XLI Incontro di Studio del Ce.S.E.T.: 91-104

Mariola Chrzanowska
Monika Krawiec

Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences
E-mail:
mariola_chrzanowska@sggw.pl
krawiec.monika@gmail.com

Application of some multidimensional comparative analysis methods to investigate secondary real estate market in Warsaw

The paper aims at application of multidimensional comparative analysis methods to investigating Warsaw real estate secondary market. The research is based on a data base covering more than 300 apartments offered in the fourth quarter of 2010. There were considered basic characteristics such as the price per 1 square meter, apartment area, number of rooms, floor the apartment is located on, number of storeys in a building and the year of building completion. Application of two alternative methods: Ward's method and k-means method allowed to divide the set of apartments into 5 clusters. The results obtained may be interesting for both individual investors and real estate agents as they help to filter the set of apartments and to find offers similar to the apartment with a prespecified characteristics.

Key words: *real estate market, clustering, Ward's method, k-means method*

1. Introduction

Real estate market is an important element of financial systems. Although it is a subject to the same rules as other markets in economy, at the same time it is a market of specific character. Collecting data on the market is difficult and costly and data on transactions is usually confidential. Markets of this kind are regional (local) markets of differentiated attractiveness. In Poland Warsaw real estate market is the most attractive, then Cracow, Wrocław and Tri City markets. Even those single markets are not homogenous as some districts are considered more attractive, other – less attractive.

Numerous institutions are interested in investigating and analyzing the real estate market. They follow the tendencies in supply and demand. Application of proper methods and correct selection of diagnostic variables affect significantly the quality and reliability of analysis. For the purpose of real estate market analysis, quantitative methods seem to be exactly useful. Usually there are applied some statistical methods (basic descriptive statistics, correlation analysis, statistical indices), methods of spatial econometrics (geographically weighted regression, Moran statistic, kriging), data mining (regression trees, CART and QUEST procedures), taxonomic methods and artificial neural networks.

Quantitative methods applied to study real estate may be used both in appraisal of an individual real property and in mass appraisal of a given real estate market. The aim of the paper is to apply some multidimensional compara-

tive analysis methods to investigation of real estate secondary market in Warsaw. Multidimensional comparative analysis allows studying objects considering number of characteristics simultaneously. This increases the efficiency of research and analysis. In the case of investigating real estate market, linear arrangement methods and clustering methods are applied the most often. From the set of numerous clustering methods, Ward's method and k-means method have been chosen by authors. The research is based on data on more than 300 apartments offered in the fourth quarter of 2010 in Warsaw.

2. Empirical data and research method

Warsaw, being the capital of Poland, is the largest residential market within the group of large Central European cities. It is located on both banks of Vistula River which is the reason of some transportation problems as there are only 3 bridges connecting both parts of the city. Warsaw, covering the area of 51.724 hectares which is 1,45% of area of Poland, is divided into 18 districts. 1.720.398 inhabitants constitute the city population with average density of population equal to 3,326 people per 1 square kilometer. The highest density of population is observed in Ochota district and Praga Południe, the lowest in Wilanów and Wawer. The city centre is dominated by office buildings where many firms and concerns are located, including the Warsaw Stock Exchange. As Warsaw is a country centre of financial, insurance and consulting services, still numerous job opportunities start up here inclining people to come to the city. The inflow creates the heightened demand on the real estate market and the city is still developing, especially with respect to residential and office construction. Unfortunately, transportation facilities hardly follow the process. Nevertheless, Warsaw real estate market is the most developed one in Poland. It is characterised by significant spatial differentiation with regard to both: building type and price level. Figure 1 shows average apartment prices on the secondary market in separate Warsaw districts in December 2010. At that time the most expensive district was Śródmieście and then Mokotów, the cheapest one – Wesoła. In general, the average price per 1 quadratic meter was higher in districts located on the left bank of Vistula river than in those located on the right bank.

