sciendo CzOTO 2019, volume 1, issue 1, pp. 716-723

THE SMOOTHED BOOTSTRAP FINE-TUNING

doi: 10.2478/czoto-2019-0091
Date of submission of the article to the Editor: 24/11/2018
Date of acceptance of the article by the Editor: 30/01/2019

Renata Dwornicka[1] – *orcid id: 0000-0001-6761-9623*
Andrii Goroshko[2] – *orcid id: 0000-0002-1386-2326*
Jacek Pietraszek[1] – *orcid id: 0000-0003-2851-1606*
[1]Cracow University of Technology, **Poland,** *renata.dwornicka@mech.pk.edu.pl*
[2]Khmelnytsky National University, **Ukraine**

Abstract: The bootstrap method is a well-known method to gather a full probability distribution from the dataset of a small sample. The simple bootstrap i.e. resampling from the raw dataset often leads to a significant irregularities in a shape of resulting empirical distribution due to the discontinuity of a support. The remedy for these irregularities is the smoothed bootstrap: a small random shift of source points before each resampling. This shift is controlled by specifically selected distributions. The key issue is such parameter settings of these distributions to achieve the desired characteristics of the empirical distribution. This paper describes an example of this procedure.
Keywords: smoothed bootstrap, statistics, design of experiments, numerical simulation

1. INTRODUCTION

The typical data analysis performed for a dataset obtained from an experiment uses well-known statistical formulas and expressions associated implicitly with many assumptions. These assumptions, in contrast, are not well-known. The most often met ones are: not very small sample size and a gaussian distribution. The latter allows you to provide asymptotic relationships that are convenient to calculate expected values and intervals/regions of confidence. The appropriately large sample size enables a reliable assessment of the previously assumed normality of the distribution.

This undoubtedly convenient scheme, however, is often disturbed by the imposed limitation of the sample size. The small size of an experimental sample is usually caused by resource limitations, the most often financial ones. In such situation, two solutions are possible:

a) a use of traditional analytical expressions with a belief that inevitable errors are acceptably small,
b) a use of the Monte-Carlo re-sampling scheme based on raw data i.e. the data-driven bootstrap method (Efron, 1979; Shao, 1995).

The first approach is often used, because it is schematic, politically safe and can always be justified by the argument that "we have always done it in this manner". In

the case of more advanced users, the bootstrap method is used more and more often. However, these users observed a significant irregularities in shapes of resampled distributions, which are caused by discontinuities in supports of distribution functions. These discontinuities are related to the fact that the sampled distribution is created by assigning mass to each point of the original data set. This leads to a pair of a density function (Fig. 1), that is the sum of the Dirac deltas, and a cumulative distribution (Fig. 2), that is a compound of several Heaviside's functions (unit step functions).

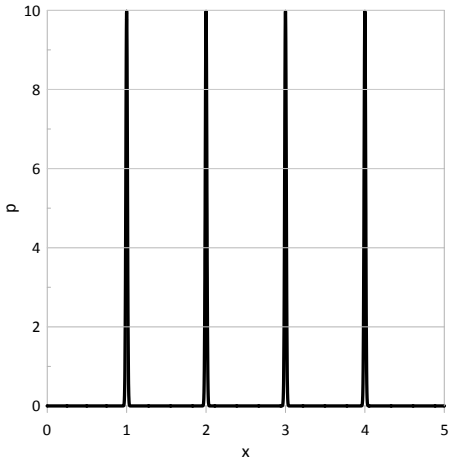

Fig. 1. An example of the simple bootstrap density function

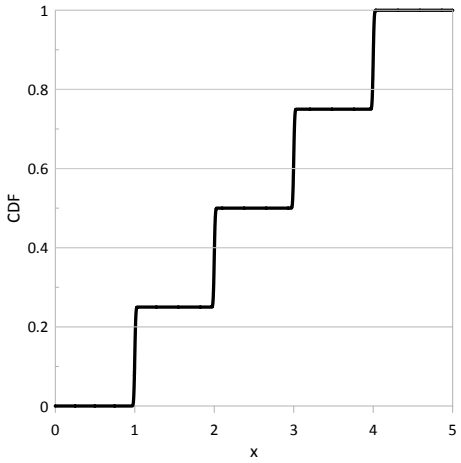

Fig. 2. An example of the simple bootstrap cumulative distribution function

The idea of the smoothed bootstrap is to replace the Dirac deltas with distributions (Fig. 3) that have limited local supports and allow required smoothness of functions (Fig. 4). It means that resampling is not limited strictly to the original dataset, but final points, based on sources (and their neighborhoods, in smoothed variant), randomly taken from the original dataset may slightly differ from original ones.

