Gaps and Overlaps of Urban Housing Sub-market: Hard Clustering and Fuzzy Clustering Approaches

Laura Gabrielli, Salvatore Giuffrida and Maria Rosa Trovato

Abstract It has long been argued that the housing market is spatially subdivided within an urban area. The argument has important implications for explaining how the housing market works and describing the distinctiveness of each housing submarkets, having determined, a priori, its segmentation. The most commonly used method for identifying housing submarkets is based on cluster analysis, although hedonic analysis has been extensively used. The hedonic analysis is used to derive dimensionality of the housing market by estimating what attributes are significant factors influencing housing price. Those attributes or variables can then be used for cluster analysis. The paper proposes an analysis of the real estate market in San Cristoforo, Catania, trying to integrate two different clustering analysis approaches to defining its possible submarkets articulation. The first one is a hard clustering approach using the K-means method and hypothesizing different numbers of clusters. The second one can be considered a verification of the previous results: a fuzzy algorithm is applied to obtain the fuzzy set membership degree of each data point to housing submarkets defined within the examined urban area. The comparison between the results coming from the two different approaches suggests some reflections about the use of these powerful techniques for integrating the knowledge of the complex and multi-layered real estate markets in the urban recovery policies.

Keywords Real estate market analysis · Market segmentation · Urban renewal · Fuzzy clustering · Cluster analysis · Knowledge discovery

L. Gabrielli (✉)
Department of Architecture, University of Ferrara, Ferrara, Italy
e-mail: laura.gabrielli@unife.it

S. Giuffrida · M.R. Trovato
Department of Civil Engineering and Architecture,
University of Catania, Catania, Italy
e-mail: sgiuffrida@dica.unict.it

M.R. Trovato
e-mail: mrtrovato@dica.unict.it

S. Stanghellini et al. (eds.), *Appraisal: From Theory to Practice*,
Green Energy and Technology, DOI 10.1007/978-3-319-49676-4_15

1 Introduction

Nowadays, many factors affect the accuracy, completeness, and reliability of the appraisals in complex urban contexts getting through transformation processes, arising generalized expectation about the increase in housing prices. Some of these factors can be considered the typical effect of the financial crisis—due to the credit crunch—and of the consequent economic crisis. The former is responsible for the reduction of loans granted to householders; the latter caused the decline of the employment opportunities, of mobility, of the demand for rental houses, of the tenants' solvency—both in residential and in the directional segments due, in particular, to the suspension of numerous professional businesses.

The general uncertainty in the real economy arises monetary hoarding; similarly, the uncertainty of the real estate investment success increases the property hoarding propensity and the related "housing market viscosity".

Thus, the fall in housing prices has not been the only relevant consequence of this crisis, whose most negative effect has been the paralysis of the transactions: owners do not sell and potential purchasers cannot buy. In such a situation: the natural or physiological transactions (such as sales for the current purchase) become very difficult and result in losses; conversely, the artificial or pathological transactions (such as purchases for the future sale) become easier and give rise to probable capital gains.

In the event of significant market increasing inactivity, prices do not reflect values: prices are reduced to mere conjuncture facts whose relevance only concerns single transactions and involves individual action and point of view. Values, instead, are structural phenomena regarding the urban policies in the perspective of the balance of the conditions of the different districts in a complex and heterogeneous city. Thus, prices reflect the positive perspective, which is "how things are" while values reflect the normative one, which is "how things should be".

The second group of issues affecting the values in complex urban contexts are: the weak relationship between characteristics of properties and prices; the difficulty of aggregating many features in just a few significant attributes; the possibility to describe systems of very different individual preferences with a single pattern; the correspondence between the internal consistence of the clustering pattern and the external correspondence with prices; the significance of the asking prices; the prospect of using the values (here meant as attributes) as an effective basis for the regulation of prices in the two cases of local and global taxation.

Finally, the convergence of the effects of the economic crisis and the urban context complexity highlights one of the major issues of the debate in the valuation discipline: the basic distinction between value and price, including on real estate market. In such a doubly uncertain situation, in fact, the conjuncture prices do not reproduce the value they represent for contractors: owners, who can delay the sale, value their assets more than the market prices; potential buyers, who can delay the purchase, wait for further declines in prices and a general improvement in the economic and urban outlook.

This situation encourages, in housing market appraisals, the segmentation techniques to apply a regulatory approach based on the analysis of the values and verified by the experience of the prices. As a result, we can assume: prices as the asking price observed in the housing market analysis; value as: a) the set of attributes associated with each property; the fair value (or monetary measure) of each property based on the attributes: in such perspective, prices become the weaker foundation than the value one in a typical valuation pattern.

Therefore, a global analysis of the urban real estate market is not possible unless the segments expressing significantly the characteristics of the properties in their specific contexts are indicated and individuated. The proposed study deals with the analysis of the real estate market in the quarter of San Cristoforo in Catania, trying to integrate different approaches defining a possible articulation in the submarkets (Bourassa et al. 2003).

The first approach takes into account the sample as a whole and tries to describe a first approximation relation between asking price and the aggregate quality index, an overall score aggregating the 28 main characteristics describing each property. This analysis represents the complexity of this housing market by a sort of clustering "by nature", namely without taking into account the scores. It provides an initial hypothesis of classification of the cases and delimitation of the segments, taking into account the ranges of prices registered in the different classes of the characteristics.

The second approach consists of an in-depth analysis basing on three different clustering hypotheses, from three to four or five clusters. In this case, according to the basic principles of the proposed technique, each element only belongs to a cluster.

