ACTA SCIENTIARUM POLONORUM Acta Sci. Pol., Administratio Locorum 19(2) 2020, 109–118.
http://administratiolocorum.uwm.edu.pl plISSN 1644-0749 eISSN 2450-0771 DOI: 10.31648/aspal.4608

ORIGINAL PAPER Accepted: 11.02.2020

CHOSEN STATISTICAL METHODS FOR THE DETECTION OF OUTLIERS IN REAL ESTATE MARKET ANALYSIS

Beata Śpiewak[1], Anna Barańska[2]

[1] ORCID: 0000-0002-9567-3114
[2] ORCID: 0000-0002-6426-4115
[1,2]Department of Integrated Geodesy and Cartography
AGH University of Science and Technology
al. A. Mickiewicza 30, 30-059 Kraków, **Poland**

ABSTRACT

The paper contains the comparison of mechanism of two separately constructed statistical methods for the detection of outliers in real estate market analysis. For this purpose, databases with various types of real estate from local markets were created. Then the estimation of parameters of functional models describing dependencies prevailing on the examined markets was carried out. Subsequently, statistical tools called Baarda's method and model residual analysis were used to detect outliers in the collected datasets. The last stage was a comparison of the obtained results of the parameters' estimation of the analyzed models and the measures of their quality, before and after the removal of outliers. The obtained results indicate that algorithms of chosen statistical methods, detecting outliers, allow to eliminate a smaller number of them, at the same time obtaining an improvement of the parameters of the functional model and its adjustment to the analyzed dataset. Therefore the conclusion is that a simple statistical method, which is the study of the occurrence of cases deviating from the functional model based on the analysis of residues can generate the same results as the use of a much more complicated algorithm as the one proposed by Baarda.

Key words: statistical methods, real estate market analysis, outliers, model residual analysis, Baarda's method

INTRODUCTION

Statistics knows a lot of methods of detecting outliers which differ in their calculation algorithms (Śpiewak 2018). Some of them have more complicated formula, others less. A question arises – is it possible to formulate criteria which allow to detect outliers from functional model matched to the observations so that the number of removed observations were as low as possible with the best values of the determination coefficient and the residue variance estimator? This gives the premise for searching for such kind of criteria for the selection of optimal statistical methods that look for gross errors in the analyzed databases, among others, depending on functional model used, type of property and number of properties.

The main purpose of research was comparison of two separately constructed methods which are model residual analysis and Baarda's method. Model residual analysis is widely known method of detection of outliers. Baarda's method is less popular and it is often used for looking for gross errors, e.g. in geodetic

✉spiewak@agh.edu.pl

network adjustment. The one thing which is common for analyzed methods is the fact that searching for outliers is preceded by building a model of multiple regression.

Results from conducted research may be helpful in different fields of science where functional model matched to data is built, for example in forecasting trends of currency price in economical studies, in environmental engineering for rating of air or water quality or in medical science for predicting patients' comfort after taking medicines in treating specific illness.

In the article research was conducted on seven datasets which were information about various types of real estate, sold on local markets. Subsequently, for every analyzed database, the basic characteristics of the random variable were determined, e.i. the average price, standard deviation and the median. Then the outliers, from the functional linear and non-linear model matched to observations, were found by using the model residual analysis and the Baarda's method.

LITERATURE'S REVIEW

The problem of detecting outliers is undertaken in many publications from various fields of science where we deal with large datasets and where the data comes from direct measurements (e.g. field or laboratory). In the scientific literature it is common to find references to the following statistical methods of detecting outliers:

- quartile criterion (Budka et al. 2013, Głowicka-Wołoszyn et al. 2018);
- interval estimation (Wiśniewski 2009, Korir 2019);
- Mahalanobis distance (Meyers et al. 2015, Li et al 2019);
- Cook distance (Cook and Weisberg 1982, Trzęsiok 2015);
- model residual analysis (Śpiewak 2017, Walesiak 1996);
- Grubbs test (Grubbs 1969);
- Hampel test (Hampel et al. 1986);
- active methods of robust estimation: Danish method, Huber method and Hampel method (Kamiński and Nowel 1992, Huber 1981, Hampel and Ronchetti, Rousseeuw and Stahel 1986);
- passive methods of robust estimation: Baarda's method and Pope's method (Baarda 1968, Pope 1961, Prószyński and Kwaśniak 2002).