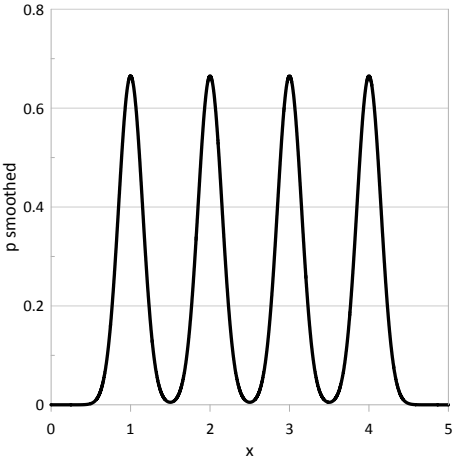

Fig. 3. An example of the smoothed bootstrap density function

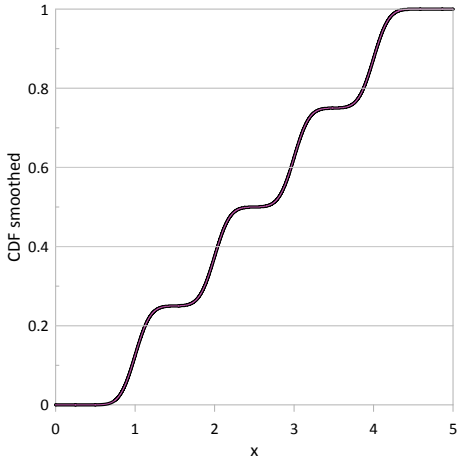

Fig. 4. An example of the smoothed bootstrap cumulative distribution function

An open question is how to choose the type of distribution and its parameters to achieve the desired characteristics of the target resampling.

In the further part of the article, the authors present the selection of parameters of deviation distributions for an exemplary dataset.

2. DATA AND METHODS

2.1. Smoothed bootstrap

The basis of a consideration is the measurement of the quantitative variable X, the distribution of which is unknown. Formally, it may be described as sampling of the random, one-dimensional and quantitative, variable X from an unknown F distribution. The gathered values x_i are stored in the dataset D of size n.

In the simple bootstrap, the empirical distribution Fb, related to the dataset and being the source for resampling, is created by putting equal mass 1/n to the each value of the dataset i.e. the cumulative distribution function has the following formula:

$$Fb(x) = \sum_{i=1}^{n} \frac{H(x - x_i)}{n} \tag{1}$$

where H is the unit step function i.e. Heaviside's function. In the smoothed bootstrap, the Heaviside's functions, shown in Eq. 1, are replaced by any distributions f_i of class C^1 with density concentrated around the dataset original values, in their nearest neighborhood:

$$Fsb(x) = \sum_{i=1}^{n} \frac{f_i(x - x_i)}{n} . \tag{2}$$

Efron compared the performances of the smoothed bootstrap resampling from Fsb and the non-smoothed bootstrap resampling from Fb using the sample correlation coefficient as an example statistics (Efron, 1982), however, he used arbitrary settings without deeper analysis.

In this paper, authors use the normal distribution with individually adjusted standard deviation as the smoothing functions i.e. the cumulative distribution function has the following formula:

$$Fsb(x) = \frac{1}{2} + \frac{\sum_{i=1}^{n} erf\left(\frac{x - x_i}{s_i\sqrt{2}}\right)}{2n} . \tag{3}$$

where erf is well-known Gauss error function.

In all bootstrap simulations, the limit of 10000 iterations was set.

2.2. Comparison criteria

The basis of a consideration is the measurement of the quantitative, one-dimensional variable X. The bootstrapped statistics is mean calculated from triplicates and bounds of its 95% confidence interval. Additionally, the maximum gap between neighboring resampled values is identified. Based on the median and range of variation, a modified coefficient of variation was calculated.

2.3. Source dataset

The source data were taken from the biotechnological investigation conducted on enhanced accumulation of harpagide and its derivatives in *in vitro* cultures *of Melittis melissophyllum* plant (Skrzypczak-Pietraszek et al., 2018). The investigation was done as a designed experiment with four controlled factors: harvesting time and possible supplementation of three different chemical additives. The volumetric density of harpagide in biomass was an observed outcome. The measurements were done in triplicate for each treatment.