A different perspective has been assumed in the third approach based on a fuzzy clustering pattern aimed at identifying the natural overlaps of the different clusters and any possible gap, in the case in which some properties cannot be included in any cluster due to its extreme inconsistency with the sample.

2 San Cristoforo Neighbourhood in Catania and the Real Estate Market Survey

San Cristoforo is part of the "Centro" Municipality (the first of the ten municipalities of the "Comune" of Catania consists of), comprising the quarters of Antico Corso, San Berillo, Civita, and Fortino. It constitutes an urban sub-system characterised by a significant functional, typological and social articulation that permeates its real estate assets.

The quarter is delimited by SS. Maria Assunta Street—Concordia Street axis on the South, Plebiscito arch on the North, the harbour area on the East and Acquicella Street on the West. The northern and southern boundary areas are the most interesting regarding urban quality and vitality. In particular, Plebiscito Street still preserves most of its original urban character, as the quarter was constructed after the 1693 earthquake, in an area outside the ancient city walls, specifically assigned for a new expansion and the reconstruction of the urban centre (Dato 1983). The quarter has a surface of ca. 0.87 km^2, with a very high building density.

The real estate sample is formed by 58 properties comprised in the residential segment. The analysis has been carried out basing on 28 characteristics, aggregated into 6 groups: as follows (Forte 1968).

The attributes are expressed in a standard scale ranging from 1 to 5 representing the lowest and the highest quality conditions. Table 1 shows the sample and the values of the aggregated characteristics.

3 Methods and Procedures

3.1 Cluster Analysis

The Cluster Analysis is a multivariate method, which aims at classifying of observations into a number of different groups based on a set of measured variables. The degree of association between two objects belonging to the same group is maximal, but if they belong to a different group, it is minimal. The cluster analysis helps to identify groups and their structures within the data and analyse those groups of similar observation rather than individual data. Moreover, cluster analysis portrays relationship not revealed otherwise within the observed data, developing taxonomies.

There is a number of different approached, which can be used in order to identify clusters in a dataset: hierarchical and partitional algorithms (Jardine and Sibson 1968). Hierarchical methods can be either agglomerative or divisive. Agglomerative hierarchical clustering starts with every observation (object, subject) being a cluster into itself. At successive steps, the two most similar clusters are merged, and this is done continuously until all data are in one cluster. The problem of this approach is to find the optimum number of clusters between all the solutions. In divisive clustering, all subjects start in one cluster and end with everyone in just one cluster. Agglomerative methods are more popular and are used more often in clustering, even if once a cluster is formed, it cannot be split but only combined with other clusters. The most frequently used methods for combining clusters at each stage, defining the distance between clusters, are single linkage, complete linkage, average linkage between groups, average linkage within groups Ward's method, among the others.

The partitional algorithms decompose the whole dataset into smaller clusters, where the analyst predetermines the number of the resulting clusters. The partitioning-based clustering methods use an iterative method and based on a distance measure it updates the cluster of each object. The most used partition-based clustering algorithms are the K-means, the K-medoids, Clara, among the others.

Table 1 Synthesis of the market survey: prices and values

id	Address		Floor	Rooms	Surface	Asking price	k_{e1}	k_{e2}	k_i	k_t	k_{a1}	k_{a2}	k*
1	plaia	169	0–1	6.6	250	€240.000	3.6	3.0	3.3	2.5	3.2	3.4	2.8
2	plaia	167	0–1	11.2	300	€250.000	3.3	3.0	2.8	2.0	2.4	2.7	2.3
3	plaia	134	0	2.6	50	€28.000	3.3	2.6	1.6	2.8	2.8	2.3	2.3
4	plaia	63	1	2.6	70	€80.000	2.8	2.6	2.7	4.0	3.8	3.7	3.2
5	ortolani	5	0–1	6.3	130	€240.000	2.8	3.0	3.1	4.0	4.0	3.3	3.4
6	del principe	177	8	3.8	80	€110.000	2.5	2.4	1.7	2.0	1.9	2.0	1.9
7	del principe	177	6	3.3	90	€90.000	2.5	2.4	1.7	2.0	1.9	2.0	1.9
8	del principe	177	5	2.8	65	€90.000	2.5	2.4	1.7	2.0	1.9	2.0	1.9
9	del principe	117	3	3.8	80	€105.000	3.0	3.0	3.3	4.0	3.8	3.3	3.4
10	del faro	5	0	2.5	50	€80.000	2.5	2.4	1.4	2.0	2.3	2.3	1.9
11	ss. assunta	70	1	3.5	70	€120.000	2.7	3.0	2.4	2.8	2.6	2.0	2.5
12	villa sgabrosa	65	1	2.7	70	€85.000	2.6	2.6	2.3	2.8	2.6	2.0	2.5
13	del principe	37	3	2.8	65	€90.000	2.5	2.4	1.7	2.3	2.5	1.7	2.1
14	domenico tempio	24	1	5.1	140	€145.000	3.7	3.0	3.9	3.0	3.5	3.0	3.1
15	grimaldi	158	0	2.9	45	€48.000	2.6	2.6	1.2	1.8	1.8	1.7	1.7
16	plebiscito	597	4	6.0	130	€190.000	3.5	3.4	3.3	3.5	3.0	3.3	3.2
17	plebiscito	597	6	7.3	105	€199.000	3.5	3.4	3.6	3.3	3.0	3.0	3.2
18	plebiscito	529	2	4.6	85	€85.000	3.5	3.4	2.7	2.8	2.5	3.1	2.7
19	plebiscito	519	5	6.1	100	€185.000	3.3	3.4	3.6	3.5	3.4	2.9	3.3
20	plebiscito	415	5	4.1	100	€185.000	3.3	3.4	3.3	3.3	2.7	3.0	3.1
21	plebiscito	397	1	4.2	90	€115.000	3.3	3.4	3.1	3.5	3.0	3.7	3.2
22	plebiscito	393	0–1	2.8	60	€50.000	3.3	3.0	2.4	2.3	2.5	2.4	2.4
23	plebiscito	345	1	4.1	90	€190.000	3.3	3.4	3.1	3.5	3.0	3.7	3.2
24	plebiscito	309	0–1	8.8	170	€160.000	3.3	3.4	3.1	2.8	2.5	3.4	2.8