However there has been no publication in which an attempt has been made to formulate criteria for selecting statistical methods to search for outliers in the analyzed datasets. This work is an excerpt from an attempt to determine such criteria based on two selected methods: model residual analysis and Baarda's method.

Model residual analysis is widely applied in many research from different fields of science, not only to find outliers, e.g. Dąbrowski and Adamczyk (2010) about global attributes utilization in predicting and forecasting a real property market value, Bittner (2007) about construction of multiple regression model in real estate valuation, Plonsky and Ghanbar (2018) about multiple regression in L2 research or Świderski et al. (2018) about road safety level in Poland. Model residual analysis is simple tool to detect outliers from model matched to examined datasets. Much more advanced algorithm is Baarda's method which relies on statistical test verifying hypothesis about lack of outliers among observations. This method, more complicated than model residual analysis, is often used in economic analyses (Orwat 2006, Dehnel and Gołata 2010, Majewska 2011) and in geodetic calculations (Kamiński and Nowel 1992, Wiśniewski 2009, Huber 1981). The main question is which of them gives better results in context of number of detected outliers in relation to increase of the determination coefficient and decrease of the standard estimation error. To answer to above questions, the research was conducted on seven datasets containing information about real estate sold in local markets in short time (up to six months).

✉spiewak@agh.edu.pl

METHODS AND MATERIALS

Model Residual Analysis

Residuals from the model reflect the differences between the values of the explained variable, observed and predicted by model. A well-matched model is characterized by small residuals for typical observations and large ones for outliers. Identifications of outliers based on model residuals can be performed using standardized residual values (Walesiak 1996), rejecting those observations which distance from the regression hyperplane is greater than the doubled residual standard deviation (Śpiewak 2017).

Baarda's Method

The basis of passive robust estimation methods are statistical tests that allow, after estimating the model parameters by the method of least squares, to determine which observations may be suspected to be gross errors. This group of methods includes Baarda's method (Kamiński and Nowel 1992, Proszyński and Kwaśniak 2002, Śpiewak 2018) where the zero hypothesis in form "there are no outliers in the examined dataset" is verified. To perform statistical test, the suitable test statistics which comes from *chi-square* distribution should be calculated. If test statistics is in critical area then the zero hypothesis is rejected in favor of the alternative hypothesis at given significance level what means that there is at least one outlier in dataset.

RESEARCH MATERIAL

The calculations were carried out on several datasets that contained information about real estate sold on local markets (in the cities of Sieradz, Skarżysko-Kamienna, Kraków, Rzeszów, Przeworsk and Busko-Zdrój) in a short period (up to six months). Each of the land properties designated in the local plan for single-family housing was described by following features: topography, sunlight, access to public transport, location, shape, fashion, basic function, utilities, road type, surroundings and land area. Housing properties have been assigned the following attributes: usable area, the area of the appendant rooms, number of rooms, storey, location, communication, sourroundings, building condition, elevator, balcony, greenery/recreation, monitoring, window exhibition, building technology and parking possibilities. Then, for each dataset, parameters of linear and non-linear functional models were estimated, describing the relationships prevailing in the examined markets. Statistical tools as Baarda's method and residual analysis were used to detect outliers in the analyzed datasets. The last stage was a comparison of the obtained results of estimation parameters of the analyzed models and their quality measures, before and after removal of outliers. Table 1 contains the basic characteristics of the examined datasets:

- $C_{med.}$ – average unit price;
- Me – median of unit price;
- σ_{n-1} – prices' standard deviation;
- R^2 – determination coefficient;
- σ_0 – standard estimation error;
- V – variation coefficient;
- n – number of observations;
- o – the number of removed outliers;
- R^2 - R_0^2 the difference between the value of the determination coefficient after and before the application of statistical methods searching outliers;
- o/n – the percentage of atypical observations in relation to the initial number of properties in the database no. 1. Results for database no. 2–7 are contained in the appendix 1 (Tab. appx. 1).

The following abbreviations have been used in the Table 1 and in the figures:

- LM-B – linear model before removal of outliers,
- LM-AR – linear model after application residual analysis,
- LM-MB – linear model after application Baarda's method,
- NM-B – non-linear model before removal outliers,
- NM-AR – non-linear model after application residual analysis,
- NM-MB – non-linear model after application Baarda's method,
- AR – residual analysis,
- MB – Baarda's method.

