sciendo

SIMULATIVE VERIFICATION OF THE POSSIBILITY OF USING MULTIPLE REGRESSION MODELS FOR REAL ESTATE APPRAISAL[1]

Sebastian Kokot
Faculty of Economics and Management
University of Szczecin
e-mail: sebastian.kokot@usz.edu.pl

Sebastian Gnat
Faculty of Economics and Management
University of Szczecin
e-mail: sebastian.gnat@usz.edu.pl

Abstract

The possibility of using multiple regression models in real estate valuation is the subject of disputes, both in theory and in practice. Econometric modelling is a difficult process, since a number of issues of substantive and numerical nature occur during that process. Modern technologies enable quick and easy model estimation with the use of virtually any quality of data. Naturally, it provokes property appraisers to use such models in the practice of real property valuation, particularly in mass appraisal, frequently without taking those issues into account. Consequently, the models obtained and applied in practice turn out to be of poor quality and, objectively speaking, should not serve as the basis for determining real estate value. The specificity of the real estate market and of the real properties themselves as objects traded in that market additionally exert a negative impact on the quality of the obtained models.

In this article, the authors present the results of research which involved a simulation of various types of disturbances of a model artificially developed database of real estate prices and attributes as well as their impact on the quality of estimated models. The research will make it possible to answer the question of the degree and type of disturbances that are permissible in the functioning of a real estate market if the estimated models are to still satisfy the qualitative requirements defined for them, and thereby produce accurate valuation results. A model database will be disturbed by the deviation of prices from model prices and by reducing its size. Radom generators were used to obtain database disturbances.

Key words: *real estate valuation, mass valuation, method of statistical analysis of the market, multiple regression models*.

JEL Classification: *C51, R52*.

Citation Kokot S., Gnat S., 2019, *Simulative Verification of the Possibility of Using Multiple Regression Models for Real Estate Appraisal*, Real Estate Management and Valuation, vol. 27, no. 3, pp. 109-123.

DOI: 10.2478/remav-2019-0029

[1] The research was conducted within the project financed by the National Science Centre, Project No 2017/25/B/HS4/01813.

1. Introduction

Econometric models based on multiple regression constitute one of the most frequent ways of employing the statistical analysis of the market in practice within the scope of a comparative approach to real estate valuation (FORYŚ, GACA 2018). A statistical analysis of the market is one of three methods of real estate valuation in the comparative approach sanctioned by legal regulations– by the secondary legislation to the Land Management Act, namely the ordinance of the Council of Ministers on real estate valuation and appraisal report preparation. Formally, in pursuance of the ordinance, when using the method of a statistical analysis of the market, a set of transaction prices is adopted, which are suitable for determining the values of representative real estate, referred to in Article 161.2 of the act. A real estate value is determined with the use of methods applied in statistical analyses. The question regarding the possibility of using the method of statistical market analysis for real estate valuation conducted by property appraisal experts on a regular basis still remains disputable in itself, since some of the appraisers' circles believe that the reference made in the above-invoked ordinance to the provision of the act concerning universal real estate taxation disqualifies the objective method from the practice of valuations carried out for purposes other than universal taxation (HOZER, KOKOT, DOSZYŃ 2018). However, the opposite opinion is equally common. According to such a view, the method of statistical market analysis is a fully-fledged valuation method, which can always be used, provided that the application of particular statistical methods is justified by methodological and substantive considerations and leads to credible results, on the condition that the set of the selected transaction prices originates from the territory of a municipality, and in the event of an insufficient number of transactions, from the territory of adjacent municipalities and provided that the remaining conditions for the application of a comparative approach are satisfied, i.e. knowledge of the transaction prices and the attributes of similar real estate (cf. section 4.1 of the ordinance) and provided that suitable price sources are used (section 5 of the ordinance). Consequently, some property appraisers employ the method of statistical market analysis in professional practice, as previously mentioned, most typically with the use of econometric models based on multiple regression. These types of models are applied especially in the field of so-called mass valuation, in which using a uniform approach, a large number of real properties of the same kind are valued simultaneously and for the same purpose, and with the expectation of obtaining cohesive results (cf. HOZER, KOKOT, KUŹMIŃSKI 2002; KURYJ 2007; TELEGA, BOJAR, ADAMCZEWSKI 2002). Over the years, there have been many disputes as to the legitimacy of use of these models for real estate valuation and their adequacy for describing the phenomena occurring in the real estate market, including the ones observed in price trade, while a number of property appraisers and scientists examining the issues of real estate market point to numerous problems related to their use (PRYSTUPA 2000; PAWLUKOWICZ 2001; PAWLUKOWICZ 2002; ADAMCZEWSKI 2006; BITNER 2007; PAWLUKOWICZ 2007; PARZYCH 2009; BARAŃSKA 2010A; BARAŃSKA 2010B; LIGAS 2010; KOKOT, DOSZYŃ 2011; ZURADA, LEVITAN, GUAN 2011; ZBYROWSKI 2012; PARZYCH, CZAJA 2015; DOSZYŃ, GNAT 2017; BIEDA 2018; GACA 2018). In some of the above-mentioned works, as well as in the practice of valuation, unsuccessful regression models can be found. However, the very fact of discovering a defective model does not provide the grounds for claiming that it cannot serve as a useful tool for determining real estate market value (PAWLUKOWICZ 2007; GACA 2017). A separate trend in research concerns the comparisons of multiple regressions models and other types of models used in the real estate valuation, including the ones based on machine learning (YOO, IM, WAGNER 2012; ANTIPOV, POKRYSHEVSKAYA 2012; ČEH et al. 2018). It appears that, in many cases, the results obtained through the application of precisely those models, based on multiple regression, produce better results than e.g. models based on other approaches (ZURADA, LEVITAN, GUAN 2011). Therefore, a study was undertaken in this paper that involves a simulation of specific disturbances in the functioning of the real property market understood as various levels of reflection (or non-reflection) of real estate market attributes by transaction prices, and valuing suitable price models with the use of multiple regression in thus disturbed conditions. The undertaken research aims to provide the basis for answering the questions of:

- whether regression models give accurate valuation results in a properly functioning market?
- how significant disturbances in market functioning can be accepted for regression models to continue producing accurate valuation results?
- what type of disturbances affect the quality of valued models?

For the purpose of this paper, the term of "a properly functioning market" is understood as a situation in which real estate prices constitute a real reflection of the proper state of market attributes of such real estate. The term of correct valuation results, in turn, is understood as high formal estimation of a model in terms of the obtained verification measures, i.e. a high degree of model fit to input data (R^2), a low degree of volatility coefficient, statistical significance of structural parameters of a model (measured with a t-Student test) and logical interpretation of structural parameters (the occurrence of coincidence effect). The presented measures do not exhaust the possibility of evaluating the quality of regression models.

