Application of the survival trees for estimation of the influence of determinants on probability of exit from the registered unemployment

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

Abstract

Survival trees are very useful regression tools for modelling of relations between the survival time and the vector of covariates. They are the example of the recursive binary partitioning. The aim of this method is creation of homogeneous subsets with respect to analysed response variables. Tree based methods, due to their non-parametric nature and flexibility, have become very popular in the last decades as an alternative to the traditional proportional hazard model. In the presented research, the conditional inference trees were used. It is the non-parametric class of regression trees that can be applied for all regression types. The goal of the analysis was the assessment of the influence of gender, age and education on the probability of exit from the registered unemployment. Due to the existence of censored observations, survival analysis methods were applied. The Kaplan-Meier estimator was used for estimation of the survival function of homogeneous groups in each terminal node. The splitting criterion was the statistic of the log-rank test, which is used for comparison of survival distribution for various groups. The two most numerous forms of de-registration were considered – finding a job and removal from the register for reasons attributable to the unemployed person. These forms helped to distinguish the subgroups of persons with the highest and lowest probability of de-registration to work, and resignation from the mediation of the labour office.

Keywords: *survival trees, Kaplan-Meier estimator, log-rank test, registered unemployment*
JEL Classification: C38, C41, J64
DOI: 10.14659/SEMF.2018.01.03

1 Introduction

Labour market analyses generally focus on the persons that leave the unemployment by finding a job. The statistical data collected in the poviat labour offices is the rich source of information about other causes of de-registration. There are about fifty of them: retirement, receiving pension, going abroad for period longer than 30 days, change of residence, death, granting pre-retirement allowance and many others. Finding a job is the most frequent cause of de-registration. Removal due to lack of readiness to work is the second one. It happens in

[1] Corresponding author: University of Szczecin/Institute of Econometrics and Statistics, Department of Operations Research and Applied Mathematics in Economics, ul. Mickiewicza 64, 71-101 Szczecin, beatus@wneiz.pl.
[2] University of Szczecin/Institute of Econometrics and Statistics, Department of Operations Research and Applied Mathematics in Economics, ul. Mickiewicza 64, 71-101 Szczecin, krzysztof.dmytrow@usz.edu.pl.

case of refusal of accepting proposed employment or absence in the labour office in due time. In years 2008-2014 removal constituted from 27% do 32% of all de-registrations in Poland (Fig. 2). Many unemployed people do not inform the labour office about finding a job thinking that it is their employer's responsibility. Formally, they should do it within the week since finding a job. The labour office pays premiums for the unemployed before removing them from the register. The trial to decrease the scale of this occurrence is punishing the unemployed people that were removed because of their fault. The punishment is the difficulty of reclaiming the status of the unemployed person, hence the right to the health insurance and benefit.

The goal of the research was the assessment of the influence of gender, age and education on the probability of exit from the registered unemployment. The two most numerous forms of de-registration were considered – finding a job and removal from the register for reasons attributable to the unemployed person. Due to the existence of censored observations, survival analysis methods were applied.

Situation on the Polish labour market in last years has improved. It can be observed by decreasing unemployment rate (Fig. 1). In the era of globalisation it is connected to the general situation on the world market and particularly in the European Union. As the analyses show, processes on the Polish labour market are similar to these on the Slovak and Hungarian markets (Hadaś-Dyduch et al., 2016). Unemployment is influenced by the social policy of the country, realised among the other things, by the labour offices. Activisation activities addressed to groups of people threatened by the unemployment influence the labour market positively. However, expanded system of benefits for the unemployed people may increase the time of job searching (Bieszk-Stolorz and Markowicz, 2015).

