Application of the classification trees for separation of groups of persons threatened by the long-term unemployment

Beata Bieszk-Stolorz[1], Krzysztof Dmytrów[2]

Abstract

In order to conduct the effective labour market policy, identification of groups of persons threatened by the long-term unemployment is essential. The goal of the research is separation of these groups of persons and designation which long-term unemployed persons are more often de-registered for work and which ones resign from mediation of the labour office. Classification trees were used for separation of these groups. Characteristics that divided the unemployed persons were: gender, age, education, seniority and the number of registrations in the office. In the first stage of the research the groups of persons at greater risk of the long-term unemployment in comparison to other groups were separated. The de-registrations to work and removals for reasons attributable to the unemployed person were the most common. These two forms of de-registration among the long-term unemployed persons were analysed in the second stage. The individual data for persons de-registered from the Poviat Labour Office in Szczecin in years 2013–2017 were used. About 1/3 of all the analysed persons were sampled to the training dataset, and obtained results were applied for the test dataset. The quality of the classification was evaluated by means of accuracy, sensitivity and precision.

Keywords: *long-term unemployment, de-registration forms, classification trees, evaluation of classification*
JEL Classification: *C38, J64*

1. Introduction

In order to lead the efficient labour market policy, the identification of persons threatened by the long-term unemployment is essential. The share of these people in the total registered unemployment in 2017 in Poland and the Zachodniopomorskie voivodeship was almost 55% and in Szczecin – less than 47%. They are the persons looking for employment actively, but in vain. Their competences, professional education or age make it difficult for them to find a job offer and they are covered by various activation programmes. The effectiveness of these programmes is territorially diversified (Bieszk-Stolorz and Dmytrów, 2018b). The registered unemployment rate in Poland and the long-term unemployment rate has decreased since 2013 (fig. 1). Processes on the Polish labour market are similar to those on the Slovak and Hungarian ones (Hadaś-Dyduch et al., 2016). In the Visegrad Group countries during the financial crisis 2007–2009 especially young persons were threatened by the long-term unemployment and it had great impact on their future professional life (Pavelka, 2016). The unemployment is affected by the social policy of a country, carried out, inter alia, by the labour offices.

[1] Corresponding author: University of Szczecin/Institute of Econometrics and Statistics, Department of Operations Research and Applied Mathematics in Economics, 64 Mickiewicza St., 71–101 Szczecin, beata.bieszk-stolorz@usz.edu.pl.

[2] University of Szczecin/Institute of Econometrics and Statistics, Department of Operations Research and Applied Mathematics in Economics, 64 Mickiewicza St. 71–101 Szczecin, krzysztof.dmytrow@usz.edu.pl.

Activation programmes directed to the threatened groups of persons have a positive impact on the labour market. They should also be focused on the social integration and formation of the long-term unemployment persons' social skills (Fervers, 2018). However, extensive system of unemployment benefits may lead to extension of the unemployment duration (Bieszk-Stolorz and Markowicz, 2015).

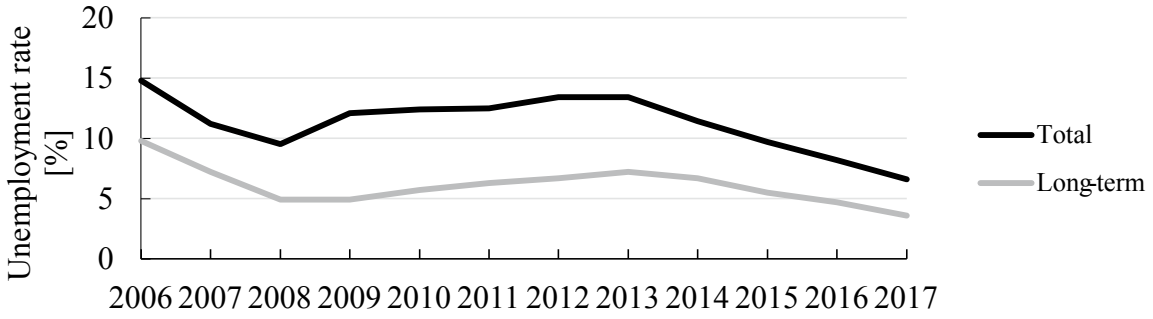

Fig. 1. Registered total and long-term unemployment rates in Poland in years 2006–2017