Real estate researchers do not undertake studies of a simulative nature often, particularly for the purpose of verifying the very methods used for analyses. A simulation is a type of experiment understood as a scientific method. In the most general terms, it involves developing a model reflecting a certain phenomenon and conducting experiments by manipulating an independent variable. A simulation is considered to be a specific research method. It is frequently placed between studying physically existing objects (experimenting with real systems) and an analytical study (building mathematical, formal models). Only three categories of objectives are assigned to a simulation: predictive, identifying and rationalizing objectives. Simulations in science were originally used as early as in the 1940's. Initially they served to solve mathematical models. In those times, digital machines were still unknown, thus simulations were highly simplified. Along with a dynamic development of information technologies, simulation research was commonly conducted with computers. In the most general terms, a computer simulation is a computer program that enables the behaviour of an abstract model of an analysed system to be imitated. The results of a study of a similar concept, though of a significantly smaller scope, were presented in an article titled: *The impact of real estate market ineffectiveness on the precision of real estate value description with the use of linear models of multiple regression* (MIESZEK, DZIADOSZ 2011). Furthermore, the article contains a postulate for deeper analyses, which are hereby provided. There are also studies in which models are valued on the basis of various scopes of empirical data (e.g. BARAŃSKA, MICHALIK 2014).

2. Econometric models of multiple regression as a tool of real estate valuation

When discussing the possibility of employing econometric methods, or even more broadly – statistical methods - in any sphere, including in real estate valuation, one needs to be aware that this may be done exclusively when certain conditions for their application have been met. Otherwise the results obtained through their application will be incorrect. Such basic conditions include (PAWŁOWSKI 1980):

- an examined regularity being constant,
- examined phenomena being measurable,
- a sub-set of significant factors being distinguishable out of the factors that affect the shaping of the examined phenomenon,
- the existence of suitable statistical data regarding the examined phenomena.

The examined phenomena ought to have a mass character, thus they need to be sufficiently numerous for the results to be referred to the entire set - in real estate valuation - to the entire local real estate market. Only the satisfaction of these conditions provides the grounds for attempts at building econometric models. Drawing conclusions on the basis of singular events does not constitute statistical conclusions, and such conclusions do not constitute statistical trends.

An econometric model is an equation (or a system of equations) presenting a stochastic relationship between core variables characterizing the examined economic phenomenon (ed. HOZER 1997). Models used in real estate valuation constitute mathematical and statistical forms of recording the relationships between a real property value (explained variable) and the factors shaping it (explaining variables). The methodological problems of econometric individual valuation are discussed, inter alia, in the article titled "Econometric real estate valuation and the Szczecin Algorithm of Real Estate Mass Appraisal – an econometric approach" (DOSZYŃ 2012, DOSZYŃ, HOZER 2017). Models making use of multiple regression need to be considered as a traditional approach to large scale real estate valuation (mass appraisal) (MARK, GOLDBERG 1998). As had already been emphasized in the introduction, methods based on econometric models of regression also belong to the most frequently undertaken attempts at implementing the method of a statistical analysis of the market (see, inter alia PAWLUKOWICZ 2001). However, the results obtained with them are typically not satisfactory, examples of which could include, inter alia, the inadequacy of signs next to the structural

parameters of models for logically or even intuitively understood relationships between particular explaining variables and an explained variable, as well as the results of interactions between the variables (LARSEN, PETERSON 1988; MARK, GOLDBERG 1988; PRYUSTUPA 2000; LIMSOMBUNCHAI, GAN, LEE 2004). This is caused by difficulties in satisfying a number of formal conditions already set at the stage of model construction, and these lie at the foundation of the broadly understood imperfection of the real estate market (see KUCHARSKA- STASIAK 2006; KUCHARSKA- STASIAK 2016). Irrespectively of the above, failure to account for spatial relationships between individual observations is considered to be a serious defect of multiple regression models, which consequently manifests itself in the autocorrelation of model residuals (DUBIN, PACE, THIBODEAU 1999). Spatial autocorrelation is perceived as the effect of the poor specification of a regression model, which can be caused by incomplete or a lack of spatial data, i.e. the data that take into account the spatial relationship of recorded transaction prices (JAHANSHIRI, BUYONG, SHARIFF 2011).

The process of building an econometric model was described in literature on multiple occasions (inter alia PWAŁOWSKI 1980; ed. HOZER 1997). Overall, it can be divided into the following stages: specification, estimation, verification. Presently, the theory and practice of econometrics are highly successful in dealing with the two last stages (estimation, verification). There are many methods of econometric model estimation, and currently their application comes down to the use of a suitable command in an econometric software package. Furthermore, there are many tests verifying the properties of model parameter estimators, correctness of the analytical form, or the properties of a component of a chance pattern. The least developed area involves model specification, which comes down to selecting a set of explaining variables and choosing an analytical (functional) form of a model. The problem of the accurate specification of an econometric model is a broader issue, concerning economics as a whole scientific discipline. Econometrics often lacks good (universally acceptable) theories that make it possible to determine which factors impact the examined phenomenon (HOZER, KOKOT, DOSZYŃ 2018).

An econometric model based on multiple regression is typically represented by the following equation:

$$Y = a_1X_1 + a_2X_2 + \cdots + a_kX_k + U \quad (1)$$

where:

Y – dependent variable – value of real estate, where transaction prices are used most often as observations,

$X_1, X_2, \ldots, X_k$ – independent variables – adequately expressed market attributes of real estate,

$a_1, a_2, \ldots, a_k$ – structural parameters of a model, which are estimated with an adequate method (e.g. the least squares method),

U – random component of a model, i.e. the part of the explained variable value that is not explained by the explaining variables.

At the age when computing algorithms are universal, the construction of econometric models has become seemingly easy. The stage of painstaking and complicated mathematical calculations aiming to estimate structural parameters, and then to estimate a set of measures intended for model evaluation, has been reduced to virtually a fraction of a second, simultaneously eliminating calculation errors. This means that it is sufficient to enter data into a computer regarding the recorded real estate prices along with its attributes and a model is ready. In order to calculate the market value of real estate, it is enough to substitute properly quantified states of valued real estate attributes for individual explaining variables. The ease and speed of calculations makes the use of this tool by property appraisers in their professional practice highly tempting. However, is real property value modelling really that easy? Basic problems encountered in building a regression model intended for real estate valuation have been briefly outlined below.

1. An adequate choice of explaining variables. In order to complete this stage, knowledge of the subject matter (expert knowledge) is required, i.e. the knowledge of trends existing in a local real estate market. Unfortunately, in practice, we often follow common patterns and, in a way, we use standard sets of attributes. This leads to a situation in which we account for attributes which do not in fact have any impact on transaction prices, or whose impact is decidedly smaller than assumed. Fragmentary research proves that appraisers have a tendency to exaggerate the weights of individual attributes (KOKOT, BAS 2016). Moreover, the issue of the quality of information on real property, i.e. database quality, is of exceptional significance. The number of transactions,

particularly in small local markets in a given period, may occur to be too low for a constructed econometric model to be of sufficient quality for real estate valuation. The attributes of real estate typically assume a qualitative nature, thus they are in fact immeasurable. Their presence may be determined, but defining their intensity is difficult and highly subjective. It is probable that two property appraisers will assign different attributes and different states of such attributes to the same real estate. As a result, the information in those appraisers' data bases will differ, although objectively these are the same real properties. This problem concerns all valuation methods, not only econometric models.