✉spiewak@agh.edu.pl

Table 1. The extract of summary of obtained results

Database 1 contains information about land properties										
Method	C_{medium} [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B	52.80	41.67	53.16	0.60	33.97	0.64	73			
LM-AR	44.17	34.15	39.86	0.83	16.65	0.38	45	28	0.23	0.38
LM-MB	48.25	42.08	41.43	0.62	25.79	0.53	64	9	0.02	0.12
NM-B	52.80	41.67	53.16	0.78	28.54	0.23	73			
NM-AR	46.21	39.61	50.11	0.86	25.12	0.54	67	6	0.08	0.08
NM-MB	47.69	39.79	49.80	0.85	26.83	0.56	66	7	0.07	0.10

Source: own study

There are figures which present:
- a change of determination coefficient value R^2 compared to its initial value R^2_0. including the percentage of observations considered outliers (Fig. 1. Fig. 2.)
- a change of standard estimation error σ_0 compared to its initial value $\sigma_{0\text{pocz}}$. including the percentage of observations considered outliers (Fig. 3. Fig. 4).
- a change of R^2 value and σ_0 value and percentage of outliers after application Baarda's method and residual analysis in linear (Fig. 5) and non-linear model (Fig. 6).

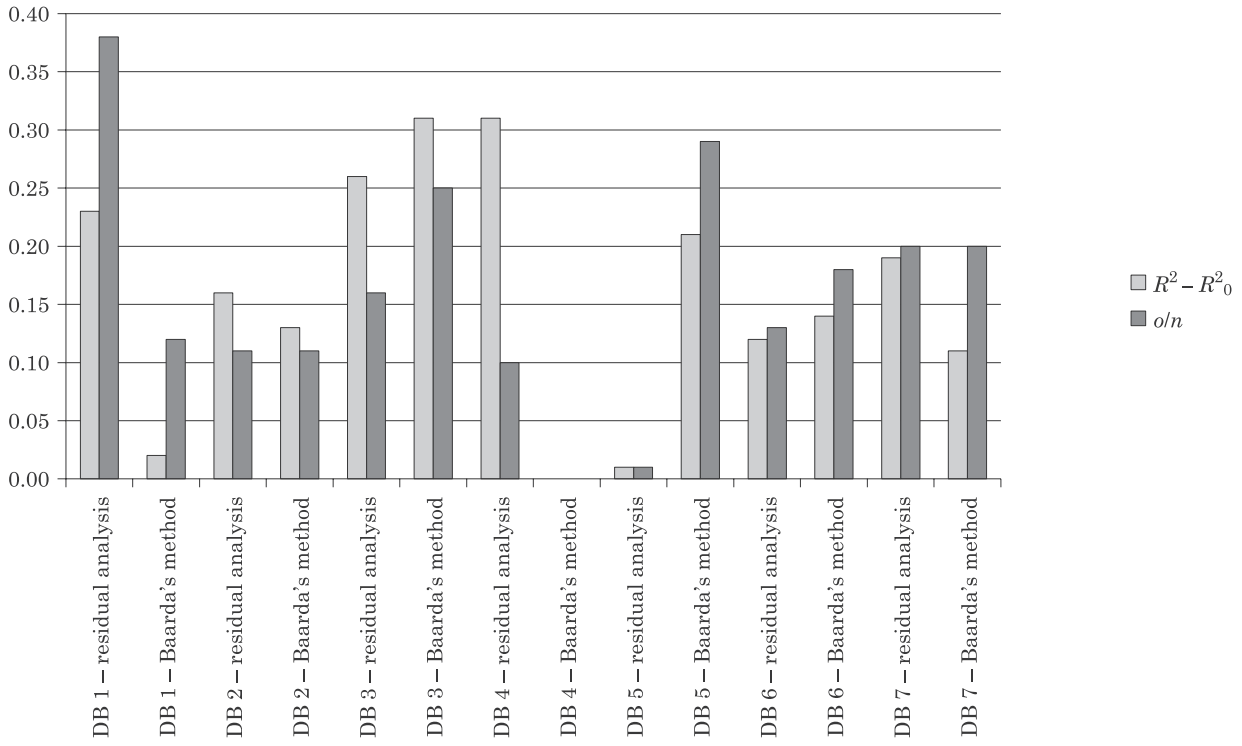

Fig. 1. A change of the value of the determination coefficient in a linear model
Source: own study