The research is based on data on 309 apartments offered in the fourth quarter of 2010 in Warsaw. Figure 2 shows the structure of the data set. Apartments located in Ursynów district constituted the largest group (15%), then those in Praga Południe (11%), Białołęka (10%), Wola (9%), Bemowo and Mokotów (8%). Apartments located in Rembertów had the smallest share of 1%.

The data basis used for the purpose of the research comprises the information on apartments described by the following characteristics: transaction price of an apartment, living area, standard of equipment, number of the floor where the apartment is located, number of floors in a building, number of rooms, year in which the building was raised, number of bedrooms, kitchen type, existence of balcony, existence of basement, existence of lift, neighbourhood, technical state,

Figure 1. Mean apartment prices in PLN per 1 m^2 on the secondary market in Warsaw – December 2010.

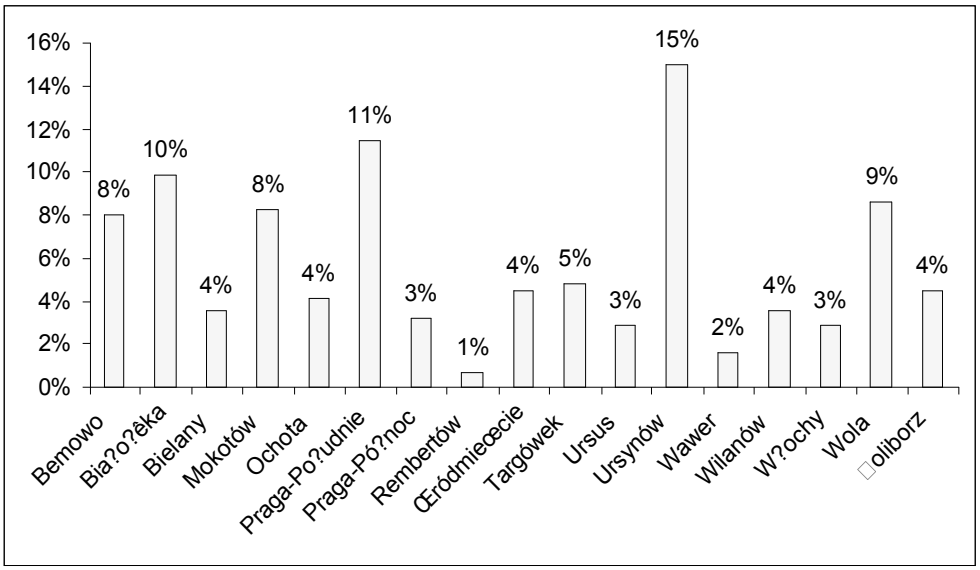
Source: www.warszawacity.com.

access to the public transport, parking place or garage, technology of building construction etc.

As in multivariate comparative analysis it is important to ensure that the final diagnostic variables are comparable, methods of normalization are used. There are many normalization procedures, but standardization is the one used most often. Thus, the first step of the research was variables standardization which was done in a classic manner: (*value of characteristics – arithmetic average)/standard deviation*). This process made possible comparison of all available data. With regard to weak differentiation of considered features, some characteristics were excluded from the research. Finally, from the set of 18 variables only 5 characteristics were taken into account: price for 1 square meter, living area, floor number, number of storeys, number of rooms and the year when building was completed.

In order to analyze collected empirical data, two alternative methods were applied: Ward's method and k-means method. The former one belongs to hierar-

Figure 2. Structure of analysed data on apartments with regard to the district.



Source: own elaboration.

chical clustering methods. Their unquestionable advantage is the fact they follow one procedure, clustering results are given in a form of series of classifications (it makes possible to control the classification process), classification results may be presented graphically in a form of dendrogram, showing subsequent linkages between classes (Gatnar, Walesiak 2004).