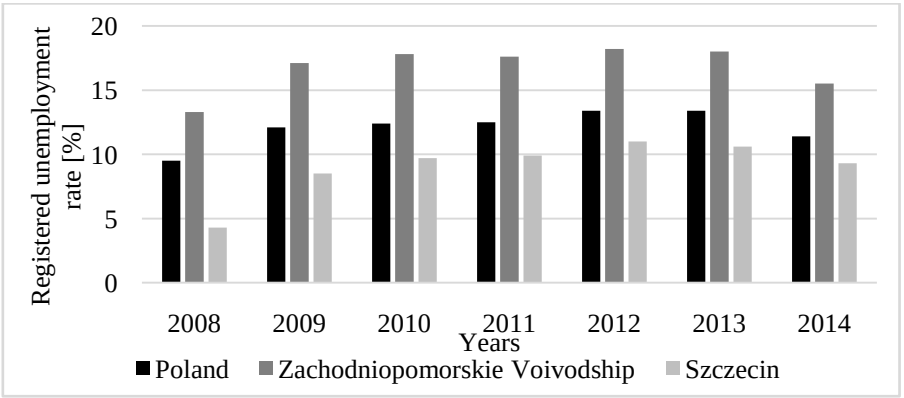

Fig. 1. Registered unemployment rate in Poland, Zachodniopomorskie Voivodship and Szczecin in years 2008-2014.

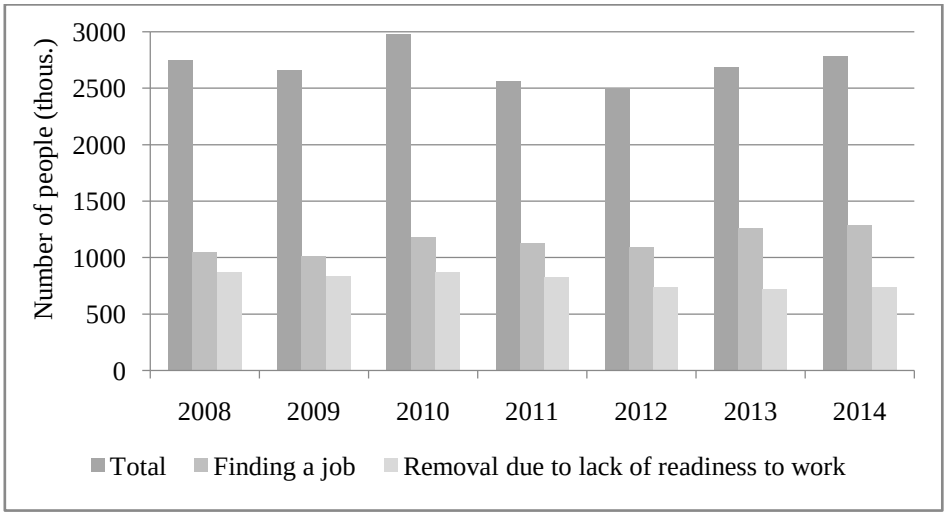

Fig. 2. Number of de-registrations from labour offices in Poland in years 2008-2014.

2 Data used in the research

In the study, anonymous individual data obtained from the Poviat Labour Office (PUP) in Szczecin (Poland) was used. The study covered 22078 unemployed individuals registered in 2013 and observed by the end of 2014. The event that terminated each observation was the moment of de-registration form the labour office list. Time *T* since the moment of registration until the moment of de-registration because of specific cause was analysed.

The two types of events terminating observations were considered: finding a job and removal due to lack of readiness to work. De-registrations due to other causes and observations that did not end with event before the end of 2014 (1856 observations) were considered as censored data. In the analysed period almost 44% of registered unemployed people found a job. They constituted the most numerous group. Slightly smaller group (almost 41%) were people that were absent in the labour office in due time or did not accept the job offer (removal). The size of each group is presented in Table 1.

The job-finding (Job) consists of three main subgroups: finding a job or another form of employment, taking up a government subsidised form of employment and entrepreneurial activity. The Removal from register category includes the unemployed individual's reluctance to cooperate with the labour office and have been removed from the register through their own fault or on their own request. The remaining causes of de-registration are less numerous and, as previous research showed, each of them had a marginal effect on the probability of de-registration (Bieszk-Stolorz, 2017).

Table 1. Structure of analysed unemployed people.