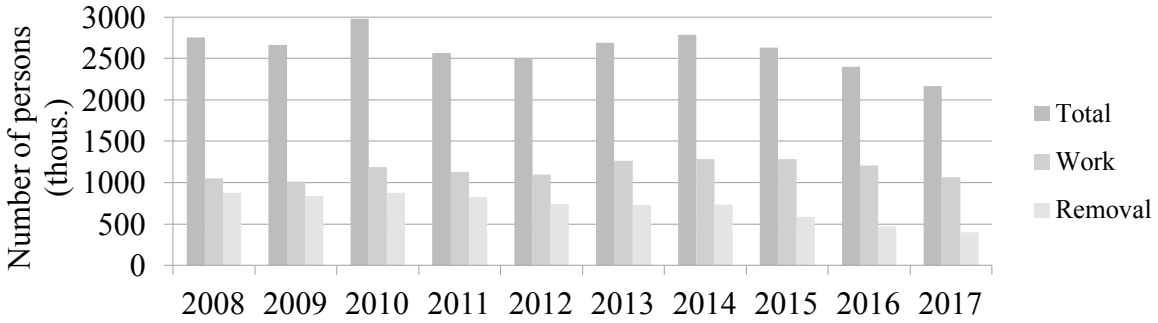

Fig. 2. Number of de-registrations from the labour offices in Poland in years 2008–2017

Not all unemployed are interested in co-operation with the office. Work is the most frequent cause of de-registration from the office. Removal due to lack of readiness for work is the second one. In the years 2006–2017 the lack of readiness for work constituted from 19% to 32% of all de-registrations in Poland (fig. 2, on the basis of data from Yearbooks of Labour Statistics). Some of the unemployed persons do not inform the office about finding employment. Hitherto performed analyses of the labour market in Szczecin indicated that gender did not significantly affect the probability of finding employment, but men were more intensively removed from the register. On the contrary, education and age of the unemployed persons strongly afffected both de-registrations to work or removals (Bieszk-Stolorz and Dmytrów, 2018a). The European Union data from the research on the income and living conditions (EU-SILC) for 24 European countries (2005–2012) indicate that the risk of the long-term unemployment increases particularly for persons with low qualifications and professions, single parents, immigrants and the disabled persons. Women, older persons and permanently employed are less affected by the short-term unemployment, but more by the long-term one (Heidenreich, 2015).

The goal of the research is separation (by means of the classification trees) of groups of persons threatened by the long-term unemployment and designation which long-term unemployed persons are more often de-registered to work and which ones – removed from the register.

2. Data used in the research

The anonymous data about the de-registered unemployed persons from the Poviat Labour Office in Szczecin in 2013–2017 was used in the research. It contains information about: gender, age, education, seniority, number of subsequent registrations, unemployment duration and cause of de-registration. The long-term unemployed persons were separated. They consisted of 22%, 22%, 20%, 17% and 14% of all registered unemployed in subsequent years, respectively (table 1). Data from the labour office contains dozens of causes of de-registrations. Among them, three groups were separated: work, removal and other causes. Other causes were far less numerous and, as earlier researches indicate, each of them separately had only marginal impact on the probability of de-registration (Bieszk-Stolorz, 2017).

Table 1. Number of unemployed persons de-registered from the Poviat Labour Office in Szczecin in 2013–2017

Year	Unemployed total (long-term)	Including	
		Work	Removal
2013	23971 (5336)	11058 (2160)	10692 (2604)
2014	24723 (5518)	11119 (2203)	11392 (2652)
2015	25881 (5266)	11201 (1960)	12672 (2695)
2016	23734 (4091)	10181 (1342)	11718 (2193)
2017	19931 (2706)	8031 (810)	10157 (1320)

In years 2013–2017 on the average 44% of persons registered in the Poviat Labour Office took job and 48% resigned from the co-operation with the labour office. On the contrary, among the long-term unemployed persons this proportion was far more disadvantageous: only 36% of them took job and even 50% of them were removed from the register.

