Magdalena Okupniak[33], Leszek Wanat[34], Elżbieta Mikołajczak[35], Łukasz Sarniak[36], Agata Dolacińska-Śróda[37]

SELECTED OUTLIER IDENTIFICATION METHODS IN THE WOOD-BASED SECTOR COMPETITIVENESS POTENTIAL RESEARCH

Abstract: Research in effect analyzing data leads to observations, according to which some values strongly exceed the others. It may cause estimation inefficiency or mistakes in descriptive statistics results. Hence, the use and development of outlier identification methods, which are a part of robust statistics, should be considered necessary. The aim of the article is to define selected outlier identification methods and present the results of using them in the wood-based sector competitiveness potential research.

Key words: outlier, outlier identification methods, wood-based sector, competitiveness potential, hat matrix, leverage values.

INTRODUCTION

The wood market is a particular market. It consists of two complementary sectors of the economy: forestry and arboriculture. Their mutual relations are shaped both by the market mechanism as well as the institutional determinants remaining in connection with the paradigm of sustainable development coming from forestry sciences.

In wood market research, the traditional approach to economics is met with an integral approach. The forest-wood sector is a significant part of the natural economy, and its analysis refers to the achievements of forestry sciences, especially the economics of arboriculture. Undertaking attempts to explain competitive phenomena and tendencies occurring in the wood industry is a remarkable and, simultaneously, difficult challenge, requiring (to capture significant trends) accurate identification and subsequent elimination of outliers.

When we talk about outliers, we usually mean unusual observations, which are different from the others. There are many sources of their formation. Two categories of atypical observations are distinguished in the literature: - incorrectly given ones, which may be improperly recorded during data collection, coding or are deliberately changed, and - observations correctly given, clearly standing out from the others.

The latter usually come from a different population than the key part of the observation, called the core. Hence, the whole set of data can be divided into the core and outliers. Stanisław Heilpern underlines, that there is a lack of a clear and precise outlier definition (Heilpern 2005, p. 65). In table 1 we present selected definitions.

The manuscript uses a proprietary approach, taking into account the identification of observations of extreme values, differing significantly from the majority of observations in the studied population, and, simultaneously, subjecting this choice to a critical descriptive analysis. On this basis, the selection and justification of the possible inclusion or elimination of the identified observations from further research and the inference process were made.

[33]*Poznan University of Life Sciences, Department of Finance and Accounting, ul. Wojska Polskiego 28, 60-637 Poznań, magdalena.okupniak@up.poznan.pl;*
[34]*Collegium Da Vinci in Poznań, Faculty of Social Sciences, ul. T. Kutrzeby 10, 61-719 Poznań, leszek.wanat@cdv.pl, corresponding author.*
[35]*Poznan University of Life Sciences, Department of Economics and Wood Industry Management, ul. Wojska Polskiego 38-42, 60-627 Poznań, emikolaj@up.poznan.pl;*
[36]*Poznan University of Life Sciences, Department of Finance and Accounting, ul. Wojska Polskiego 28, 60-637 Poznań, lukasz.sarniak@up.poznan.pl;*
[37]*The European Academy of Hotel Management and Catering Industry in Poznań, agata.sroda@wshig.poznan.pl;*

