www.degruyter.com/view/j/remav

DE G DE GRUYTER OPEN

SEARCH FOR DISSIMILARITY FACTORS FOR NOMINALLY INDISCERNIBLE FACILITIES

Jacek Zyga, PhD
Faculty of Engineering and Architecture
Lublin University of Technology
e-mail: j.zyga@pollub.pl

Abstract

The similarity between premises, statutorily defined as the comparability of the compiled objects, is, in practice, identified by the statement of identity or the proximity of evaluation of the selected features describing the complied objects. Independent of the nature of these features (qualitative or quantitative), as well as of the method of identifying the proximity of their prices, the quest process for similar premises ends in the compilation of these premises into a group, which continues to remain inconsistent to some extent. The inevitable heterogeneity of the prices of premises summarized in this way leads to the suspicion that other features which had not been taken into account during the current stages of analysis exist. The identification of these features can significantly improve the quality of the valuation process based on the selected premises.

The presented article discusses a method of identifying the plausible number of indiscernible factors influencing the differentiation of unitary prices in an analyzed set of premises, in the event when the collected information on these premises does not provide sufficient knowledge on the reasons behind such a differentiation. The result of the performed research can, in practice, be applied to program the procedures for the data search, and hence reduce the costs related to the acquisition of information on the premises (e.g. from the real estate market).

Key words: *similarity, data models, data selection*.

JEL Classification: R15, C18, C38, C51.

Citation: Zyga J., 2015, *Search for Dissimilarity Factors for Nominally Indiscernible Facilities*, Real Estate Management and Valuation, Vol. 23, No. 3, pp. 65-72.

DOI: 10.1515/remav-2015-0026

1. Introduction

In the course of the analyses of data sets containing the descriptions and prices of the properties on the real estate market, the repetitive phenomenon was noticed of the characteristic grouping of the premises contained in them into subsets formally defined as equivalent groups of certain (selected) parameters. The selection of properties using an increasing number of dependent features/variables (exogenous) results in subsets whose elements become so formally similar (corresponding dependent variables assume the same values) that such a subset, in the case of its indiscernible elements, due to the dependent variables, makes it impossible to find the answer to the question of the reasons behind the variability of endogenous elements (response variable).

From the formal point of view, the outlined problem constitutes a case of searching for a set of attributes together with their corresponding values which would complement the knowledge of the starting set of information on the prices of selected market properties, so that the collected data could be recognized as an information system (in accordance with the definition of Z. Pawlak), constituting the starting point for further analyses. This step is different with regard to the majority of econometric methods which are based on the elimination of the dependent variables or their proper aggregation. The real estate market does not create the right conditions for collecting the information in abundance

(so that there is something to be eliminated), which provides the basis for the creation of dependent variables.

2. Databases on prices vs the information system

In order to discuss the information system, apart from the set of Y elements (e.g. the properties represented by their prices or the corresponding unitary prices), it is necessary to define and determine the attribute sets of A premises as well as the values of V attributes (PAWLAK 1983 p.16). Only on the basis of such sets and their elements, the described relationship $\rho(Y,A)$, $\rho: Y \times A \rightarrow V$ can be developed. All four mentioned elements together contribute to the information system:

$$S=<Y, A, V, \rho(y,a)> \quad (1)$$

Its simplest example is a typical table summarizing the prices and the features of properties which are, in this case, the dwellings presented in the table below:

Table 1

The data set describing a part of the real estate market as an example of an information system

Date	Area	Address *	Floor	Net price	Usable area	Number of rooms	Year of construction	Number of floors
2004-01-05	27	Wyżynna -	4th floor	1 535.35	49.5	3	1989	5
2004-01-09	27	Szmaragdowa -	2nd floor	2 291.67	36	2	1989	5
2004-01-10	29	Gliniana -	ground floor	1 852.68	58.24	4	1976	5
2004-01-19	19	Zimowa -	3rd floor	1 669.45	59.9	4	1983	5
2004-01-28	10	Nowy Świat -	1st floor	1 964.43	48.36	4	1976	5
2004-01-29	10	Wyścigowa -	1st floor	857.49	48.98	4	1974	5
2004-02-10	14	Koryznowej -	1st floor	1 388.46	32.41	3	1972	11
2004-02-18	30	Różana -	ground floor	1 862.89	67.1	4	1984	5
2004-02-20	5	Lipińskiego -	1st floor	1 994.10	57.67	5	1978	12

*numbers of buildings and premises were eliminated from the addresses

Source: self-study on the basis of the data from the Register of Real Estate Prices and Values of the Lublin City Office.