3. Methodology of the research

The research was conducted in two stages. In the first stage homogeneous groups of the unemployed persons particularly threatened by the long-term unemployment (or those in which the ratio of the long-term unemployed was higher than for all population) were separated. In the second stage only the long-term unemployed persons were analysed. Only the ones who took job or were removed from the register were selected. These persons constituted the vast majority of the long-term unemployed persons. Their share decreased from 89% in 2013 to 79% in 2017 (table 1). Later on, the homogeneous groups of the long-term unemployed

persons de-registered to work and removed more frequently than the whole population were separated.

The classification was done by means of the classification trees. Among available algorithms, the C&RT (Łapczyński, 2010) was used. Heterogeneity of obtained subsets was analysed by means of the χ^2 statistics (Gatnar, 2001). When constructing the tree, the two aspects should be considered (Capelli and Zhang, 2007): data division, or tree growing and pruning the tree. The more the data is divided, the more homogeneous groups will be obtained, which may result in *overfitting*, which decreases the predictive capacities of the model. The *tree pruning* methods are applied in order to prevent its extensive growth. The tree growing can be stopped if the decrease of heterogeneity of obtained subsets is smaller than certain, assumed value, if we assume certain, minimal group size, at which the division may occur or the maximum tree depth (Mudunuru, 2016). In the research it was assumed that the division will not occur if the size of the subset is less than 100.

In the first stage of the research the unemployment duration (t) until the moment of de-registration was the dependent variable. It was encoded as follows: if $t < 12$ months then the variable took the value 0 and 1 otherwise (in such case it was assumed that analysed person was long-time unemployed). Independent variables in both stages were encoded as follows:

- gender: woman – 1, man – 0;
- age (years): 18–24 – 1, 25–34 – 2; 35–44 – 3, 45–54 – 4, 55–59 – 5, 60–64 – 6;
- education: at most lower secondary – 1, basic vocational – 2, general secondary – 3, vocational secondary – 4, higher – 5;
- seniority: without seniority – 0, with seniority – 1;
- number of subsequent registrations: first– 0, subsequent – 1.

In the second stage the cause of de-registration was the dependent variable. It was encoded as follows: work – 1, removal – 0.

In both stages the whole population of de-registered unemployed persons was divided into the training (about 1/3 of the whole population) and test datasets. In every year the classification trees were constructed for the training set and on the basis of groups obtained for the training dataset, the groups of the unemployed persons particularly threatened by the long-term unemployment (in the first stage) and groups of persons de-registered to work and removed (in the second stage) were separated in the test dataset. The quality of classification was evaluated for the test dataset. It was based on the confusion matrix (table 2) (Fawcett, 2006).

Table 2. Confusion matrix

condition	**actual negative**	**actual positive**
predicted negative	true negative (*TN*)	false negative (*FN*)
predicted positive	false positive (*FP*)	true positive (*TP*)

On the basis of the confusion matrix the following measures were calculated:

1. accuracy:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2. sensitivity:

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

3. precision:

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

4. the F_1 measure:

$$F_1 = \frac{ZTP}{ZTP + FP + FN} \tag{4}$$

Measure (1) is the basic measure of accuracy of classification. From the point of view of the research goal (selection of groups of persons particularly threatened by the long-term unemployment in the first stage and groups of long-term unemployed persons de-registered to work and removed in the second stage) measures (2) and (3) are more important. If we increase the sensitivity, the precision will be adversely affected and vice versa. The F_1 score, being the harmonic mean of sensitivity and precision, measures balance between them. Assessing the quality of classification, more attention was devoted to sensitivity – when classifying too many objects as positive, we will certainly be able to cover more of the real positives, however it will also bring more false positive objects. The reason for attaching more importance to the sensitivity is the fact that if the labour office wants to counteract the long-term unemployment, when the precision is relatively low, substantial part of funds will reach persons who do not need it, but, on the contrary, larger part of persons who actually need help will receive it.