All hierarchical procedures may be described by the use of one general scheme. Differences between separate agglomeration methods are related to different understanding of distances[1] between clusters. In central clustering procedure, the starting point is **D** matrix of distances between the objects P_1, P_2,…,P_n being classified. Each object forms a separate cluster. Thus, there are given n clusters G_1, …,G_n.

1. Having given the matrix of distances between clusters G_1, …,G_n:

$$\mathbf{D} = [d_{ij}] \qquad (i,j = 1,2,\ldots,n), \tag{1}$$

the smallest element is set. In other words: one is looking for the pair of clusters least distant from each other:

$$d_{pq} = \min_{i,\,j}\{d_{ij}\} \qquad (i,j = 1,2,\ldots,n), \quad p<q. \tag{2}$$

[1] An important element of a clustering algorithm is the distance measure between data points. Several types of distances are usually applied, such as the Euclidean distance, Spearman distance, Minkowski distance or Czebyszev distance (Zyga 2011).

2. Clusters G_p and G_q are being linked into one new cluster with the number p:

$$G_p := G_p \cup G_q \tag{3}$$

3. The row and column with the numbers q are removed from matrix **D** and one substitutes $n:=n$-1.
4. Distances $\boldsymbol{d}_{pj}(j=1, 2,\ldots, n)$ between the G_p cluster and all other clusters are calculated in compliance to the chosen method. Then $\boldsymbol{d}_{pj}$, values substitute the p-th row in **D** matrix (p-th column is replaced by $\boldsymbol{d}_{jp}$ elements).
5. Steps 1 – 4 are repeated until all objects create one cluster (e.g. when $n=1$).

After each iteration of hierarchical procedures one receives the division of objects being compared into a smaller number of clusters. Moreover, after each iteration one receives modified matrix of distances between clusters. A general formula for calculating distances while combining the clusters G_p i G_q into a new cluster in hierarchical clustering is given by:

$$d_{pj} = a_p d_{pj} + a_q d_{qj} + b d_{pq} + c\left|d_{pj} - d_{qj}\right| \tag{4}$$

or:

$$d_{pj}^2 = a_p d_{pj}^2 + a_q d_{qj}^2 + b d_{pq}^2 + c\left|d_{pj}^2 - d_{qj}^2\right| \tag{5}.$$

Symbols a_p, a_q, b, c are transformation parameters specific for different clustering methods. In Ward's method we have:

$$a_p = \frac{n_i + n_p}{n_i + n_p + n_q} \tag{6},$$

$$a_q = \frac{n_i + n_q}{n_i + n_p + n_q} \tag{7},$$

$$b = -\frac{n_i}{n_i + n_p + n_q} \tag{8},$$

and $c = 0$.

In Ward's method the distance between clusters is the difference between sums of squares of deviations for separate units from the centroids of groups where those points belong to (Ostasiewicz 1999).

Another method applied in the research is the k-means method. This technique proposed in 1950s is an iterative procedure for partition of a population into k groups in order to minimize within-cluster variance. The method is based on the following function:

$$\sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{m} \left(z_{ij} - v_{kj}\right)^2 \tag{9}$$

where: z_{ij}- normalised value of i-th object of j-th variable,
v_{kj} - j-th component of vector of position measures calculated for objects belonging to the k-th class,
$i \in C$ means that i-th object belongs to the k-th class.

Formula (9) may be also given by:

$$\sum_{k=1}^{K} \sum_{i \in C_k} \left(d_{ik}\right)^2 \tag{10}$$

where: d_{ik} – distance betwe0en the i-th object and the vector of position measures calculated on the base of objects belonging to the k-thclass.

Function (10), as well as (9), is the sum of squares of distances between objects and vectors of position measures of classes those objects belong to. As classes should contain the most similar objects, distances between objects belonging to those classes as well as distances of objects from vectors of position measures of classes the objects belong to – should be minimal. Thus in order to obtain optimal classification, minimum of function (9) should be found.