(continued)

Table 1 (continued)

id	Address		Floor	Rooms	Surface	Asking price	k_{e1}	k_{e2}	k_i	k_t	k_{a1}	k_{a2}	k*
25	plebiscito	246	3	6.2	100	€130.000	3.3	3.0	3.1	3.5	3.0	3.3	3.1
26	s. m. delle salette	45	0–1	5.9	110	€90.000	2.8	3.0	2.0	2.8	2.8	3.0	2.5
27	s. m. delle salette	40	0	4.0	90	€110.000	2.9	3.0	1.4	3.8	2.7	2.9	2.7
28	s. m. delle salette	38	1–2	5.8	120	€150.000	2.9	3.0	1.9	3.0	2.7	2.6	2.5
29	s. di giacomo	44	1	2.6	65	€48.000	2.3	2.6	1.7	1.0	1.2	1.7	1.4
30	reitano	1	2	3.7	80	€150.000	2.8	3.0	2.7	3.0	3.2	3.4	2.8
31	plebiscito	148	1	2.6	40	€65.000	3.3	3.4	3.0	3.0	2.7	3.0	2.9
32	plebiscito	119	1	4.7	100	€160.000	3.3	3.4	2.8	2.3	2.9	2.4	2.5
33	grimaldi	14	6	6.2	140	€260.000	3.1	3.0	4.6	3.5	3.5	3.4	3.5
34	fornai	27	1	2.6	55	€70.000	2.6	2.4	1.8	2.8	2.4	2.7	2.3
35	g. zurria	37	1	8.4	200	€240.000	3.4	3.4	3.3	3.0	2.7	3.0	3.0
36	gentile	22	2	4.9	130	€140.000	3.0	3.0	2.3	2.3	2.4	2.4	2.3
37	scuto	32	1	1.5	35	€59.000	2.6	3.0	2.0	2.3	2.3	2.0	2.2
38	cristoforo colombo	94	2	3.7	80	€145.000	3.1	2.6	2.9	2.8	2.4	2.4	2.6
39	domenico tempio	30	1	2.9	60	€78.000	3.6	3.0	2.6	3.0	2.8	2.7	2.8
40	domenico tempio	30	1	2.6	65	€85.000	3.6	3.0	2.7	3.0	2.8	3.0	2.8
41	della concordia	68	2–3	5.5	120	€145.000	3.5	3.0	3.4	3.8	3.5	3.3	3.3
42	della concordia	70	4	4.9	110	€170.000	3.5	3.0	3.6	3.0	3.0	2.7	3.0
43	de lorenzo	200	1	2.8	45	€55.000	2.4	2.0	1.5	2.8	2.2	2.3	2.1
44	mulino a vento	210	1	3.6	80	€115.000	2.2	2.0	2.6	2.8	2.7	3.0	2.4
45	belfiore	210	0–1	4.1	80	€75.000	2.3	2.0	2.1	3.0	2.9	3.0	2.4
46	belfiore	218	0	1.5	30	€25.000	2.3	2.0	2.1	2.0	2.2	2.4	2.0
47	della concordia	126 A	1	3.6	70	€90.000	2.8	2.6	2.1	2.0	2.0	2.0	2.1
48	tripoli	47	0–1	7.8	240	€230.000	1.6	2.0	2.0	2.3	2.2	2.1	2.0

(continued)

Table 1 (continued)

id	Address		Floor	Rooms	Surface	Asking price	k_{e1}	k_{e2}	k_i	k_t	k_{a1}	k_{a2}	k*
49	velis	28	1	4.1	60	€80.000	3.6	3.0	2.6	4.0	4.0	3.3	3.3
50	piombai	11	0–1–2	4.1	80	€75.000	3.6	3.0	1.7	2.8	2.2	2.3	2.4
51	zuccarelli	15	0	1.5	40	€40.000	3.3	3.0	1.4	2.3	2.5	2.0	2.1
52	cordai	97	0	3.0	55	€40.000	2.4	2.6	1.8	2.3	2.4	2.3	2.1
53	cordai	131	1	3.5	90	€70.000	2.4	2.6	2.4	3.0	2.5	2.3	2.5
54	delle margherite	30	2	3.5	50	€55.000	2.4	2.6	2.7	3.0	3.5	3.0	2.7
55	mulino a vento	116	0	4.5	100	€67.000	2.6	2.6	1.7	1.0	1.0	1.0	1.4
56	del principe	142	2	4.7	100	€50.000	2.8	3.0	3.0	3.0	3.1	3.3	2.8
57	alogna	26	1	2.6	50	€60.000	2.5	2.6	2.1	2.8	2.6	3.0	2.4
58	ortolani	35	2	5.9	120	€150.000	2.8	3.0	3.9	3.3	3.3	2.7	3.2