Table 1. Review and discussion of selected outlier definitions

Type	Type identification	Description	Sources
Type 1	the tail of the statistical distribution	"An outlier is a data value that lies in the tail of the statistical distribution of a set of data values. In the distribution of raw data, outliers are often regarded as more likely to be incorrect. In contrast, an inlier is an erroneous data value which actually lies in the interior of a statistical distribution, making it difficult to distinguish from good data values" (Eurostat glossary 2017).	Eurostat glossary; Statistical themes (2017)
Type 2	separated from others, from a different population	"In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is a fault. Such values are called outliers" (European Commision 2007, p. 119).	European Commision (2007)
Type 3	they are clearly different from others	"Often in a data set we can differentiate values which clearly different from the others. These we name outliers. Their impact on the results of the methods used, especially classical, closely related to parametric models, is usually quite significant" (Ostasiewicz 1999).	Ostasiewicz (1999)
Type 4	it differs from the pattern, a determined by the majority	"Outlier observation is such an element of the sample that in some way differs from the pattern determined by the majority of the sample elements. (...) Or we define them in terms of the observation position in the sample without referring to the random mechanism that the sample has generated, or assume a certain model of outlying" (Kosiorowski 2012, p. 27).	Kosiorowski (2012)
Type 5	observation far away from others (extreme)	"Extreme observation can be an outlier, when is appropriately far from the others, i.e. when its value is too big or too small, even if it is an extreme value" (Heilpern 2005, p.45-46)	Heilpern (2005)
Type 6	a small part of extreme data, when most have skew distribution	"Outliers are a frequent concern in surveys with quantitative variables like household budget surveys or business surveys on production or turnover. A relatively small fraction of the data has extreme values in one or several variables. Often these extreme values occur when the bulk of the data has already a markedly skew distribution" (Hulliger 2016).	Hulliger (2006)

Source: Authors' own elaboration

THE IDEA OF RESEARCH AND METHOD

While developing the research scenario, the experience of other studies concerning the forest-wood sector was used, in which problems emerged with the identification and elimination of outliers. In addition, references were made to the manuscripts in which proposals for such organization and research optimization were formulated to reduce or eliminate the impact of outliers if possible (Wysocki 2010; Ostasiewicz 2012; Popek and Wanat 2014; Kusiak *et al.* 2017; Paluš *et al.* 2017). The following theoretical approach was proposed in the study:

1. **One-dimensional statistical tests.** The outliers observed in the studies in an unequal way affect the values of the estimation of regression parameters estimated with the least-squares method or the degree of adjustment of the regression hyper-plane to the observations. After separating them from the remaining observations, they significantly change the parameters and sizes of the modelling residues. However, it should be considered that not always the outlier observation

must be part of another population and we cannot reject it in every case. Homogeneity tests of the sample in the one-dimensional case allow the identification of outliers. They make it possible to decide whether the considered observation violates the homogeneity of a sample derived from a normal distribution.

2. **Multi-dimensional-Welsch measure**. In this part of the article we will discuss the methodology that allows for multidimensional identification of outliers and, in particular, influential observations.

For the general form of the linear regression model, which is expressed by the following formula:

$$Y = X\beta + e \tag{1}$$

we can define a square **hat matrix** H, its dimension is n×n (Ostasiewicz 1999, p. 334):

$$H = X(X^T X)^{-1} X^T \tag{2}$$

It should be noted here that the matrix X is a matrix of explanatory variables values, its dimension is $n \times (p+1)$, where p is the number of auxiliary variables considered in the given regression model, and n is the sample size.

Hat matrix is symmetric and idempotent. Its diagonal elements (h_i, i=1,…,n) are named leverage values and they describe the impact of individual observations on the assessment of the parameters of the regression model and meet inequality $\frac{1}{n} \leq h_i \leq 1$.

The strength of the effect of observations increases proportionally to the value of the leverage value. A method of inference are selected - based on the analysis of the descriptive study, formulated according to the main hypothesis: direct or reverse.

The question arises, from which moment - for which value of the influential value, the limit value can be determined, which will allow the assessment of the observation to be influential or not?

Such thresholds were proposed by Hoaglin, Welsch and Velleman.

They are the following (Ostasiewicz 1999; 2012):

a) Hoaglin and Welsch: $h_H = \frac{2p}{n}$, (3)

b) Velleman and Welsch: $h_V = \frac{3p}{n}$, (4)

The classical inference procedure proceeds in such a way that if the observed value of the leverage value is greater than the threshold value h_H or h_V, then we consider the given observation to be influential. In the presented study, the reverse hypothesis was verified, because the analyzed source matrix did not include individual values of each index, but aggregated competitiveness measures (Ostasiewicz 2012). As a result of a descriptive analysis, influential values were determined indicating the competitive ability of the analyzed countries (and their wood markets), the lowest indications for "leverage value" were adopted successively (Popek and Wanat 2014).