Group	Total	of which	
		Job	Removal
Total	22078	9678	8965
Gender			
Women (*K* or 1)	9770	4836	3264
Men (*M* or 0)	12308	4842	5701
Age			
18-24 (W_1)	4148	1506	2257
25-34 (W_2)	7356	3614	2966
35-44 (W_3)	4259	1869	1734
45-54 (W_4)	3497	1642	1214
55-59 (W_5)	2185	837	629
60-64 (W_6)	633	210	165
Education			
At most lower secondary (S_1)	5123	1410	2932
Basic vocational (S_2)	5016	1968	2220
General secondary (S_3)	2859	1226	1223
Vocational secondary (S_4)	4086	1943	1415
Higher (S_5)	4994	3131	1175

3 Research methodology

Survival analysis, commonly applied in demography and medicine, is more and more often applied in the social and economic phenomena, e.g. in labour market analysis. In this manner, the economic activity of the population (Landmesser, 2013) or duration of unemployment (Bieszk-Stolorz, 2013; Bieszk-Stolorz and Markowicz, 2012) can be analysed. The duration time at given state (duration of the company activity, duration of unemployment, duration of debt payment) is the random variable *T*. The basis for such analysis is the survival function, defined as follows:

$$S(t) = P(t > T) = 1 - F(t) \tag{1}$$

where *T* represents duration and *F*(*t*) – the cumulative distribution function of the random variable *T*. The most widely used estimator of the survival function is the Kaplan-Meier estimator (Kaplan and Meier, 1958):

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \quad (2)$$

where d_j is the number of events at the moment t_j and n_j is the number of individuals at risk by the moment t_j. The survival function $S(t)$ specifies the probability that the event will not occur at least by the time t. Depending on the defined event, sometimes it is more convenient to analyse the cumulative distribution function $F(t)$ which expresses the probability for the event to occur at most by the time t. When the event is defined as de-registering then the survival function estimator specifies the probability of remaining in the labour office register, while the estimator of the cumulative distribution function designates the probability of de-registering. In this case, d_j was the number of de-registrations due to particular cause at the moment t_j (finding a job or removal due to lack of readiness to work). In case of de-registration to work it is desirable that the survival curves are low-lying, while for the removal it is the opposite.

Analysed community can be divided into groups with respect to specific features, survival function for each group and the significance of differences between these functions can be estimated. Because distributions of the duration were unknown, the non-parametric tests, based on the rank order, were used. Unfortunately, there are no commonly accepted methods of selection of test at given situation. Most of them yield reliable results only for large samples, while effectiveness of these tests for small samples is less recognised. For comparison of two survival curves, the log-rank test is commonly used (Kleinbaum and Klein, 2005). It is used for verification of hypothesis H_0: $S_1(t) = S_2(t)$ stating that the survival curves for both groups are the same versus the hypothesis H_1: $S_1(t) \neq S_2(t)$ stating that they are not the same. Assuming that the null hypothesis is true, the test statistics is chi-square distributed with one degree of freedom. This test has the highest power, when the difference between the hazard functions for single subgroups is constant in time (Landmesser, 2013). Initial analysis with use of the function $\ln(-\ln S(t))$ and certain limitations resulting from assumptions for other tests confirmed the validity of application of the log-rank test in the research. In order to divide the analysed community into homogeneous groups with respect to shape of the survival curves, the survival trees are very useful tools. They are the subgroup of the so-called conditional inference trees. The idea of binary partitioning is used in construction of these trees. They has recently become popular in comparison with other methods (for example the discriminant analysis) because less assumptions are required and they can deal with various data structures (Al-Nachawati et al., 2010; Bou-Hamad et al., 2009;. Zhou and McArdle, 2015; LeBlanc and Crowley, 1993). Construction of any tree is

connected with two aspects (Cappelli and Zhang, 2007): partitioning the data, or tree growing and pruning the tree in order to make it shorter and increase the clarity of results.