2. The manner in which qualitative explaining variables are taken into account in an econometric model. Let us consider an attribute of "real estate environment", for example. Let us assume that this attribute may appear in three states: 0 – unfavourable, 1 - neutral, 2 – favourable. By incorporating an attribute thus defined into an econometric model as one of the explaining variables, i.e. adopting its three values: 0, 1, 2, we risk losing information on distances between individual states, since we assume that they are equal. Meanwhile, the market may perceive them differently. It is worth adding that such an approach is even found in numerous scientific publications where econometric valuation models are constructed (see, inter alia, SAWIŁOW 1995; ŹRÓBEK, BEŁEJ 2000; DACKO 2000; LIPIETA 2000; ZADUMIŃSKA, SZTAUDYNGER 2001). The values of 0, 1, and 2 ought to be treated only as "codes" signifying the occurrence of a given state. They indicate nothing in terms of "by how much", or "how many times" a given attribute state is better than the other. What is more, such a perspective of explaining variables typically results in their strong collinearity, which leads to a catalysis effect and to other negative numerical consequences during model estimation. Consequently, we can obtain a model that is formally well fitted to the data (with a high R^2 determination coefficient), but this fit results from the collinearity of explaining variables, and not from a good relation between explaining variables and an explained variable. In order to prevent this, each state of an attribute ought to be taken into consideration as a separate dummy variable. Such a variable signals whether or not a given attribute state occurs. However, taking into account explaining variables in such a way results in generating a large number of explaining variables, which in turn translates into a requirement of ensuring a sufficiently high number of observations (inter alia: ACZEL 2011; MADDALA 2006). It is a frequently applied method, involving a switch from an ordinal scale to a ratio scale. It is not entirely proper, because it assumes a linear shift between the states of market attributes. A more appropriate approach calls for a transformation of each attribute into a k-1 dummy variable (where k is a number of variants of a given attribute). Yet, owing to the fact that linear transitions were used in an initial, hypothetical data set, the variables were kept in a ratio scale. The use of dummy variables does not change the research results in relation to the presented approach (the finding was further confirmed with calculations performed during the research). Furthermore, precisely this way of treating the states of real estate attributes coincides with how property appraisers proceed in the valuations conducted with a comparative approach.
3. The analytical form of a model (type of a functional relationship). Linear models constitute a predominant majority of real estate valuation models found both in literature and in practice. Model creators do not usually even take into account the option that a relationship between explaining variables and prices can be other than linear. Yet, it may appear that attributes are related to real estate value in a nonlinear manner. Frequently, the variability of real estate attributes and prices is insignificant in itself. This may cause linear models to approximate the relationship well enough only in a given data set. Nevertheless, this does not mean that the actual relationships are linear. When such types of linear models are extrapolated, which may occur in the process of valuation, it turns out that the relationships are not linear after all. When detecting the type of a functional relationship we typically need extensive data sets, which are not always at our disposal. Additionally, it is important for real properties to be varied, since otherwise, insufficient variability of explaining variables will not permit the type of a relationship to be determined.
4. Properties of a random component. A random component reflects the factors not accounted for in a model as explaining variables. In econometric valuation models, a random component is usually heteroscedastic and a spatial autocorrelation will occur. Heteroscedasticity of a random component (instability of its variance) results from real estate diversity. Real properties are not homogeneous.

Spatial autocorrelation, in turn, is related to the fact that typically "better" properties border with "better" ones, while "worse" properties are adjacent to "worse" ones. Heteroscedasticity and autocorrelation decrease the efficacy of estimators (evaluations of structural model parameters), i.e. in other words, the results of model parameter estimation deteriorate.

5. The above-described problems of a substantive and numerical nature are compounded by the conditions of price modelling, resulting from the specificity of the researched area, namely the real estate market. The nature of the real estate market is extremely distant from a theoretical category of a perfect market, since it is characterized by, inter alia, complete heterogeneity of products, significant participants' ignorance of the phenomena occurring in the market, a frequent trend of participants following irrational reasons, a relatively low number of buyers and sellers, and significant entry and exit barriers (cf. inter alia KUCHARSKA-STASIAK 2006; NOWAK, SKOTARCZAK 2013; KUCHARSKA-STASIAK 2016).

The proper construction and use of econometric models of real estate valuation requires very good knowledge of econometrics, both in theory and in practice, as well as understanding the specificity of the real estate market itself as an area in which those models are to be applied. The construction of a model does not merely involve "entering" numbers into a computer and reading the results. The process of building a model requires resolving a number of issues each time, such as choosing variables, entering explaining variables accurately into a model, defining a functional form, choosing an estimation method, evaluating model quality (its verification). Building an econometric model requires a good – reliable and credible – data base, sufficiently numerous, without diverging observations, with the attributes of a suitably high degree of variability.

Universal availably of calculation packages makes it easy to learn valuation and, in general, to use econometric models. However, without proper in-depth studies it is difficult to comprehend them from the numerical perspective. This refers equally to the conditions of applying the methods of model estimation as well as the ability to assess the quality of an estimated model, and in particular the statistical significance of its structural parameters, autocorrelation of model residuals, or the so-called catalysis effect. All this is compounded by specific features of the real estate market, hindering the process of model estimation; such features may include the relatively low number of transactions, a subjective perception of real estate attributes and difficulties in their standardization, as well as emotional factors frequently accompanying buyers in decision-making, and thereby significantly affecting the prices they pay.

3. Research method

For the purpose of conducting the study, it was assumed *apriori* that the conditions listed in section 2 for using econometric methods have been fulfilled. The study has a statistical character, the variables (market attributes) were quantified and it was assumed that they constitute significant factors affecting price levels. The study was conducted with the use of a model, artificially developed data base of real estate prices and attributes. The data base was built under the assumption that the obtained prices are influenced by 5 market attributes marked as I, II, III, IV, V, of which each one is expressed in three states – the weakest, average and strongest – A, B, C respectively. Distances between individual states are equal, i.e. "B" state is better than "A" state by the same degree it is better than "C" state. Each combination of attribute states occurs in the data base only once and, at the same time, the data base contains all the possible attribute states. The result is a number of observations (potential or theoretical transactions) equal to $3^5 = 243$. The following weights were assigned to the attributes:

- Attribute I – 10%,
- Attribute II – 10%,
- Attribute III – 20%,
- Attribute IV – 20%,
- Attribute V – 40%.

