www.degruyter.com/view/j/remav

DE G DE GRUYTER OPEN

USE OF STATISTICAL MODELS FOR SIMULATING TRANSACTIONS ON THE REAL ESTATE MARKET

Radosław Cellmer, assoc. prof., PhD
The Faculty of Geodesy, Geospatial and Civil Engineering
University of Warmia and Mazury in Olsztyn
e-mail: rcellmer@uwm.edu.pl

Katarzyna Szczepankowska, M. Sc.
The Faculty of Geodesy, Geospatial and Civil Engineering
University of Warmia and Mazury in Olsztyn
e-mail: katarzyna.szczepankowska@uwm.edu.pl

Abstract

The regularities and relations between real estate prices and the factors that shape them may be presented in the form of statistical models, thanks to which the diagnosis and prediction of prices is possible. A formal description of empirical observation presented in the form of regressive models also offers a possibility for creating certain phenomena in a virtual dimension. Market phenomena cannot be fully described with the use of determinist models, which clarify only a part of price variation. The predicted price is, in this situation, a special case of implementing a random function. Assuming that other implementations are also possible, regressive models may constitute a basis for simulation, which results in the procurement of a future image of the market.

Simulation may refer both to real estate prices and transaction prices. The basis for price simulation may be familiarity with the structure of the analyzed market data. Assuming that this structure has a static character, simulation of real estate prices is performed on the basis of familiarity with the probability distribution and a generator of random numbers. The basis for price simulation is familiarity with model parameters and probability distribution of the random factor.

The study presents the core and theoretical description of a transaction simulation on the real estate market, as well as the results of an experiment regarding transaction prices of office real estate located within the area of the city of Olsztyn. The result of the study is a collection of virtual real properties with known features and simulated prices, constituting a reflection of market processes which may take place in the near future. Comparison between the simulated characteristic and actual transactions in turn allows the correctness of the description of reality by the model to be verified.

Key words: *real estate market, multiple regression, simulation.*

JEL Classification: C15, C53, R32

Citation: Cellmer R., Szczepankowska K., 2015, *Use of Statistical Models for Simulating Transactions on the Real Estate Market*, Real Estate Management and Valuation, Vol. 23, No. 2, pp. 99-108.

DOI: 10.1515/remav-2015-0019.

1. Introduction

The real estate market is a very difficult phenomenon for mathematical mapping on account of the multidimensionality of processes occurring in it. Transaction prices of real estate are mainly shaped by supply and demand, whose sources depend on demographic, economic or spatial conditions. During

the course of constructing a model of the real estate market, attention is usually focused on its selected elements and mutual relations. Therefore, mathematical models of the real estate market most often refer to relations and dependencies between transaction prices (or offer prices) and certain measurable factors which influence such prices. The models, even though they sometimes significantly simplify reality, are successfully used to determine the value or identify and diagnose the dependencies on the market. Models which successfully completed substantive and statistical verification may constitute a basis for estimating value as the most probable price. It is necessary to note that, in this respect, we are dealing with single point estimation of expected value. In reality, we may be dealing with multiple variants of price implementation, which, as a variable, may be subject to specific distribution. Familiarity with such distribution, thus the probability of the occurrence of a specific price (or price in a specific range), creates the possibility of performing simulation. The generated virtual prices will constitute a set of certain possible implementations of the response variable, indicating one of the many possible states of the real estate market in the future.

The main objective of the conducted study is to indicate the possibility of transaction simulation on the basis of relatively simple statistical models. These models may be treated not only as prediction or diagnostic models, but also, having taken into account proper assumptions, as simulation models.

Methods of statistical analysis rely on certain characteristic parameters of random variables represented by individual prices of real estate and their specific physical, legal or economic characteristics (CZAJA, LIGAS 2010). Analysis of the relations and dependencies between variables characterizing the real property market often makes use of classic linear multiple regression models (ISAKSON 1998, CZAJA 2001, BENJAMIN ET AL. 2004, SIRMANS et al. 2004, ADAMCZEWSKI 2006, BITNER 2007, BARAŃSKA 2010, SAWIŁOW 2010). The characteristic feature of regressive models is the possibility of detecting dependencies between the observed phenomena, examining the strength of such dependencies and forecasting phenomena on the basis of observing the size of other phenomena. The condition for applying the regression model does not have to be initial ascertainment of the fact of the existence of any dependency between variables, whereas one of the objectives for constructing such a model may be the urge to check whether such dependency takes place. A proper regression model may be applied on the condition that it is possible to obtain a sufficient amount of data from a specific real estate market in order to obtain the desired level of model materiality. The conditions enabling the application of a model of, e.g. multiple regression, include the existence of a proper specification theory of a set of significant explanatory variables, the condition of their measurability, high variability, and the occurrence of a strong dependency between the explanatory variables and the response variable in the absence of mutual relations between explanatory variables (HOZER et al. 2002). For example, multiple regression models, as linear models of the real estate market, are built for the needs of determining the value of real estate assuming that the relation between the market price of real estate and the transaction price is of a cause-and-effect nature.