It should be pointed out that, in general, it is required to determine all the possible relationships $\rho(y,a)$ or, in other words, all the values of the attributes should be included in the table. This creates the possibility of the complete analysis of the dependencies between response and dependent variables. The first of the observed obstacles preventing such an analysis from being carried out is the disruption of the response variable correlation with a selected dependent variable. This disruption can be reflected in the assumption or identification of the same value of the relationship $\boldsymbol{\rho(y,a)}$ of a V domain for a certain part of attributes from the A set, which subsequently leads to the so-called non-discernibility of variables in relation to the specified attributes.

In accordance with the definition (PAWLAK 1983), the premises $y_1, y_2 \in Y$ are indiscernible when in an S system due to the attribute $a \in A$ only when $\boldsymbol{\rho_{y1}(a)= \rho_{y2}(a)}$. Nevertheless, the consequence of the non-discernibility is the decrease of the correlation coefficients of response and dependent variables, which in effect has an impact on the selection of the latter in the process of developing the econometric model, based on the S system.

Non-discernibility of properties (and in reference to real estate market analyses – prices of properties (usually unitary)) leads to confusion resulting from the lack of possibilities for identifying the cause-and-effect relations between the effect (price) and at least alleged reasons. An extreme case of such a phenomenon is the price $\boldsymbol{Y}$, which is not accompanied by any explanation. Sets $\boldsymbol{A}$ and $\boldsymbol{V}$ are thus empty.

3. The problem of data shortage

The case outlined above, against all appearances quite common, leads a lot of analysts to the conclusion that the lack of variability in relationship $\rho(y,a)$, in relation to certain attributes - corresponding variables of an exogenous nature, authorizes the element of the Y set to be treated as homogeneous and subjected to alterations solely due to random factors. In effect, the elements of the Y set are subjected to simple statistical tests together with the assessment that a good characteristic of this set is a simple arithmetic mean, despite the fact that such a set may be heterogeneous, which is rarely suggested by the visible range of extreme values of its elements. However, it is a simplification equalizing the identification of certain information shortages with information on the lack of evidence of the value variability of certain variables.

Observing the above, the question was raised whether it is possible to identify the variables of the properties which in the light of acquired information remain indiscernible? And further: in relation to the fact that the non-discernibility can, in particular cases, result from the lack of any information enabling the development of dependent variables in the created information system – whether it is possible to identify variability of premises solely on the basis of the analysis on information of exogenous nature related to them?

In the studied literature related to both the area of real estate management in particular, and to the econometric and information system research, no solutions or even tackling the similar problem were found. An issue addressed more commonly is the problem of interdependence between the recognized attributes (ARMSTRONG (1974); ORŁOWSKA (1980); HOZER et al. (2002); CZAJA, LIGAS (2010); BARAŃSKA (2004)) as well as the issue of careful and deliberate reduction of dependent variables (PAWLAK (1983); SKOWRON (1990); BAZAN et al. (1998); BEYNON (2001); BAGOZZI (2012); KUKUŁA (2004)) so as to reveal or highlight the essence of the analyzed issue. Such problems have been intensively analyzed for seventy years. Thus it should be borne in mind that the issues listed above constitute only a small fraction of the subject literature, from which it can be inferred that the analysts of various phenomena were frequently inspired by the abundance of information. After the publication of Z. Pawlak on the theory of fuzzy sets (PAWLAK 1982), a new direction of investigation was opened, which accounts for the incompleteness of data (Orłowska E., Skowron A, Chien-Chung C., Grzymala-Busse W.), yet the situation of their shortage was treated as a special case (CHIEN-CHUNG, GRZYMALA-BUSSE 1998).