Model transaction prices of the remaining simulated real properties were determined with the given market weights, amount adjustments and a real estate transaction price of the lowest variants of all market attributes. The type of real estate and attribute types are not named here. The examined set is a purely hypothetical set of an ideal functional relationship, which will be used to achieve the research objective, i.e. determining the impact of the degree of disturbance of an ideal, functional relationship between prices and variants on the quality of a linear regression model. Model real estate

prices of particular attribute states were determined in such a way so that the price of the real estate of the weakest states of all attributes would amount to 100 units, the real estate of the best states of all attributes – 200 units, while the prices of real estate of particular combinations of average attribute states were determined in such a way so that they would strictly correspond to those combinations. Since it was assumed that the distances between attribute states are equal, individual attribute states were quantified by assuming the value of 0 for the weakest state, the value of 1 for the average state, and the value of 2 for the best state. The adopted method of attribute quantification for the purpose of the conducted research is permissible, since as was indicated in the second part of the paper, such presentation of attribute description is possible in a situation when distances between attribute states are equal. The issue of collinearity of variables, in turn, does not occur on account of the assumed manner of data base construction. Fragments of a model data base are presented in Table 1.

Table 1

Fragments of a model data base

No.	Attribute I (weight 10%)	Attribute II (weight 10%)	Attribute III (weight 20%)	Attribute IV (weight 20%)	Attribute V (weight 40%)	Price
1	0	0	0	0	0	100
2	0	0	0	0	1	120
3	0	0	0	0	2	140
4	0	0	0	1	0	110
5	0	0	0	1	1	130
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
40	0	1	1	1	0	125
41	0	1	1	1	1	145
42	0	1	1	1	2	165
43	0	1	1	2	0	135
44	0	1	1	2	1	155
45	0	1	1	2	2	175
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
140	1	2	0	1	1	145
141	1	2	0	1	2	165
142	1	2	0	2	0	135
143	1	2	0	2	1	155
144	1	2	0	2	2	175
145	1	2	1	0	0	125
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
239	2	2	2	1	1	170
240	2	2	2	1	2	190
241	2	2	2	2	0	160
242	2	2	2	2	1	180
243	2	2	2	2	2	200

Source: own study.

The study comprised two stages. The first one provides for a random adjustment of model transaction prices. The adjustment involves a change of each transaction price by a random coefficient from the assumed range of, e.g. ⟨−10%, 10%⟩. After price disturbance, the estimation of a multiple regression model follows and a statistical evaluation of its usefulness takes place. Several disturbance ranges, from 10% to 60%, will be assumed in the study. For each of the disturbance levels, a price adjustment will be performed 5000 times, which means 5000 estimated equations. This approach aims to eliminate random results that could occur with an insufficient number of experiment repetitions. Altogether with 6 disturbance levels in the first stage of the study, 30001 regression equations were estimated (one equation for an undisturbed set). During the second stage, an additional dimension was added, which involves a random reduction in the number of transactions in the set. This was done in order to counter a possible accusation frequently levelled against the use of multiple repression models of insufficient sample representativeness (set of transactions). For the previously employed ranges of disturbance, models were estimated with a set numbering between 10% and 100% of an initial transaction set size by a 5% step, which means 19 different set sizes. As a result 19*7 = 133 combinations were created. 1000 models were estimated for each of them. Therefore, the second stage entails the estimation of more than 100 thousand models of multiple regression models. The measures used in the estimation of each of 1000 models in each of 133 combinations were averaged for the purpose of visualization. A determination coefficient, volatility coefficient, and statistical significance of model structural parameters were employed in order to assess the possibility of using linear regression models on data sets diverging degrees from an ideal set in various.

4. Simulation results

As previously mentioned, the first stage of the study involved various degrees of disturbance of transaction prices. The prices from the model set were disturbed with a random coefficient within 6 ranges: ⟨−10%, 10%⟩, ⟨−20%, 20%⟩, ⟨−30%, 30%⟩, ⟨−40%, 40%⟩, ⟨−50%, 50%⟩, ⟨−60%, 60%⟩. For each of the 6 ranges, the disturbance was conducted 5000 times. Figure 1 presents R^2 determination coefficient for the estimated equations.

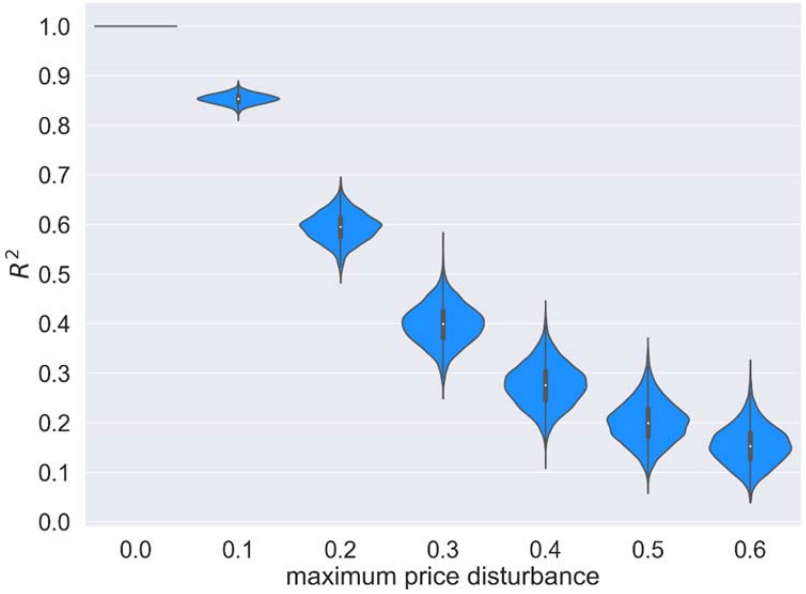

Fig 1. Kernel density estimations (KDE) of determination coefficients of the models estimated with various maximum levels of disturbance of model transaction prices. *Source*: own elaboration.

A "violin" plot, which combines the characteristics of box plots and distributions, was used for the visualization of over 30 thousand coefficients of determination. Inside each "violin" there is a box plot with a median, marked as a more brightly coloured spot. The shape of the "violin" itself is the shape of the distribution of probability density for 5000 coefficients of determination. A horizontal line obtained in the case of a lack of any disturbance, thus with a model data base, means that only one result was obtained, a result which obviously equals R^2 = 100%. In the remaining cases, fairly symmetrical distributions with some diverging values were obtained. An increase of a maximum price disturbance led to a decrease in the median of determination coefficients; however, the decrease

was not linear, but exponential in nature with a negative exponent. Initially, a rise of the maximum level of disturbance gives a strong decrease in the average coefficient of determination. Greater maximum levels of disturbance are related to a further drop in average R^2, however, the drop is already weaker. When analyzing the already obtained distributions, one can notice that, along with growing possible disturbances, they are becoming more and more dispersed, which translates into the growth of variability coefficients defined for the R^2 measure. Table 2 presents the measures of the structure for the analyzed coefficient of determination.

Table 2

Selected measures of the structures of determination coefficients of estimated regression models

maximum price disturbance	min	max	average	median	standard deviation	coefficient of variability
0	100%	100%	100%	100%	0%	0%
0.1	81.6%	88.9%	85.4%	85.4%	1.0%	1.2%
0.2	45.8%	70.4%	59.4%	59.5%	3.0%	5.0%
0.3	23.1%	58.5%	39.8%	39.8%	4.1%	10.3%
0.4	12.6%	44.2%	27.7%	27.6%	4.4%	16.0%
0.5	6.0%	36.7%	20.1%	19.9%	4.2%	21.1%
0.6	4.2%	30.5%	15.4%	15.2%	4.0%	26.0%

Source: own study.