Components of vector of position measures of k-th class (k=1, 2, …, K), providing optimal classification are given by:

$$v_{kj} = \frac{1}{n_k} \sum_{i \in C_k} z_{ij} \tag{11}$$

where n_k – number of objects belonging to the k-th class.

In practice k-means method is applied in compliance with the following algorithm. First, number of classes and initial partition into K classes are set. It may be set in any manner: randomly or according to the researcher's intuition. If there are no premises, preliminary hierarchical analysis of a sample obtained in a result of drawing from a large set can be done. The analysis may be carried out several times on several samples in order to assess the stability of solution obtained. After setting the number of clusters, one may calculate averages for all classifying variables for all clusters. Then, use them as starting points for k-means analysis. Further on, the algorithm is iterative. In each iteration the following takes place:

1. vectors of means (centroids) are calculated for each of classes according to formula (11) on the base of classification obtained in previous iteration or from initial classification;
2. distances of each object from the vector of means for all classes are calculated;
3. the new classification is set by assigning each object to the closest class, e.g. to the class for which distance from the vector of means is the smallest.

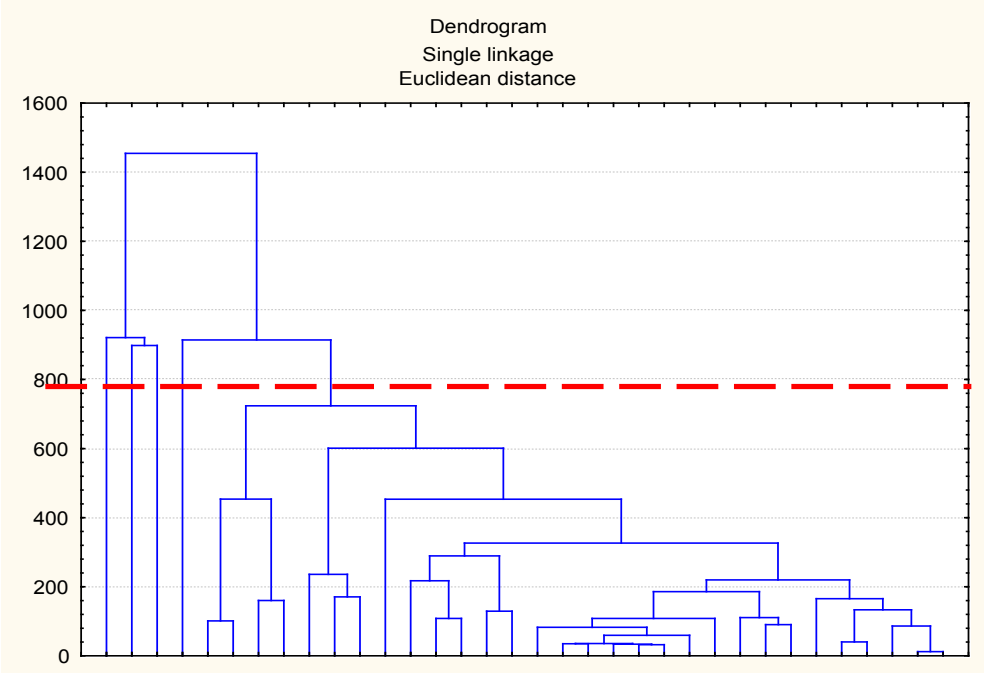
The procedure is to be continued until the classifications obtained in two iterations are the same (Dziechciarz 2002).

3. Research results

At the beginning the Ward's method was used in order to set a number of clusters. Numerous simulations confirm that within all agglomerative methods,

this is the method recognizing the best the group structure (Grabiński, Sokołowski 1980). With respect to the large number of objects (309), their simultaneous analysis is impossible as the matrix of distances would be unclear. Moreover, it would be impossible to interpret a dendrogram prepared for such a large number of objects. Thus, following (Denkowska at al. 2008) there was made a series of drawings of 10% of real objects each time. The series of dendrograms obtained revealed that all real estates should be divided into 5 groups. An exemplary dendrogram is displayed in Figure 3.