4. The analysis of indiscernible variables case

While performing the analyses of a number of information sets on real estate prices, a regularity was observed that can be defined as the characteristic grouping of the objects intosubsets, depending on the criteria type of their ranking. It is demonstrated in Graphs No. 1 and 2. In Graph No. 1, a disordered collection of the unitary prices of dwellings (premises) is shown, sold in the years 2004-2013 in Lublin but updated in accordance with average annual prices to the price level from December 2013. The collection contains 1,728 records obtained from the Register of Real Estate Prices and Values of the Lublin City Office, covering data analogous to that presented in Tab. 1.

It is observed that the ranking of prices (updated) according to certain available criteria (e.g. according to the year of the building in which the premises are situated was constructed) results in the characteristic grouping of the same data which form (respectively in groups) a graph of the tangent function.

This, in turn, gave rise to the consideration on the selection of the group of records containing various prices still indiscernible due to the value of the remaining attributes. In this way, a sample group of premises situated in one building was obtained (identical construction year, number of floors) which cover the same number of rooms, and with the same usable area (39±0.5 m^2). The above collection, revealing the transaction prices and corresponding updated prices, is presented in Figure 3.

The Figure 4 depicts the collection of the updated prices in its characteristic tangent shape, resulting from the ranking in accordance with the increase of variable value. The values of the known attributes of premises represented by the above-mentioned prices are shown in Tab. 2.

The distinguished set of eight dwellings is an example of indiscernible premises (in the terminology of The Real Estate Management Act: similar). All the premises are described (apart from the variables "Floor" and "Usable Area") by the same values of particular attributes, and the correlation of the "Net price" variable with the "Floor" and "Usable Area" variables is negligible (Tab.

3). Despite such high compliance in the evaluation of all known attributes of the collected premises, their unitary prices (updated only for one date (31 December 2013)) are surprisingly discrepant. Their standard derivation (547.55 PLN/m²) amounts to 12% of their average value (4,570 PLN/m2), and the indicator of relative dispersion is:

Table 2

Collection of attributes of the selected premises

Area	Address	Floor	Net price	Usable area	Number of rooms	construction year	Number of floors
29 - Rury Św.Ducha	Gliniana 27	1	4 416.73	38.49	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	1	5 291.69	38.74	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	9	4 325.44	38.84	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	6	5 374.13	38.89	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	3	2 297.09	39.18	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	9	2 424.71	39.18	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	4	5 203.05	39.4	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	4	4 822.33	39.4	3	1 975	11
29 - Rury Św.Ducha	Gliniana 27	1	4 416.73	38.49	3	1 975	11

Source: self-study on the basis of data from the Register of Real Estate Prices and Values of the Lublin City Office.

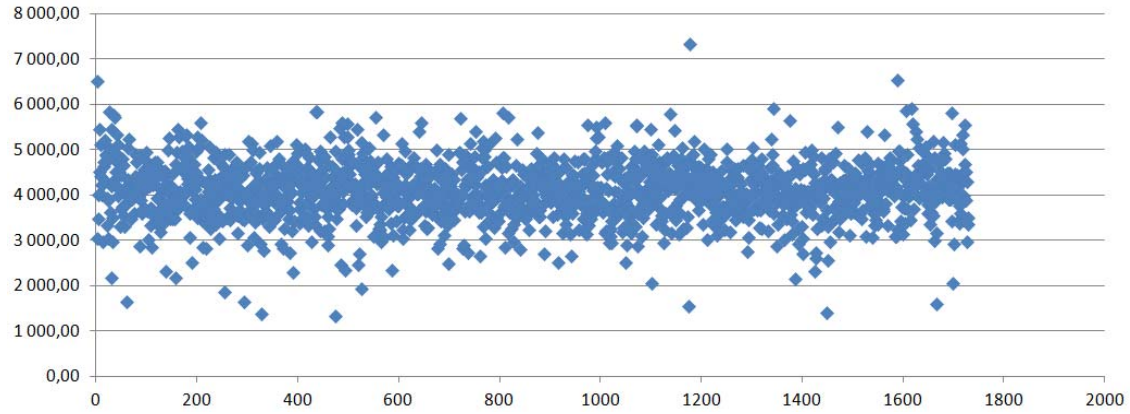

Fig. 1. Prices of dwellings in Lublin updated at the end of 2013. *Source:* own research.