Data partitioning is connected with separation of homogeneous with respect to analysed covariates groups. Splitting criterion can be based on the impurity measure or on the value of the log-rank test statistics. Partitioning occurs until the stopping criterion is reached. The necessity of the tree pruning is caused by the fact that data partitioning makes the tree very large and the overfitting occurs. Generally, the partitioning stops if the empirical significance level of the log-rank test statistics exceeds assumed value. However, for large sample size this approach is not always effective. The other criterion is defining the minimum group size, for which the partitioning may occur or the minimum group size in the terminal node. Also, the maximum tree depth may by defined (Mudunuru, 2016). Presented in the article survival trees were constructed by using the `ctree` function in the `partykit` package in R language. They were the conditional inference trees. Every observed unemployed person was described by the following triplet: $\{y_i, \delta_i, \mathbf{x}_i\}$, where y_i was the duration of registration, δ_i indicated whether the observation is censored or not (1 – uncensored, 0 – censored) and the $\mathbf{x}_i$ vector contained three covariates: gender, age and education. The duration of registration was the numerical continuous variable, censoring was the dichotomic variable. The covariates were the categorical variables. In the `ctree_control` function, two default parameters were changed: the `mincriterion` was set at 0.99 in order to set the significance level at 0.01 and by means of the `maxdepth` parameter the tree was pruned at the third level.

4 Analysis of time to de-registration to work or to removal

The analysis was conducted in two stages. The first one consisted in selection of homogeneous groups of unemployed people with respect to the probability of exit from unemployment to work. The Fig. 3 shows that gender was not the significant splitting criterion. In the first step the unemployed people were divided with respect to education – into persons with higher education and the remaining ones. In the next step, persons with at most secondary education were further divided into persons with at most lower secondary and secondary. The unemployed persons were further divided with respect to age. Finally, seven terminal nodes were obtained, their specifications are presented in Table 2. The lowest probability of exit from unemployment to work was observed for persons at the age of 60 and older with at most lower secondary education. On the other hand, the highest probability was observed for young persons (up to 35 years old) with higher education. The distribution of

time to de-registration to work is presented on the Fig. 4. It was extremely positively skewed. The largest number of people (1999 – 20.65%) found a job within first month since registration.

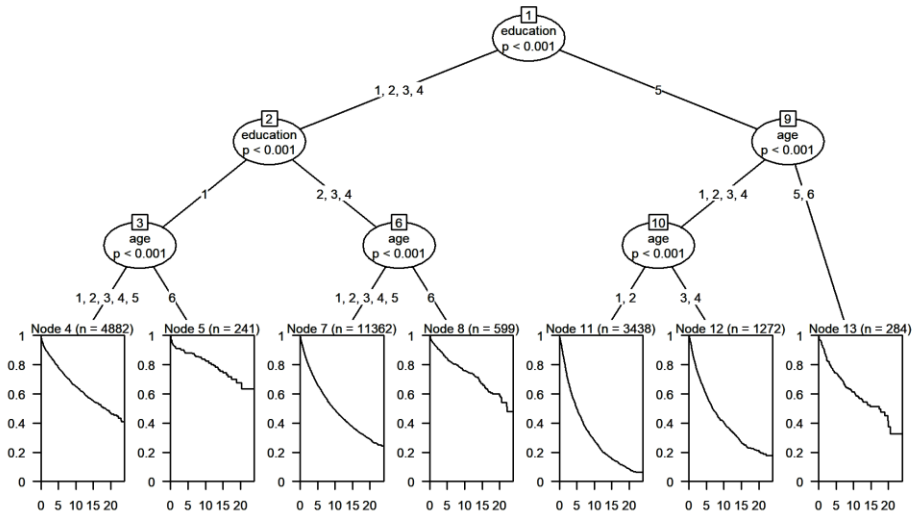

Fig. 3. Survival tree for de-registration to work.

Table 2. Homogeneous groups of unemployed persons – de-registration to work.

Specification	**Number of the terminal node**						
	4	**5**	**7**	**8**	**11**	**12**	**13**
education	S_1	S_1	S_2-S_4	S_2-S_4	S_5	S_5	S_5
age	W_1-W_5	W_6	W_1-W_5	W_6	W_1, W_2	W_3, W_4	W_5, W_6