A Location: $k_{e1}1$ centrality and settlement quality; $k_{e1}2$ functional mix; $k_{e1}3$ socio-economic mix; $k_{e1}4$ urban maintenance; $k_{e1}5$ equipment; $k_{e1}6$ facilities; $k_{e1}7$ accessibility by private transportation; $k_{e1}8$ accessibility by public transportation; $k_{e1}9$; internal access; *B* Location 2: $k_{e2}1$ micro-environmental functional features; $k_{e2}2$ micro-environmental symbolic features; *C* Intrinsic features; k_i1 panoramic quality; k_i2 view; k_i3 brightness; k_i4 exposure; k_i5 security; *D* Technology: k_t1 plants; k_t2 finishes; k_t3 maintenance status; *E* Building Architectural quality: $k_{a1}1$ usability; $k_{a1}2$ structural and plant quality; $k_{a1}2$ finishes and building technologies; $k_{a1}4$ stylistic coherence; $k_{a1}5$ decorum; $k_{a1}6$ internal coherence; *F* Property Architectural quality: $k_{a2}1$ size, distribution and usability; $k_{a2}2$ accessories and restrooms; $k_{a2}3$ finishes

A combination of the hierarchical and partitional algorithm can be used, and many other clustering techniques have been proposed during the years, especially with the spread of the use of statistical software packages.

In order to cluster our variables collected in the real estate market, in this paper we proceeded to determine how the clusters are to be formed and the number of clusters. We used both the hierarchical and non-hierarchical approaches in order to identify different groups in real estate market.

Regarding the agglomerative hierarchical algorithms, we used the Ward's methods, which looks at clustering as an analysis of variance, rather than using distance metrics of measures of association, like other approaches. It looks at clustering as an analysis of variance problem, instead of using distance metrics or measures of association. Ward's method is based on a classical sum-of-squared criterion, producing clusters that minimize within-group dispersion at each fusion (Murtagh and Legendre 2014). In this minimum variance method, the distance between two clusters is the ANOVA sum of squares between the clusters added up over all the variables. At each step, the two clusters that merge are those that result in minimizing the within-group sum of squares. This method is most appropriate for quantitative and not binary variables.

Among the partitional approaches, we used the K-mean method, which aims at grouping data into K clusters based on how close an observation is to the mean of the observations in each cluster. The method segments the data, minimizing the within-cluster variation. The steps in the process are different, consisting in assigning, randomly, each observation to a K cluster, reassign the observations to other clusters to minimize the within-cluster variation, which is the squared distance of each observation from the mean of each cluster, and, finally, repeating the process until no observation needs to be reassigned. As K-means method does not build a hierarchy (the cluster affiliation of data could change during the process), the approaches belong to the non-hierarchical clustering approaches.

To assign an observation to the closest centroid, a proximity measure must be chosen. In this case, the Euclidean distance was used to implement the K-mean approach. Another step in the method was the definition of the appropriate number of clusters, which are correlated to the quality of clusters. In this case—study, the sum of the squared error (SSE) was used and it is the sum of the squared errors between every observation and the centroid of the cluster it belongs (Krzanowski and Lai 1968). It can be used as a measure of variation within a cluster. It is possible then to compute the total sum of the squared errors. The cluster with the smallest SSE (the centroids of this clustering are a better presentation of the point in their clusters) is preferred.

The problem with the K-means method is the choice of the number of the clusters into which the observations will be divided. The initial choice of the number k is mainly subjective, and so the results can be biased by the opinion of the user. Successive runs of K-means can optimize the clustering of the observation for a different number of clusters. A comparison with the hierarchical methods could also be used.

A process for determining the optimal number of clusters is (Gabrielli et al. 2015):

- assumed the dataset X, a specific clustering algorithm and a range of number of clusters [M_{min}, M_{max}], are defined;
- the clustering algorithm is repeated from predefined values of M_{min} to M_{max};
- he clustering results (partitions P and centroids C) are obtained and then the index value for each of them are calculated;
- the cluster M is selected, for which the partition offers the best outcome according to some criteria (minimum, maximum or knee point).

3.2 Fuzzy Clustering

The aim of cluster analysis is to partition a set of objects into two or more clusters such that objects within a cluster are similar and objects in different clusters are dissimilar (Kaufman and Rousseeuw 1986). The fuzzy clustering methods (Hwang and Thill 2009) making use of the fuzzy set theory, allow us to associate a unit to groups with a certain degree of membership, expressed by a membership function which takes values in the interval [0,1]. The interest in these methods stems from the awareness that there is a certain degree of inaccuracy in the data, and then that such a method is able to represent more than a crisp method can do.

The fuzzy clustering methods are richer in information, as they provide the degree of consistency by one unit with each cluster, allowing to establish a group hierarchy (the hierarchy is given by the different degree of unit belonging to the groups) to which it may belong to the unit, by virtue of the fact that the groups are viewed as fuzzy sets. In addition, they have no claim to provide definite answers on how you added the data. Figure 1 shows an ideal situation in which the points are perfectly separated in two clusters and a situation closer to reality in which the points are distributed in such a way that is difficult to attach a point to a cluster or another.