This article uses data from a publicly available dataset attached to the article (Skrzypczak-Pietraszek et al., 2018a). For further analysis, the record (26.9, 23.5, 21.6) was used, with gaps between values (3.4, 1.9), respectively.

3. RESULTS

3.1. Classic mean and 95% confidence interval estimation

The classic, well-known formulas were used to estimate mean and 95% confidence interval bounds. The result was mean 24.0 and 95% confidence interval (17.33, 30.67).

3.2. Simple bootstrap

The record (26.9, 23.5, 21.6) was a base dataset for the simple bootstrap. Theoretically, this dataset may lead to 3^3 i.e. 27 different combinations, but only 10 different mean values are available, due to the neutrality of permutation to the statistics value. This values are as following: 21.60, 22.23, 22.87, 23.37, 23.50, 24.00, 24.63, 25.13, 25.77, 26.90. The maximum gap 1.13 is between 25.77 and 26.90. As a result, the histogram is either coarse (Fig. 5) or full of holes (Fig. 6). It is not possible to reliably determine the bounds of the confidence interval.

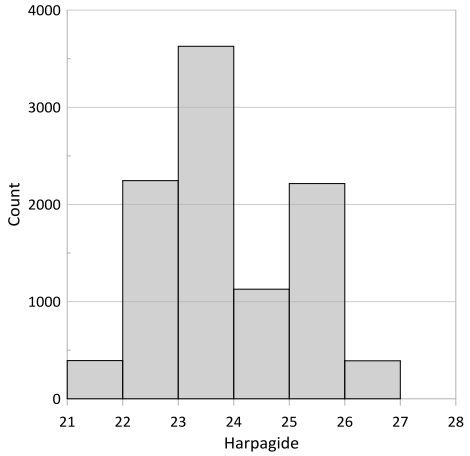

Fig. 5. The coarse histogram with bin size equal 1

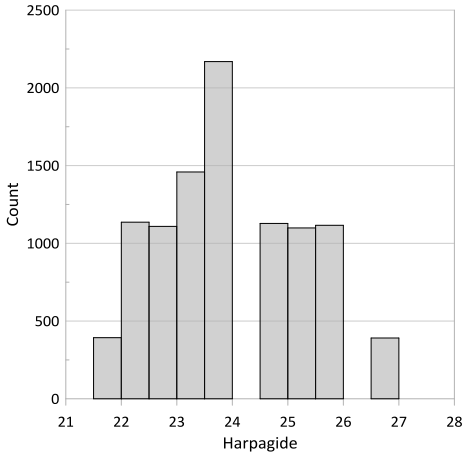

Fig. 6. The histogram with bin size 0.5

3.3. Smoothed bootstrap

The record (26.9, 23.5, 21.6) was a base dataset for the smoothed bootstrap. All standard deviations s_i (see Eq.3) were set at the same values. Seven variants were simulated at s_i values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7. The limit of bootstrap iterations was set at 10000. The mean and bounds of its 95% confidence intervals were calculated. Additionally, the maximum observed gap between resampled values was identified.

The results are presented in Table 1. The histograms for s = 0.1 and s = 0.4 are presented in Fig.7 and Fig.8. Variation of the confidence intervals bounds are shown in Fig.9 and Fig.10. Variation of the estimated mean, confidence interval bounds and the maximum gap is presented in Table 2.

Table 1
Results gathered from numerical simulations of the smoothed bootstrap (s – standard deviation for smoothing, mean – mean of resampled triplicates, ±95%CI – bound of confidence interval, max gap – maximum observed distance between neighboring resampled values)

s	mean	-95%CI	+95%CI	max gap
0.1	21.63	24.01	26.88	0.81
0.2	21.66	23.99	26.86	0.36
0.3	21.69	24.01	26.80	0.09
0.4	21.70	24.03	26.84	0.06
0.5	21.69	24.00	26.73	0.14
0.6	21.65	24.00	26.71	0.13
0.7	21.60	24.00	26.74	0.11

