RIGHT TAIL OF THE UNEMPLOYMENT DURATION IN THE CZECH REPUBLIC

ADAM ČABLA
University of Economics in Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability
W. Churchill Square 1938/4, Prague 3, Czech Republic
email: adam.cabla@vse.cz

Abstract

The aim of the paper is to compare three different approaches to modelling the distribution with supposed heavy tail using data about unemployment duration in the Czech republic – two models of spliced distributions and model with dynamic weights. All three aproaches are based on the Pickands-Balkema-de Haan theorem and incorporate the parametric distribution of the body of the probability distribution and generalized Pareto distribution used for the modelling of the right tail. The data come from the Labour Force Survey from three distinct periods - before the last economic crisis, during it and during the recovery from the crisis. The paper shows the differences between the right tails of the three periods and discussion of the used methods.

Key words: *Unemployment Duration, Labour Force Survey, Right Tail, Extreme Value Theory, Spliced Models, Dynamic Weight Mixture Model*

JEL Codes: *C24, J64*

DOI: 10.15611/amse.2017.20.05

1. Introduction

Unemployment is considered to be one of the most important topics in the economic discussion. From the statistical point of view, the unemployment is usually described using some characteristics of the rate of unemployment, e.g. in the Czech Republic there are two distinct rates – general unemployment rate provided by Czech Statistical Office and registered unemployment rate provided by Ministry of Labour and Social Affairs

Unemployment duration is considered in the term "long-term unemployment rate", which means rate of unemployment of those, who are unemployed for more than a year. OECD (2016a) calculates average unemployment duration on national levels using Labour Force Surveys.

But unemployment duration can be viewed as a random variable on its own, and we can thus describe more details or look out for the influence of demographic or other factors on the duration of unemployment. This point of view has been used many times, in the Czech Republic e.g. by Esser and Popelka (2003), Jarošová and Malá, (2005), Malá (2014) or Čabla (2012, 2015, 2016).

The aim of this paper is to built on the topic of unemployment duration as a random variable and to make a novel focus on the right tail of the probability distribution of this variable, that is usually underestimated by standard parametric models. Right tail of many probability distributions converges to the generalized Pareto distribution as is described by extreme value theory, to be more specific by Pickands-Balkema-de Haan theorem (de Haan

and Balkema, 1974, Pickands, 1975). We can use this property to model right tails on their own or to incorporate their parametric estimate with standard parametric estimates using so called spliced models.

In the paper I use data from the Labour Force Survey from three distinct time periods. These data are interval censored, which must be taken into consideration and I work only with persons that found a job during selected periods.

Chapters are organized as follows: first I specify available data, then describe basics of methodology of survival analysis, maximum likelihood estimates for interval censored data, peaks over threshold method as a part of extreme value theory, and spliced models. Last, but not least are the results in the form of final parametric models and figures.

2. Data

The aim of this chapter is to describe data from the Labour Force Sample Survey (LFSS) used in this paper. Data from LFSS are published as aggregates on annual or quarter basis. The source of some points below are guidelines in the ČSÚ (2013) and OECD (2016b).

The LFSS is a survey conducted by the Czech Statistical Office (CZSO) since December 1992 continuously throughout the year. Its main objective is to "obtain regular information about the labor market situation, enabling its analysis from various aspects, especially economic, social and demographic". Since 2002, the LFSS questionnaire has been fully harmonized with Eurostat standards and is a national modification of the Labor Force Survey (LFS). LFS is a large sample survey of households that yields quarterly results. The LFS is conducted in all countries of the European Union and the European Free Trade Association (EFTA) and provides statistics comparable across countries in line with the ILO methodology, more on Eurostat (2016).

The subject of the survey are all persons usually resident in the selected households of investigated dwellings who intend to stay in the Czech Republic for at least one year. For all persons over 14, details of their economic activity and education are collected.

The LFSS is held continuously throughout the year, with data typically summed up quarterly, and each household is surveyed once per quarter. Each household stays in a dataset for 5 consecutive inquiries, so the dataset changes every quarter by approximately 20 %.

Individuals are assigned weights determined as the proportion of the number of persons in the whole population and the number of persons of the same gender in the same age group and district of residence except for Prague, which is taken as a whole. The number of people in the entire population is projected from the end of the previous year to the middle of the current quarter by the natural increase and decrease and the migration balance in the previous year.

There are three distinct periods of time in the paper, each covering a total of five consecutive quarters. The first period begins in the fourth quarter of 2007 and ends in the fourth quarter of 2008, which is called before the crisis. The second begins in the first quarter of 2010 and ends in the first quarter of 2011 and is called during the crisis. The third period begins in the fourth quarter of 2013 and ends in the fourth quarter of 2014 and is called after the crisis.