Multiple regression models, in order to be used to predict transaction prices on the real estate market, have to be subjected to verification, both substantive and statistical, which encompasses confirmation of: the model's compliance with market data, independent (explanatory) variables of the model, and the stochastic structure of the model's assumptions (BARAŃSKA 2010). A correctly constructed statistical model constitutes a compromise between the necessity for a significant simplification of reality and the possibility of easy interpretation. When choosing the function form of the model, it is necessary to focus on the simplest mathematical formulas (ACZEL 2000, RAO et al. 2007, BITNER 2007). Therefore, it is possible to first rely on the classic additive linear model of multiple regression. This has a number of advantages, consisting primarily of the relatively easy estimation of parameters and interpretation of results.

However, classic linear models of multiple regression do not offer satisfactory results in every case (HOZER et al. 2002, PAWLUKOWICZ 2006). One of the most significant problems occurring during the application of such models is the fact that many variables are non-measurable or measurable solely on the ordinal or nominal scale (HOZER et al. 2002). It is also worth noting that the relations between market characteristics and values are not always linear, even though this problem may be partially solved by the proper selection of interval scales of explanatory variables (SAWIŁOW 2010) or the selection of a non-linear form of the model, where such a model may take the form, among others, of an exponential model or power model. Estimation of non-linear models is slightly more complex than that of linear models as such models usually require prior linearization. This may be performed by

logarithming or Taylor series expansion. Subsequently, linearization enables the application of the least square method for estimating model parameters (BARAŃSKA 2010).

Even though statistical models built on the basis of data from the real estate market are usually used to evaluate the examined phenomenon and potentially predict the response variable, they may also be used for transaction simulation. Modelling of the real estate market structures for cognitive purposes has been undertaken by many scientists; however, only a few tried to examine the market with the use of simulation modelling (e.g. DIAPPI, BOLCHI 2008; MC BREEN, GOFFETTE-NAGOT, JENSEN, 2011; VOREL 2014). The main advantage of this particular research approach is the lack of limitations regarding structure and the degree of complexity of the examined system, and the possibility of taking into account stochastic processes, thanks to which it is possible to model the actual systems with a high degree of complexity and a significant share of random factors. For researchers of real estate market behavior, where processes and relations may be random in nature, this is of key importance. So far, simulation studies in the area of the real estate market have referred to streamlining the undertaken economic decisions (e.g. BAO et al. 2012), the efficiency of investing, including maximization of income from the rental sector (MC BREEN, GOFFETTE-NAGOT, JENSEN, 2011), or the impact of environmental as well as social and economic factors on demand (VOREL 2014). Simulations were also used to analyze and examine the fluctuations of prices on the real estate market and examine the dynamics of the local real estate market; on the other hand, the issue of simulating transactions as such has rarely been discussed. Simulation modelling in this case allows for emulating the probabilistic nature of real phenomena, enabling mathematical modelling of actual processes, the results of which, on account of the complexity of the process, cannot be foreseen with the use of analytical solutions.

2. Methodology and course of study

In the course of the study, an assumption was adopted that both the transaction structure, determined on the basis of the distribution of attributes, and the structure of transaction prices have a static character in a relatively short period of time. In the course of statistical modelling of transaction prices with the use of regressive models, it is most often assumed that regularities determined on the basis of data from a certain period of time are also going to refer to the nearest future. Obviously, this is an assumption which significantly simplifies reality and may, at the same time, influence the result of the simulation. However, taking into account the dynamic character of factors shaping the market would additionally require the construction of models for trends present on the market. If the analysis encompasses a relatively short period of time, with the market being relatively stable at this time, such trends can be omitted. Assuming that the structure of values of real estate prices is relatively constant, simulation may rely on familiarity with the frequency distribution of the values of individual features.

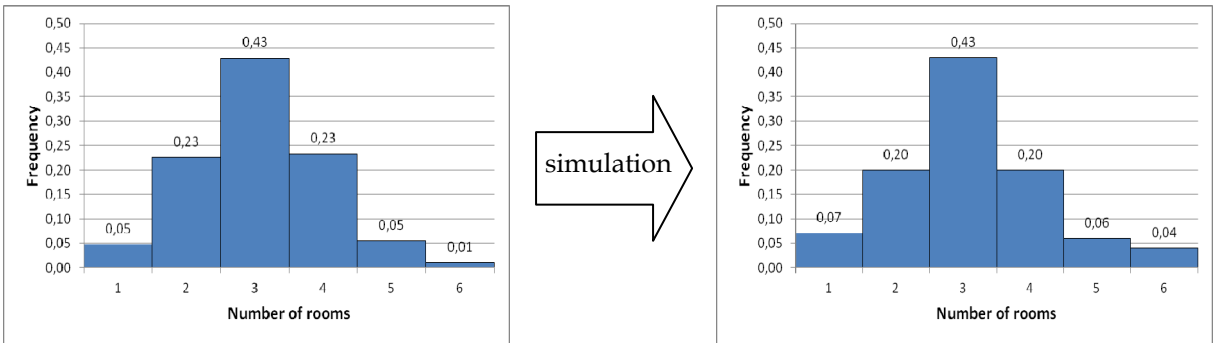
Real estate prices treated as explanatory variables should be expressed in a numerical scale, whereas the specific nature of multiple regression models requires this scale to be at least an interval scale. In the case of a discrete variable, its values occur with a specific frequency, which may in turn form the basis for determining a discrete distribution of probability. On the basis of distribution determined in this manner, it is possible to perform simulation of variable values. Generally speaking, the process of simulation in this case is going to encompass the following steps:

1) determining the frequency at which given value of the feature occurs,
2) determining discrete distribution of probability on the basis of frequency analysis,
3) generating random (or pseudo-random) numbers corresponding to the values of a feature with an adopted discrete distribution,
4) determining the frequency of occurrence of a given feature in the generated set of data.

Along with an increase in the number of simulated values, their frequency distribution is going to asymptotically aim for discrete distribution on the basis of which the simulation was made. Fig. 1 presents, as an example, a diagram of the frequency of occurrence of residential premises with a specific number of rooms on the market, as well as the same frequency graph for values generated on the basis of discrete distribution.

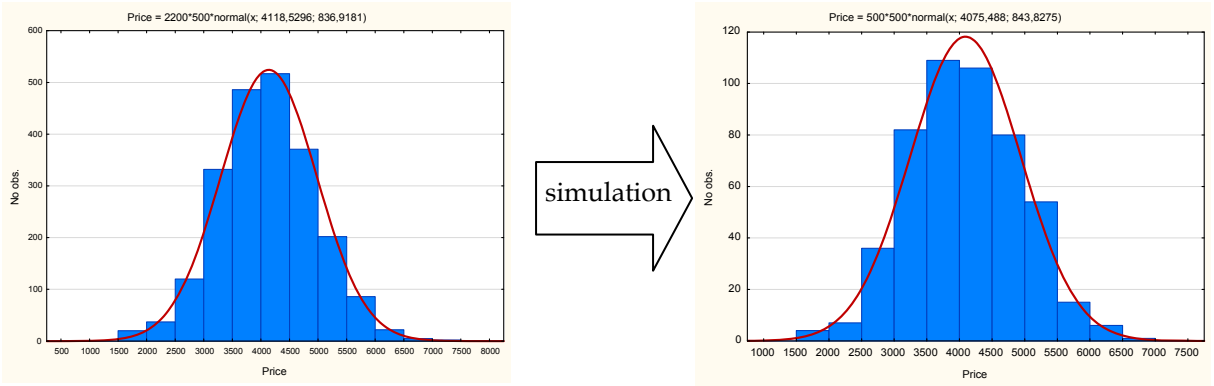
In the case when the variable has a continuous character (or quasi-continuous character), the basis for simulation may be statistical distribution, whose parameters correspond or are similar to empirical distribution parameters. An example of such a variable may be surface area or transaction price. In the course of the simulation, a generator of random numbers is used based on, e.g. normal distribution.

Figure 2 presents an example of a simulation of unit transaction prices generated on the basis of the empirical distribution of prices approximated by the density function of normal distribution. Values of individual explanatory variables generated in this manner constitute a set of features which may be assigned to virtual real estate, which is the object of trade in the simulated transactions.



number of rooms - real transactions | number of rooms - simulated transactions

Fig. 1 Example of the discreet distribution of the frequency of a given attribute value occurring as a base for simulation (example). *Source*: own study.



unit price - real transactions | unit price - simulated transactions

Fig. 2. Example of normal distribution as a base of simulation. *Source*: own study.

Statistical models may be used for the simulation of transaction prices. In this case, the sole determination of the empirical distribution of prices and approximation with the use of selected statistical distribution may turn out to be an excessive generalization; the impact of significant features shaping the prices would not be taken into account. Therefore, it is suggested that prices be simulated with the use of a regressive model (e.g. linear model of multiple regression). Such a model would have to first be positively verified – both with respect to substance and statistics. A multiple regression model contains deterministic and stochastic elements:

$$Y = \beta_0 + \sum_{i=1}^{k} X_i + \varepsilon \qquad (1)$$

where:

$\beta_0 + \sum_{i=1}^{k} X_i$ - deterministic element

ε - stochastic element with normal distribution N(0,σ)

Price simulation would take place in two stages. First of all, the price resulting from the determinist component of the model would be determined, and later, the random element would be generated with the use of the residual probability distribution. In this case, the simulation process will encompass the following steps:

1) constructing a multiple regression model (on the basis of actual transactions),
2) substantive and statistical verification of the model,
3) determining residual distribution and indicating its parameters,

4) calculating prices on the basis of the determinist component and simulating the values of features,
5) generating a residual component on the basis of the designated distribution parameters,
6) calculating simulated price as a total of the determinist and stochastic components.