Figure 3. An exemplary dendrogram for the sample of real estates from Warsaw secondary market.



Source: own elaboration.

In the second step of research, in a result of k-means method application, there were obtained 5 groups of objects. Next, mean values of separate characteristics in separate groups were compared with the use of univariate analysis of variance. The results are reported in table 1.

Results of variance analysis revealed that all characteristics included in the study discriminated groups well. In other words: in al cases (for all characteristics) one should reject the hypothesis that the mean level of characteristic was the same in all groups. Considering Euclidean distances between clusters (see Table 2) one may state that the partition is proper. The first and the fourth clusters as well as the first and the fifth clusters are relatively close to each other. The largest distance is the one between the second and the third clusters.

Table 1.Results of variance analysis testing differences of means in groups

	Intergroup variance	df	Intragroup variance	df	F-statistic	p-value.
Price per 1m^2	10,5906	4	222,5530	309	3,6761	0,006091
Area	183,0541	4	125,1916	309	112,9543	0,000000
Floor number	186,6649	4	130,1272	309	110,8136	0,000000
Number of storeys	201,8585	4	110,0712	309	141,6680	0,000000
Number of rooms	169,0598	4	141,1672	309	92,5135	0,000000
Year of building completion	75,7873	4	223,9446	309	26,1429	0,000000

Source: own calculations.

Table 2. Euclidean distances* for clusters.

	No 1	No 2	No 3	No 4	No 5
No 1	0,000000	**1,292830**	**2,288970**	**0,704210**	**0,774841**
No 2	1,137027	0,000000	**2,828375**	**0,970539**	**1,707720**
No 3	1,512934	1,681777	0,000000	**2,308412**	**2,258348**
No 4	0,839172	0,985159	1,519346	0,000000	**0,949448**
No 5	0,880251	1,306798	1,502780	0,974396	0,000000

*Note: bold type denotes squares of Euclidean distances.
Source: own calculations.

In table 3, there are reported some descriptive statistics for the cluster No 1 containing 113 apartments. On their base one may notice that the group No 1 comprises mainly not too big apartments with the average living area of 46 square meters. These apartments are rather old as they were built in the second half of 20th century. The price of apartments corresponds with the median for the whole analysed data set.

In the table 4, there is presented structure analysis for the group No 2, consisting of 53 apartments. Although their living areas are similar to those of objects from the group No 1, it is clearly visible they are located in much higher buildings. An additional analysis revealed that all apartments from the group were in buildings with lifts. Moreover, apartments from the group were built on the turn of 20th century.

While analyzing results for the group 3, comprising 14 objects, that are reported in table 5, one may notice that this group contains large apartments. Their average living area is equal to 118 square meters. All objects are not older than 10 years.

Table 3. Descriptive statistics for cluster No 1.

Cluster No 1	Average	Standard deviation	Median	Mode
Price per 1m^2	8741.00	161.16	8595	8750
Area	46.62	1.11	46	38
Floor	1.73	0.12	2	1
Number of storeys	4.04	0.13	4	2
Number of rooms	1.96	0.06	2	4
Year of building completion			1974	1960

Source: own calculations.

Table 4. Descriptive statistics for cluster No 2.

Cluster No 2	Average	Standard deviation	Median	Mode
Price per 1 m^2	8364.43	176.65	8167	8750
Area	43.38	1.49	42	38
Floor	6.49	0.21	7	5
Number of storeys	10.15	0.35	10	10
Number of rooms	1.96	2.00	2	2
Year of building completion			1980	2001

Source: own calculations.

Table 5. Descriptive statistics for cluster No 3.

Cluster No 3	Average	Standard deviation	Median	Mode
Price per 1m^2	9145.07	637.85	9700	-
Area	118.00	3.78	118	100
Floor	2.62	0.58	2	2
Number of storeys	5.38	1.30	4	3
Number of rooms	4.54	0.18	5	5
Year of building completion			2004	2001

Source: own calculations.