THE RESEARCH DESCRIPTION AND RESULTS

The analysis of the competitive position resulting from the competition process is considered in the perspective of the results obtained by competitors (Olczyk 2008). Its aim is to try to determine the place of the industry on the market, defined in comparison to competitors (Gorynia 2010), i.e. the position that the domestic industry achieved in relation to analogous branches of other national economies (Lubiński *et al.* 1995). In the study of the competitive position of the round wood market, selected synthetic measures of a resultant nature were used. In order to obtain an aggregated result, detailed sorting of wood raw material was omitted. Statistical secondary data were included and verified based on the value of exports and imports of round wood in selected countries (OECD

2012; Popek and Wanat 2014). Taking the consistency, comparability and adequacy of the available data as the starting point, the following measures of competitiveness were selected (see table 2):

- SI export specialization indicator (Specialization Indicator),
- coverage of imports by CR export (Coverage Ratio),
- intra-industry trade intensity indicator IIT (Intra-Industry Trade),
- RCA (Revealed Comparative Advantage Index).

Table 2. Comparison of competitive position indicators of selected European countries for round-wood markets, according to the substantive and statistical criterion of the OECD (2017)

Selected countries in Europe	SI Index	CR Index	IIT Index	RCA Index
Austria	2,09	11,88	0,21	-0,79
Belgium	1,15	50,64	0,67	-0,33
Czech Republik	9,31	203,34	0,66	0,34
Denmark	2,72	73,67	0,85	-0,15
Estonia	36,64	559,59	0,30	0,70
Finland	3,68	19,02	0,32	-0,68
France	2,22	223,86	0,56	0,44
Germany	0,93	57,78	0,73	-0,27
Italy	0,19	7,45	0,14	-0,86
Irleland	0,93	76,23	0,86	-0,14
Netherland	0,21	169,85	0,74	0,26
Norway	2,89	153,73	0,79	0,21
Poland	**4,12**	**165,99**	**0,75**	**0,25**
Portugal	6,87	79,31	0,88	-0,12
Slovakia	11,77	646,11	0,27	0,73
Slovenia	13,23	245,48	0,58	0,42
Spain	1,79	115,07	0,93	0,07
Switzerland	1,55	408,61	0,39	0,61
Sweden	1,56	15,24	0,26	-0,74
United Kingdom	0,46	93,36	0,97	-0,03

Source: Authors' own elaboration based on Wanat (2015), Wanat and Klus (2016)

To analyze the competitive position indicators: SI, CR, IIT and RCA, a group of 20 countries was selected after verification of statistical data: Austria, Belgium, the Czech Republic, Denmark, Estonia, Finland, France, Spain, the Netherlands, Ireland, Germany, Norway, Poland, Portugal, Slovakia, Slovenia, Switzerland, Sweden, Great Britain and Italy. Unfortunately, no coherent data were obtained for Lithuania and Latvia (non-compliance of FAO and OECD data), which is why these countries were omitted. Competitiveness measures have been calculated for selected countries.

The values of the SI export specialization index were determined, which enabled to compare the share of round timber exports of the studied country with the share of this raw material in world exports. The high values of this indicator show a strong competitiveness of the examined market.

The export coverage rate for the round wood market was also calculated. The strength of specialization in the studied area is demonstrated by the CR index values, for which indications higher than 100 were obtained (for CR given as a percentage) (Lubiński *et al.* 1995). The CR index is considered a measure of the 'internal comparative advantage', useful for researching the competitiveness of the industry (Pawlak 2013, p. 98). Next, the value of the IIT index of intra-industry trade in round wood was determined, according to the formula of Grubel and Lloyd and the

value of the index of revealed comparative advantages of RCA (Jankowska 2005). The results are summarized in Table 2.