The fuzzy clustering generalizes partition clustering methods (such as K-means and medoid) (Kaufman and Rousseeuw 1987) by allowing an individual to be partially classified into more than one cluster. In regular clustering, each individual

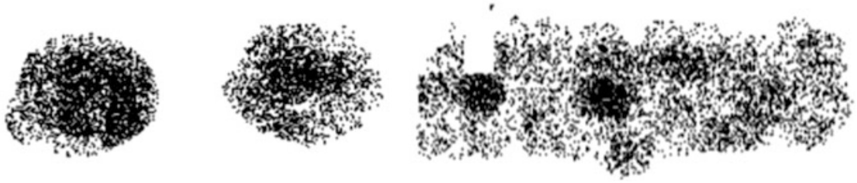

Fig. 1 Comparison between an ideal situation and a real one

is a member of only one cluster. Suppose we have K clusters and we define a set of variables $m_{i1}, m_{i2}, \ldots, m_{ik}$ that represent the probability that object i is classified into cluster k. In partition clustering algorithms, one of these values will be one and the rest will be zero. This represents the fact that these algorithms classify an individual into one and only one cluster. In fuzzy clustering, the membership is spread among all clusters. The m_{ik} can now be between zero and one, with the stipulation that the sum of their values is one. We call this a fuzzification of the cluster configuration. It has the advantage that it does not force every object into a specific cluster. It has the disadvantage that there is much more information to be interpreted.

There are different approaches in the literature to fuzzy clustering, such as hierarchical and non-hierarchical. In particular, in the case of non-hierarchical classification methods, they have the characteristic of providing directly a certain number of groups fixed a priori, through iterative procedures that seek to optimize an objective function.

In this regard, there are several algorithms, which differ in the objective function, and then adopted for the choice different iterative procedure to compute the membership degrees of the unit to the groups.

The objective function determines for each solution a measure of the error, based on the distance between the data and the representative elements of the cluster.

It seeks to minimize the following objective function, C (Kaufman and Rousseeuw 1990) defined on the basis of membership in the cluster and distances $C = \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} m_{ik}^2 m_{jk}^2 d_{ij}}{2 \sum_{j=1}^{N} m_{jk}^2}$, where m_{ik} represents the unknown membership of the object i in cluster k and d_{ij} is the dissimilarity between objects i and j. The memberships are subject to the constraints that they all must be non-negative and that the memberships for a single individual must sum to one. That is, the memberships have the same constraints that they would if they were the probabilities that an individual belongs to each group (and they may be interpreted as such).

The medoid partitioning algorithms presented to accomplish this by finding a set of representative objects called medoids. The medoid of a cluster is defined as that object for which the average dissimilarity to all other objects in the cluster is minimal. If k clusters are desired, k medoids are found. Once the medoids are found, the data are classified into the cluster of the nearest medoid.

Two algorithms are available in this procedure to perform the clustering. The first, from Spath (1985), uses random starting cluster configurations. The second, from Kaufman and Rousseeuw (1990), makes special use of silhouette statistics to help determine the appropriate number of clusters.

The fundamental value used in cluster analysis is the dissimilarity between two objects. This section discusses how the dissimilarity is computed for the various types of data. For multivariate data, a critical issue is how the distance between individual variables is combined to form the overall dissimilarity. This depends on the variable type, scaling type, and distance type that is selected.

A *brief discussion of the possible* types of variables will follow. The dissimilarity (distance) between two objects is fundamental to cluster analysis since the techniques goal is to place similar objects in the same cluster and dissimilar objects in different clusters. Unfortunately, the measurement of dissimilarity depends on the type of variable. For interval variables, the distance between objects is simply the difference in their values. However, how do you quantify the difference between males and females? Is it simply 1 − 0 = 1? How do you combine the difference between males and females with the difference in age to form an overall dissimilar? These questions will be answered in this section. This discussion follows Kaufman and Rousseeuw (1990) very closely.

Assume that you have N rows (observations), which are separated to be clustered into K groups. Each row consists of *P* variables. Two types of distance measures are available in the program: Euclidean and Manhattan.

The *Euclidean distance* d_{jk} between rows j and k is computed using $d_{jk} = \sqrt{\frac{\sum_{j=1}^{P} \delta_{jik}^2}{P}}$ and *Manhattan distance* d_{jk} between rows j and k is computed using $d_{jk} = \sqrt{\frac{\sum_{j=1}^{P} |\delta_{jik}^2|}{P}}$ where for interval, ordinal, and ratio variables $d_{jk} = z_{ij} - z_{ik}$ and for asymmetric-binary, symmetric-binary, and nominal variables $d_{jk} = \begin{cases} 1 & if\ x_{ij} \neq x_{ik} \\ 0 & if\ x_{ij} = x_{ik} \end{cases}$ with the exception that for asymmetric-binary, the variable is completely ignored (*P* is decreased by one for this row) if both x_{ij} and x_{ij} are equal to zero (the non-rare event).

The value of z_{ij} for interval, ordinal, and ratio variables is defined as $z_{ij} = \frac{x_{ij}-A}{B_i}$, where x_{ij} represents the original data value for variable i and row j and z_{ij} j represents the corresponding scale value. The scaling choice determines the values used for A_i and B_i. Type of scaling of the value A_i and B_i are: absolute value, standard deviation, range $(Min_{overj}(x_{ij})$ or $Max_{overj}(x_{ij}) - Min_{overj}(x_{ij}))$.

4 Applications and Results

4.1 Hard Cluster Analysis

In our case study we fixed the number of clusters $M_{min} = 2$ and $M_{max} = \sqrt{N}$. We use the previously defined 6 variables v.