In table 6, there are displayed results of structure analysis for the group No 4 comprising 44 objects. The structure of the group is similar to the structure of the

groups No 1 and 2 . The lower prices of apartments included in the group, suggest less attractive localization (longer distance from the city centre or right bank of the Vistula river) Those apartments are small, most of them consists of two rooms only and is located in high old buildings.

Table 6. Descriptive statistics for cluster No 4.

Cluster No 4	Average	Standard deviation	Median	Mode
Price per 1 m^2	7753.54	212.69	7586	7929
Area	52.09	1.87	50	49
Floor	1.41	0.16	1	1
Number of storeys	10.60	0.38	10	10
Number of rooms	2.50	0.11	2	2
Year of building completion			1979	1978

Source: own calculations.

Structure analysis of cluster 5 containing 90 apartments that is given in table 7, let us to conclude that apartments from the group are of similar structure as the structures of clusters No 1 and 4. This is also confirmed by Euclidean distances presented in table 2. Prices of apartments belonging to the group are a little higher than the value of median. They are located in buildings with the small number of storeys. Their living areas are rather large (circa 70 square meters).

Table 7. Descriptive statistics for cluster No 5.

Cluster No 5	Average	Standard deviation	Median	Mode
Price per 1m^2	8687.16	183.48	8585.50	11210
Area	70.02	1.23	67	70
Floor	2.34	0.17	2	1
Number of storeys	4.22	0.20	4	4
Number of rooms	3.06	0.06	3	3
Year of building completion			2001	2010

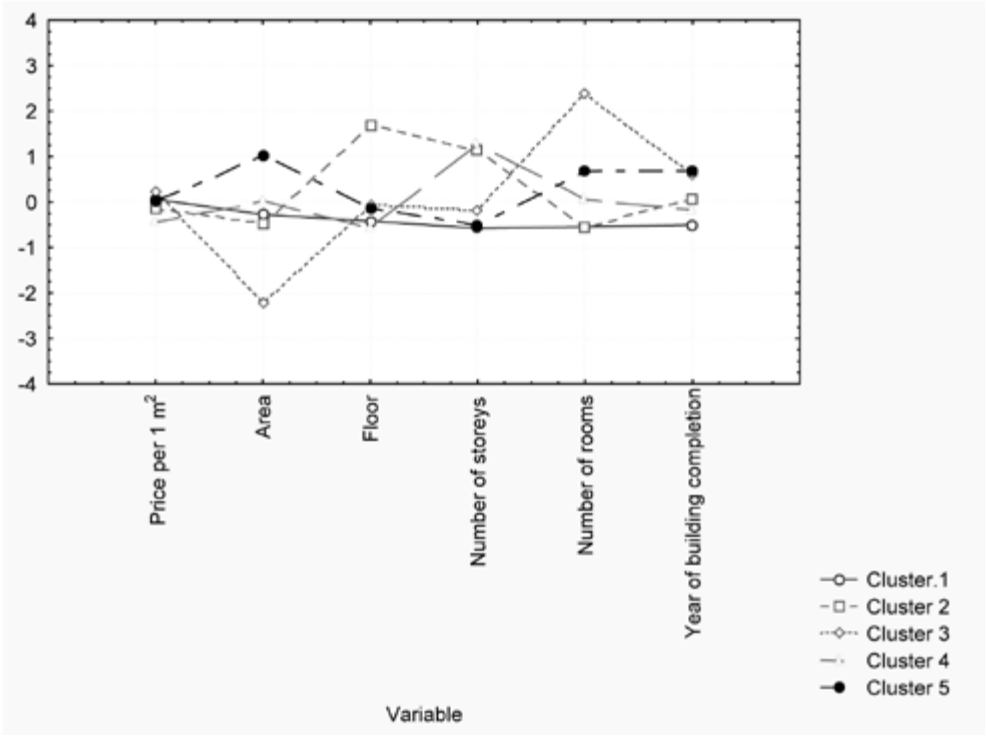
Source: own calculations.