4. Results of the analysis

Firstly, the groups of persons particularly threatened by long-term unemployment were separated (fig. 3). The quality of classification for the test dataset is presented in table 3.

Age, education and gender were the variables that divided the population of the de-registered unemployed persons with respect to the unemployment duration. In subsequent years persons threatened by long-term unemployment created various groups and it was difficult to notice a greater regularity. More frequently there were women, persons with at most secondary education and in older age groups. These unemployed persons should be particularly covered by the programmes of professional activation.

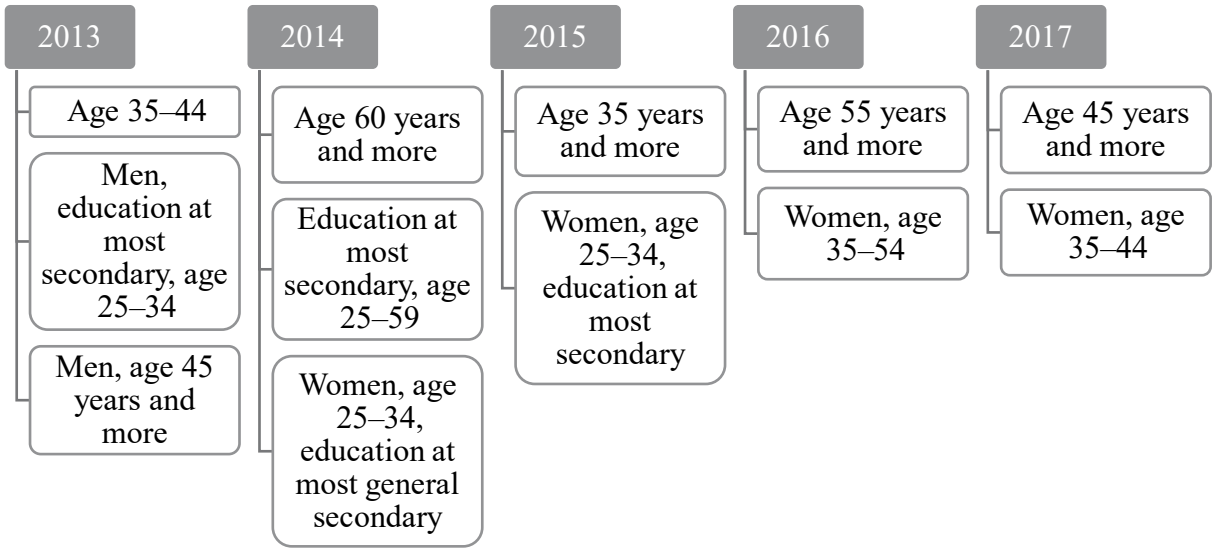

Fig. 3. Groups of persons particularly threatened by long-term unemployment in 2013–2017

Table 3. Quality of classification of persons particularly threatened by the long-term unemployment

Coefficient	2013	2014	2015	2016	2017
ACC [%]	59.88	63.23	61.55	67.20	55.36
TPR [%]	48.03	49.90	43.77	57.93	80.04
PPV [%]	27.66	62.51	54.46	27.38	20.66
F_1 [%]	35.11	55.50	48.53	37.18	32.85

The quality of classification was quite low – analysed unemployed persons constituted quite heterogeneous population. The accuracy was between 55.4% in 2017 and 67% in 2016. The most important from the point of view of social policy sensitivity was generally weak – only in two last years it was better (80% in 2017). Precision was generally at the very low level (except for the years 2014 and 2015). It means that most persons that need help will receive it. High sensitivity means that in 2017 80% of long-term unemployed persons who actually need help will receive it. The second stage of the research was classification of the long-term unemployed persons as taking job and removed due to reasons attributable to them. Selected groups of persons in subsequent years are presented on figures 4 and 5. Quality of classification for the test dataset is presented in table 4.