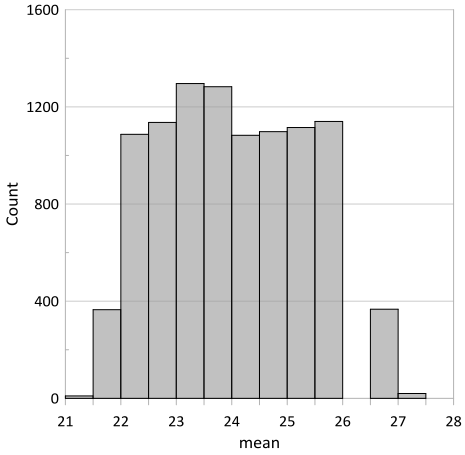

Fig. 7. The histogram for s = 0.1

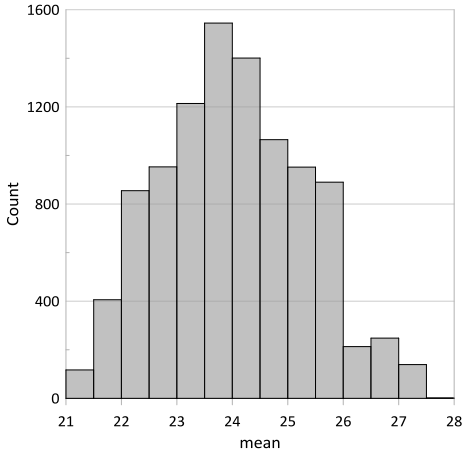

Fig. 8. The histogram for s = 0.4

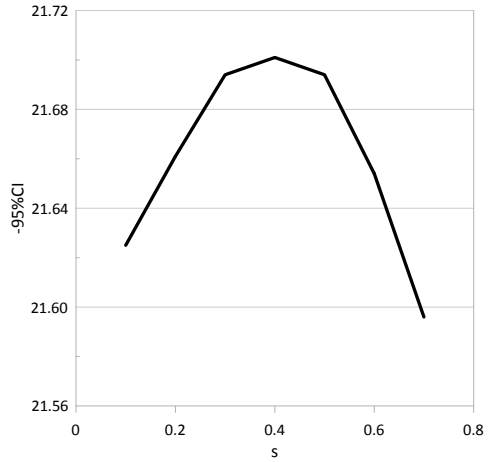

Fig. 9. The -95% bound for confidence interval

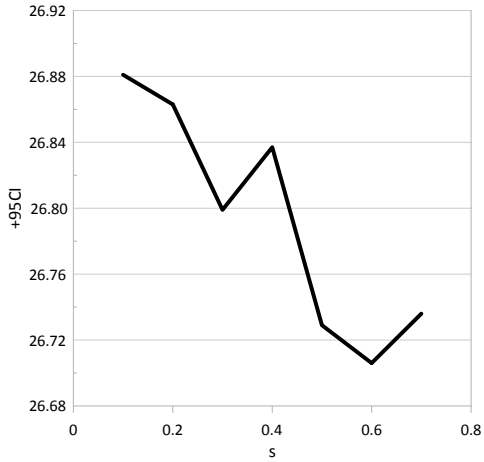

Fig. 10. The +95% bound for confidence interval

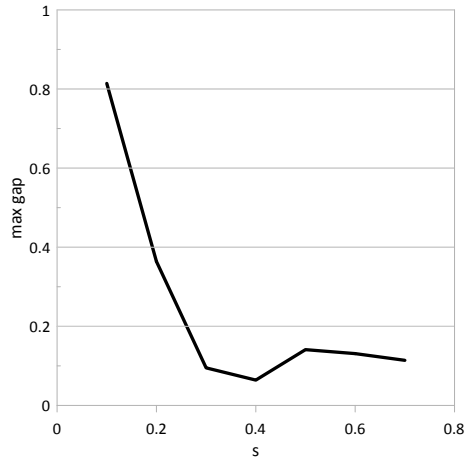

Fig. 11. Maximum gap in smooth bootstrap

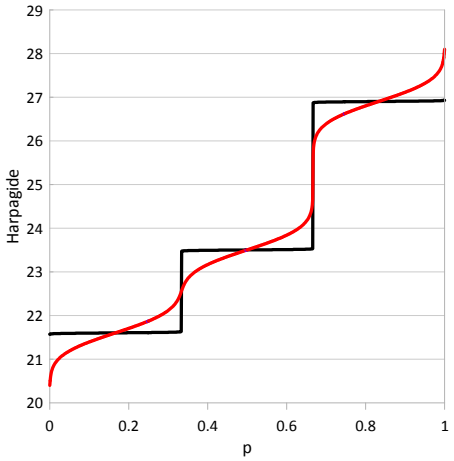

Fig. 12. Inverse cumulative functions used in the simple bootstrap (step function, black) and the smoothed bootstrap with s = 0.4 (curved, red)