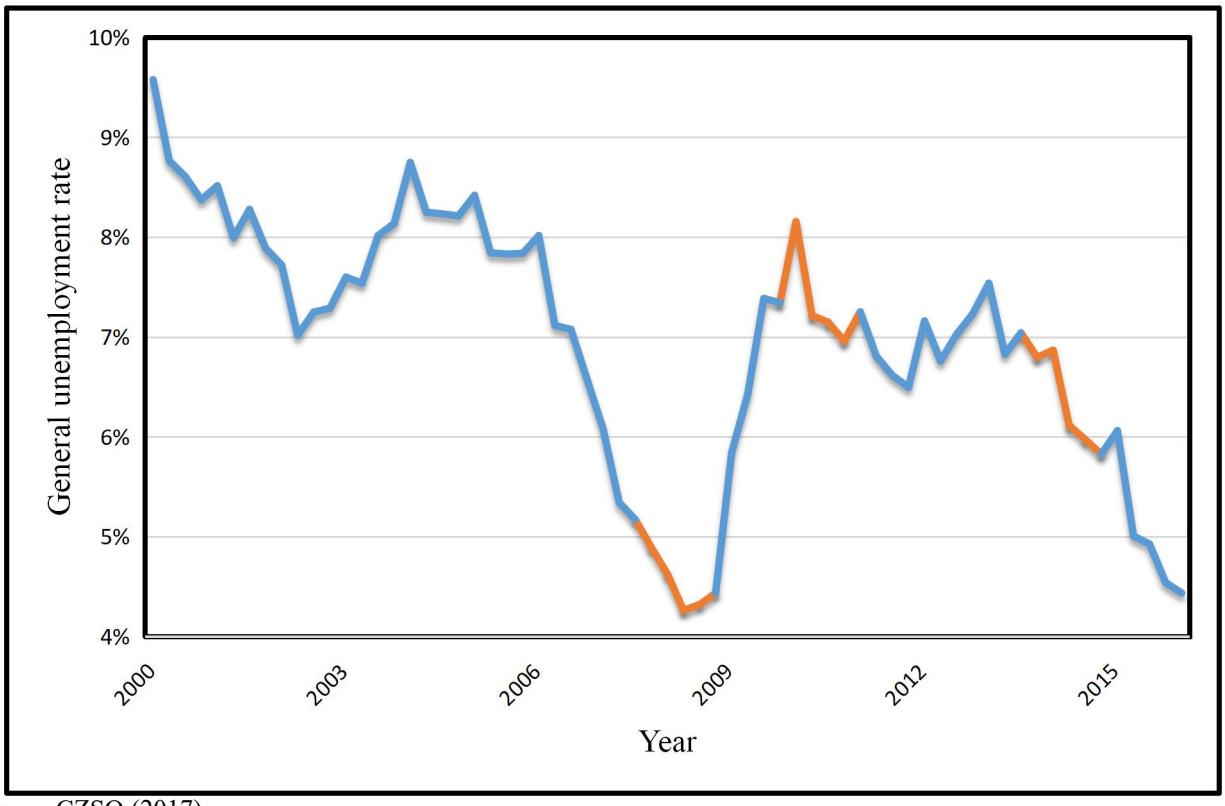
The first and second periods were selected taking into account the general unemployment rate in the Czech Republic - the first period includes the quarter with a low unemployment rate (about 4.3-5.2 %), the second one with a high unemployment rate (about 7.0 - 8.2 %). The third period was selected as a period of declining unemployment, where the general rate ranged between the above-mentioned values (5.8-6.8 %). The rate of unemployment is shown in figure 1, selected periods are red.

From datasets published each quarter as mentioned above one can find those persons, that were unemployed in one quarter and were employed in the following quarter, which means

that they found a job between the two surveys. Since for every person we know how long he/she is unemployed or employed, we can calculate the unemployment duration of any specific person that found a job. Since the unemployment and employment durations in the datasets are written in intervals, so is the resulting unemployment duration I work with. Hence we have only interval or right censored observations here. Rights censored observations are those in which we know only that the unemployment duration was larger than some specific time, which again results from the original datasets.

Using the procedure, I found out 2,675 entries of persons that found a job, but 18 of them were deleted because employment duration was higher than the time between the two surveys (generally three months), which indicates incorrect information. Out of the 2,657 correct entries 624 were before the crisis, 1,069 were during the crisis and 964 were after the crisis.

Figure 1: General unemployment rate in the Czech Republic



Source: CZSO (2017)

3. Methodology

In this chapter I outline basics of the methodology that is used in the calculations.

3.1. Survival analysis

Survival analysis is a general term covering several methods used to investigate random variable that is time to occurrence of some event. Here in the paper the event is finding a job.

The primary objectives of survival analysis are usually the estimation and interpretation of survival and/or risk functions, the comparison of different survival functions, and the relationship between explanatory variables and probability distribution of time to event.

Therefore, in the survival analysis we examine the continuous non-negative random variable T. The basic description of its probability distribution is the probability that this time will be greater than time t, therefore the survival function is defined as

$$S(t) = P(T > \mathrm{t}) = 1 - F(t) = \int_t^{\infty} f(s)ds. \tag{1}$$

Survival function is complement to cumulative distribution function F(t) and so is non-increasing, in time t = 0 is S(0) = 1 and $\lim_{t\to\infty}(S(t)) = 0$.

Another way to describe the probability distribution in the survival analysis is the risk function h(t) for which other names are used, such as conditional failure rates in the reliability analysis, mortality rate in demography, age-related failure rate in epidemiology, or the reversal of Mill rate in economics.

This function is the potential of immediate occurrence of event for a unit of time assuming, that this event has not yet occurred. Expressed by the formula it is

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}. \tag{2}$$

The risk function h(t) is non-negative and at any time t can obtain any values from zero to infinity. Its course reflects the properties of the observed phenomenon. There is a clear relationship between the risk function h (t), the survival function S (t) and the probability density (Klein, Moeschberger, 1997)

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\log S(t)}{dt}. \tag{3}$$

3.2. Interval censoring and maximum likelihood estimates

In general, we can say that the value of any observation lies in the interval (L, R], so we will assume

$$T \in (L, R], \tag{4}$$

where L ≤ R. We can distinguish the following four situations:

- If L = R, this is a complete observation with the value R.
- If L = 0 applies, it is a left-centered observation in the interval (0, R).
- If R = ∞ is true, it is a censored observation in the interval (L, ∞).
- If none of the above is valid, then it is the interval-censored observation in the interval (L, R).