The object of simulation may also be the location of a transaction. The basic assumption, similarly to the simulation of real estate attributes, refers to the static character of the market's spatial activity. This means that location of "virtual" transactions will depend on the spatial distribution of hitherto market activity. This activity will be determined with the use of kernel estimation (CELLMER, SZCZEPANKOWSKA, 2014) or in a slightly simpler manner, e.g. through division into basic fields. In such a case, the number of previous transactions in each basic field should be determined and then, on this basis, the discrete empirical distribution of frequency established. This would provide a foundation for generating random numbers characterizing the simulated number of transactions in each separated unit of spatial division.

3. Research results

The study was conducted on the basis of data derived from the local real estate market in Olsztyn (a city located in the north-eastern part of Poland). Data referred to transactions in residential premises constituting objects of ownership. In total, 2,200 samples of data were accumulated, concerning transactions which took place between 2012 and 2014. Data from the years 2012 – 2013 were used to construct a simulation model, whereas data from 2014 was used to evaluate the simulation results. The locations of transactions within the city have been presented in Fig. 3.

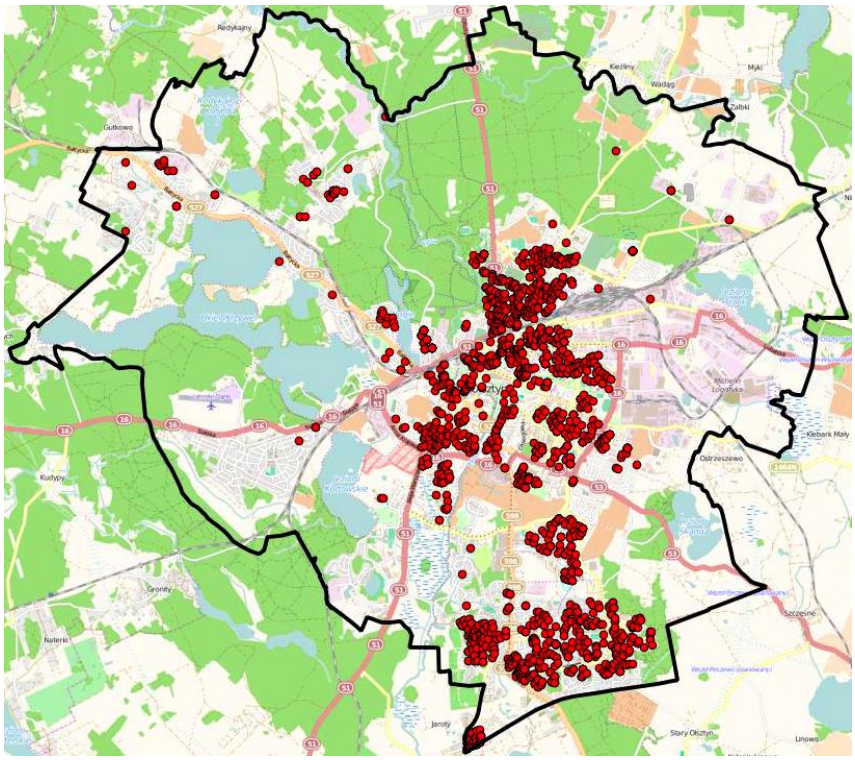


Fig. 3. The location of transactions of residential dwellings in the city of Olsztyn in 2012-2014. *Source*: own study.

The conducted analysis was aimed at illustrating the simulation process and indicating the possibilities offered by the use of a multiple regression model as a simulation model. In the course of the studies, STATISTICA and ArcGIS software were used, along with a generator of pseudo-random numbers in an Excel calculation sheet.

The explanatory variables were usable surface area expressed in square meters, location within the building, and the year of construction of the building. For the location of the floor, a three-degree scale was adopted (0 – unfavorable location, i.e. ground floor; 1 – average location; 2 – favorable 1st and 2nd floor). The year of construction was expressed using the following scale: 0: building

constructed in the 19th century; 1: pre-war building constructed in the 20th century; 2: building constructed in the 1940s; 3: building constructed in the 1950s, etc. On the basis of data prepared in this manner, estimation of the parameters of the multiple regression model was performed. The results have been presented in Table 1.

Table 1

	β	Standard Error β	t	p Level
Intercept	3502.187	69.929	50.081	< 0.001
Area	-7.399	0.955	-7.745	< 0.001
Floor	75.482	22.087	3.417	< 0.001
Age of building	152.585	6.920	22.049	< 0.001

$R^2 = 0.229$, R^2 corrected = 0.228, F = 182.53, p < 0.0001, standard error = 739.08

Source: own study.

All three explanatory variables turned out to be statistically significant, with a significance level of less than 0.001. The determination coefficient amounted to 0.23, whereas the standard estimation error (standard deviation of residuals) amounted to 739.08. It is possible to conclude that the above model is correct with respect to substance and statistics. Distribution of residuals from regression is of particular importance for transaction simulation; on account of the manner of estimation (least square method), it should correspond to the normal distribution. The histogram of residuals and the diagram of residual normality have been presented in Fig. 4.