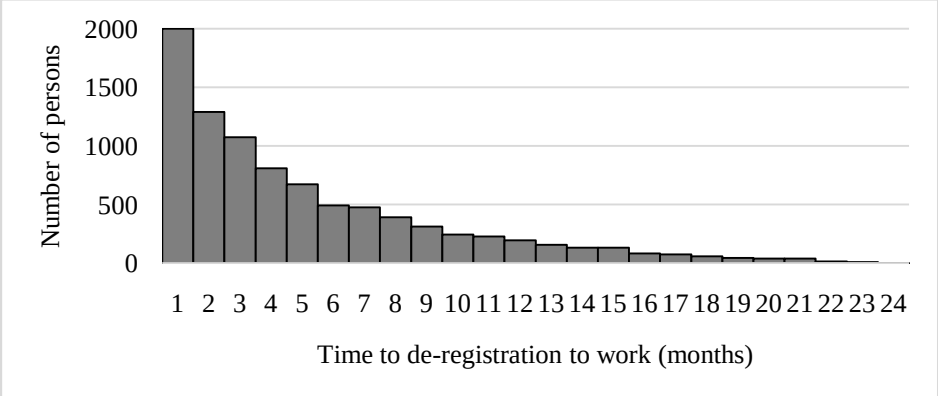

Fig. 4. Distribution of time to de-registration to work.

The second stage of the research consisted in selection of homogeneous groups of unemployed persons with respect to the probability of removal from the register. As seen on

the Fig. 5, unemployed persons removed from the register were in the first step divided with respect to age – into persons at the age up to 24 years and older. In the second step, the youngest persons were further divided with respect to education and gender. On the other hand, the oldest persons were not further divided according to age or education. It can be observed that the largest number of unemployed persons was removed from the register within the first month (values of the survival curve decreased most rapidly) and the second largest – in the fourth month. Finally, eight terminal nodes were obtained, their specifications are presented in Table 3. The lowest probability of removal from the register was observed at the age of 60 and more and the highest – the youngest males with at most lower secondary education.

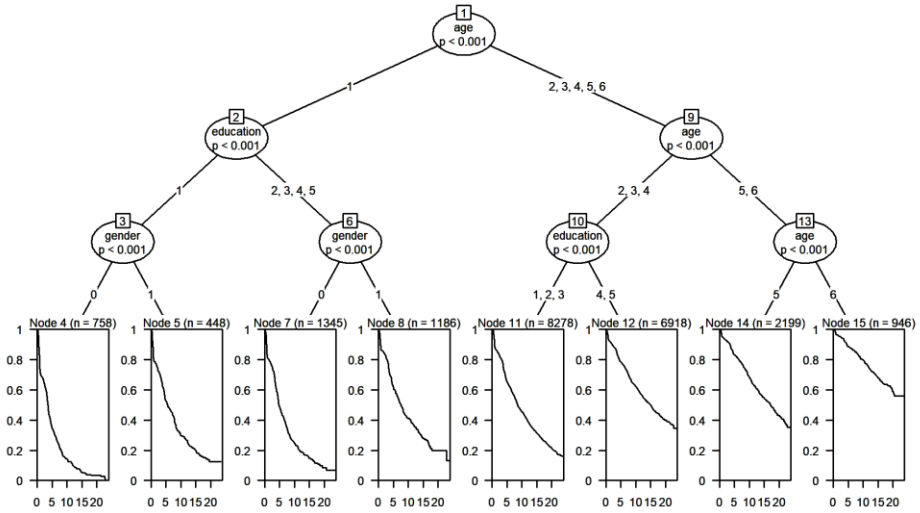

Fig. 5. Survival tree for removal from the register.

Table 3. Homogeneous groups of unemployed persons – removal from the register.

Specification	Number of the terminal node							
	4	**5**	**7**	**8**	**11**	**12**	**14**	**15**
education	S_1	S_1	S_2-S_5	S_2-S_5	S_1-S_3	S_4, S_5	S_1-S_5	S_1-S_5
age	W_1	W_1	W_1	W_1	W_2-W_4	W_2-W_4	W_5	W_6
gender	*M*	*K*	*M*	*K*	*K*, *M*	*K*, *M*	*K*, *M*	*K*, *M*

The distribution of time to removal from the register is presented on Fig. 6. It is worth noting that this distribution differs from the distribution of time to de-registration to work. It is bimodal – the largest probability of removal (2290 persons – 25.5%) was within the first

month since registration. The second largest probability of removal was in the fourth month (1148 persons – 12.8%).