Table 2
Variation of the mean statistics, confidence interval bounds and maximum gap in the smoothed bootstrap simulations (variability of the standard deviation s_i from 0.1 to 0.7)

	mean	**-95%CI**	**+95%CI**	**max gap**
minimum	23.99	21.60	26.71	0.06
maximum	24.03	21.70	26.88	0.81
range	24.00	21.66	26.80	0.13
median	0.04	0.11	0.18	0.75
range/median	0.15%	0.48%	0.65%	572%

4. DISCUSSION

As can be seen from the results obtained, the value of the standard deviation, used to the smooth bootstrap simulations, did not significantly affect the estimated mean and bounds of the confidence interval. The use of the method itself was significant, not the specific value of the standard deviation parameter. It means that the transition of a cumulative distribution function from the C^0 class to the C^1 class is important.
However, the different behavior was in the case of the shape of the distribution. The value of the standard deviation significantly changed the shape of the distribution histogram, so the parameter value selection may be important for the value generator for simulation purposes.

5. CONCLUSION

After the carried out investigations, the following conclusions can be drawn:

- the smooth bootstrap significantly improves estimation results compared to the simple bootstrap,
- the value of the standard deviation parameter responsible for smoothening the distribution is not important; it is important to change the function class from C^0 to class C^1,
- further work should check whether the shape of the additional distribution matters – the normal distribution is computationally expensive.

Smooth bootstrap, as the example of data-driven analysis i.e. a statistical analysis without additional assumption about distribution, may be useful in other areas like e.g. industrial management (Maszke et al., 2018), materials science (Pietraszek and Gadek-Moszczak, 2013; Ulewicz and Novy, 2016, Ulewicz et al., 2016; Dudek et al., 2017; Pietraszek et al., 2017; Jambor et al., 2018; Radek et al., 2018), especially supported by an image analysis (Gadek-Moszczak, 2017; Gadek-Moszczak and Matusiewicz, 2017), even in biomaterials (Gadek-Moszczak et al., 2015), hydraulic machines design (Pobedza and Sobczyk, 2013a; Pobedza and Sobczyk, 2013b; Guzowski and Sobczyk, 2014; Walczak and Sobczyk, 2014) and corrosion protection (Klimecka-Tatar, 2016). It may be also very useful in power industry (Dwornicka, 2014), chemical industry (Ulewicz and Radzyminska-Lenarcik, 2014; Gnatowski et al., 2018) or pharmaceutical and biotechnology industry (Skrzypczak-Pietraszek, 2016; Skrzypczak-Pietraszek et al., 2018b), where phytochemistry investigation (Skrzypczak-Pietraszek and Pietraszek, 2009; Skrzypczak-Pietraszek et al., 2017) are conducted with a huge random noise from individual differences in plant reactions.

REFERENCES

Dudek, A., Lisiecka, B., Ulewicz, R., 2017. *The effect of alloying method on the structure and properties of sintered stainless steel*. Arch. Metall. Mater., 62, 281-287.

Dwornicka, R., 2014. *The impact of the power plant unit start-up scheme on the pollution load*. Adv. Mater. Res.-Switz. 874, 63-69.

Dwornicka, R., Radek, N., Krawczyk, M., Osocha, P., Pobedza, J., 2017, *The Laser Textured Surfaces of the Silicon Carbide Analyzed with the Bootstrapped Tribology Model*. Metal 2017: 26th Int. Conf. Metallurgy and Materials, Ostrava, Tanger, 1252-1257.

Efron, B., 1979. *Bootstrap Methods: Another Look at the Jackknife*. Ann. Statist., 7(1), 1-26.

Efron, B., 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

Gadek-Moszczak, A., 2017. *History of stereology*. Image Anal. Stereol., 36, 151-152.

Gadek-Moszczak, A., Pietraszek, J., Jasiewicz, B., Sikorska, S., Wojnar, L., 2015. *The bootstrap approach to the comparison of two methods applied to the evaluation of the growth index in the analysis of the digital x-ray image of a bone regenerate*. New Trends in Comput. Collective Intelligence, 572, 127-136.

Gadek-Moszczak, A., Matusiewicz, P., 2017. *Polish Stereology – a Historical Review*. Image Anal. Stereol., 36, 207-221.