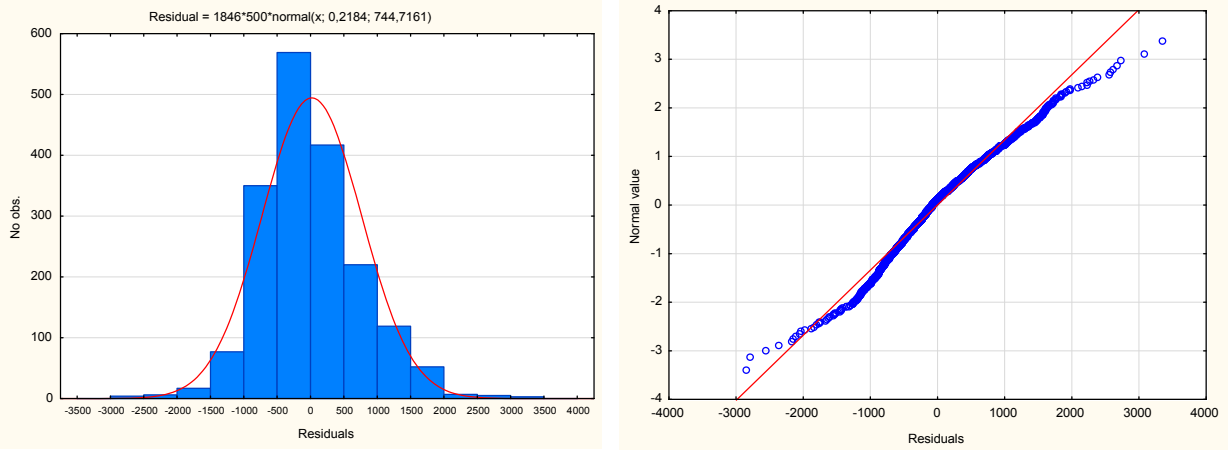


Fig. 4. The histogram of residuals and the diagram of residual normality. *Source*: own study.

Even though the distribution of residuals is not, in this case, completely compliant with normal distribution (as indicated by a non-parametric Shapiro-Wilk compliance test), it may be considered similar, obviously taking into account the aim and demonstrative nature of the conducted studies.

During the next stage of the study, data which referred to virtual transactions was generated. In total, 360 samples of data were generated, and this number corresponds to the number of transactions in the year 2014 which were used to evaluate the quality of the simulation. For each of the three adopted explanatory variables, it is necessary to determine the distribution which is later going to provide a basis for generating data. The surface area may be treated as a variable with a continuous character, whereas its distribution can be approximated most simply by the normal distribution curve. The distribution of surface area has been presented in Fig. 5.

Location on a given floor of a building and the year of construction are discrete variables. The distribution of the frequency of individual values of these variables in the form of percentages has been presented in Fig. 6.

On the basis of distribution determined in this manner, numbers were generated which correspond to the characteristics of virtual real estate in the simulated transactions. If an assumption is adopted that both the structure of transactions and the distribution of the values of characteristics has a static character, it can be expected that the distribution of individual values determined on the basis of

"virtual" real estate and the transactions which took place in 2014 will be similar. Comparison of such distribution has been presented in Fig. 7.

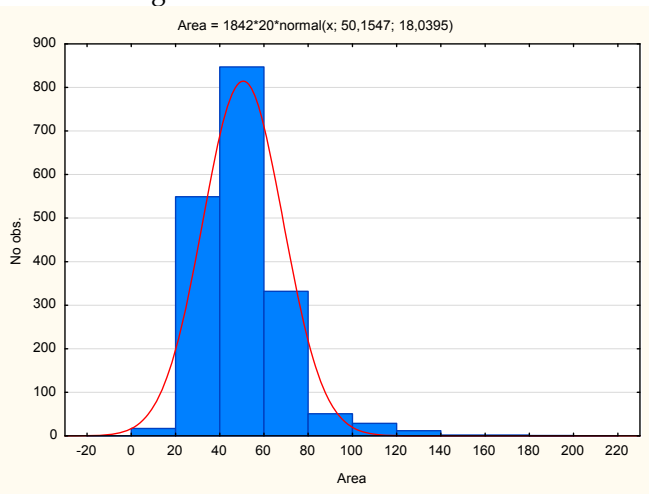


Fig. 5. The distribution of surface area in 2012-2013. *Source*: own study.