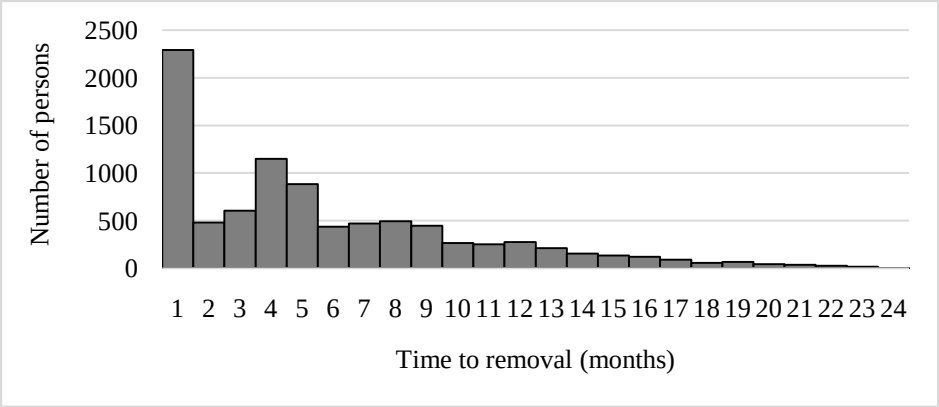

Fig. 6. Distribution of time to removal from the register.

Conclusions

On the basis on conducted analysis, the influence of gender, education and age on the probability of finding a job or removal due to lack of readiness to work was assessed. Results obtained in this research confirm the results obtained by means of other methods of survival analysis referring to the registered unemployment in Szczecin in years 2007-2011 (Bieszk-Stolorz, 2013). Generally, gender hardly influenced the probability of finding a job. On the other hand, education and age of the unemployed people were the strong determinants. In the analysed period, young people with higher education had the largest probability of finding a job. One of goals of labour market analysis is indication of groups particularly threatened by the unemployment and development of activisation programmes for them. Many of these programmes are directed to the young people. However, young and poorly educated persons were more quickly removed from the register. It indicates a lack of interests in the labour office offers such, as: participation in traineeships, trainings, providing additional equipment of workstations, etc. It is worth noting that gender of the unemployed person was important determinant of removal from the register. Males were much more often removed from the register than females.

References

Al-Nachawati, H., Ismail, M. & Almohisen, A. (2010). Tree-structured analysis of survival data and its application using SAS software. *Journal of King Saud University (Science)*, *22*, 251-255.

Bieszk-Stolorz, B. (2013). *Analiza historii zdarzeń w badaniu bezrobocia*. Volumina.pl Daniel Krzanowski, Szczecin.

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. & Markowicz, I. (2012). *Modele regresji Coxa w analizie bezrobocia*. CeDeWu, Warszawa.

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.

Bou-Hamad, I., Larocque, D., Ben-Ameur, H., Mâsse, L. C., Vitaro, F. & Tremblay, R. E. (2009). Discrete-time survival trees. *Canadian Journal of Statistics - Revue Canadienne De Statistique*, *37*(1), 17-32.

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.

Hadaś-Dyduch, M., Pietrzak, M. B. & Balcerzak, A. P. (2016). Wavelet Analysis of Unemployment Rate in Visegrad Countries. In: Kliestik, T. (ed.). *16th International Scientific Conference on Globalization and its Socio-Economic Consequences*. Zilina: University of Zilina, 595-602.

Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of American Statistical Association*, *53*, 457-481.

Kleinbaum, D. & Klein, M. (2005). *Survival Analysis. A Self-Learning Text*. Springer, New York.

Landmesser J. (2013). *Wykorzystanie metod analizy czasu trwania do badania aktywności ekonomicznej ludności w Polsce*. Wydawnictwo SGGW, Warszawa.

LeBlanc M. & Crowley J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association; 88*(422), 457-467.

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 (10.12.2017).

Zhou, Y. & McArdle, J. J. (2015). Rationale and Applications of Survival Tree and Survival Ensemble Methods. *Psychometrika*, *80*(3), 811-833.