spiewak@agh.edu.pl

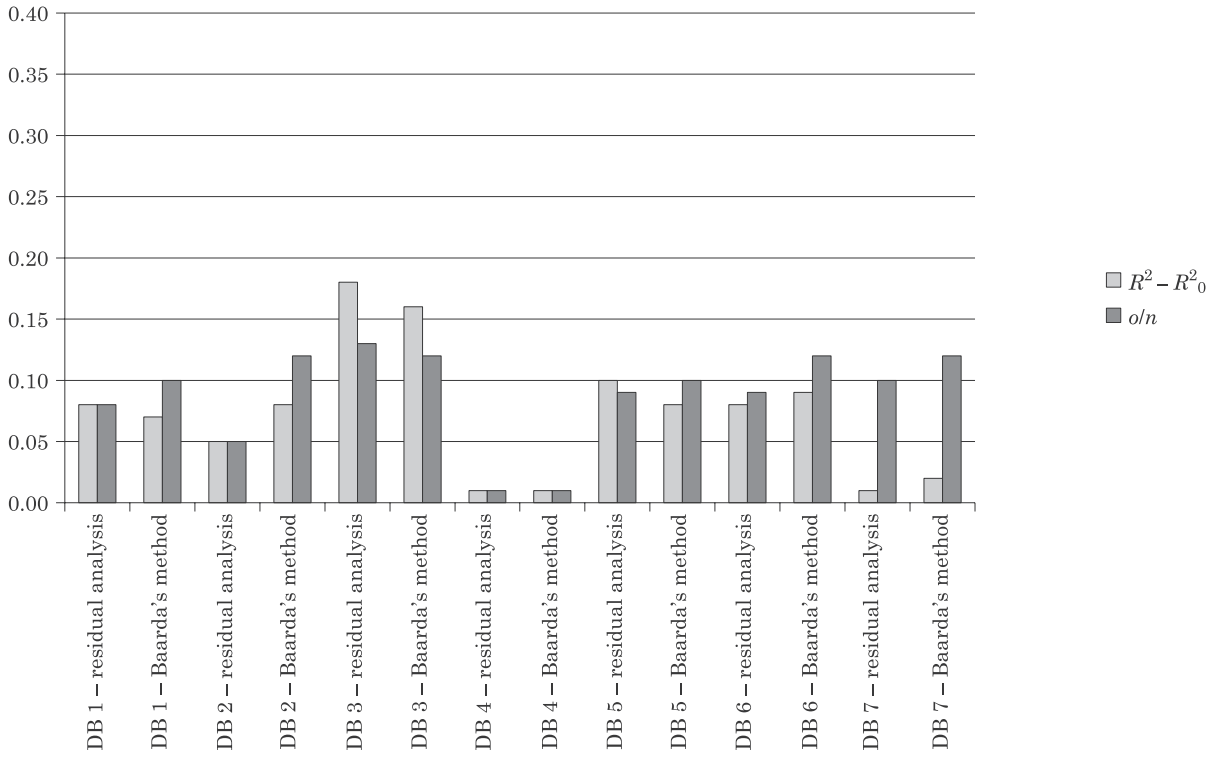

Fig. 2. A change of the value of the determination coefficient in a non-linear model
Source: own study