Additional information on each of clusters is provided by the Figure 4. It enables the qualitative analysis of each group and its summary description:

- group No 1: the oldest apartments with middle living areas and moderate prices;

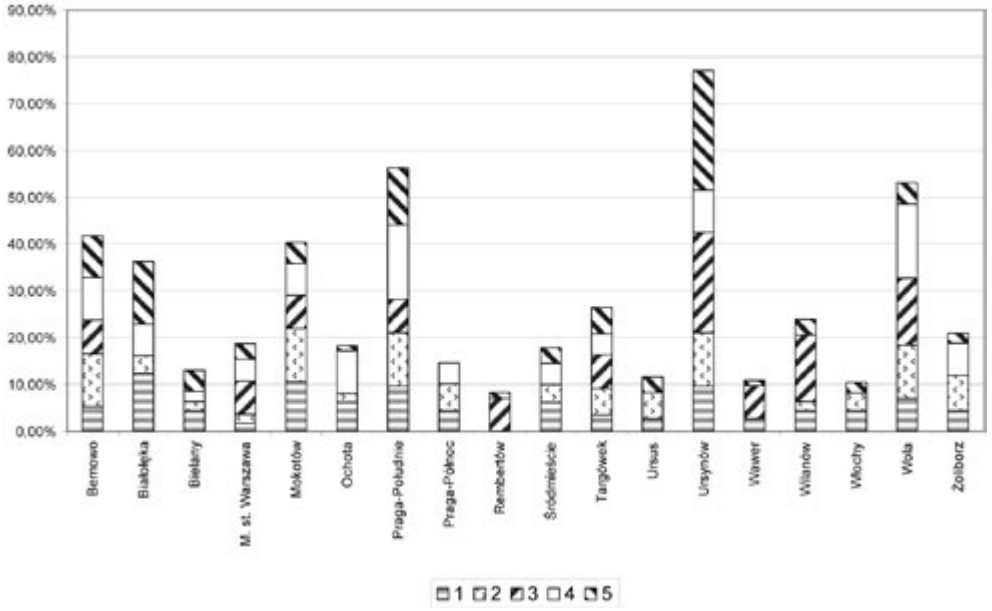
- group No 2: "middle age" apartments with low prices and middle living areas;
- group No 3: the newest apartments with high prices and large living areas;
- group No 4: old apartments with low prices and middle living areas;
- group No 5: the newest apartments with moderate prices and quite large living areas.

Figure 4. Plots of means for each cluster.



It is worth to mention that the cluster No 1 was dominated by apartments located in Białołęka district (12,39% of objects), Mokotów (10,61%), Praga Południe and Ursynów (9,73% each). In the cluster No 2 several districts had the same largest share of 11,32%. These districts are: Bemowo, Mokotów, Praga Południe, Wola and Ursynów. Ursynów district had also the largest share in the cluster No 3 (21,43%). In the cluster No 4, the largest shares were that of Praga Południe and Wola districts (15,91% each). Finally, in the cluster No 5, again Ursynów had the largest share (25,56%) – almost two times higher than the second share of Białołęka (13,33%). The detailed data, presented in the appendix, is illustrated in Figure 5.

Figure 5. Analysis of cluster structure with respect to apartment localisation.



4. Concluding remarks

Taxonomic methods are widely applied in numerous research areas such as marketing, biology, insurance, city-planning or capital market analysis. The aim of the paper was the application of some methods of multidimensional comparative analysis to investigating Warsaw real estate market. The study was based on data on more than 300 apartments offered on the secondary market in the fourth quarter of 2010. To investigate the data two alternative methods were applied: Ward's method and k-means method.

