity index values v in the matrix c by m in Buffalo-Niagara Falls MSA

c / m	1	1.1	1.2	**1.3**	1.4	1.5	1.6	1.7	1.8
2	0.47	0.46	0.44	8.098	10.4	12.54	14.4	16.1	17.5
3	0.41	0.39	0.35	**0.339**	10.8	12.91	14.8	16.4	17.8
4	0.78	0.71	0.61	0.524	1.32	6.884	7.48	8.04	8.56
5	0.56	0.56	0.59	0.612	0.47	**0.34**	0.65	0.69	0.72
6	0.62	0.76	1.02	0.817	0.69	1.339	1.41	1.48	1.56
7	0.88	0.69	0.69	0.602	0.62	0.952	2.44	2.63	2.83
8	0.6	0.59	0.57	0.523	0.4	0.738	0.89	1.24	1.29
9	0.97	0.62	0.48	0.487	0.85	1.402	1.42	1.83	1.86
10	0.71	0.6	0.66	0.587	0.59	1.347	1.51	1.69	1.82
c*	3	3	3	3	8	5	5	5	5

The optimal fuzzy partitioning given c^* and m^* produces the cluster means v_i and the membership degree matrix u_{ik}. The cluster means v_i is visualized as parallel coordinate plot in Figure 6. A vector of cluster means values of each housing submarket in *y*-axis is graphed as three separate lines across eight dimensions in *x*-axis.

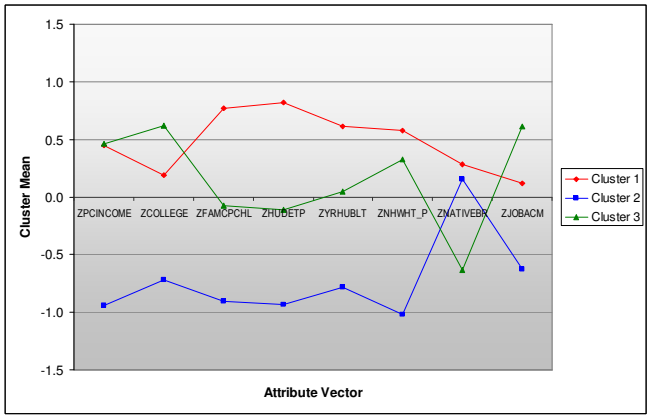


Figure 6 Cluster means from the optimal fuzzy partitioning of the Buffalo housing market where c^*=3, m^*=1.3, p=8, n=294

Cluster 1 (shaded in red, marked by a circle, and adequately named "Settled Suburbia") in Buffalo is largely characterized by moderate to high income, relatively high proportion of married couples with children, recently built detached houses, Caucasian residents, long residency, and moderate job accessibility. Figure 7A maps the fuzzy set membership degree of each census tract to cluster 1. A majority of census tracts located in outlying areas of the metropolitan area are classified to cluster 1 in a large extent. Census tracts' membership degrees to cluster 1 decrease toward the central city.

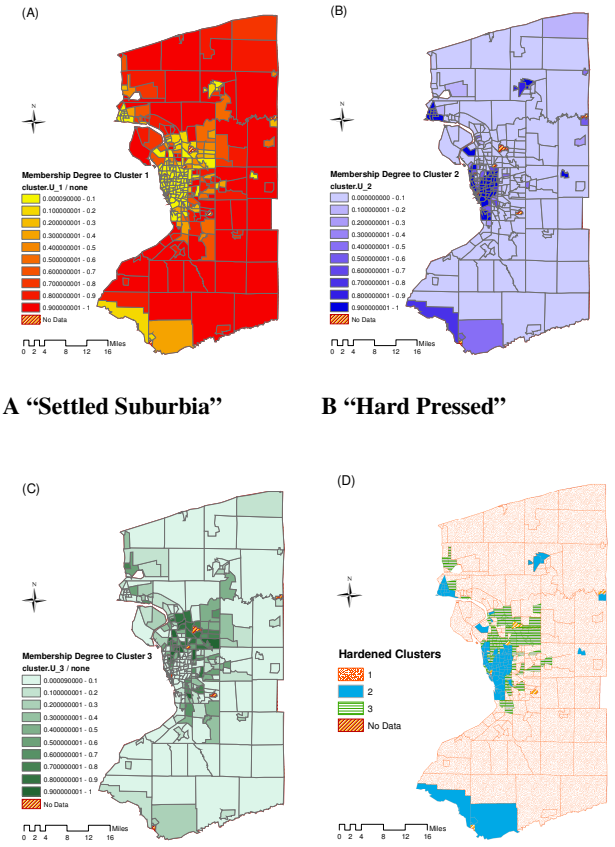
Cluster 2 (shaded in blue, marked by a rectangle, and adequately named "Hard Pressed") shows the opposite case of cluster 1. Residents living in cluster 2 are more likely to be poor, non-white, single, with lack of higher education. They are also likely to live in residential areas largely composed of old attached houses, and lack access to job opportunities. The spatial pattern of the membership degree to cluster 2 is highly clustered, as shown in Figure 7B. Census tracts with high membership degrees to cluster 2 are concentrated distinctively in the main street corridor linking downtown Buffalo to SUNY South Campus (where the 7-mile light railway lies), the west side of the City of Buffalo, Lackawana (south of Buffalo), the City of Niagara Falls, the northwest side of Lockport, and the Indian Reservation of Cattaraugus.