From the definition above we can say, that interval censored observations are those, for which we can say that the event occurred after some specific time larger than 0 and prior to the different specific time.

The contribution of the i-th uncensored observation to the likelihood is equal to the probability density at the point, the contribution of censored observation is then, based upon the condition of non-informativness based on discussions in Peto (1973) and Kiefer and Wolfowitz (1956) function of the parameter vector θ and is equal to the integer of probability density over interval (Li, Ri), i.e.

$$L_i(\boldsymbol{\theta}) = \int_{L_i}^{R_i} f(t, \boldsymbol{\theta})dt = F(R_i, \boldsymbol{\theta}) - F(L_i, \boldsymbol{\theta}) = S(L_i, \boldsymbol{\theta}) - S(R_i, \boldsymbol{\theta}). \tag{5}$$

The likelihood function is the product of the contributions of individual observations by type of censorship. From equation 5 it can be deduced that the contribution of the right censored observation to the likelihood is equal to S(Li) and the contribution of the left centered observation is 1 – S (Ri). Besides the likelihood function, we can calculate with its natural logarithm $l(\boldsymbol{\theta}) = \ln[L(\boldsymbol{\theta})]$.

For interval censored data we can build maximum likelihood nonparametric estimates, from which the most commonly used is Turnbull estimate (Turnbull, 1976) with the adjustment from Gentleman and Geyer (1994). This estimate is an EM algorithm.

Estimate of parametric distributions are done via straightforward numeric maximization. For comparison I use Akaike information criterion (Akaike, 1974) and Bayesan information criterion (Schwarz, 1978), both of which compensates the likelihood of the estimate for the number of estimated parameters, penalizing more complicated models.

3.3. Peaks over threshold

Extreme value theory is the branch of mathematical statistics, which deals firstly with the stochastic behavior of the sampling maxima or the minima of independent, identically distributed random variables. Secondly, it studies the stochastic behavior of independent, identically distributed random variables whose value exceeds a sufficiently high value of threshold u.

Peaks over threshold analysis (POT) is the second of the two types of extreme value analysis, and works with all values exceeding a certain sufficiently high threshold u. We usually consider the T – u values, i.e. the differences between the value of the random variable T and the threshold u. According to Balkema and de Haan (1974) and Pickands (1975), Pickand-Balkema-de Haan theorem states:

Let us consider independent, identically distributed random variables X with a generally unknown distribution function F. Random values for which X > u have a conditional excess distribution function

$$F_u(y) = P(X - u \le y | X > u) \text{ for } 0 \le \varpi_F - u, \tag{6}$$

Where u is the threshold, y = x – u are the excess values, and ω_F is the right end-point of the distribution function F. For a large class of distribution functions we can state,

$$F_u(y) \to Z_{\xi,\mu,\sigma} = 1 - \left(1 + \xi \frac{x - \mu^*}{\sigma}\right)^{-\frac{1}{\xi}} \qquad \text{for } u \to \infty. \tag{7}$$

Z is the distribution function of generalized Pareto distribution. For the parameter ξ = 0, the distribution converges to exponential distribution

$$Z(x) = 1 - \exp\left(-\frac{x - \mu^*}{\sigma}\right). \tag{8}$$

Two decisions are important in the peaks over threshold method. Firstly, determining the threshold value u and secondly determining the estimation method.

Determining the threshold value u is usually done using ad hoc graphical procedures, see e.g. Tanaka and Takara (2002), or in relation to the spliced models Scarrott and MacDonald (2012). From these procedures, a graph of the parameter estimate ξ or sample mean excess for different threshold values can be used for interval censored data. Instead of these, I use the logarithm of the value of the likelihood for different thresholds.

These methods are based on two considerations – on the one hand, the higher the threshold chosen, the more likely the distribution of the right tail will converge to the generalized Pareto distribution. On the other hand, the lower the threshold will be, the more observations will be used to estimate the generalized Pareto distribution´s parameters, so their estimate will be more accurate.

Various methods are used for estimating the parameters of the generalized Pareto distribution, the most important being the method of maximum likelihood discussed in Smith (1985), the probability-weighted moment method and the regression method based on the graph of mean excess value against the value of u. Discussion of the characteristics of these estimates is, for example, in Castilo and Hadi (1997).

Smith (1985) states that the maximum likelihood method is unlikely to lead to the estimate if $|\xi| > 1$ and does not have the usual asymptotic properties for $|\xi| > 0.5$. For $|\xi| \leq 0.5$, the estimate is asymptotically normal.

For censored data, the properties of individual estimates are not well explored. The article of Lin and Wang (2007), based on a simulation study, states that for censored data, maximum likelihood estimates are more reliable than moment estimates. Estimates in this work are maximum likelihood.

3.4. Spliced models

Using the extreme value theory, we can estimate the right tail of the distribution by the generalized Pareto distribution. We can use this fact to refine the estimate of the probability distribution by spliced models.

The spliced model is a combination of an estimate of the distribution function of the bulk of distribution and a parametric estimate of the distribution function of its right tail. The distribution functions of the bulk can be estimated both non-parametrically or parametrically. This model has been developed in recent years, particularly in insurance and risk analysis, see e.g. chapter 3 in Dey and Yan (2016) or Chapter 1.5 in Peters and Shevchenko (2015).