With a maximum interference of 10%, an average fit of the models is equal to slightly more than 85%. The diversification of the obtained coefficients of determination is not substantial (V_S = 1.2%). In turn, the coefficients of determination for a maximum disturbance equalling 60% range between 4 and 30%. The fits are not only low, but also fairly strongly diversified. Even for 20% of maximum disturbance, the coefficient of determination ranging 46 to 70% means that high and repeatable results of model fits to transaction prices can be expected. The model of multiple regression requires data featuring a strong linear relationship between the explaining variables and the explained variable. Failure to satisfy the requirement quickly leads to unacceptably low values of the coefficients of determination that are also dependent on random factors. Figure 2 presents the development of variability coefficients (V_{S_e}) for the estimated models.

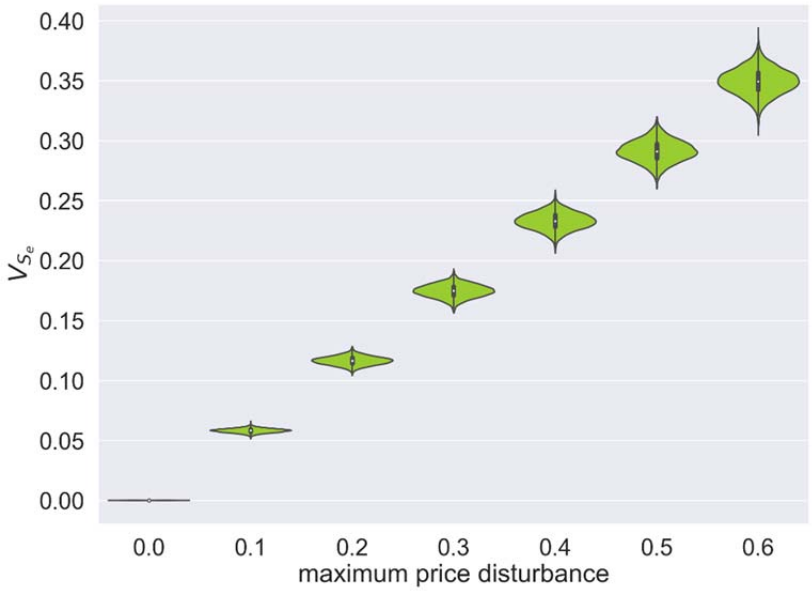

Fig. 2. Kernel density estimations (KDE) of volatility coefficients of models estimated with various maximum levels of disturbance of model transaction prices. *Source*: own elaboration.

With a maximum disturbance of 10%, the average random variability coefficient for 5000 models was equal to 5.83%. This is by how much theoretical transaction prices differed on average from "real" prices. This value can be deemed as an acceptable or even a perfect one. Still, for the subsequent threshold of a maximum disturbance, V_{S_e} amounted to 11.65%. Such a deviation cannot be accepted in every case. With higher maximum disturbances of model transaction prices, average deviations of theoretical values from real ones reach levels that are beyond acceptable. From the presented results a similar conclusion can be deduced as in the analysis of determination coefficients. Only a slight deviation from given models yields modelling results at a satisfactory level.

However, since it appears that only multiple regression models can be used successfully for strong regularities, in the subsequent stage of the conducted simulation, it will be determined whether the size of a sample (a set of data regarding market transactions) has a significant impact on model quality. It was said on many occasions that in order to obtain a good econometric model, a large number of transaction data is required. As had previously been mentioned, a complete transaction set numbered 243 observations. It will be reduced in a random manner. Between 10% and 100% of the available transactions will be used for modelling. Each such randomly drawn sample will be subject to analogous price disturbance as in the first stage. Figure 3 presents average coefficients of determination for 1000 models estimated for each combination of maximum disturbance and set size of transaction prices.

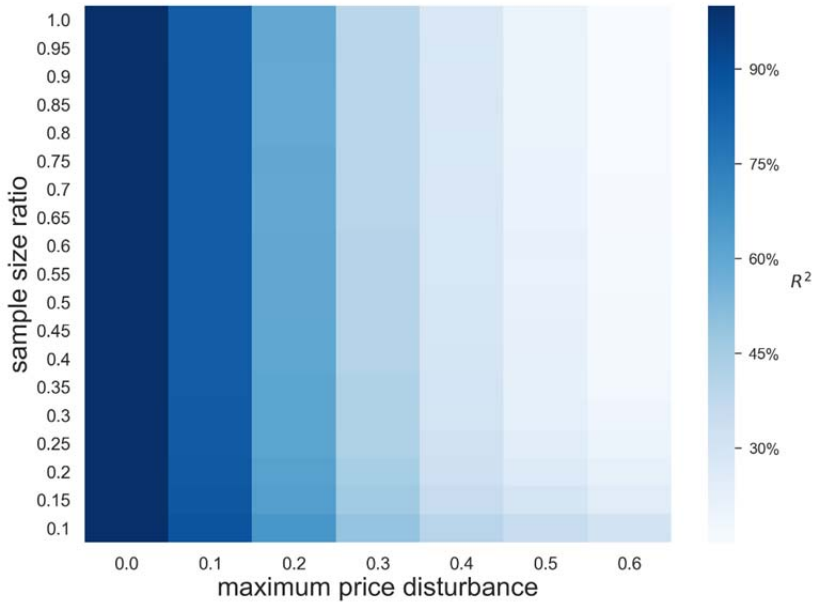

Fig. 3. Average coefficients of determination with various maximum levels of disturbance of model transaction prices and various sample sizes. *Source*: own elaboration.

It is not surprising, as was demonstrated in the first stage, that the greater the disturbance, the weaker the model fit. A more interesting observation concerns changes in the fit depending on the sample size. It turned out that the size of a sample has a slight influence on model fit. Although an increase of a maximum disturbance from 10 to 20% causes a drop in average fit from 85% to 59%, the extreme values of average coefficients of fit for 10% of maximum disturbance are equal to 85 and 88%, while for 20% - 59 and 66% respectively. On account of the number of chosen levels, the models were obtaining a higher fit with smaller samples. Figure 4 presents the development of the average of variability coefficients calculated on the basis of a group of various sample sizes for individual maximum disturbances. The figure demonstrates that the degree of V_{S_e} changes for individual maximum disturbances of transaction prices is not significant.

Another method of assessing the impact of a sample size and the scale of imperfect representation of market attribute states by prices involved an analysis of statistical significance of the estimation of model structural parameters. Figure 5 represents an average number of such insignificant estimations.

In this case a sample size has a substantial impact on whether the estimations of structural parameters are statistically significant (the adopted level of significance is 0.05). In the case of slight price disturbance – maximum 10% - no insignificant estimations were obtained only for a sample of

55%, i.e. 134 observations (or more). This is a fairly significant number in terms of market analysis and finding over 100 similar real estate transactions. With a maximum disturbance equal to 20%, even the entire sample did not provide certainty that all the estimations of structural parameters will be statistically significant.