Multidimensional classification of investigated apartments obtained by the use of cluster analysis provided interesting quantitative and qualitative information on Warsaw real estate secondary market. The information is useful from both cognitive and practical points of view. The agglomerative procedure resulted in constituting classes of objects the most similar to each other with regard to the characteristics considered in the study and at the same time the most different from objects in other clusters. Finally, there were selected 5 clusters. Their identification and description allowed showing the phenomena of differentiation of objects from a given district and to reveal relationships among objects and their attributes as well as formulate conclusions valuable from the point of view of rational real estate management.

The results obtained may be an useful source of information for both individual investors and real estate agents, who may efficiently use them in order to raise

the standard of customer service as the methods allow to form quickly a group of apartments satisfying individual needs and preferences of interested clients. They can help to filter a set of offered apartments or to find the class of similar objects for the purpose of real estate appraisal. Nevertheless, the study presented in the paper is of preliminary character and the research is to be continued taken into consideration other apartment characteristics (not included in a present investigation) as well as other taxonomic methods of analysis.

Bibliography

Denkowska S., Salamaga M., Sit S., Sokołowski A. (2008). A taxonomic analysis of apartments sold in Poland in the period of from January 2004 to October 2007. Świat Nieruchomości 3/2008 (in Polish).

Dziechciarz J. (2002). *Econometrics*. Wrocław Economic University Pub. (in Polish).

Gatnar E., Walesiak M. (2004). *Methods of multidimensional statistical analysis in market research*. Wrocław Economic University Pub. (in Polish).

Grabiński T., Sokołowski A. (1980). The effectiveness of some single identification procedures. *Signal Processing: theories and applications*. Kunt M., De Coulon (ed.), North-Holland Publishing Company, EURASIP, Amsterdam.

Ostasiewicz W. (1999). *Statistical methods of data analysis*. Wrocław Economic University Pub. (in Polish).

Zyga J. (2011). Recognizing of real estate similarity. *Journal of the Polish Real Estate Scientific Society.* Vol. 19. No 4. pp. 141-157. (in Polish).

Appendix

Detailed structure of groups with respect to apartments localisation

District	Group					Sum
	1	2	3	4	5	
Bemowo	5,31%	11,32%	7,14%	9,09%	8,89%	7,96%
Białołęka	12,39%	3,77%	0,00%	6,82%	13,33%	9,87%
Bielany	4,42%	1,89%	0,00%	2,27%	4,44%	3,50%
M. st. Warszawa	1,77%	1,89%	7,14%	4,55%	3,33%	2,87%
Mokotów	10,62%	11,32%	7,14%	6,82%	4,44%	8,28%
Ochota	6,19%	1,89%	0,00%	9,09%	1,11%	4,14%
Praga-Południe	9,73%	11,32%	7,14%	15,91%	12,22%	11,46%
Praga-Północ	4,42%	5,66%	0,00%	4,55%	0,00%	3,18%
Rembertów	0,00%	0,00%	7,14%	0,00%	1,11%	0,64%
Śródmieście	6,19%	3,77%	0,00%	4,55%	3,33%	4,46%

(Continued)

District	Group					Sum
	1	2	3	4	5	
Targówek	3,54%	5,66%	7,14%	4,55%	5,56%	4,78%
Ursus	2,65%	5,66%	0,00%	0,00%	3,33%	2,87%
Ursynów	9,73%	11,32%	21,43%	9,09%	25,56%	14,97%
Wawer	2,65%	0,00%	7,14%	0,00%	1,11%	1,59%
Wilanów	4,42%	1,89%	14,29%	0,00%	3,33%	3,50%
Włochy	4,42%	3,77%	0,00%	0,00%	2,22%	2,87%
Wola	7,08%	11,32%	14,29%	15,91%	4,44%	8,60%
Żoliborz	4,42%	7,55%	0,00%	6,82%	2,22%	4,46%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Source: own calculations.