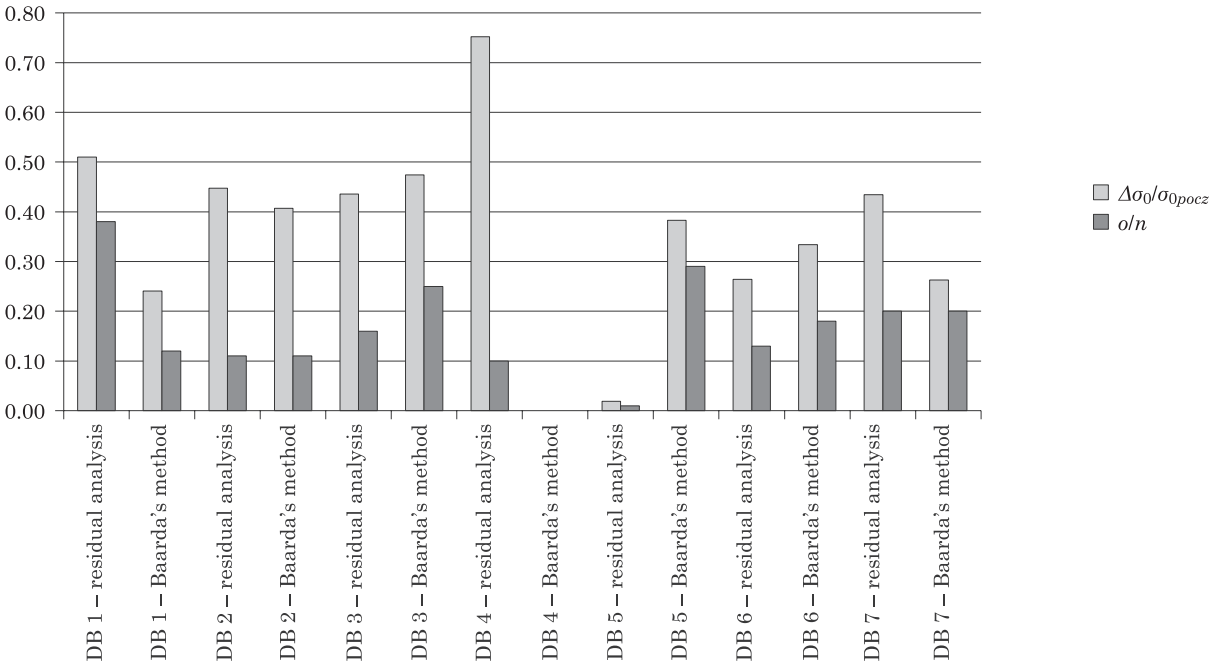

Fig. 3. A change of the value of the standard estimation error in the linear model
Source: own study

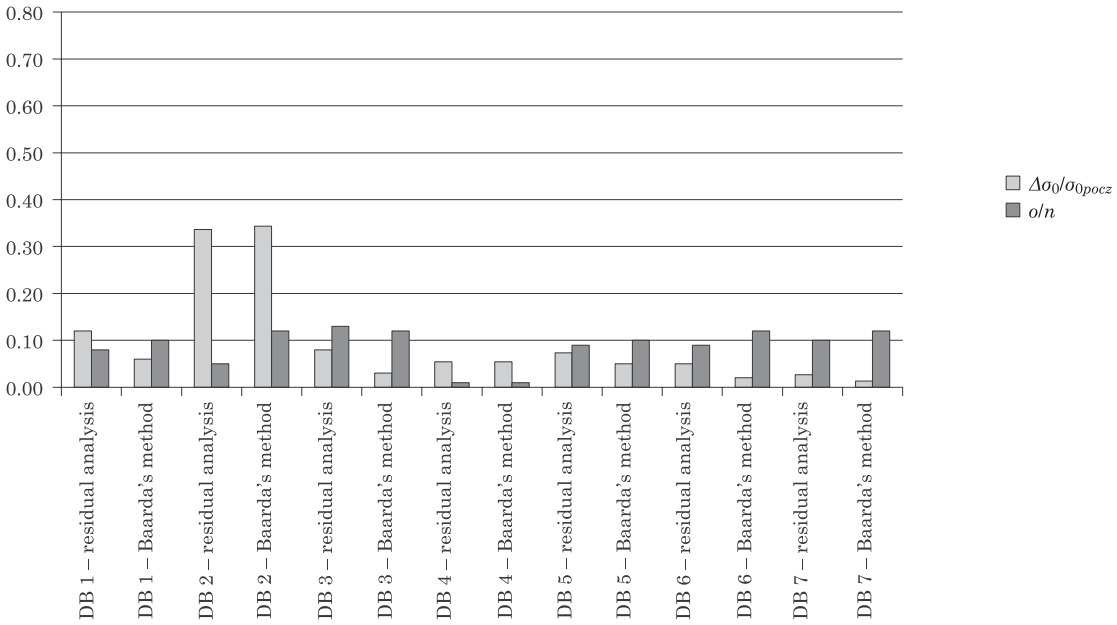

Fig. 4. A change of the value of the standard estimation error in the non-linear model
Source: own study