In case of the long-term unemployed persons de-registered to work and removed from the register, the following features: education, seniority, number of registrations (in 2013) and gender and age (both in 2015) were the ones that divided the population. Generally, long-term unemployed persons that were de-registered to work more frequently were the persons with at least secondary education and with seniority. In 2015 there were women. On the contrary, long-term unemployed persons removed from the register were the persons with lower level

of education without seniority. In 2015, as in the case of de-registration to work, gender was also the significant feature that divided the population. On the contrary to the de-registration to work, in 2015 men were more frequently removed than women.

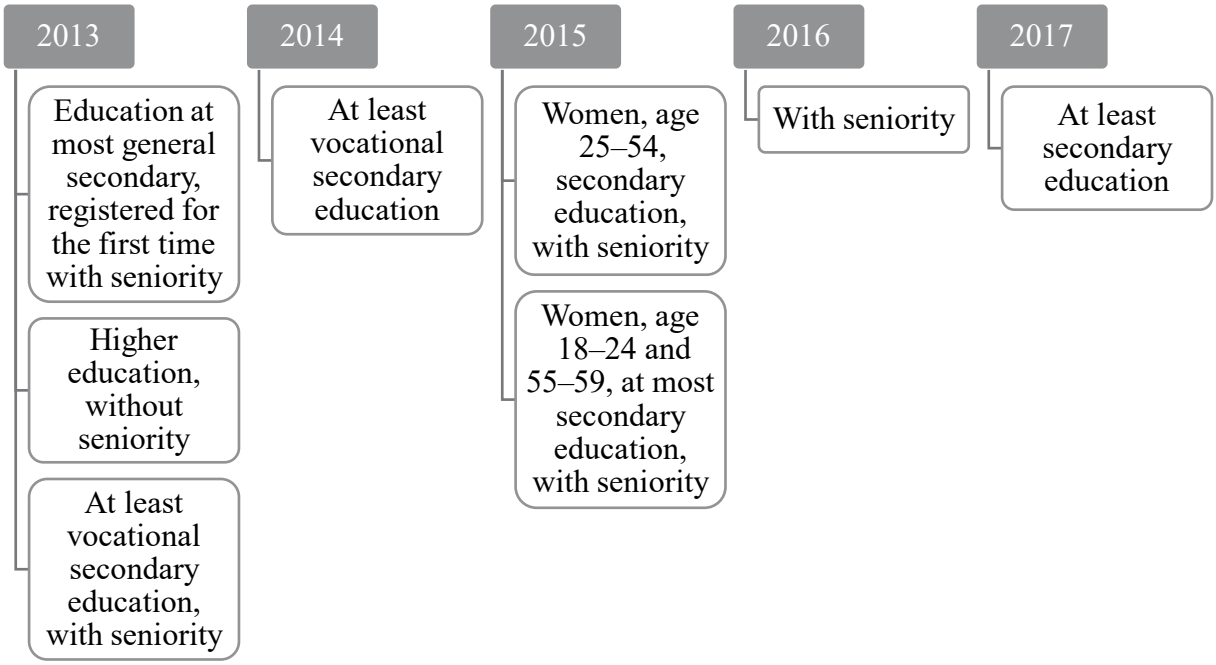

Fig. 4. Groups of long-term unemployed persons de-registered to work in 2013–2017