The following assumptions were made to determine influential indications, referring to threshold values:

$h_H = 0{,}4$
$h_V = 0{,}6$

Based on them, calculations were made, the results of which are summarized in Table 3.

Table 3. Identification of influential values and elimination of outliers in the study of the competitive position of round-wood markets in selected European countries

Classification	State	Leverage value
Type 6. (**outliers that have no impact**)	*Estonia*	*0,815*
	Slovakia	*0,599*
Type 5. (minimal impact)	Switzerland	0,339
	Italy	0,282
Type 4. (low impact)	Austria	0,228
	Sweden	0,198
	Finland	0,172
Type 3. (average impact)	Portugal	0,143
	United Kingdom	0,134
	Slovenia	0,129
	Spain	0,120
Type 2. (significant impact)	Netherlands	0,106
	Ireland	0,103
	Denmark	0,101
	France	0,097
	Czech Republik	0,093
Type 1. (high impact)	Norway	0,089
	Germany	0,085
	Belgium	0,083
	Poland	**0,082**

Source: Authors' own elaboration

It is worth noting that the proposed method of identifying and then eliminating outliers allows to determine the competitive position of the surveyed European countries, in this case taking into account their primary wood market (round wood) (Wanat and Klus 2015), with a greater precision.

The method gives an additional opportunity to capture and verify some of the characteristics that are not easily discernible using traditional methods. This way (see Table 3), a completely different description of the Estonian and Slovakian markets was noticed (Kaputa *et al.* 2016), whose competitiveness is probably not determined by the classic competitive potential, but it is necessary to seek other additional development factors (in the study, these countries were identified as so-called "non-influential").

Three so-called 'middle' groups (type 2,3 and 4), selected by a descriptive method may be of interest (Ostasiewicz 2012), although other agglomeration methods can be used as well (Wysocki 2010; Popek and Wanat 2014). An important observation is the identification of Poland among the influential leaders of the competitiveness ranking of round wood markets, alongside Germany and Norway. It is necessary to additionally verify the observed influential position of Belgium, about which in this classification decided not so much the market potential as the relative balance between

the internal market and international trade. This indicates different than the so-called resource model of competitiveness, which characterizes Poland.

CONCLUSIONS

The identification and classification of the competitive position in selected European countries, from the point of view of round wood markets in these countries, are usually carried out using traditional methods. This often leads to incomplete knowledge, sometimes to incomprehensible conclusions from outcomes with outlier traits.

For these reasons, the use of both traditional mathematical methods and indicators of competitiveness should be enriched with methods for identifying and eliminating outliers, then with agglomeration methods, and above all with contextual statistical inference, supported by descriptive analysis. This is particularly important when the subject of research is an industry based on natural resources, in this case on round wood.

The research conducted on the basis of a selected case study, which concerned the identification of the competitive position of the Polish round wood market, is also a recommendation for the industry policy. The opportunity of an economic growth, which is the position and potential of the domestic wood market, should be taken into consideration as one of the priorities of the Polish development policy.