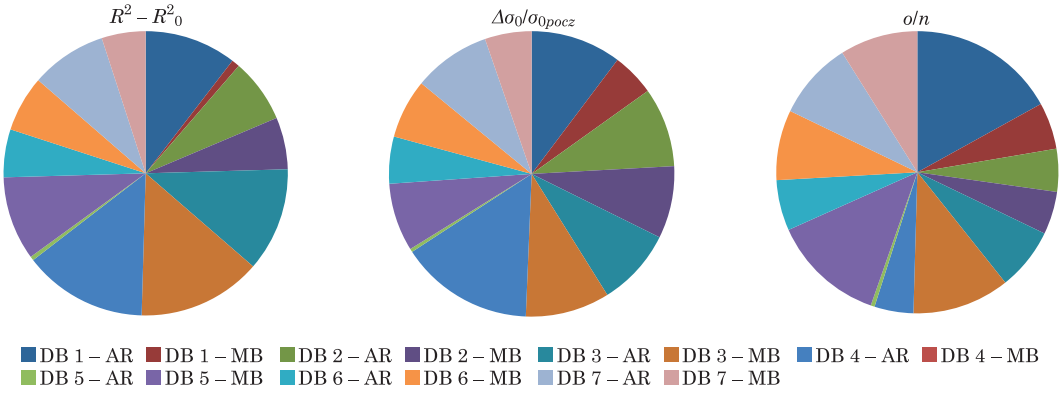

Fig. 5. Results after application residual analysis and Baarda's method in linear model
Source: own study

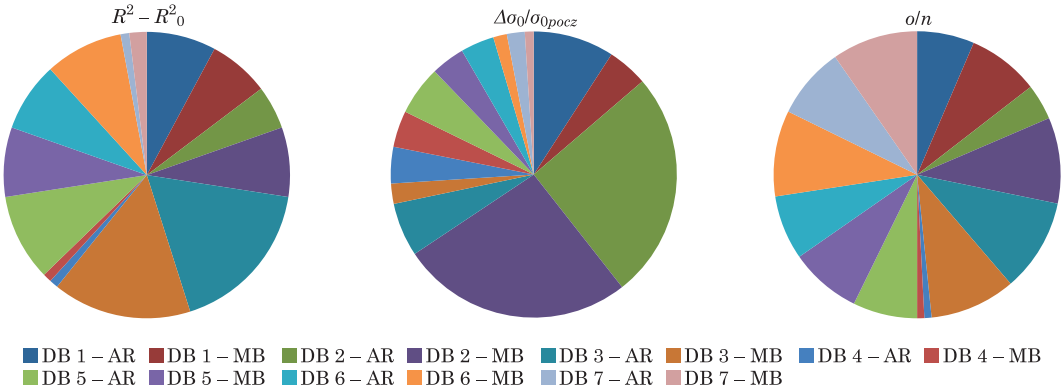

Fig. 6. Results after application residual analysis and Baarda's method in non-linear model
Source: own study

spiewak@agh.edu.pl

CONCLUSIONS

In the examined datasets. both tested algorithms perform the function of searching outliers from a linear or non-linear model matched to the observations. The exception is database 4 (Tab. appx. 1) where the Baarda's method did not detect cases departing from the linear model (Tab. 1, Fig. 1). Removal large amount of data does not always cause a significant increase of determination coefficient value and reduce the value of the residue variance estimator, e.g. in database 1 after using Baarda's method for a linear model. In a few cases, model residual analysis gives better results than Baarda's method because the smaller number of outliers is rejected when the increase of determination coefficient is at the similar level, e.g. in datasets 3 and 6 (Fig. 1) and databases 1, 6 and 7 (Fig. 2). A similar effect was achieved for decreasing the value of the residue variance estimator, e.g. in databases 3 and 6 in the linear model (Fig. 3). In the non-linear model, both methods cause change of the σ_0 value at a similar level with a comparable percentage of outliers (Fig. 4). Therefore the conclusion is that a simple statistical method, which is the study of the occurrence of cases deviating from the functional model based on the analysis of residues, generates the similar results as the use of a much more complicated algorithm as the one proposed by Baarda.

It is important to underline that both algorithms used in this article detected outliers from model well matched to data. We obtained increase of the value of the determination coefficient and decrease of standard estimation error but eliminated observations may also be valuable information from analyzed markets (from another point of view, regarding unusual behavior).

REFERENCES

Baarda, W. (1968). A testing procedure for use in geodetic networks. Publications on Geodesy, New Series, Netherlands Geodetic Commision 2(5).

Bittner, A. (2007). Construction of the multiple regression model in real estate valuation. Acta Sci. Pol. Administratio Locorum 6(4), 59–66.

Budka, A., Kayzer, D., Piotruczuk, K., Szoszkiewicz, K. (2013). testing procedures for detection of observation influential for river assessment. Infrastructure and Ecology of Rural Areas 3(II), 85–95.

Cook, R., Weisberg, S. (1982). Residues and influence in regression, Chapman & Hall, Nowy Jork.

Czekaj, T. (2006). Outliers and influential observations in regression analysis. Analysis of profitability of material factors of production in farms. Annals of the Polish Association of Agricultural and Agribusiness Economists 5, 11–15.