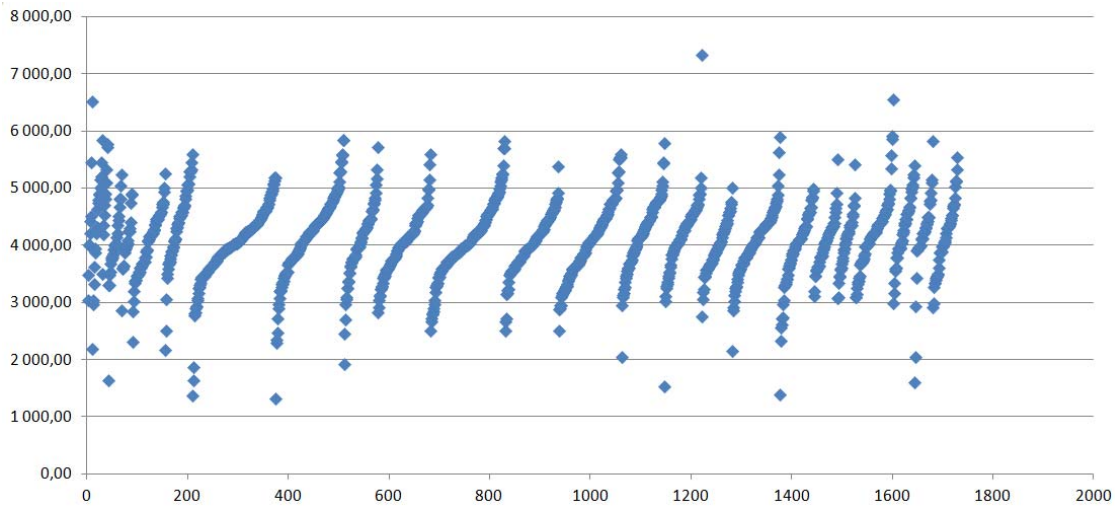

Fig. 2. Prices of dwellings in Lublin updated at the end of 2013 arranged in an ascending order according to the variable "year of construction". *Source*: own research.

Table 3

The correlation coefficient of the dependent variables with the response variable "net price"

Area	Address	Floor	Usable area	Number of rooms	Construction year	Number of floors.
-	-	-0.35	0.20	-	-	-

Source: own research.

$$Sc = 2\frac{Y_{\max} - Y_{\min}}{Y_{\max} + Y_{\min}} = 0.35 \quad (2)$$

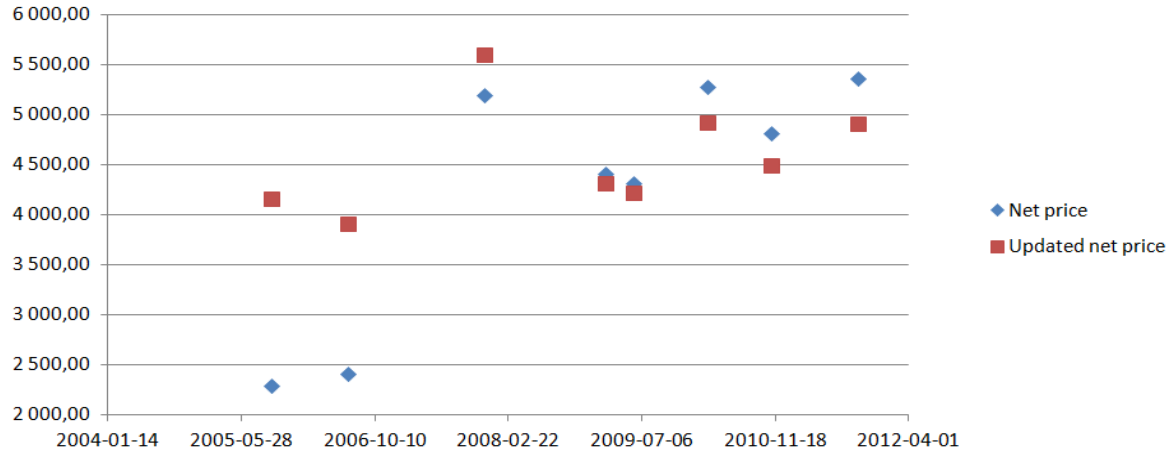

Fig. 3. Prices of selected dwellings in a building on Gliniana 27 Street in Lublin ranked according to dates. *Source*: own research.