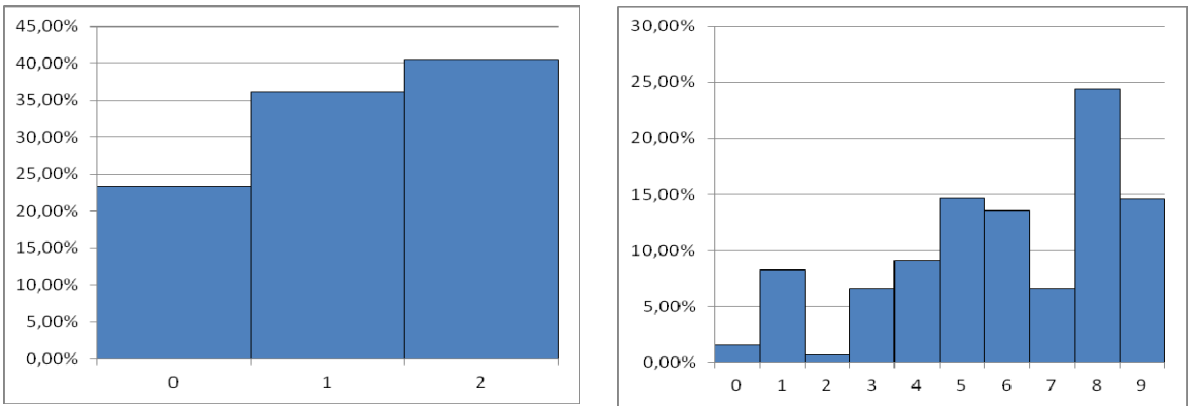


Fig. 6. Frequency distribution of the location within a building and the age of building specified on the base of transactions from 2012-2013. *Source*: own study.

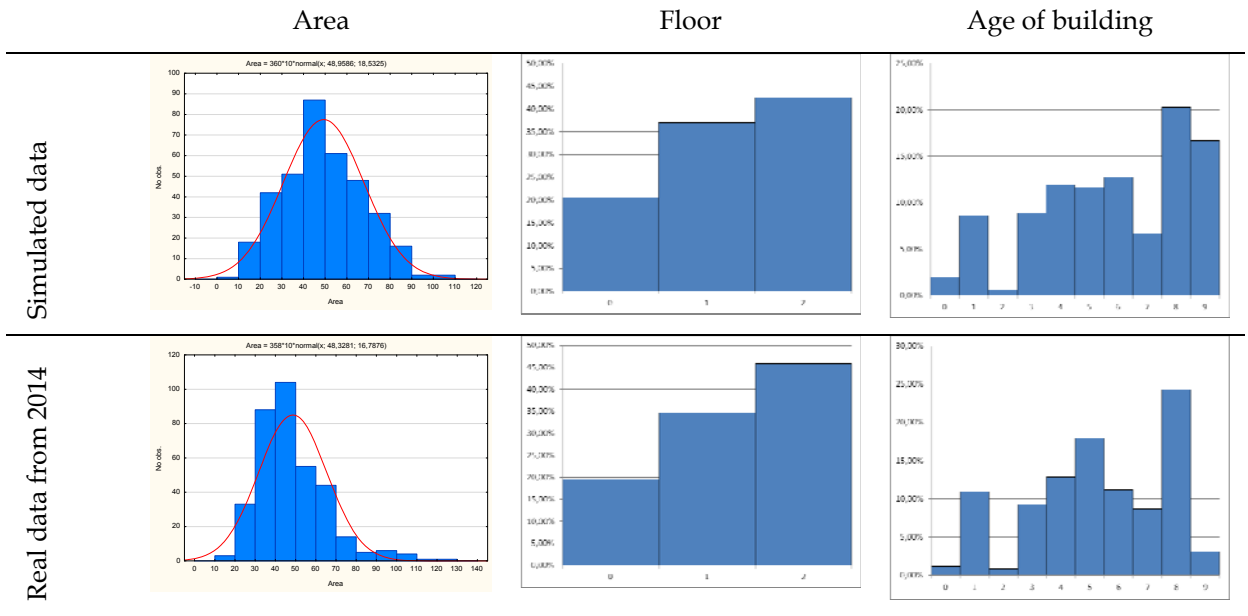


Fig. 7. Comparison of distribution characteristics based on simulated data and actual transactions from 2014. *Source*: own study.

The multiple regression model was used for price simulation. On the basis of generated real estate features, a price prediction was made and, subsequently, a random component was added to each price, generated on the basis of residual distribution. In this way, a set of simulated transactions was obtained, whose internal structure should correspond to the structure of transactions on the basis of which the model was built. In order to verify the adopted assumptions and the course of proceedings, data from 2014 were used, on the basis of which the regression model was built. Subsequently, the model was compared with the model built on the basis of simulated transactions (Table 2).

Table 2

	Data	β	Std. error β	t	p level
Intercept	simulation	3455.357	162.901	21.211	< 0.001
	transaction	3554.073	146.704	24.226	< 0.001
Area	simulation	-7.665	2.171	-3.530	< 0.001
	transaction	-7.926	2.194	-3.611	< 0.001
Floor	simulation	20.866	54.487	0.383	0.702
	transaction	-10.944	48.025	-0.228	0.820
Age of Building	simulation	170.178	16.170	10.524	< 0.001
	transaction	175.743	15.499	11.339	< 0.001
R^2	simulation		0.274		
	transaction		0.279		
Standard Error	simulation		783.340		
	transaction		692.542		

Source: own study.

The study shows significant similarity of parameters between the regression models built on the basis of simulated and actual transactions. Apart from the location within a building (floor), which turned out to be statistically insignificant in both models, the remaining parameters do not differ significantly. The value of the determination coefficient in both cases is also very similar, and the value of the standard estimation error is comparable. This means that transaction simulation on the basis of a multiple regression model, to a large degree, reflects the reality of the real estate market.