Denote the distribution function for the bulk, where θb denotes the vector of the parameters of the bulk of the distribution, and the distribution function of the right tail, where θu denotes the vector of parameters of the right tail. The tail fraction is $\phi_u = 1 - F(u|\boldsymbol{\theta}_\mathbf{b})$, as in MacDonald et al. (2011) and we can write the distribution function of the spliced model

$$F(x|\boldsymbol{\theta}_\mathbf{b},\boldsymbol{\theta}_\mathbf{u},\phi_u) = \begin{cases} \dfrac{1-\phi_u}{F(u|\boldsymbol{\theta}_\mathbf{b})} F(x|\boldsymbol{\theta}_\mathbf{b}) & \text{for } x \leq u \\ (1-\phi_u) + \phi_u Z(x|\boldsymbol{\theta}_\mathbf{u}) & \text{for } x > u \end{cases} \quad (9)$$

This mixture of non-overlapping distributions is called by Dey, Yan (2016) a standard model. In the standard model, the fraction on the first line normalizes the distribution function so that the threshold corresponds to the empirical distribution function, and at this point follows the estimate of the right end.

If we lay $\phi_u = 1 - F(u|\boldsymbol{\theta}_\mathbf{b})$, we use the estimate of the bulk to directly estimate the quantile from which the estimate of the right tail follows. In this paper, I choose this second procedure because it is not possible to find an estimate of quantiles in the Turnbull (non-parametric) estimate of the empirical distribution function without pronouncing a certain probability distribution within the estimated intervals.

The first procedure (the first model) I use to estimate spliced model is as follows:

1. Values in the dataset are sorted by the mean of L_i and R_i.

2. The threshold u is set at the 95% quantile of the parametric estimate based on all the values.
3. Other parameters of generalized Pareto distribution are estimated via maximum likelihood estimate
4. Based on this estimate the mean excess value is estimated as

$$E(X) = u^* + \frac{\sigma}{1-\xi}, \text{ pro } \xi < 1. \tag{10}$$

For $\xi \geq 1$ mean excess value is not defined.
5. The values of ξ and mean excess value are written down.
6. Percentage of the quantile used as a threshold is reduced by one percentage point and the algorithm is repeated till the point, where logarithm of the likelihood is clearly decreasing.
7. The resulting estimate is the one, where the logarithm of the likelihood is maximal. If the value of the ξ is not statistically different from 0, the exponential distribution is used as in equation 8.

The second procedure I use is direct maximization of the likelihood of the distribution described by the equation 9, where all parameters are estimated at once including the tail fraction $\phi_u = 1 - F(u|\boldsymbol{\theta}_\mathbf{b})$.

The third procedure I use is not actually spliced model but it aims the similar way. It is the dynamic weighted mixture model and was first described in the context of extreme value theory by Frigessi et al. (2002). The data are simultaneously estimated by log-logistic distribution (or any other suitable distribution) and generalized Pareto distribution and there is a function that shifts the weight between the two, so there is a rising weight on generalized Pareto distribution as the T grows. The density function is

$$v(t) = \frac{\big(1 - p(t;\boldsymbol{\theta})\big) f(t;\boldsymbol{\beta}) + p(t;\boldsymbol{\theta}) z\big(t;\xi,\sigma\big)}{Z(\boldsymbol{\theta},\boldsymbol{\beta},\xi,\sigma)}, \tag{10}$$

where f(t;β) is probability density function chosen for the small values, z(t, ξ, σ) is probability density function for generalized Pareto distribution, p(t;θ) is increasing dynamic weight with values (0, 1) and Z(θ, β, ξ, σ) is normalizing constant that ensures, that the integer of the probability density function v(t) is equal to 1.

As a weight I use distribution function of Cauchy distribution in line with Frigessi et al. (2002)

$$p(t,\boldsymbol{\theta}) = \frac{1}{2} + \frac{1}{\pi} arctg\left(\frac{t-\mu}{\sigma}\right), \quad \boldsymbol{\theta} = (\mu,\sigma), \quad \mu,\sigma > 0. \tag{11}$$

This chosen weight ensures the important properties: is increasing and in the interval (0, 1). Parameter μ is median of Cauchy distribution; it is the value, at which the weight is a half for bulk distribution and a half for generalized Pareto distribution.

If Heaviside function is chosen as a weight, this dynamic model is equal to the standard model described by equation 9, see Scarrott a MacDonald (2012).

4. Results

In this chapter, I present the estimates of the spliced models of the unemployment duration. For each time period, I estimate a total of 6 models, comparing them using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Starting model is standard parametric estimate using log-logistic distribution in line with Čabla (2016).

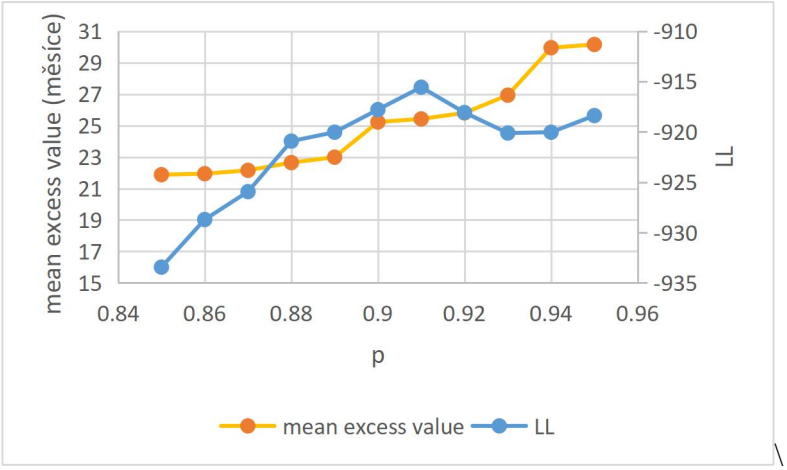
The first model is a log-logistic estimate of the entire distribution and the substitution of its right tail by generalized Pareto distribution from a threshold determined by the sequence

described in chapter 3.4 (LL + GPD). Figures 2 through 4 show logarithm of likelihood of the models dependent on the quantile selected as a threshold for three studied periods.