We used the data for knee point detection in order to detect the proper number of clusters. The knee point in the graphs indicates the optimal number of clusters, even if the recognition of the knee points is not that easy. The maximum value and the minimum values are the most straightforward points to identify. Some other indices are monotonous, so it is not clear what the optimum value for the number of clusters. We used some validation indexes such as: SSB/SSW (the ratio between the sum-of-squares between clusters and the sum-of-squares within cluster); WB (the ratio between the sum-of-squares within cluster and the sum-of-squares between clusters, multiply by M, the number of the cluster); RSQ (the ratio between the sum-of-squares within cluster and the sum of the latter with the sum-of-squares between clusters). It was possible then to identify a range of numbers of the optimal cluster, which means that the optimal number of clusters is not defined. The option between 3 and 5 clusters are considered in the following analysis.

The three different options (3, 4 and 5 clusters) are commented hereafter. Using the K-mean method, the solution with 3 clusters has 26, 20 e 12 observations in the cluster no. 1, 2 and 3, respectively. Cluster n. 1 is quite different from cluster n. 2 and very different from cluster n. 3, while in cluster 2 and 3 the variables are not so different. The variables that have the greatest impact on clustering are k_i and k_{a2}, while both k_e variables have a small impact on clusters. In cluster n. 1 all properties with high quality and good characteristics are grouped together. In cluster n. 3, the observations show poor quality, especially for k_i, k_{a1} and k_{a2}. If the Ward's method is used replacing the K-mean, we obtain the same results. The two approaches are very consistent in clustering the data. Only two properties, no. 37 and 52, in Ward's method move from cluster 2 to cluster 3, which is plausible as the two clusters show the smallest difference between them.

The second hypothesis has 4 clusters with 16, 11, 20 and 11 cases respectively. In this scenario, the variable, which has the greater impact on clustering, is k_{a2}, while the k_{e1} and k_{e2} show a small value, and so impact, as measured by the value of the F-ratio.

The groups n. 1 and n. 4 have all properties with a high value of the variables (almost all mean >3): while the group n. 1 has a high value of variables k_{e1}, k_{e2} and k_i, group n. 4 shows the high value of the remaining variables, namely k_t, k_{a1} and k_{a2}. The group n. 2 have a very low value of almost all variable (around 1), meaning the poor quality of the characteristics of the properties included in the group. The group n. 3 has medium level characteristics and it is collocated between the groups 1–4 (high quality) and group 2 (poor quality). Again, using the Ward's method, the results is quite robust. In this case, the group n. 3, even though it retains the same characteristics of group 4 obtained with the K-mean method, it has a lower number of observations and therefore few cases are included (5 cases rather than 11).

In the third scenario, with 5 clusters (of 20, 12, 11, 3 and 12 cases each) shows less difference. The group n. 2 of the previous scenario divided into group 4 and 5 in this scenario, all the other changing only a bit (even if the cluster 1 here was the cluster 4 in the previous scenario, and so the cluster 2 here was cluster 1 in the previous situation). In this last test, the better discriminators between observations are k_i and k_{a2}, which seems to be the most significant variables to cluster in all the hypothesis analysed. The market and its demand seem to appreciate particularly the characteristics intrinsic and the ones linked to the property asset unit. The splitting of cluster n. 2 of the previous scenario, which was the one with the poorest characteristics, generates the cluster 4 and 5. The cluster n. 4 has only three observations, and they are very poor quality properties (as maintenance, location, view, etc.) and so less attractiveness for the market.

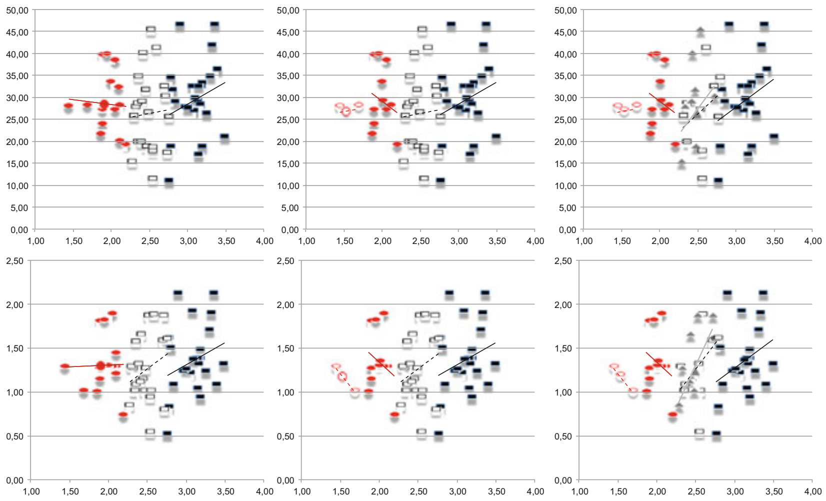

Fig. 2 Unit prices/overall value relation for each of the three hard clustering hypotheses

The mean of the ks are about 1.5, or below that number. Only k_e s are >2. Similarly, cluster 5 has buildings whose characteristics' mean is around 2. Group 1 and 2, which were group 4 and 1, respectively, show a higher value of the characteristics: in cluster 1 all the means are >3. The most similar clusters are the n. 3 and n. 5, which show little distance in their centroid and the means of the variables used for clustering. The cluster n. 5 shows smaller values of the k_i characteristic in comparison to the cases included in the 3 cluster ($k_i < 2$). The Ward's method, in this last application, differs from the—means, despite the fact that group 4 and 5 are identical. In Ward's method, the group n. 1 is a very small cluster represented by very top properties, with $k_t = 4$.