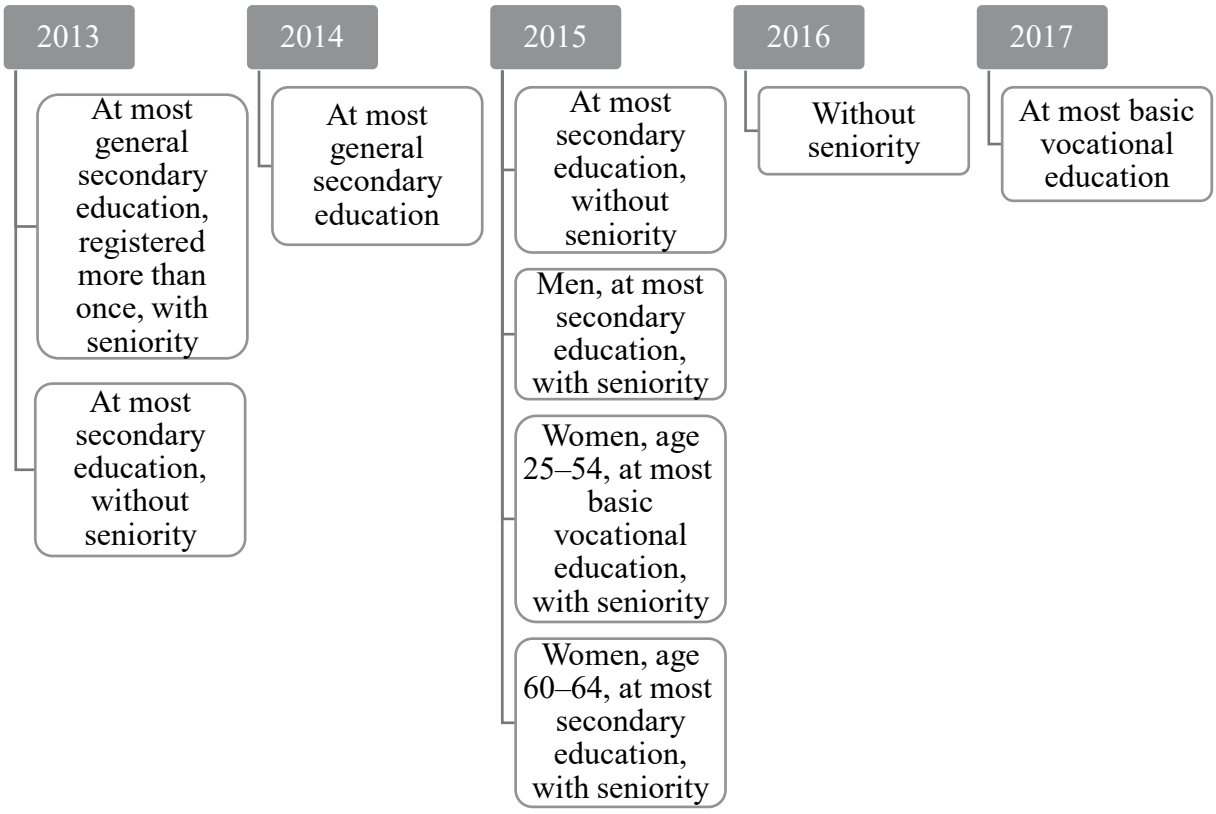

Fig. 5. Groups of long-term unemployed persons removed from the register in 2013–2017

Table 4. Quality of classification of long-term unemployed persons de-registered to work and removed from the register

Coefficient	2013	2014	2015	2016	2017
ACC [%]	63.42	63.23	61.55	52.33	56.95
TPR (work)/*PPV* (removal) [%]	49.58	49.90	43.77	79.69	69.50
PPV (work)/*TPR* (removal) [%]	61.90	62.51	54.46	42.93	45.80
F_1 [%]	55.06	55.50	48.53	55.80	55.21

Similarly, as in the first stage, also in case of the long-term unemployed persons de-registered to work or removed from the register, the population was highly heterogeneous. The accuracy of the classification was between 52% in 2016 and 63.4% in 2013. In years 2013–2015 the precision (for de-registration to work, in case of removal it was the sensitivity) was at the higher level and in the remaining years the sensitivity (for de-registration to work, in case of removal it was the precision) was higher. The F_1 measure, being the harmonic mean of those two, was between 48.5% in 2015 and 55.8% in 2016.