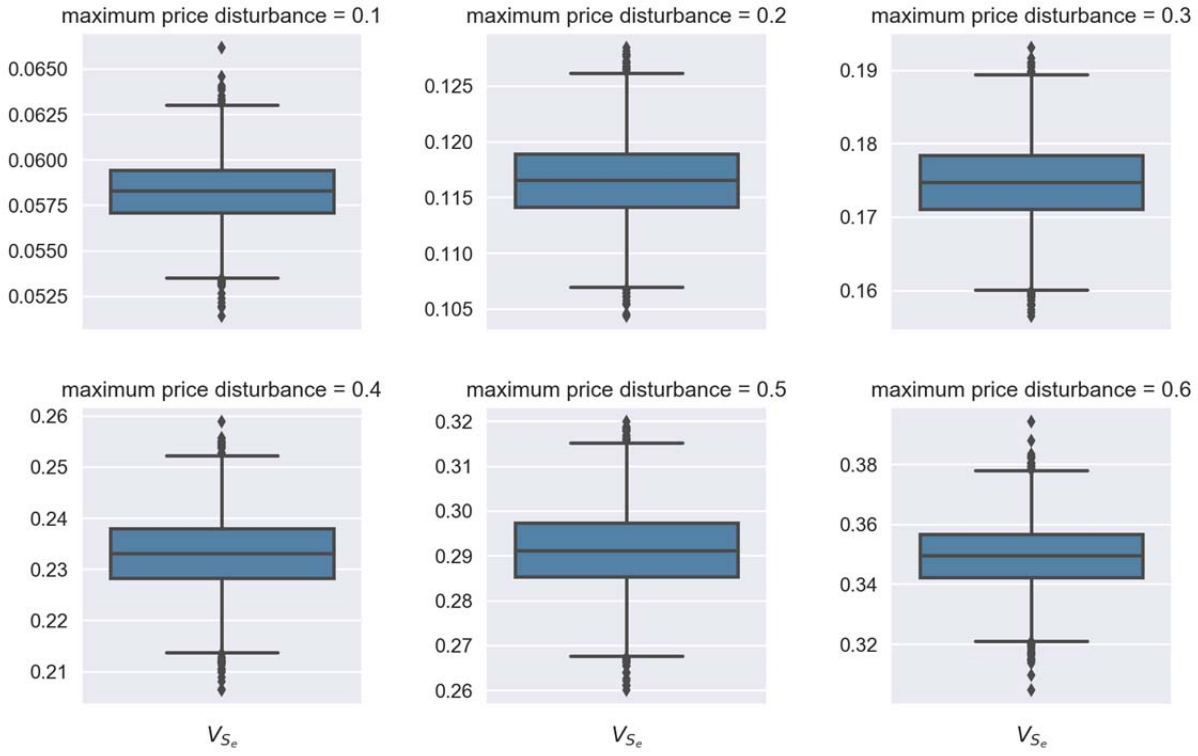

Fig. 4. Box plots of volatility coefficients with various maximum disturbance levels of model transaction prices. *Source*: own elaboration.

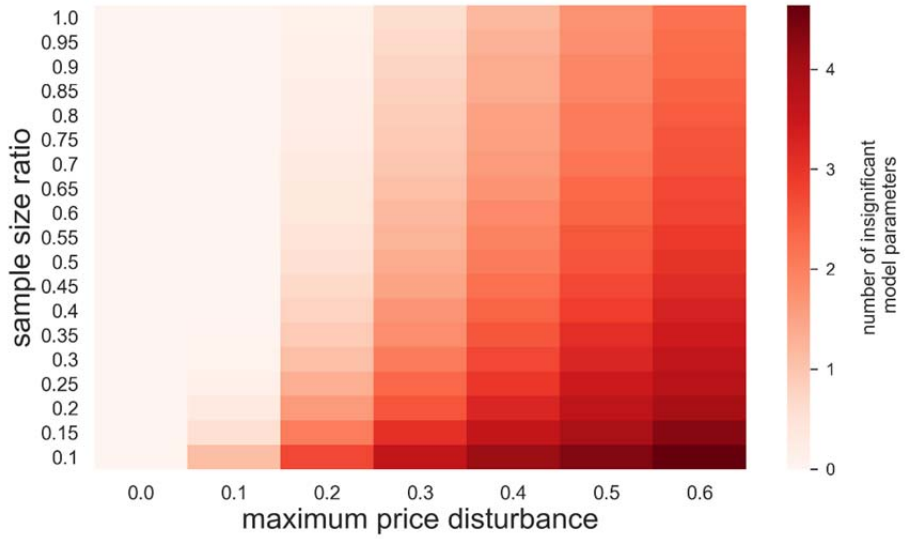

Fig. 5. Average number of insignificant estimations of structural parameters for various maximum disturbance levels of model transaction prices and various sample sizes. *Source*: own elaboration.

The last criterion of estimating regression models was the examination of the lack of coincidence, therefore, a situation in which the evaluation of a structural parameter would feature a negative sign, which would prove a decrease of a theoretical price with a shift to a higher (more favourable) state of a market attribute. Figure 6 represents the results of that criterion.

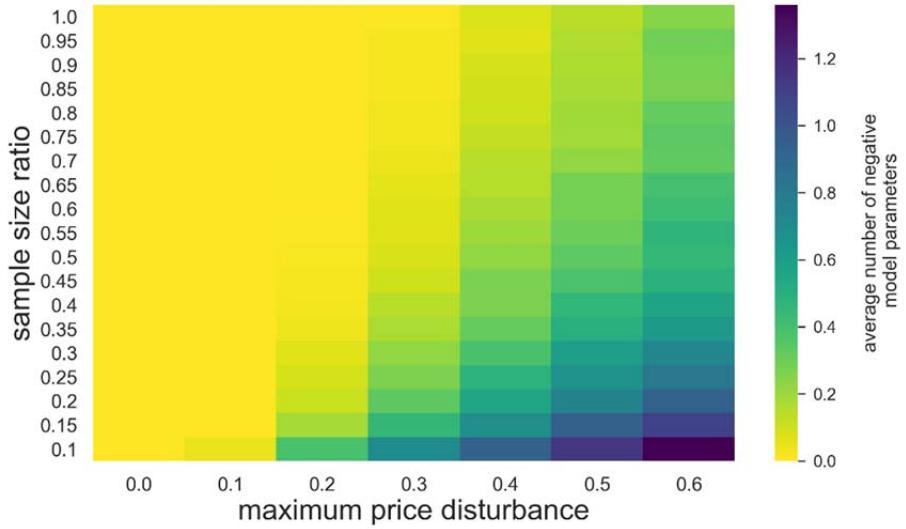

Fig. 6. Average number of negative model parameters with various maximum disturbance levels of model transaction prices and various sample sizes.
Source: own elaboration.