The graphs in Fig. 2 also show that in all three hypotheses of clustering segmentation for aggregated value (k*) is respected: assuming the four-cluster hypothesis, the third cluster is divided consistently into two groups, resulting in a fourth cluster comprising the three elements of limited value. Assuming the five-cluster hypothesis, the second cluster is still divided into two groups, giving rise to a fifth cluster of an intermediate value between the second and fourth groups.

4.2 Fuzzy-Cluster Analysis

As far as consistent, further subdivisions into four and five clusters do not add crucial information for the segmentation of the sample. This is confirmed by the fuzzy clustering analysis that strengthens the results obtained so far by changing the composition of the clusters previously delimited.

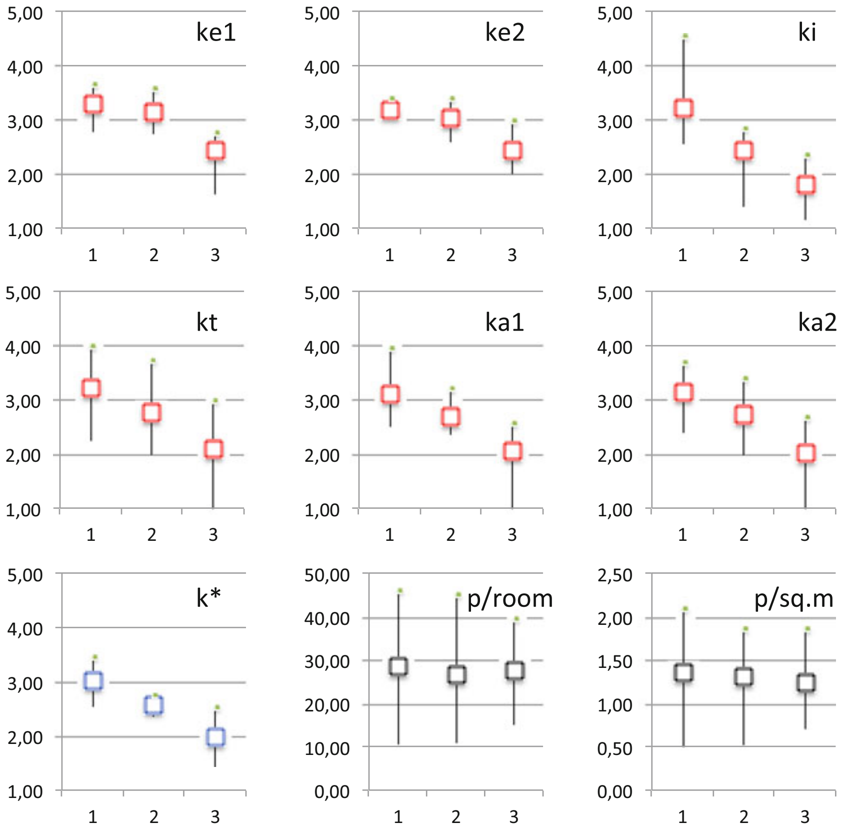

Fig. 3 Comparison of the three fuzzy clusters by minimum, average and maximum values and unit prices

The algorithm associates each element the degree of membership to each cluster. As a result, in a scenario of *strong clustering*, each element belongs to the cluster for which the degree of membership is higher, and no element can be ruled out.

Fuzzy logic "weakens" this hypothesis by selecting the elements that most reasonably can be excluded from the sample (gap) and those that may belong to two clusters (overlap). This selection can be made by requiring that any element whose three cluster membership degree is below the threshold-gap, should be excluded not belonging significantly to any cluster. Moreover, the elements that have a degree of membership to two clusters above the threshold-overlap will be included in both the clusters regardless of which is the greater degree of membership. The first test (gap-test) is a condition of admission to the second (overlap-test), as it is possible that the gap and overlap conditions occur simultaneously. Therefore, having established that at least one of three degrees of membership exceeds threshold-gap, it is checked if the element exceeds the threshold-overlap on two degrees of membership.

By imposing a threshold-gap equal to 0.38 and a threshold-overlap of 0.41, it is possible to get a segmentation in which the first cluster contains 25 elements, the second 15 and third 17. 8 elements are excluded (gap) while 7 elements belong to two clusters at the same time. The results are shown in the graphs of Fig. 3 in which we see how the clusters are well defined with respect to the value of all the characteristics taken separately and with to their aggregate value k^*, while this distinction is less marked than the unit prices. The same can be done by comparing, in Fig. 3, fluctuations in values—individual and aggregate—and unit prices: fluctuations in values distinguish very clearly some clusters; the price movements do not provide significant elements to discriminate the clusters, despite the number of elements contained therein is significantly different.

5 Discussions and Conclusions

The final verification concerns the consistency of the proposed clustering and the urban shape regarding the location of the elements belonging to the different clusters.

1. The hard-clustering pattern provides the following distribution: a group dislocated along the main axes and the other two internal (more characterized by technologic and architectural homogeneity) is outlined in the 3-cluster segmentation. The detachment of the fourth cluster does not add any significant information to the subdivision while the passage to five clusters reveals a subdivision of the second cluster basing on the architectural characteristic and independent from the urban location.
2. The fuzzy clustering analysis pattern, applied in just a 3-clusters hypothesis, provides a more strong and consistent distribution of the whole sample as displayed in Fig. 4.

The first cluster, comprising the best properties, is mostly located along the main roads and the elements are well characterized from every point of view. The second cluster has a good location but lower technological and architectural features. The third one comprises the properties locate in the internal areas with the worst characteristics from all the points of views. Figure 4 also shows the position of the gap/overlap-properties.