Conclusions

An attempt of classification of persons threatened by long-term unemployment and the long-term unemployed persons de-registered to work and removed from the register due to reasons attributable to them in the Poviat Labour Office in Szczecin in years 2013–2017 was made in the article. Conducted analyses indicated that the analysed population of the unemployed persons was highly heterogeneous, which made the quality of classification quite weak – only in last two years it improved, especially the sensitivity. In the first stage of the research, persons particularly threatened by long-term unemployment were classified with respect to age, education and gender. Persons with lower levels of education were more threatened by long-term unemployment. Also, more often there were women. Despite the insufficient quality of the classification, it was possible to separate the groups of persons to which the help within the counteracting the long-term unemployment could be directed. In the second stage of the research, also the values of the features characterising the long-term unemployed persons de-registered to work and removed from the register were specified (also despite the insufficient quality of the classification). In this case the persons were divided by the following variables: education, seniority and gender (in 2015). Generally persons with better education were de-registered to work, while persons with lower education were more frequently removed from the register. Identification of groups of persons threatened by long-term unemployment on the local labour market is important due to creation of effective labour market policy.

References

Bieszk-Stolorz, B., & Markowicz, I. (2015). Influence of Unemployment Benefit on Duration of Registered Unemployment Spells. *Equilibrium. Quarterly Journal of Economics and Economic Policy,* 10(3), 167–183.

Bieszk-Stolorz, B. (2017). Cumulative Incidence Function in Studies on the Duration of the Unemployment Exit Process, *Folia Oeconomica Stetinensia* 17(1), 138–150.

Bieszk-Stolorz, B., & Dmytrów, K. (2018a). Application of the Survival Trees for Estimation of the Propensity to Accepting a Job and Resignation from the Labour Office Mediation by the Long-Term Unemployed People. In: Nermend, K., Łatuszyńska, M. (eds.), *Problems, Methods and Tools in Experimental and Behavioral Economics. CMEE 2017*. Springer Proceedings in Business and Economics. Springer, Cham, 141–154.

Bieszk-Stolorz, B., & Dmytrów, K. (2018b). Efektywność form aktywizacji zawodowej w przekroju wojewódzkim. *Wiadomości Statystyczne,* 12(691), 57–74.

Cappelli, C., & Zhang, H. (2007). Survival Trees. In: Härdle, W.,·Mori, Y., Vieu, P. (eds.), *Statistical Methods for Biostatistics and Related Fields*. Springer-Verlag, Berlin, 167–179.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 27, 861–874.

Fervers, L. (2018). Can public employment schemes break the negative spiral of long-term unemployment, social exclusion and loss of skills? Evidence from Germany. *Journal of Economic Psychology,* 67, 18–33.

Gatnar, E. (2001). *Nieparametryczna metoda dyskryminacji regresji*. Wydawnictwo Naukowe PWN, Warszawa.

Hadaś-Dyduch, M., Pietrzak, M.B., & Balcerzak, A.P. (2016). Wavelet Analysis of Unemployment Rate in Visegrad Countries. *Globalization and Its Socio-Economic Consequences, 16th International Scientific Conference, Conference Proceedings, University of Zilina, The Faculty of Operation and Economics of Transport and Communication, Department of Economics, 5th–6th October 2016*, 595–602.

Heidenreich, M. (2015). The end of the honeymoon: The increasing differentiation of (long-term) unemployment risks in Europe. *Journal of European Social Policy,* 25(4), 393–413.

Łapczyński, M. (2011). *Drzewa klasyfikacyjne i regresyjne w badaniach marketingowych*. UEK Kraków.

Mudunuru, V.R. (2016). Modeling and Survival Analysis of Breast Cancer: A Statistical, Artificial Neural Network, and Decision Tree Approach. *Graduate Theses and Dissertations*. http://scholarcommons.usf.edu/etd/6120 (19.01.2019).

Pavelka, T. (2016). Long-term unemployment in Visegrad countries. In: Loster, T., Pavelka, T. (eds.), *Proceedings of the 10th International Days of Statistics and Economics*, 1408–1415.