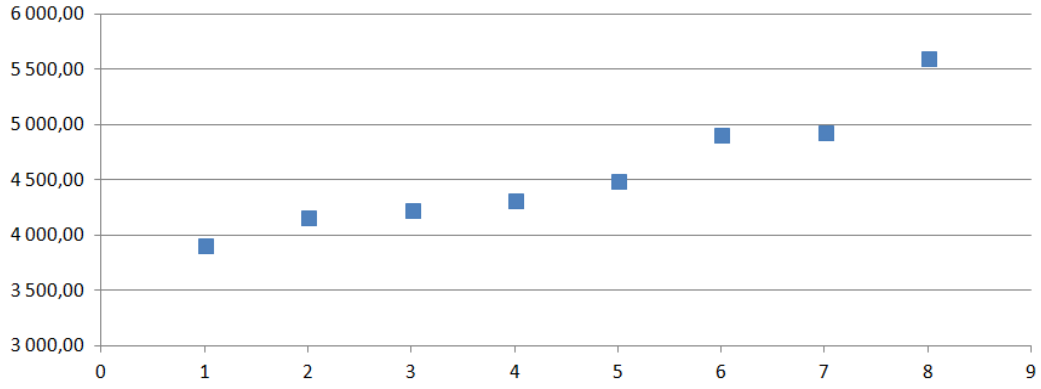

Fig. 4. Prices of selected dwellings in a building on Gliniana 27 Street in Lublin ranked according to prices. *Source*: own research.

In order to identify the potential factors responsible for the observed variability of the "Net price" variable, an interpretation of the price distribution ranked in ascending order was performed (see Graph No. 4) and it was assumed that the increase of the variable value can be related to the order of particular values in a set ranked in ascending order. Due to the analogy to the principles of the scree test applied to the factor analysis, the inflection point of the graph was selected (Item 5) assuming that, in the part of the set from items 1 to 5, a different increase in the value of the response variable was noticed, linear in relation to the factor of order. Similarly, in the part of the set from items 5 to 8, Table 3 and Figure 5 demonstrate the analysis of the slopes of the line approximating a particular subset.

Table 4

Analysis of line slopes approximating data in the subsets of the selected premises

No.	Raw values Y ('updated net price')	Normalized values Y_{norm}
1	3916.099	0.85691
2	4162.915	0.91091
3	4226.586	0.92485
4	4315.789	0.94437
5	4493.094	0.98316
6	4910.506	1.07450
7	4930.409	1.07885
8	5604.925	1.22645
Average Y_{avg}	4 570.04	1.00000
Deviation stand. σ_Y	547.55	0.120
σ_Y / Y_{avg}	0.120	0.120
The slope factor "a" of line approximating at appropriate subsets		
a1-8	212.9561	0.046598
a1-5	130.6863	0.028596
a5-8	335.5396	0.073422
The ratio of slope factors "a"		
a5-8/ a1-5	1.00	1.00
a1-5/ a1-8	0.61	0.61
a5-8/ a1-8	1.58	1.58
a5-8/ a1-5	2.57	2.57

Source: own research.