The second model is a concurrent estimate of the log-logistic and generalized Pareto distribution with the threshold (LL + GPD + μ*) and log-logistic bulk with exponential right tail (LL + EXP + μ*).

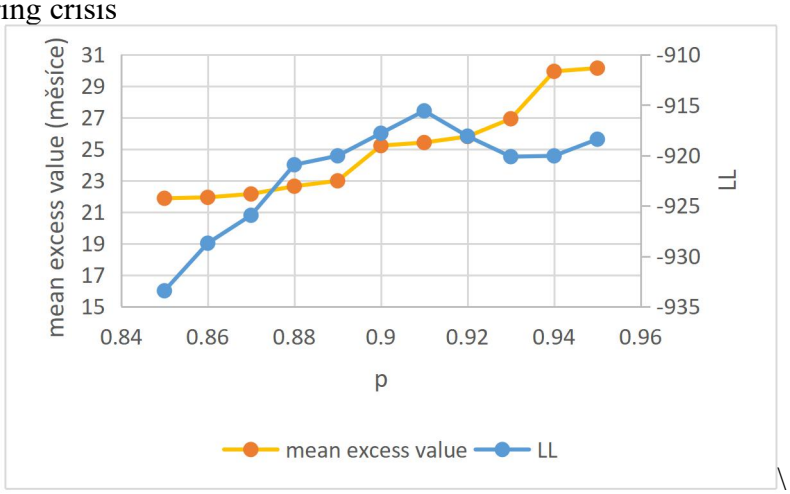
The third model is a mixture of log-logistic and generalized Pareto distribution with the dynamic weight represented by the distribution function of the Cauchy distribution (LL + GPD + C). The right tail distribution function can also be exponential (LL + EXP + C).

Figure 2: Mean excess value and logarithm of likelihood function (LL) for different quantiles as a threshold, before crisis



Source: the author's work

Figure 3: Mean excess value and logarithm of likelihood function (LL) for different quantiles as a threshold, during crisis



Source: the author's work

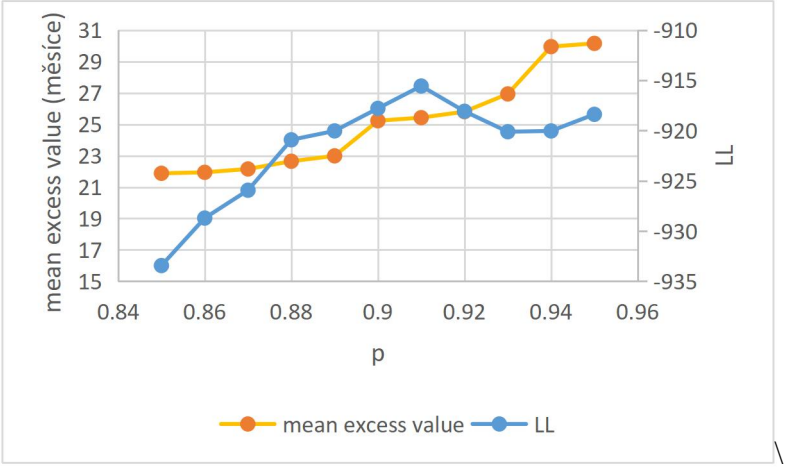
In the period before crisis, in the search for the threshold used for generalized Pareto distribution, the likelihood of the model is highest at 91 % quantile. So the threshold has a value of 20.972 months.

In the second model, the estimation of parameter ξ is not statistically significant, which is in consistent with the comparison using AIC and BIC.

In the third model there is again a statistically insignificant estimate of parameter ξ, so the table gives an estimate with exponential distribution as well as the one with the generalized Pareto distribution.

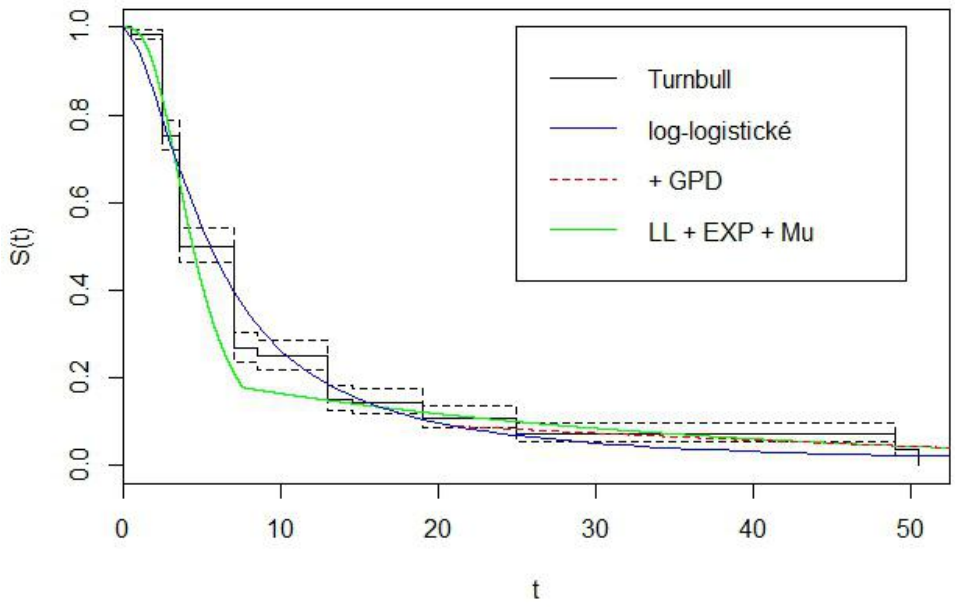
All six estimated models are compared in the table 1, where they are sorted by AIC in ascending order. As the best model, we can designate the second model with exponential distribution used for the right tail. This model is fully specified in the table 2, where α and β are parameters of log-logistic distribution, other parameters are those of exponential distribution. Figure 5 depicts four different estimates, note the different estimate of LL+EXP+ μ*.