During the study, simulation of the spatial distribution of transactions was also conducted. The area of study was covered with basic fields in the form of squares with sides measuring 500 m. Areas where the probability of the occurrence of a transaction regarding residential premises amounts to zero or is close to zero were excluded from analyses (e.g. undeveloped areas, forests, lakes, industrial areas). Discrete distribution of frequency determined on the basis of transactions from the years 2012–2013 was used for simulation. The spatial distribution of frequency has been presented in Fig. 8

The conducted study shows that significant similarity exists between the distribution in space of simulated and actual transactions.

It is necessary to note that in the course of the study, one simulation process was conducted, whereas in practice there may be many more simulated variants. In the case of examining continuous variables, the number of such variants is never-ending. Along with an increase in the number of simulated transactions, their structure and prices are going to asymptotically head towards distributions on the basis of which simulations were made.

The results of the simulation, as well as the spatial distribution of the quantity of actual transactions conducted in 2014, have been presented in Fig. 9.

4. Conclusions

Simulation of transactions on the real estate market allows for determining possible future tendencies being shaped within its structure. In this case, the bases for simulation are models of the structure of features and prices of real estate, as well as models of the dependencies between them. An indispensable condition for the proper evaluation of possible future situations is the understanding and description, in the form of a formalized model, of the condition of the examined phenomenon at the present time and in the future. When analyzing transaction prices, the simplest model seems to be the multiple regression model, which allows for the evaluation of dependencies between the features

of real estate and transaction prices. The conducted study has shown that it may function not only as a diagnostic or prediction model, but also as a simulation model.

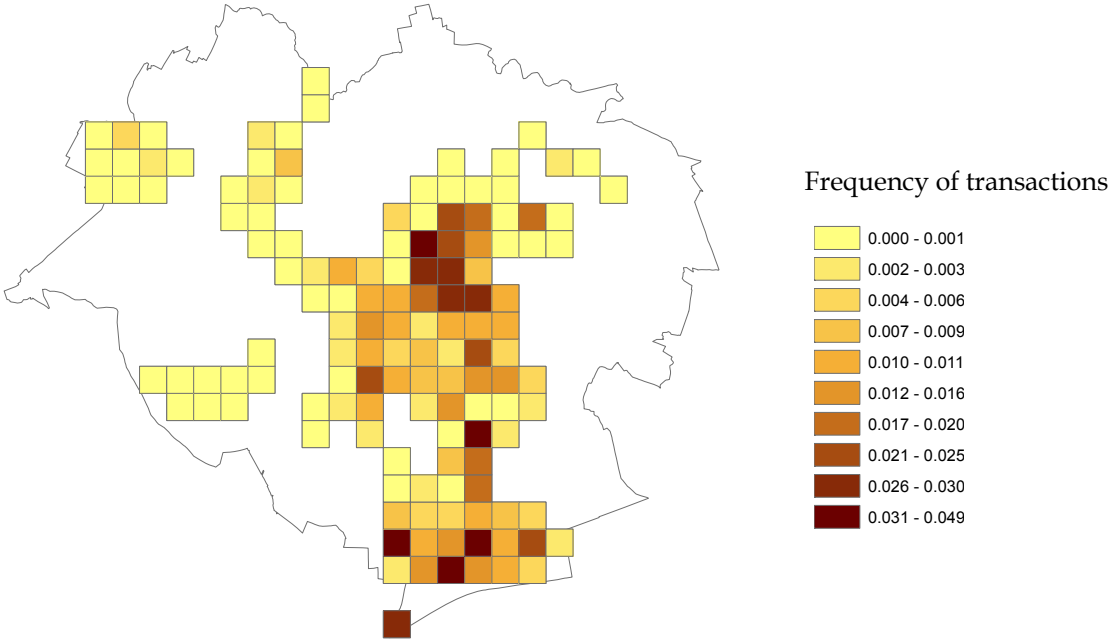


Fig. 8. Spatial distribution of the frequency of residential dwelling transactions in the years 2012 - 2013. *Source*: own study.