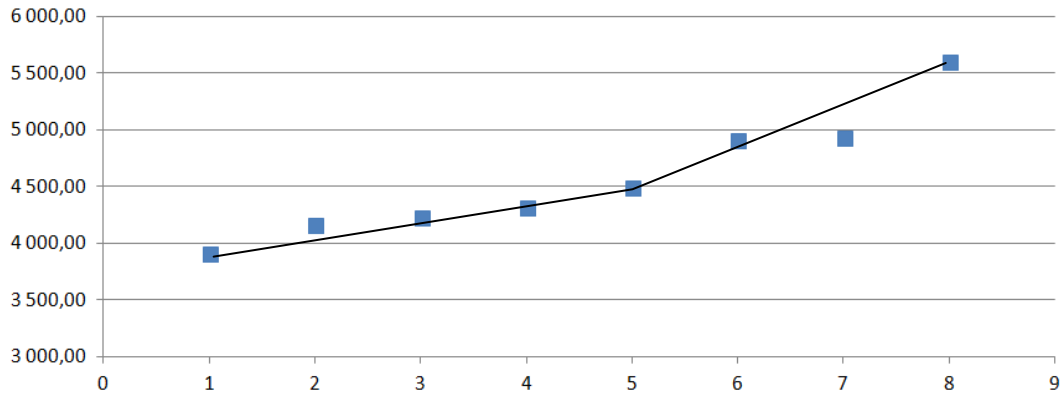

Fig. 5. Analysis of line slopes approximating data in the subtests of the selected premises. *Source*: own research.

Therefore, the following complementary table for the Y set was developed, related to the relationship ***ρ(y,a)*** with the identified attributes A: (Factor 1, Factor 2)

Table 5

Proposed values of the factors of the relationship ***ρ(y,a)*** for the identified attributes

No.	constant	Factor 1	Factor 2
1	1	1	0
2	1	2	0

3	1	3	0
4	1	4	0
5	1	5	0
6	1	5	1
7	1	5	2
8	1	5	3

Source: own research.

The above table constitutes a description of the dependencies of the attribute values with the response variable. In the case when there are no other assumptions, such relations are based on the linear combinations of attributes ***V***, additionally assuming the lack of impact of the attributes defining the non-discernibility of the response variable.

$$Y = A \cdot V + \varepsilon \tag{3}$$

Table 6

Estimation of the numerical values of attributes ***A*** for the tested model

Constant	Factor 1	Factor 2
3834.9	128.6	337.0

Source: own research.

The quality of the proposed model of the "Net price"(***Y***) variable was evaluated by determining the following measures: coefficient of determination R^2, mean squared error MSE, as well as Hellwig's coefficient of the integral capacity of information in comparison with the values of these measures for other alternative models of the "net price" variable. Their compilation is can be found in Table 6.

References to the proposed model were the following concepts of price modelling:

1. Horizontal line const. = Yavg (price average).
2. Slope line connecting the end to the start of data (No. 1 and No. 8).
3. Line dependency in relation to the two variables: Factor 1 = "No." and Factor 2 = "floor".
4. Line dependency in relation to one variable: Factor 1 = "floor".

Table 7

The comparison of price models

Variant	MSE	H	R2
Model proposed in Tab. 5	**65.01**	**0.943**	**0.962**
Horizontal line const. = Yavg (price average)	332.35	-	-
Slope line connecting the end to the start of data (No. 1 and No. 8)	169.62	0.91	0.91
Factor 1 = "No." and Factor 2 = "floor"	93.73	0.70	0.92
Factor 1 = "floor"	311.55	0.12	0.12

Source: own research.

The summary presented in Table 7 proves that, among the proposed variants of variable ***Y*** modelling representing the unitary prices (updated) of a certain group of premises recognized as practically indiscernible (similar), the first model, which was described in Table 5 and developed on the basis of observations of the variability of the element of the ***Y*** vector, is the best. This shows that the thesis on the homogeneity of the set elements represented by the ***Y*** variable (unitary prices of premises) is true, and that two unspecified factors are responsible for the ***Y*** variable. Their numerical representatives are obviously open (see Tab. 6), although at this stage of analysis, they do not have the interpretation of a market.

The rejection of the thesis on the non-discernibility of prices (elements of the ***Y*** set) indicates that the current recognition of the group of qualities of premises represented by the prices is not sufficient enough, and suggests that basing the estimation of an analogous facility (premises) on, e.g. simple

arithmetical mean of the collected prices, can lead to incorrect results. Hence, there arises the suggestion of taking on additional research activities aiming at detecting and identifying the factors recognized as determining the response variable (unitary prices) and at interpreting them from the point of view of an appropriate theory underlying the model construction of this variable, and consequently to measure this factor and compliment the initial model by the subsequent dependent variables.