Figure 4: Mean excess value and logarithm of likelihood function (LL) for different quantiles as a threshold, after crisis



Source: the author's work

Figure 5: Turnbull (black), log-logistic (blue), LL+GPD (blue and red) and LL+EXP+μ* estimates, before the crisis.



Source: the author's work

In the period during crisis, in the search for the threshold used for generalized Pareto distribution, the likelihood of the model is highest at 97% quantile, but the estimate of ξ is

below – 3, so I do not use this value and use the second highest value, which is obtained at 95% quantile. So the threshold has a value of 26.934 months.

Table 1: Comparing estimates, sorted by AIC, before crisis

Model	LL	no. of parameters	AIC	BIC
LL+ EXP + μ*	-874.29	4	1756.58	1774.33
LL + GPD + μ*	-873.87	5	1757.74	1779.92
LL + GPD	-915.57	4	1839.14	1856.89
LL + EXP + C	-915.24	5	1840.48	1862.66
LL + GPD + C	-922.47	6	1856.94	1883.56
LL	-935.06	2	1874.12	1882.99

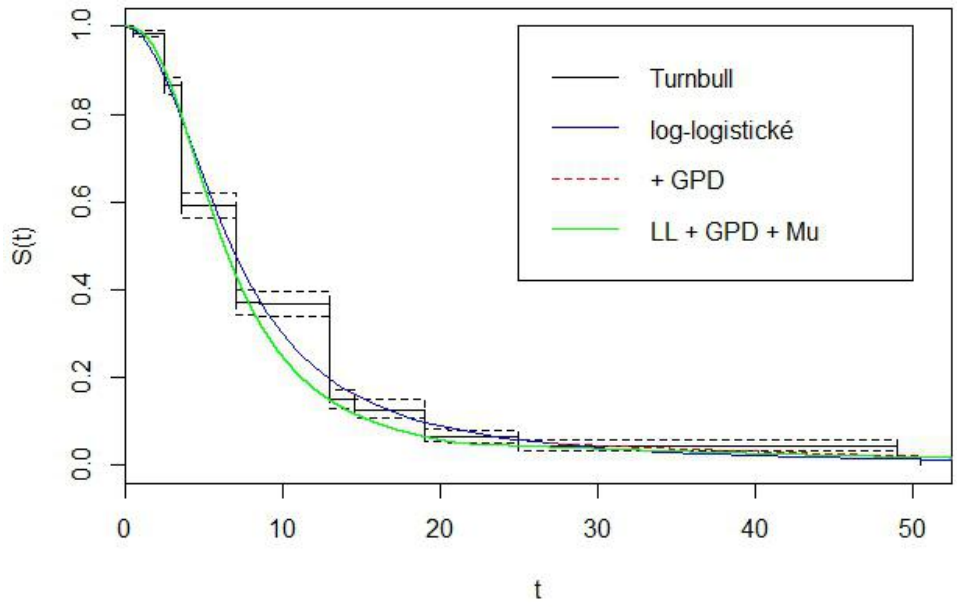
Source: the author's work

Table 2: estimate of LL + EXP + μ* model before crisis

Parameter	Estimate	s.e.	Z	p-value
α	4.393	0.7543	5.82	< 0.0001
β	2.817	0.0904	31.16	< 0.0001
μ*	7.568	1.1353	6.67	< 0.0001
σ	30.396	8.3430	3.64	0.0001

Source: the author's work

Figure 6: Turnbull (black), log-logistic (blue), LL+GPD (blue and red) and LL+EXP+μ* estimates, during the crisis.



Source: the author's work

In the third model there is again a statistically insignificant estimate of parameter ξ, so the table gives an estimate with exponential distribution as an addition to the one with the generalized Pareto distribution. I designate this model as LL + EXP + C.

All six estimated models are compared in the table 3, where they are sorted by AIC in ascending order. As the best model, we can designate the second model with generalized Pareto distribution used for the right tail. This model is fully specified in the table 4. Figure 6 depicts four different estimates.

Table 3: Comparing estimates, sorted by AIC, during crisis

Model	LL	no. of parameters	AIC	BIC
LL + GPD + μ*	-1558.00	5	3126.00	3150.87
LL + EXP + μ*	-1560.45	4	3128.90	3148.80
LL	-1567.57	2	3139.14	3149.09
LL + GPD	-1566.98	4	3141.96	3161.86
LL + EXP + C	-1567.25	5	3144.50	3169.37
LL + GPD + C	-1606.04	6	3224.08	3253.93

Source: the author's work

Table 4: estimate of LL + GPD + μ* model during crisis

Parameter	Estimate	s.e.	Z	p-value
α	6.267	0.9423	6.65	< 0.0001
β	2.394	0.1012	23.66	< 0.0001
ξ	– 0.251	0.1423	– 1.76	0.0389
μ*	21.408	1.4323	14.95	< 0.0001
σ	33.174	5.4324	6.11	< 0.0001

Source: the author's work

Table 5: Comparing estimates, sorted by AIC, after crisis

Model	LL	no. of parameters	AIC	BIC
LL + EXP + μ*	-1196.36	4	2400.72	2420.20
LL + GPD + μ*	-1248.53	5	2507.06	2531.42
LL + GPD	-1434.68	4	2877.36	2896.84
LL + EXP + C	-1437.11	5	2884.22	2908.58
LL	-1449.37	2	2902.74	2912.48
LL + GPD + C	-1479.31	6	2970.62	2999.85

Source: the author's work

Table 6: estimate of LL + EXP + μ* model after crisis

Parameter	Estimate	s.e.	Z	p-value
α	3.414	0.8535	4.00	< 0.0001
β	9.766	0.5643	17.31	< 0.0001
μ*	3.436	1.2324	2.79	0.0027
σ	14.766	3.2324	4.57	< 0.0001

Source: the author's work

In the period after crisis, in the search for the threshold used for generalized Pareto distribution, the likelihood of the model is highest at 91% quantile. So the threshold has a value of 21.522 months.