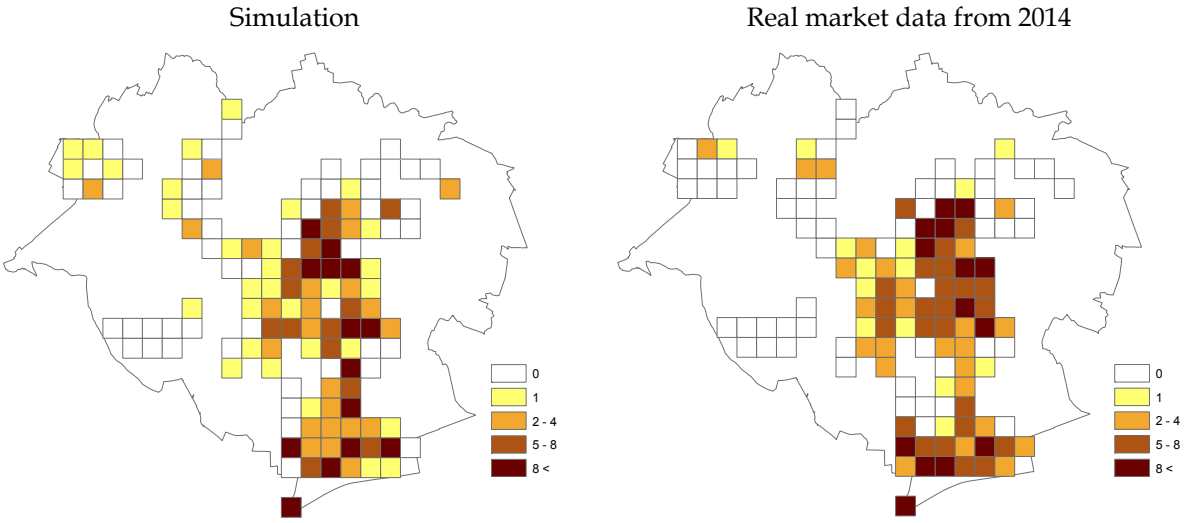


Fig. 9. Spatial distribution of the quantity of simulated and actual transactions. *Source*: own study.

5. References

ACZEL A. D., 2000, *Statystyka w zarządzaniu* (Statistics in management), PWN Warszawa

ADAMCZEWSKI Z., 2006, *Elementy modelowania matematycznego w wycenie nieruchomości. Podejście porównawcze* (Elements of mathematical modeling in real estate appraisal. Comparative approach), Oficyna Wydawnicza Politechniki Warszawskiej

BAO H., CHONG, A.Y.L, WANG H., WANG L., HUANG Y., 2012, *Quantitative Decision making in land banking: A case study on China's Real Estate Developers via Monte Carlo Simulation*, International Journal of Strategic Property Management, Vol. 16(4), 355-269.

BARAŃSKA A., 2010, *Statystyczne metody analizy i weryfikacji proponowanych algorytmów wyceny nieruchomości* (Statistical methods of analysis and veryfication proposed algorithms of valuation), Rozprawy i Monografie, Wydawnictwa AGH, Kraków.

BENJAMIN J. D., RANDALL S. GUTTERY R. S., SIRMANS C. F., 2004, *Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation*, Journal of Real Estate Practice and Education, Vol. 7, No. 1, pp. 65-77

BITNER A., 2007, *Konstrukcja modelu regresji wielorakiej przy wycenie nieruchomości* (The construction of the multiple regression model for the valuation of real estate), Acta Scientiarum Polonorum, Administratio Locorum, No. 6(4), 59-66.

CELLMER R., SZCZEPANKOWSKA K. 2014, *Simulation modeling in a real estate market*, 9th International Conference "Environmental Engineering" Vilnius Gediminas Technical University, May 22-23, http://dx.doi.org/10.3846/enviro.2014.113

CZAJA J., 2001, *Metody szacowania wartości rynkowej i katastralnej* (Methods for estimating the market and cadastral value), Komp-System, Kraków

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

DIAPPI L., BOLCHI P.,2008 *Smith's Rent gap Theory and Local Real Estate Dynamics: A Multi-agent Model*. Computers, Environment and Urban Systems, 32(1), pp. 6-18

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

ISAKSON H. R., 1998, *The Review of Real Estate Appraisals Using Multiple Regression Analysis*, Journal of Real Estate Research, Vol. 15, Issue 2, pp. 177-190

MC BREEN J., GOFFETTE-NAGOT F., JENSEN P., 2011, *Information and Search on the Housing Market: An Agent-based Model*, ERSA conference papers ersa11p1395, European Regional Science Association.

PAWLUKOWICZ R., 2006, *Użyteczność modeli ekonometrycznych w wycenie nieruchomości – polskie i zagraniczne opinie* (The utility of econometric models in the valuation of real estate - Polish and foreign opinions), Zeszyty Naukowe Uniwersytetu Szczecińskiego Nr 450, Prace Katedry Ekonometrii i Statystyki, No. 17, pp. 453-466

RAO C. R., TOUTENBURG H., SHALABH, NEUMANN C., 2007, *Linear Models and Generalizations: Least Squares and Alternatives*, Springer-Verlag, New York

SAWIŁOW E., 2010, *Problematyka określania wartości nieruchomości metodą analizy statystycznej rynku* (The problems of qualifying the value of real estate with the method of the statistical analysis of the market), Studia i Materiały Towarzystwa Naukowego Nieruchomości, Vol. 8, No. 1, pp. 21-32

SIRMANS G. S., MACPHERSON D. A., ZIETZ E. N., 2005, *The Composition of Hedonic Pricing Models*, Journal of Real Estate Literature, Vol. 13, No. 1, pp. 3-46.

VOREL J., 2014, *Residential location choice modelling: a micro-simulation approach*, AUC Geographica, 49, No. 1, pp. 83–97