In the presented case, it was stated that the evaluation of the technical condition and equipment of particular premises can be assumed as the identified factors. Regrettably, due to the lack of consent of the owners of premises in question to participate in a survey enabling the inventory of these features, further verification of the results shown was not possible.

4. Conclusions

Although the test was performed on a narrow scope of data, its results shall be recognized as positive, especially for analysts of the real estate market. The conducted research shows that there is an alternative to the construction of an econometric model using the methods of data elimination or aggregation. This is significant especially in conditions of extremely limited information, enabling the creation of dependent variables, and in cases when already collected dependent variables suggest, contrary to the variability of the response variable, the non-discernibility of the gathered elements. The variability of the response variable can be the reason for questioning the thesis of the non-discernibility of the gathered elements of the developed information system, and thus, may help establish at least its potential sources (dissimilarity factors). In practice, the outcome of the conducted research can serve to develop procedures of intensive search for data (as opposed to extensive reliance on the analyses of large data sets collected from the stock), and thus to reduce the costs of acquiring information on facilities (e.g. from the real estate market).

5. References

ARMSTRONG W., 1974, *Dependency structures of data base relationships*. Proc, IFIP Congress, Stockholm.

BAGOZZI R.; YI Y., 2012, *Specification, evaluation, and interpretation of structural equation models*, Journal of the Academy of Marketing Science, Vol. 40, No. 1, pp. 8–34

BARAŃSKA A., 2004, *Wybór cech nieruchomości do modelowania matematycznego wartości rynkowej na przykładzie kilku baz nieruchomości gruntowych, (Selection of real estate features for mathematical modelling of the market value exemplified by several databases of ground real estates),* Geodezja, Vol. 10. No. 1. pp. 31-38

BAZAN J.G., NGUYEN H.S., NGUYEN T.T., SKOWRON A., STEPANIUK J., 1998, *Synthesis of decision rules for object classification, W: Incomplete information: rough set analysis*, Red. E. Orłowska, Studies in fuzziness and soft computing, vol.13, Physica-Verlag, pp. 23-34.

BEYNON M., 2001, *Reducts within the variable precision rough sets model: A further investigation*, European Journal of Operational Research, Vol. 134, Issue 3 pp. 592-605.

CHIEN-CHUNG C., GRZYMALA-BUSSE W., 1998, *On the Lower Boundaries in Learning Rules from examples, W: Incomplete information: rough set analysis*, Red. E. Orłowska, Studies in fuzziness and soft computing, vol.13, Physica-Verlag, pp. 23-34.

CZAJA J., LIGAS M., 2010, *Zaawansowane metody analizy statystycznej rynku nieruchomości, (Advanced statistical methods for the analysis of real estate market),* Studia i materiały Towarzystwa Naukowego Nieruchomości, Vol. 18. No. 1. pp. 7-20

HOZER J., KOKOT S., KUŹMIŃSKI W., 2002, *Metody analizy statystycznej rynku w wycenie nieruchomości, (Methods of statistical market analysis in real estate valuation),* Wyd. PFSRM,. Warszawa.

KUKUŁA K., 2004, *Badania operacyjne w przykładach i zadaniach, (Operational research in the examples and tasks),* PWN, Warszawa.

PAWLAK Z., 1982, *Rough sets*, International Journal of Computer and Information Sciences, Vol. 11, No. 5, pp. 341-356.

PAWLAK Z., 1983, *Systemy informacyjne. Podstawy teoretyczne, (Information systems. Theoretical basis),* WNT, Warszawa.

ORŁOWSKA E., 1980, Dependency of attributes in information systems. ICS PAS Report, Warszawa.

SKOWRON A., 1990, *The rough sets theory and evidence theory*, Fundamenta Informaticae, pp. 245-262.

Rejestr Cen i Wartości prowadzony przez Urząd Miasta w Lublinie (Register of Real Estate Prices and Values kept by the Lublin City Office).