REFERENCES

1. European Commission (2007). Handbook on Data Quality Assessment Methods and Tools, Eurostat. Wiesbaden.
2. Eurostat, Forestry statistics (2017). Economic accounts for forestry and logging - values at current prices. Forestry. Published on-line: http://ec.europa.eu/eurostat/statistics-explained/index.php/Forestry_statistics, accessed: 30.09.2017.
3. FAOSTAT (2017). Database on-line: http://faostat.fao.org/ , accessed: 30.09.2017.
4. Gorynia, M. (2010): Teoretyczne aspekty konkurencyjności. In: Gorynia M., Łaźniewska E. (eds.), Kompendium wiedzy o konkurencyjności. Warszawa: PWN.
5. Heilpern, S. (2005). Nietypowe realizacje jednowymiarowych zmiennych losowych. In: Bąk A. (eds.). Prace naukowe Akademii Ekonomicznej we Wrocławiu, 097.
6. Hulliger, B. (2006). Variance Estimation for Complex Surveys in the Presence of Outliers, Seattle: Joint Statistical Meeting.
7. Jankowska, B. (2005). Międzynarodowa konkurencyjność branży na przykładzie polskiej branży budowlanej w latach 1994-2001. Poznań: Wydawnictwo Akademii Ekonomicznej.
8. Kaputa, V., Paluš, H., Vlosky, R. (2016). Barriers for Wood Processing Companies to Enter Foreign Markets: a Case Study in Slovakia. European Journal of Wood and Wood Products, 74(1), 109-122.
9. Kosiorowski, D. (2012). Wstęp do statystyki odpornej. Kurs z wykorzystaniem środowiska R, Kraków: Wydawnictwo Uniwersytetu Ekonomicznego.
10. Kusiak, W., Mikołajczak, E., Molińska-Glura, M., Moliński, K., Biszof, A., Wanat, L. (2017). Innovative approach to traditional case studies: Economic, social and ergonomic aspects of wooden church pews functionality - the case of Poland. In: D. Jelačic (eds.) Innovations in forestry, wood processing and furniture manufacturing, (p.253-264). Zagreb: WoodEMA, i.a.
11. Lubiński, M., Michalski, T., Misala, J. (1995). Międzynarodowa konkurencyjność gospodarki. Pojęcia i sposób mierzenia. Warszawa: Instytut Rozwoju i Studiów Strategicznych.
12. Mikołajczak, E. (2011). Ekonomiczne aspekty przerobu odpadów drzewnych na paliwa ekologiczne. Poznań: Wyd. UP.
13. OECD, http://stats.oecd.org/, accessed: 30.09.2017.

14. Olczyk, M. (2008). Konkurencyjność. Teoria i Praktyka. Na przykładzie polskiego eksportu artykułów przemysłowych na unijny rynek w latach 1995-2006. Poznań: Wyd. CeDeWu.
15. Ostasiewicz, W. (eds.), (1999). Statystyczne metody analizy danych (p. 312-351).Wrocław: Wydawnictwo Akademii Ekonomicznej im. Oskara Langego.
16. Ostasiewicz, W. (2012). Myślenie statystyczne. Warszawa: Wydawnictwo Wolters Kluwer Business.
17. Paluš, H., Parobek, J., Vlosky, R. P., Motik, D., Oblak, L., Jošt, M., ... & Wanat, L. (2017). The status of chain-of-custody certification in the countries of Central and South Europe. European Journal of Wood and Wood Products, 1(12), 699-710.
18. Popek, M., Wanat, L. (2014). Price Versus Non-Price Factors of Sector Competitiveness: Case Study of the Round Wood Market in Poland. Intercathedra, 30(2), 71-77.
19. Pawlak, K. (2013). Międzynarodowa zdolność konkurencyjna sektora rolno-spożywczego krajów Unii Europejskiej, 448. Poznań: Wyd. UP.
20. Wanat, L. (2015). Rynek drzewny w Polsce – potencjał i pozycja konkurencyjna (unpublished doctoral dissertation). Poznań: Wyd. UP.
21. Wanat, L., Klus, S. (2015). Sytuacja konkurencyjna branży i mezoekonomiczne aspekty polityki sektorowej państwa na przykładzie rynku drzewnego w Polsce, Rynek-Społeczeństwo-Kultura, 1 (13), 41-45.
22. Wysocki F. (2010): Metody taksonomiczne w rozpoznawaniu typów ekonomicznych rolnictwa i obszarów wiejskich. Poznań: Wyd. UP.
23. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Outlier.