The results are similar as in the case of examining the statistical significance of estimations of structural parameters. The greater the estimations and the smaller the size sample, the greater the chance for the lack of coincidence.

5. Summary

The conducted simulations demonstrate that multiple regression models may constitute a good tool for real estate valuation only in the conditions of a developed, well-functioning real estate market, i.e. in a situation when real estate prices are well reflected by market attributes and with sufficiently numerous sets of transactions. In cases when transaction prices fall within a narrow range for market conditions, between 90% and 110% of model prices, an average R^2 coefficient of determination is equal to 85.4%. The extension of the price variability range, in such a way that they oscillated between 80% and 120% of model prices, resulted in its reduction to as much as 59.4%, which is already an unacceptable value from the perspective of valuation accuracy and credibility. A model of such multiple regression gives satisfactory results only in a situation when the data feature a strong linear relationship between the explaining variables and an explained variable (real estate prices and market attributes). Failure to satisfy that requirement quickly leads to unacceptably low values of the coefficients of determination, which are dependent on random factors. A reduction in sample size exerts a decidedly smaller impact on the quality of the obtained models in terms of the determination coefficient and variability coefficient. Even a relatively small size may constitute the grounds for estimating a well-fitted model, however, on condition that transaction prices are strictly related to real estate attributes (undisturbed). Unfortunately, the reduction of the sample size has negative effects on the statistical significance of structural model parameters. The general conclusion from the conducted research is that, in order to obtain a relatively good model of real estate value, i.e. one featuring R^2 of over 80%, V_{S_e} of less than 6%, with statistifaclly significant structural parameters not lacking the coincidence effect, price disturbance in relation to model prices ought not to be greater than 10%, while a real property sample should reflect at least 20% of all the possible combinations of states of market attributes.

The research was conducted within the framework of a project financed by the National Science Centre, Project No. 2017/25/B/HS4/01813.

6. References

ACZEL A.D., 2011, *Statystyka w zarządzaniu* (Statistics in management), PWN, Warszawa.
ADAMCZEWSKI Z., 2006, *Elementy modelowania matematycznego w wycenie nieruchomości* (Elements of mathematical modelling in real estate valuation), Oficyna Wydawnicza Politechniki Warszawskiej, Warsaw.

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

BARAŃSKA A., 2010a, *Modele multiplikatywne w procesie wyceny nieruchomości* (Multiplicative models in the property valuation proces), Studia i Materiały Towarzystwa Naukowego Nieruchomości, vol 18, nr 1, Olsztyn, pp. 65-82.

BARAŃSKA A., 2010b, *Statystyczne metody analizy i weryfikacji proponowanych algorytmów wyceny nieruchomości* (Statistical methods of analysis and verification of proposed property valuation algorithms), Rozprawy i monografie, v. 214, Wydawnictwa AGH, Kraków.

BARAŃSKA A., MICHALIK S., 2014, *Variants of modeling of dwelling market value,* Real Estate Management and Valuation, vol. 22, no. 3, pp. 28-35.

BIEDA A., 2018, *Conditional Model of Real Estate Valuation for Land Located in Different Land Use Zones*, Real Estate Management and Valuation, vol. 26, no. 1, pp. 122-130.

BITNER A., 2007, *Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości* (Construction of the multiple regression model in real estate valuation), Acta Scientiarum Polonorum, Administratio Locorum, no 17(3), pp.59-66.

ČEH M., KILIBARDA M., LISEC A., BAJAT B., 2018, *Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS*, International Journal of Geo-Information, 7(5).

DACKO M., 2000, *Zastosowanie regresji wielokrotnej w szacowaniu nieruchomości w arkuszu kalkulacyjnym Microsoft Excell* (Application of multiple regression in property appraisal in Microsoft Excell spreadsheet), Wycena no 2, Educaterra, Olsztyn.

DOSZYŃ M., 2012, *Ekonometryczna wycena nieruchomości* (Econometric valuation of real estate), Metody Ilościowe w Ekonomii, Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania US, no 26, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, pp. 41-52.

DOSZYŃ M., HOZER J., 2017, *Szczeciński algorytm masowej wyceny nieruchomości – podejście ekonometryczne* (Szczecin's algorithm of mass valuation of real estate - econometric approach), Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania US, no 50/1, pp. 19 – 30.

DOSZYŃ M., GNAT S., 2017, *Econometric Identification of the Impact of Real Estate Characteristics Based on Predictive and Studentized Residuals*, Real Estate Management and Valuation, vol. 25, no. 1, pp. 84-92.

DUBIN R., PACE R. K., THIBODEAU T. G., 1999, *Spatial autoregression techniques for real estate data*. Journal of Real Estate Literature, 7, 79-96.

Ekonometria (Econometrics), ed. Hozer J., 1997, Katedra Ekonometrii i Statystyki Uniwersytetu Szczecińskiego, Stowarzyszenie Pomoc i Rozwój, Szczecin.

FORYŚ I., GACA R., 2018, *Intuitive Methods Versus Analytical Methods in Real Estate Valuation: Preferences of Polish Real Estate Appraisers: Computational Methods in Experimental Economics*, in Problems, Methods and Tools in Experimental and Behavioral Economics, (CMEE) 2017 Conference.

GACA R., 2017, *Metody statystyczne i modele ekonometryczne w wycenie nieruchomości. Czy metoda wyceny powinna być adekwatna do charakteru badanego zjawiska?* (Statistical methods and econometric models in real estate valuation. Should the valuation method be adequate to the nature of the investigated phenomeon?), Rzeczoznawca Majątkowy, no 2 (94).

GACA R., 2018, *Metody statystyczne i modele ekonometryczne w wycenie nieruchomości. Analiza zbioru nieruchomości podobnych* (Statistical methods and econometric models in real estate valuation. Analysis of comparable property set), Rzeczoznawca Majątkowy, no 1 (97).

HOZER J., 2001, *Regresja wieloraka a wycena nieruchomości* (Multiple regression and property valuation), Rzeczoznawca Majątkowy, no 2, Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych, Warszawa, pp. 13-14.

HOZER J., KOKOT S., DOSZYŃ M., 2018, *Wycena nieruchomości a wybrane metody statystyczne* (Valuation of real estate and selected statistical methods), Rzeczoznawca Majątkowy, no 3 (99), Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych, Warszawa, pp. 9-17.

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

JAHANSHIRI E., BUYONG T., SHARIFF A.R.M., 2011, *A Review of Property Mass Valuation Models*. Pertanika Journal of Science and Technology, vol. 19, pp. 23 – 30.

KAWA A., FUKS K., JANUSZEWSKI P., 2016, *Symulacja komputerowa jako metoda badań w naukach o zarządzaniu* (Computer simulation as a research method in management sciences), Studia Oeconomica Posnaniensia, vol. 4, no. 1, pp. 109-127.

KOKOT S., BAS M., 2016, *Postrzeganie cech rynkowych przez rzeczoznawców majątkowych, pośredników w obrocie i nabywców nieruchomości* (Perception of market characteristics by property valuers, agents and buyers of real estate), Studia i Prace Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Szczecińskiego, no 45, tom I, Metody Ilościowe w Ekonomii, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, pp 355—369.