Dąbrowski, J., Adamczyk, T. (2010). Global attributes utilization in predicting and forecasting a real property market value. Acta Sci. Pol. Administratio Locorum 9(2), 47–58.

Dehnel, G., Gołata, E. (2010). On robust estimators for a polish business survey, Cracow Review of Economics and Management 10, 107–121.

Głowicka-Wołoszyn, R., Wysocki, F. (2018). The problem of identifying of development levels in constructing synthetic fatures. Research Papers of Wrocław University of Economics 508, 56–65.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. Technometrics 11(1), 1–21.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J, Stahel, W.A. (1986). Robust statistics. the approach based on influence functions. Wiley and Sons, Nowy Jork.

Huber, P.J. (1981). Robust statistics, John Wiley and Sons, Nowy Jork.

Kamiński, W., Nowel, K. (1992). Analysis of chosen robust methods of geodetic observation adjustment. Geodesy and Cartography 41(3–4), 183–195.

Koronacki, J., Mielniczuk, J. (2001). Statistics for student of technical and natural sciences. Scientific and Publishing House, Warsaw.

Korir, B.C, Kinyanjui, J.K. (2019). Parametric interval estimation of the greeta distribution. International Journal of Statistics and Probability 8(1), s. 1–15.

Li, X., Deng, S., Li, L., Jiang, Y. (2019). Outlier detection based on robust mahalanobis distance and its application. Open Journal of Statistics 9, 15–26.

Meyers, L.S, Gamst, G., Guarino, A.J. (2006). Applied multivariate research. Sage, Nowy Jork.

Pope, A.J. (1976). The statistics of residues and the detection of outliers. NOAA Technical Report. NOS 65 NGS 1, U.S. Dept. of Commerce, Rockville, Md.

✉spiewak@agh.edu.pl

Orwat, A. (2006). Sample applications of robust estimation in modelling financial time series. Scientific Journal of Warsaw University of Life Sciences-SGGW, Warsaw, pp. 279–288.

Plonksy, L., Ghanbar, H. (2018). Multiple regression in l2 research. A methodological synthesis and guide to interpreting r^2 values. The Modern Language Journal, https://www.researchgate.net/profile/Hessameddin_Ghanbar/publication/327923945_Multiple_Regression_in_L2_Research_A_Methodological_Synthesis_and_Guide_to_Interpreting_R_2_Values/links/5dcba869a6fdcc57504406cd/Multiple-Regression-in-L2-Research-A-Methodological-Synthesis-and-Guide-to-Interpreting-R-2-Values.pdf, access: 19.11.2019.

Prószyński, W., Kwaśniak, M. (2002). Reliability of geodetic networks, Warsaw Polytechnic Publishing House.

Śpiewak, B. (2017). Ocena poprawności doboru metod statystycznych jako narzędzi analizy rynku nieruchomości, w: Wybrane problemy rynku nieruchomości i gospodarowania przestrzenią (The appraisal of correct choice of statistical methods as tools of property market analysis, in: Chosen problems of real estate and space management). Eds. I. Foryś, Jan Kazak. Towarzystwo Naukowe Nieruchomości, pp. 89–101, http://tnn.org.pl/tnn/publik/25/Monografia_2017.pdf, access: 9.10.2019.

Śpiewak, B. (2018). Application of passive methods of robust estimation: baarda and pope's in real estate market analysis. Real Estate Management and Valuation 26(1), 5–15.

Świderski, A., Borucka, A., Skoczynski, P. (2018) Characterisistics and assessment of the road safety level in poland with multiple regression model. Proceedings of 22nd International Scientific Conference. Transport Means 2018, https://www.researchgate.net/profile/Anna_Borucka3/publication/330811276_Characteristics_and_Assessment_of_the_Road_Safety_Level_in_Poland_with_Multiple_Regression_Model/links/5c54a08f458515a4c7502a94/Characteristics-and-Assessment-of-the-Road-Safety-Level-in-Poland-with-Multiple-Regression-Model.pdf, access: 18.11.2019.

Trzęsiok, J. (2015). Robustness for outliers of selected nonparametric regression models. Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach 227, 75–84.

Walesiak, M. (1996). Methods of marketing data analysis. Wydawnictwo Naukowe PWN, Warszawa.

Wiśniewski, Z. (2009). Compensatory calculation in geodesy (with examples), Warmia and Mazury Publishing House, Olsztyn.