All six estimated models are compared in the table 5, where they are sorted by AIC in ascending order. As the best model, we can designate the second model with exponential distribution used for the right tail. This model is fully specified in the table 6. Figure 7

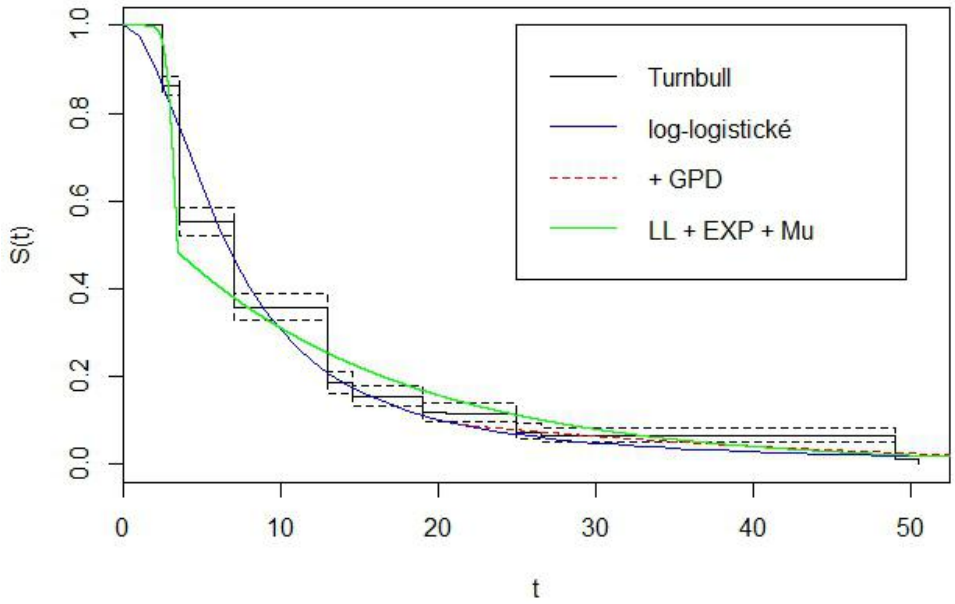
It can be seen from the estimates above that in the pre-crisis and post-crisis period the estimation of the right tail of the unemployment duration can be improved using extreme value theory, preferably by direct estimate of the model from the equation 9.

Right tails tend to exponential distribution, which is applied before the crisis from about 7.5 months, while after the crisis it has been from less than 3.5 months. During the crisis, the estimate of the right tail is slightly improved using generalized Pareto distribution being applied only after the twentieth month.

The use of the mixture with dynamic weight did not appear to be very appropriate on these data, the distribution of the tail dominates only from the 57th, 78th and 54th month respectively. These estimates, however, are in line with the knowledge gained from the other models, that the right tail was not much underestimated during the crisis by log-logistic distribution.

Graphical comparison shows that before and after the economic crisis the spliced model with the highest likelihood distinctly differs from other parametric and non-parametric estimates.

Figure 7: Turnbull (black), log-logistic (blue), LL+GPD (blue and red) and LL+EXP+μ* estimates, after the crisis.



Source: the author's work

5. Conclusion

In this paper I introduced the spliced models and used this class of models to improve the estimates of the unemployment duration from the data from Labour Force Survey from three distinct periods. I compared three different approaches in spliced models and compared them to standard parametric estimate. From the results in chapter 4 it is clear, that spliced model, where bulk, right tail and tail fraction are estimated at once, improves likelihood of the estimate of given datasets.

In the period during the crisis the other two spliced models did not bring better estimates, in terms of likelihood, than the log-logistic model. This can be interpreted so that log-logistic estimate has not underestimated right tail in this period. In the other two periods, the first model, where right tail continues from the estimate of whole distribution, also gave better results than log-logistic distribution. The model of mixture with dynamic weights did not fare much better than log-logistic distribution and its usefulness in this area remains doubtful.

Concept of spliced models helps to better understand behavior of long-term unemployment duration. Since in all models the estimate of ξ was not far away from 0 or was equal to 0, aka exponential distribution, the hazard function is near constant or constant. This would mean, that immediate potential of obtaining a new job after some period of time is not influenced by the time lapsed (the memorylessness of exponential distribution). This period of time was much prolonged during the crisis and was shortest after the crisis. The exact point from which this starts remains for the further investigation and its determination is dependent on the model used.

Also the difference of survival functions of the spliced model with the highest likelihood and other estimates warrants further investigation.

References

[1] Akaike, H. 1974. A New Look at the Statistical Model Identifiaction. IEEE Translations of Automatic Control. 19 (6), 716 – 723. doi:10.1109/TAC.1974.1100705.

[2] Castillo, E. Hadi, A. 1997. Fitting the Generalized Pareto Distribution to Data. Journal of the American Statistical Association, 92(440), 1609-1620.