KOKOT S., DOSZYŃ M., 2011, *Ekonometryczna wycena nieruchomości w aspekcie twierdzenia Frischa – Waugha – Stone'a* (Econometric valuation of real estate in terms of Frisch-Waugh-Stone's claim), Studia i Materiały Towarzystwa Naukowego Nieruchomości, vol. 19 no 3, Olsztyn, pp. 49-58.

KUCHARSKA_STASIAK E., 2006, *Nieruchomość w gospodarce rynkowej* (Real estate in a market economy), Wydawnictwo Naukowe PWN, Warszawa.

KUCHARSKA-STASIAK E., 2016, *Ekonomiczny wymiar nieruchomości* (Economic dimension of real estate), Wydawnictwo Naukowe PWN, Warszawa.

KURYJ J., 2007, *Metodyka wyceny masowej nieruchomości na bazie aktualnych przepisów prawnych* (Methodology of mass valuation of real estate on current legal regulations), Wycena, no 4 (81). Educaterra, Olsztyn, pp. 50-85.

LARSEN J.E., PETERSON M.O., 1998, *Correcting for Errors in Statistical Appraisal Equations*. The Real Estate Appraiser and Analyst. No 54 (3), pp. 45–49.

LIGAS M., 2010, *Metody statystyczne w wycenie nieruchomości* (Statistical Methods in Real Estate Valuation), Studia i Materiały Towarzystwa Naukowego Nieruchomości, Vol. 18, No. 1, pp. 49-64.

LIMSOMBUNCHAI V., GAN C., LEE M., 2004, *House Price Prediction: Hedonic Price Model Vs. Artificial Neural Network*. American Journal of Applied Sciences, 2004, No 1 (3), 193–201.

LIPETA A., 2000, *Model ekonometryczny ze zmiennymi jakościowymi opisujący ceny mieszkań* (Econometric model with quality variables describing housing prices), Wiadomości statystyczne, no 8, GUS, Warszawa, pp. 10-20.

MADDALA G.S., 2006, Ekonometria (Econometrics), PWN, Warszawa.

MARK J., GOLDBERG M., 1998, *Multiple Regression Analysis and Mass Assessment: A Review of the Issues*. Appraisal Journal, Vol. 56 no. 1, pp. 89–109.

MIESZEK W., DZIADOSZ A., 2011, Budownictwo i Inżynieria Środowiska, no 2, Oficyna Wydawnicza Politechniki Białostockiej, Białystok, pp. 589-594.

NOWAK M, SKOTARCZAK T., 2013, *Podstawy gospodarowania nieruchomościami* (Basics of real estate management), CeDeWu, Warszawa.

PARZYCH P., 2009, *Modele estymacji wartości rynkowej lub katastralnej nieruchomości zurbanizowanych, rolnych i leśnych* (Estimation Models of the Market of Cadastral Value of Urbanized, Agricultural and Forest Estates), AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków.

PARZYCH P., CZAJA J., 2015, *Szacowanie rynkowej wartości nieruchomości*, AGH, Kraków.

PAWLUKOWICZ R., 2001, *Przegląd propozycji określania wartości rynkowej nieruchomości z wykorzystaniem modeli ekonometrycznych* (Review of proposals to determine the market value of real estate using econometric models), Zeszyty Naukowe Uniwersytetu Szczecińskiego, Prace Katedry Ekonometrii i Statystyki, „Mikroekonometria w Teorii i Praktyce", no 320, Szczecin, pp.315-334.

PAWLUKOWICZ R., 2002, *Polskie koncepcje zastosowań analizy regresji wielorakiej w wycenie rynkowej nieruchomości* (Polish concepts of applications of multiple regression analysis in real estate market valuation), Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Ekonometria. Tom 10, Nr 950. Zastosowania metod Ilościowych, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, pp. 209-224.

PAWLUKOWICZ R., 2007, *Użyteczność modeli ekonometrycznych w wycenie nieruchomości* (The usefulness of econometric models in real estate valuation), Zeszyty Naukowe Uniwersytetu Szczecińskiego, Prace Katedry Ekonometrii i Statystyki, „Metody ilościowe w ekonomii, no 450, Szczecin, pp. 453-466.

PAWŁOWSKI Z., 1980, *Ekonometria* (Econometrics), Państwowe Wydawnictwo Naukowe PWN, Warszawa.

PRYSTUPA M., 2000, *O potrzebie dalszych prac nad zastosowaniem regresji wielorakiej w wycenie nieruchomości* (On the necessity of further work on the application of multiple regression in

property valuationi). Rzeczoznawca Majątkowy, no 4, Polska Federacja Stowarzyszeń Rzeczoznawców Majątkowych, Warszawa, pp. 16-17.

Rozporządzenie Rady Ministrów z dnia 21 września 2004 r. *w sprawie wyceny nieruchomości i sporządzania operatu szacunkowego* (Ordinance of the Council of Ministers of 21 September 2004 on the valuation of real estate and the preparation of an appraisal report).

SAWIŁOW E., 1995, *Próba matematycznego modelowania wartości gruntów na terenach zurbanizowanych* (An attempt at mathematical modelling of land values in urban areas), Materiały IV Krajowej Konferencji Rzeczoznawców Majątkowych, Stowarzyszenie Rzeczoznawców Majątkowych we Wrocławiu i Polska Federacja Rzeczoznawców Majątkowych, Wrocław – Warszawa.

TELEGA T., BOJAR Z., ADAMCZEWSKI Z., 2002, *Wytyczne przeprowadzenia powszechnej taksacji nieruchomości* (Guidelines for carrying out general property taxation), Przegląd Geodezyjny, no 6, pp. 6-11.

Ustawa z dnia 21 sierpnia 1998 r. *o gospodarce nieruchomościami* (Act of 21 August 1998 on real estate management).

YOO S., IM. J., WAGNER J. E., 2012, *Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County*, NY. Landscape and Urban Planning, 107(3), pp. 293-306.

ZADUMIŃSKA M., SZTAUDYNGER J.J., 2001, *Wykorzystanie modelu ekonometrycznego do wyceny nieruchomości* (Use of econometric model for property valuation), Wiadomości statystyczne, no 1, GUS, Warszawa, pp. 6-13.

ZBYROWSKI R., 2012, *Szacowanie wartości nieruchomości mieszkaniowych na podstawie modeli czasowo-przestrzennych* (Estimation of the value of residential properties on the basis of time-spatial models), Roczniki Kolegium Analiz Ekonomicznych nr 27, Szkoła Główna Handlowa, Warszaqwa, pp. 101-112.

ZURADA J., LEVITAN A.S., GUAN J., 2011, *A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context*. Journal of Real Estate Research, Vol. 33, No 3, pp. 349-387.

ŹRÓBEK S., BEŁEJ M., 2000, *Podejście porównawcze w szacowaniu nieruchomości* (Comparative approach to property appraisal), Educaterra, Olsztyn.