✉spiewak@agh.edu.pl

APPENDIX 1

Table appx. 1. Summary of obtained results

Database 2 contains information about land properties										
Method	$C_{med.}$ [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B	39.44	18.66	37.82	0.77	18.27	0.46	65	–	–	–
LM-AR	36.78	13.20	38.19	0.93	10.10	0.27	58	7	0.16	0.11
LM-MB	34.42	13.20	34.31	0.90	10.84	0.31	58	7	0.13	0.11
NM-B	39.44	18.66	37.82	0.74	18.27	0.46	65	–	–	–
NM-AR	31.29	15.25	33.12	0.79	12.12	0.39	62	3	0.05	0.05
NM-MB	29.56	17.26	28.76	0.82	11.99	0.41	57	8	0.08	0.12
Database 3 contains information about land properties										
Method	$C_{med.}$ [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B	50.22	31.12	43.33	0.46	32.21	0.64	69	–	–	–
LM-AR	41.18	27.38	33.82	0.72	18.18	0.44	58	11	0.26	0.16
LM-MB	40.90	26.00	34.63	0.77	16.94	0.41	52	17	0.31	0.25
NM-B	50.22	31.12	43.33	0.68	27.43	0.55	69	–	–	–
NM-AR	48.40	27.89	35.61	0.86	25.25	0.53	60	9	0.18	0.13
NM-MB	47.12	26.61	34.84	0.84	26.54	0.56	61	8	0.16	0.12
Database 4 contains information about land properties										
Method	$C_{med.}$ [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B	51.49	31.66	92.30	0.43	70.20	1.36	86	–	–	–
LM-AR	36.83	30.00	33.60	0.74	17.41	0.47	77	9	0.31	0.10
LM-MB	51.49	31.66	92.30	0.43	70.20	1.36	86	0	0.00	0.00
NM-B	51.49	31.66	32.30	0.87	19.64	0.38	86	–	–	–
NM-AR	46.29	33.69	31.12	0.88	18.58	0.40	85	1	0.01	0.01
NM-MB	46.29	33.69	31.12	0.88	18.58	0.40	85	1	0.01	0.01
Database 5 contains information about land properties										
Method	$C_{med.}$ [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B*	55.06	54.73	34.14	0.65	20.34	0.37	185	–	–	–
LM-AR	54.93	54.73	34.18	0.66	19.96	0.36	183	2	0.01	0.01
LM-MB	68.75	56.07	32.61	0.85	12.55	0.18	131	54	0.21	0.29
NM-B	55.06	54.73	34.14	0.72	18.30	0.33	185	–	–	–
NM-AR	49.13	50.74	31.55	0.82	16.96	0.35	168	17	0.10	0.09
NM-MB	51.38	52.05	36.89	0.80	17.33	0.34	166	19	0.08	0.10

✉spiewak@agh.edu.pl

cont. Table appx. 1.

Database 6 contains information about flat properties										
Method	$C_{med.}$ [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B	3180.41	3204.23	659.54	0.73	341.84	0.11	77	–	–	–
LM-AR	3217.86	3203.70	662.54	0.86	251.50	0.08	67	10	0.12	0.13
LM-MB	3155.20	3158.71	627.65	0.87	227.75	0.07	63	14	0.14	0.18
NM-B	3180.41	3204.23	659.54	0.68	351.37	0.11	77	–	–	–
NM-AR	2985.12	3101.25	625.99	0.76	333.89	0.11	70	7	0.08	0.09
NM-MB	2874.02	3000.85	648.12	0.77	344.27	0.12	68	9	0.09	0.12
Database 7 contains information about flat properties										
Method	$C_{med.}$ [PLN/m^2]	Me [PLN/m^2]	σ_{n-1} [PLN/m^2]	R^2	σ_0 [PLN/m^2]	V	n	o	$R^2 - R_0^2$	o/n
LM-B	2575.52	2432.25	871.68	0.73	454.71	0.18	101	–	–	–
LM-AR	2585.13	2427.70	921.38	0.92	257.11	0.10	81	20	0.19	0.20
LM-MB	2650.25	2447.23	844.68	0.84	335.11	0.13	81	20	0.11	0.20
NM-B	2575.52	2432.25	871.68	0.83	324.71	0.13	101	–	–	–
NM-AR	2521.08	2432.25	856.47	0.84	316.12	0.13	91	10	0.01	0.10
NM-MB	2516.60	2469.87	870.64	0.85	320.43	0.13	89	12	0.02	0.12

*Explanations as in Table 1
Source: own study

✉spiewak@agh.edu.pl