Cluster 3 (shaded in green, marked by a triangle, and adequately named "Prosperous Young Professionals") shares some characteristics with cluster 1; those are income, education and ethnic composition (Figure 6). But Cluster 3 is differentiated from Cluster 1 by less proportion of detached housing units. Cluster 3 is distinct from other two clusters, being characterized by high job accessibility (proximity to interstate highway) and lower length of residence (high proportion of foreign born near SUNY North Campus). A map of the membership degrees to cluster 3 is shown in Figure 7C.

From maps of fuzzy set membership degrees to different submarkets, it can be seen that some residential neighborhoods

are highly homogeneous (e.g., Main Street corridor), while others are rather heterogeneous (e.g., Amherst area around SUNY North Campus). For example, Amherst has relatively fuzzy (i.e., 0.5 rather than 0.9) membership degrees to both Cluster 1 and Cluster 3; it shares some characteristics of Cluster 1 and Cluster 3 simultaneously, which can be better examined by comparing between Figure 7A and Figure 7C.

Hardened clusters are plotted in Figure 7D where census tracts exclusively belong to one of the three clusters based on the maximum membership degree. The hardened cluster map summarizes exclusive membership degrees to three different submarkets. The other three maps (Figures 7A, 7B, and 7C) provide auxiliary information on variation in membership degrees to three submarkets that is not revealed in the map of hardened membership.



A "Settled Suburbia" **B "Hard Pressed"**

C "Prosperous Young Professional" **D Hardened Clusters**

Figure 7 Choropleth maps of fuzzy set membership degree to three housing submarkets in the Buffalo-Niagara Falls MSA

4.3 Does Fuzzy Clustering Outperform Hard Clustering?

Now the question is "does fuzzy clustering improve classifying housing market?" Addressing this question requires reference data considered to represent ground truth. Unlike other kinds of classification tasks, housing submarkets are not likely to be observed to bare eyes, but rather in the eye of experts. Groundtruthing can be more appropriately done through ethnographic methods. However, the goal of this paper is to evaluate the performance of fuzzy clustering in statistical terms. This task is supported by sufficiently large degree of freedom.

To achieve this goal, we compare the sum of squared errors that results from exclusive (or hard) partitioning and fuzzy partitioning, respectively. The sum of squared errors is the weighted sum of intra-cluster variations:

$$\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^2 \left\|x_k - v_i\right\|_A^2 \qquad \text{(Eq. 8)}$$

Equation 8 will be denoted by J_2 for simplicity hereafter. Large J_2 means that the clustering algorithm yields less compact clusters. Obviously, it is not desirable to have a large J_2 statistic. Good clustering algorithms should produce clusters as compact as possible. Readers can be reminded that the objective of the clustering task is equivalent to reducing J_2.[3]

As suggested earlier, exclusive clustering can be seen as a special case of fuzzy clustering. Fuzzy clustering is reduced to hard clustering if the fuzziness exponent m is set to 1. Therefore, a hard J_2 is computed from the optimal fuzzy partitioning when m equals 1. A fuzzy J_2 comes from the optimal fuzzy partitioning given m^*. To control for the effect of other parameters, the two J_2 statistics are calculated with the same set of parameters (except for m), including the number of optimal clusters c^*, cut-off threshold ε_L, and so on. The descriptive statistics for both hard and fuzzy J_2 denoted by j2_hcm and j2_fcm, respectively is shown in Figure 8.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	j2_hcm	1026.546	85	3848.268377	417.4033
	j2_fcm	745.7332	85	3022.266891	327.8109

Figure 8 Descriptive statistics of evaluation measures

A paired samples test is conducted to evaluate if the mean of fuzzy J_2 is significantly different (smaller thus more desirable) from that of hard J_2 . The mean of the hard J_2 statistic is 1026.546 while the fuzzy J_2 mean is 745.7332. T-statistics for the mean difference of 280.8133 is 2.828 with a p-value of .006. The significance test shows that the sum of squared errors resulting from fuzzy clustering methods is significantly less than that resulting from a hard clustering method. The test confirms that fuzzy clustering outperforms hard clustering.

5. CONCLUSIONS

This paper demonstrates the potential of fuzzy clustering for delineating housing submarkets. Fuzzy c-means algorithm was extended such that parameter values c and m can be chosen in a non-arbitrary way. The algorithm maps fuzzy set membership degrees of census tracts to different housing submarkets per metropolitan area.

[3] Indeed, it is same as the objective function of fuzzy clustering algorithms shown in Eq. 1, except for the exponent of the membership degree.

The results derived from fuzzy clustering and hard clustering were compared. The significance test indicates that fuzzy clustering yields more compact clusters than hard clustering while other factors are controlled for. The study supports the premise that fuzzy clustering can enrich the method of (housing) market segmentation.