[3] Čabla, A. 2012. Unemployment duration in the Czech Republic. Prague 13.09.2012 – 15.09.2012. In: The 6th International Days of Statistics and Economics, Conference Proceedings. Praha, 2012, pp. 257 – 267. ISBN 978-80-86175-86-7.

[4] Čabla, A. 2015 Unemployment Duration in the Czech Republic After the Economic Crisis. In: Applications of Mathematics and Statistics in Economics – AMSE [CD ROM]. Jindřichův Hradec, 02.09.2015 – 06.09.2015. Praha : Oeconomica Publishing House, 2015. 11 s. ISBN 978-80-245-2099-5.

[5] Čabla, A. 2016 Minimal Adequate Model of Unemployment Duration in the Post-Crisis Czech Republic. Statistika. roč. 96, č. 1, s. 50–62. ISSN 0322-788X.

[6] ČSÚ. 2013. Pokyny pro tazatele a krajské garanty pro rok 2014: Výběrové šetření pracovních sil. Praha.

[7] De Haan, L. Balkema, A.A. 1974. Residual Life Time at Great Age. The Annals of Probability, 2.

[8] Dey, D., Yan, J. 2016 Extreme Value Modeling and Risk Analysis: Methods and Applications. Boca Raton, FL: CRC Press, Taylor. ISBN 1498701299.

[9] Esser, M., Popelka, J..2003 Analysis of Factors Influencing Time of Unemploment Using Survival Time Analysis. Bratislava 04.12.2003 – 05.12.2003. In: Výpočtová štatistika. Bratislava: Slovenská štatistická a demografická spoločnosť, 2003, s. 250–254. ISBN 80-88946-29-8.

[10] EUROSTAT. 2016. European Union Labour Force Survey (EU LFS). Eurostat [online]. 2016 [cit. 2016-10-10]. Dostupné z: http://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey
[11] Frigessi, A., Haug, O. Havard, R. 2002. A dynamic mixture model for unsupervised tail estimation without threshold selection, Extremes, 5, 219–235.
[12] Gentleman, R. a C.J. Geyer. 1994. Maximum Likelihood for Interval Censored Data: Consistency and Computation. Biometrica. 81(3).
[13] Jarošová, E., Malá I. 2005 "Modelling time of unemployment in the Czech Republic." In: KOVÁČOVÁ, M. Proceedings of the 4th International Conference APLIMAT 2005. 2005. Bratislava: Slovak University of Technology. ISBN 978-80-969264-2-8.
[14] Kiefer, J., Wolfowitz, J. 1956 Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters, Ann. Math. Statist. 27.
[15] Klein, J. P., Moeschberger, M. L. 1997. Survival Analysis : Techniques for Censored and Truncated Data. New York : Springer-Verlag. 502 s. ISBN 0-387-94829-5.
[16] Lin, Ch. Wang, W. 2007. Estimation for the generalized pareto distribution with censored data. Communications in Statistics - Simulation and Computation DOI: 10.1080/03610910008813660.
[17] Macdonald, A., Scarrott, C.J., LeeD., Darlow B., Reale M., Russell G. 2011 A Flexible Extreme Value Mixture Model. Computational Statistics & Data Analysis. 56(6), 21.
[18] Malá, I. 2014 The Use of Finite Mixture Model for Describing Differences in Unemployment Duration. In: AMSE [CD ROM]. Jerzmanovice, 27.08.2014 – 31.08.2014. Wroclaw : Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu. s. 164–172. ISBN 978-83-7695-421-9.
[19] OECD. 2016a. Average Duration of Unemployment. OECD.Stat [online]. [cit. 2016-10-10) Dostupné z: https://stats.oecd.org/Index.aspx?DataSetCode=AVD_DUR
[20] OECD. 2016b. LABOUR FORCE STATISTICS IN OECD COUNTRIES: SOURCES, COVERAGE AND DEFINITIONS [online]. 39 s. [cit. 2016-10-10]. Dostupné z: http://www.oecd.org/els/emp/LFSNOTES_SOURCES.pdf.
[21] Peters, G. W., Shevchenko, P.V. 2015 Advances in heavy tailed risk modeling: a handbook of operational risk. Hoboken, New Jersey: Wiley. Wiley handbooks in financial engineering and econometrics. ISBN 978-1-118-90953-9.
[22] Peto, R. 1973. Experimental Survival Curves for Interval-Censored Data. Applied Statistics. 22.
[23] Pickands, J. 1975. Statistical Inference Using Extreme Order Statistics. The Annals of Statistics, 3.
[24] R CORE TEAM. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
[25] Scarrott, C., MacDonald A. 2012. A Review of Extreme Value Threshold Estimation and Uncertainity Quantification. REVSTAT – Statistical Journal. 10(1), 28.
[26] Schwarz, G.E. 1978. Estimating the Dimension of a Model. Annals of Statistics. 6(2). 461 – 464.
[27] Smith, R. 1985. Maximum Likelihood Estimation in a Class of Nonregular Cases. Biometrika, 72(1), 67-90.
[28] Tanaka, S. Takara, A. 2002. A study on threshold selection in POT analysis of extreme floods. The Extremes of the Extremes: Extraordinary Floods. Oxfordshire, UK: IAHS Publ. 299 - 304.
[29] Therneau, T. 2015 A Package for Survival Analysis in S. version 2.38. http://CRAN.R-project.org/package=survival.

[30]Turnbull, B.W. 1976. The Empirical Distribution Function with Arbitrary Grouped, Censored and Truncated Data. Journal of the Royal Statistical Society: Series B., 69.