Despite the results of the two processes converge towards a definition of segments altogether consistent, the extension of the method to a more flexible approach allowed, through iterative displacement of *thresholds gap and overlap*, to get the best and most fitting configuration of the segments. This is due to the flexibility of a process that allows overcoming the constraint of separation and admits the possibility of multiple memberships.

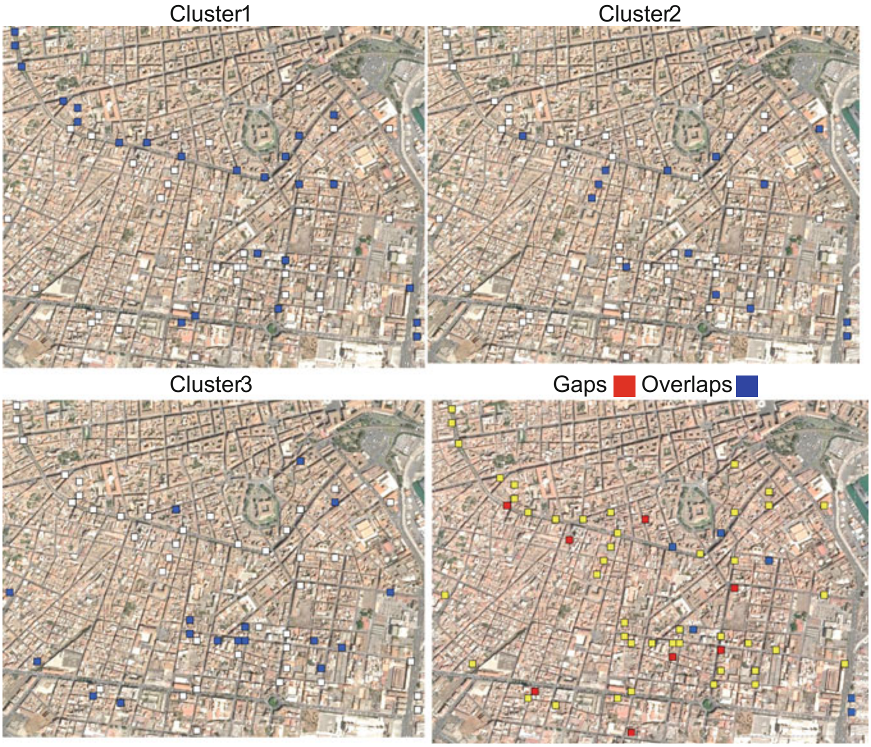

Fig. 4 Spatial distribution of the elements of the three fuzzy clusters and location of the gaps and overlaps

The fact that the strong consistency of the characteristics of the segments does not match the same consistency in prices indicates that the complexity of the context is not represented by expectations about prices, especially during a rarefaction of the transactions.

In this sense, the check of relations of similarity can be very helpful in the negotiated urban transformation processes, in which the value, rather than the price, assumes in the internalization of positive and negative externalities.

In such a perspective, the case has highlighted the complementary nature of value, specific, concrete, and that of the price, general and abstract. The price has the function of making homogeneous combinations of heterogeneous values although substitutable, by defining more easily observable preferences systems in active and transparent markets. Instead, in very articulated, complex, opaque and episodic markets, value and price are made independent from each other and, at worst, indifferent, giving rise to "semantic gaps and overlaps". As a result, from a physiological condition for which very different values are substitutable as they correspond to only one price, the pathological condition prevails whereby the same value can have very different prices. In the latter case, the value consistently measured and represented in its syntactic structure at the level of semantic chains formed by urban areas, becomes again the real foundation for the realignment of the system of administered prices in the context of local taxation (Equalization processes) and global (Land Register).

Acknowledgments Laura Gabrielli edited par. 2.1, 3.1 and 4; Salvatore Giuffrida edited Introduction, par. 1, 3.2 and 4; Maria Rosa Trovato edited par. 2.2, 3.2 and 4.

References

Bourassa SC, Hoesli M, Peng VS (2003) Do housing submarkets really matter? J Hous Econ 12:12–28

Dato G (1983) La città di Catania. Forma e struttura 1693–1833. Officina Edizioni, Roma

Forte C (1968) Elementi di estimo urbano. Etas Kompass, Milano

Hwang S, Thill JC (2009) Delineating urban housing submarket with fuzzy clustering. Environ Plan B Plan Des 36:865–882

Gabrielli L, Giuffrida S, Trovato MR (2015) From surface to core: a multi-layer approach for the real estate market analysis of a central area in Catania. In: Gervasi O et al (eds) Computational science and its applications (ICCSA 2015), vol III. Springer, Berlin, pp 284–300

Jardine N, Sibson R (1968) The construction of hierarchic and non-hierarchic classifications. Comput J 1:177–184

Krzanowski W, Lai Y (1968) A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 44(1):23–34

Kaufman L, Rousseeuw PJ (1986) Clustering large data sets (with discussion).In: Gelsema ES, Kanal LN (eds) Pattern recognition in practice II. North-Holland, Amsterdam, pp 425–437

Kaufman L, Rousseeuw PJ (1987) Clustering by means of medoids. In: Dodge Y (ed) Statistical data analysis based on the L1 norm. North-Holland, Amsterdam, pp 405–416

Kaufman L, Rousseeuw PJ (1990) Finding groups in data. Wiley, New York

Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 31:274–295

Späth H (1985) Cluster dissection and analysis: theory, FORTRAN programs, examples, Ellis Horwood Ltd. Wiley, Chichester