The empirical tests on the optimal number of clusters c^* and optimal fuzziness exponent m^* suggest that the Xie-Beni index is most robust among the sample index tested. The work presented in this paper will benefit most from further studies in cluster validity index given that the choice of index remains open in the extended algorithm (or Algorithm 2).

Future research would be to incorporate spatial weighting into fuzzy clustering algorithms. The adjustment can be made to Algorithm 2 in a way that a new fuzzy set membership of instance k, U_k' is updated as

$$U_k' = (1-a)U_k + a\frac{1}{A}\sum_{l}^{n} w_{kl}U_l \qquad \text{(Eq. 9)}$$

where l (not to be confused with iterators in Algorithm 2) denotes other instances (not same as k), w_{kl} measures the spatial interaction between instance k and l, and a represents relative importance of surrounding areas of instance k on a new fuzzy set membership value. The higher a is, the stronger the effect of neighboring areas is. Initial experimentations over a range of a [0, 0.5] indicate that validity index values increase with a value of a. It is not surprising given that spatial weighting makes partitioning fuzzier, leading to less compact clusters. Determining to which degree neighboring areas influence U_k' (that is, a), and how to define neighboring areas (that is, w_{kl}) depends on scale and data structure. Image segmentation field suggests the use of kernel-based method [17]; spatial statistics field recommends the use of spatial autocorrelation statistics [6]. The marriage between pattern recognition and spatial statistics deserves further attention [16].

This study is not without limitations. The methodology can be refined further by noting how parameters c, m, and a interact with each other, and how the interaction affects clustering results[4]. The use of better indicators (e.g., average SAT score as an indicator of school quality) and crime rate in a finer spatial resolution will improve results significantly. Qualitative research techniques (such as interview with local realtors or long-time residents) can complement statistical validation methods presented in this paper.

6. REFERENCES

[1] Bezdek JC, 1974a, Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biolology* 1:57–71.

[2] Bezdek JC, 1974b, Cluster validity with fuzzy sets. *Journal of Cybernet* 3: 58–72.

[3] Bezdek JC, 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press.

[4] Bezdek JC, 1984, FCM: the fuzzy *c*-means clustering algorithm. *Computers and Geosciences* 1(2-3): 191-203.

[5] Cadwallader M, 1995, *Urban Geography: An Analytical Approach*. Pearson Education.

[6] Feng Z and Flowerdew R, 1998, Fuzzy geodemographics: a contribution from fuzzy clustering method. In Carver S (Ed.), *Innovations in GIS 5*. London: Taylor and Francis, pp.119-127

[7] Follain DJ and Jimenez E. 1985, Estimating the demand for housing characteristics: A survey and critique. *Regional Science and Urban Economics* 15:77-107.

[8] Fotheringham AS and O'Kelly ME, 1989, *Spatial interaction models: formulations and applications.* Dordrecht; Boston: Kluwer Academic Publishers.

[9] Hansen W, 1959, How accessibility shapes land use. *Journal of the American Institute of Planners* 25: 73–76.

[10] Kim D-W, Lee KH, and Lee D, 2004, On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition* 37: 2009-2025.

[11] Lancaster KJ, 1971, *Consumer Demand: A New Approach*. New York: Columbia University Press.

[12] Mortgage Bankers Association, 2004, U.S. Homeownership rate hits record high. Retrieved March 2007 from http://www.allbusiness.com/personal-finance/real-estate-mortgage-loans/771203-1.html

[13] Muth RF and Goodman AC, 1989, *The Economics of Housing Markets*. New York: Harwood Academic Publishers.

[14] Pal NR and Bezdek JC, 1995, On Cluster Validity for the Fuzzy *c*-Means Model. *IEEE Transactions on Fuzzy Systems* 3(3): 370-379.

[15] Rogerson PA, 2001, *Statistical Methods for Geography*. Sage Publications.

[16] Shekhar S, Zhang P, Huang Y, and Vatsavai R, 2004, Trend in Spatial Data Mining. In: *Data Mining: Next Generation Challenges and Future Directions* by Kargupta H, Joshi A, Sivakumar K, and Yesha Y (eds.), AAAI/MIT Press

[17] Tolias YA and Panas SM, 1998, Image segmentation by a fuzzy clustering algorithm using adaptive spatially constrained membership functions. *IEEE Transactions on Systems, Mans, and Cybernetics – Part A* 28(3): 359-369.

[18] Tsekouras GE and Sarimveis H, 2004, A new approach for measuring the validity of the fuzzy *c*-means algorithm. *Advances in Engineering Software* 35(8/9): 567-575.

[19] Watkins CA, 2001, The definition and identification of housing submarkets. *Environment and Planning A* 33(12): 2235-2253.

[20] Witten IH and Frank E, 2000, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems

[21] Xie XL and Beni G, 1991, A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(8): 841-847.

[4] The idea is indebted to Luc Anselin when the paper was presented in 52nd Annual North American Meetings of the Regional Science Association International in November 2005.