Gnatowski, A., Ulewicz, M., Chyra, M., 2018. *Analysis of Changes in Thermomechanical Properties and Structure of Polyamide Modified with Fly Ash from Biomass Combustion*. Journal of Polymers and the Environment, 26, 647-654.

Guzowski A, Sobczyk A., 2014. *Reconstruction of Hydrostatic Drive and Control System Dedicated for Small Mobile Platform*. ASME – Fluid Power Systems Technology, 8th FPNI Ph.D Symposium on Fluid Power, art.V001T05A012.

Jambor, M., Ulewicz, R., Novy, F., Bokuvka, O., Trsko, L., Mician, M., Harmaniak, D., 2018. *Evolution of Microstructure in the Heat Affected Zone of S960MC GMAW Weld*. Mat. Res. Proc., 5, 78-83.

Klimecka-Tatar, D., 2016. *Electrochemical characteristics of titanium for dental implants in case of the electroless surface modification*. Arch. Metall. Mater., 61, 923-926.

Maszke, A., Dwornicka, R., Ulewicz, R., 2018. *Problems in the implementation of the lean concept at a steel works – Case study*. MATEC Web Conf., 183, art. 01014.

Pietraszek, J., Gadek-Moszczak, A., 2013. *The smooth bootstrap approach to the distribution of a shape in the ferritic stainless steel AISI 434L powders*. Solid State Phenom., 197, 162-167.

Pietraszek, J., Dwornicka, R., Krawczyk, M., Kołomycki, M., 2017. *The non-parametric approach to the quantification of the uncertainty in the design of experiments modelling*. UNCECOMP 2017, NTU of Athens, 598-604.

Pobedza, J., Sobczyk, A., 2013a. *Modern Coating Used in High Pressure Water Hydraulic Components*. Key Engineering Materials, 542, 143-155.

Pobedza, J., Sobczyk, A., 2013b. *Properties of High Pressure Water Hydraulic Components with Modern Coatings*. Adv. Mat. Res.-Switz., 849, 100-107.

Radek, N., Szczotok, A., Gadek-Moszczak, A., Dwornicka, R., Broncek, J., Pietraszek, J., 2018. *The impact of laser processing parameters on the properties of electro-spark deposited coatings*. Arch. Metall. Mater., 63, 809-816.

Shao, J., Tu, D., 1995. *The Jackknife and Bootstrap*. Springer, New York.

Skrzypczak-Pietraszek, E., 2016. *High production of flavonoids and phenolic acids for pharmaceutical purposes in Vitex agnus castus L. shoot culture*. New Biotechnol., 33, S155-S155.

Skrzypczak-Pietraszek, E., Pietraszek, J., 2009. *Phenolic acids in in vitro cultures of Exacum affine Balf. f*. Acta Biol. Cracov. Bot., 51, 62-62.

Skrzypczak-Pietraszek, E., Kwiecien, I., Goldyn, A., Pietraszek, J., 2017. *HPLC-DAD analysis of arbutin produced from hydroquinone in a biotransformation process in Origanum majorana L. shoot culture*. Phytochem. Lett., 20, 443-448.

Skrzypczak-Pietraszek, E., Reiss, K., Zmudzki, P., Pietraszek, J., 2018a. *Enhanced accumulation of harpagide and 8-O-acetyl-harpagide in Melittis melissophyllum L. agitated shoot cultures analyzed by UPLC-MS/MS*. PLoS ONE 13, e0202556.

Skrzypczak-Pietraszek, E., Piska, K., Pietraszek, J., 2018b. *Enhanced production of the pharmaceutically important polyphenolic compounds in Vitex agnus castus L. shoot cultures by precursor feeding strategy*. Eng. Life Sci., 18, 287-297.

Walczak, P., Sobczyk A., 2014. *Simulation of water hydraulic control system of francis turbine*. ASME – Fluid Power Systems Technology, 8th FPNI Ph.D Symposium on Fluid Power, art. V001T04A001.

Ulewicz, M., Radzyminska-Lenarcik, E., 2014. *Application of Polymer and Supported Membranes with 1-Decyl-4-Methylimidazole for Pertraction of Transition Metal Ions*. Separation Science and Technology, 49, 1713-1721.

Ulewicz, R., Jelonek, D., Mazur, M., 2016. *Implementation of logic flow in planning and production control*. Management and Production Engineering Review, 7, 89-94.

Ulewicz, R., Novy, F. R., 2016. *The influence of the surface condition on the fatigue properties of structural steel.* Journal of the Balkan Tribological Association, 22, 1147-1155.