MODEL FOR DISTRIBUTION OF WAGES

MICHAL VRABEC, LUBOŠ MAREK
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill Sq. 4, Prague, Czech Republic
e-mail: vrabec@vse.cz, marek@vse.cz

Abstract

This article deals with estimation of the probability distribution of wages in the Czech Republic. We work with annual time series data over the years 1995 – 2014. Data are available for the entire Czech Republic and are further divided by gender, age (3 groups), regions (14 groups) and education (6 groups). The lognormal distribution is mostly used in modelling the wage distribution, To specify the future wage density function, it is necessary to know the values of its parameters. The estimation of parameters is normally based on past empirical observations. However, if we want to predict the future density function, we do not have any usable data for the process of parameter estimation. In the paper we propose a procedure to solve this problem. Because we know the history of the parameters, we apply a trend analysis to their historical time series. When we obtain a suitable model, we are able to estimate the future values of the parameters. Then we can specify the future wage density function. The entire analysis is performed in the EasyFit software.

Key words: *probability distribution, wage distribution model, lognormal distribution.*

1. Introduction

One of the most discussed statistical characteristic is the average wage. There is an ongoing debate about the suitability of the average as a measure of the wage level. There are proposals to replace the average by median, trimmed mean, Winsdorised mean etc. In our opinion, to solve this problem, it is necessary to work with the entire wage distribution.

All currently used methods describe the current state only. We want to construct a model of the future shape of the probability density. This model could be used for further analysis - hypothesis testing, confidence intervals, etc. Therefore, the main aim of our article is the building of the future shape of the wage density function.

We work with an interval frequency distribution table of wages. Our data are observed over years 1995 – 2015. We work with very detailed data. Our interval frequency tables have the interval-length of 500 CZK. The sample size is great – over two million observations. Data are available for the entire Czech Republic, and are further divided by gender, age (3 groups), regions (14 groups) and education (6 groups). In the first step, we create a model for year 2014 which is based on years 1995 – 2013. In the second step, we build a forecast model for the year 2015 (data from 2015 are not yet available). We use time series analysis to estimate the parameters of the future distribution. The choice of the theoretical distribution and parameter estimates are performed in the EasyFit program, version 5.5. The whole data preparation was made in MS Excel. The EasyFit program has a great offer of probability distributions. The number of included continuous distributions is 65. Similar problems solve authors Bartošová and Longford (2014) or Malá (2012; 2015). None of them, however, did not have such a large dataset and did not test such a large number of probability distributions.

2. Data set

Our data are in the form of an interval frequency distribution table and are obtained from the Czech wage and personnel consultant firm Trexima, s. r. o. (http://www.trexima.cz). A small subset is at Table 1. The length of interval is 500 CZK.

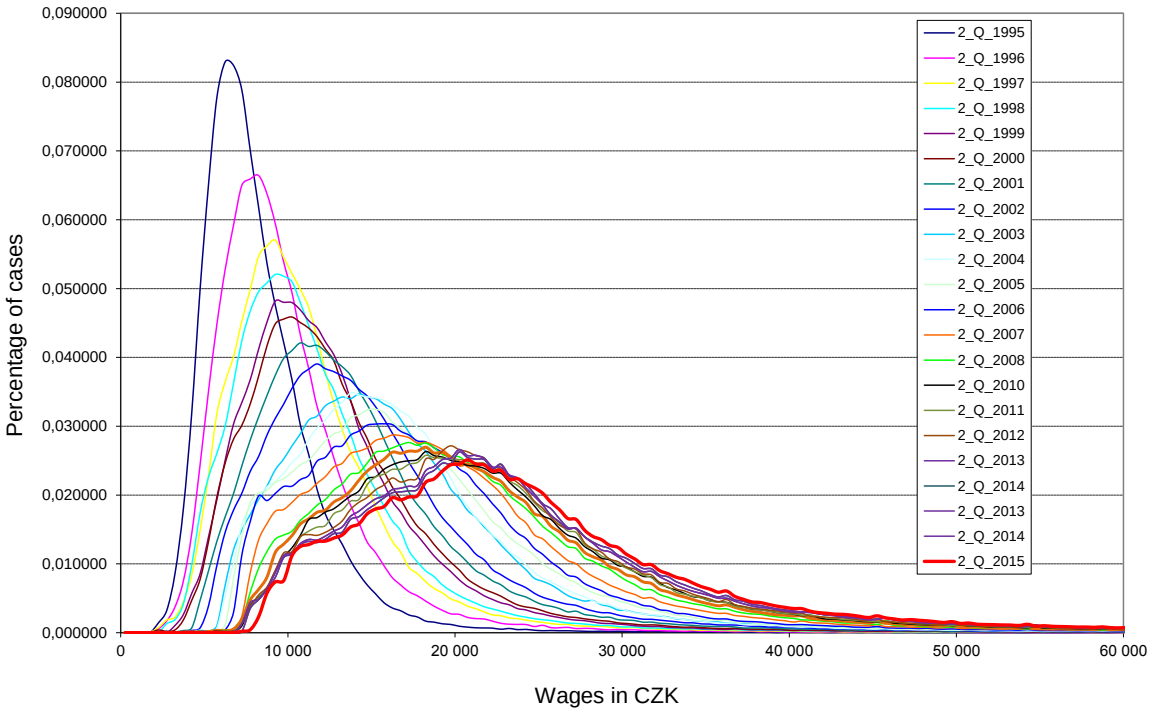
Table 1: Example of the data set

Lower bound	Upper bound	Frequency (Number of cases)
…	…	…
25,000 CZK	25,500 CZK	42,504
25,500 CZK	26,000 CZK	39,710
26,000 CZK	26,500 CZK	37,522
26,500 CZK	27,000 CZK	34,836
…	…	…

Source: the authors based on the data from Trexima.

The frequency polygons (empirical counterparts of density functions) for different time periods ranging from Q2/1995 to Q2/2015 on a quarterly basis have the form that is displayed in Figure 1.

Figure 1: Frequency polygons for each quarter from Q2/1995 to Q2/2015



Source: the authors.

We can see that the basic statistical characteristics are changing over time – mean and variability are growing, skewness is more significant and kurtosis keeps getting smaller. More in articles of Marek (2010; 2013).

We know the basic statistical characteristics for each year and they are exemplified in Table 2.

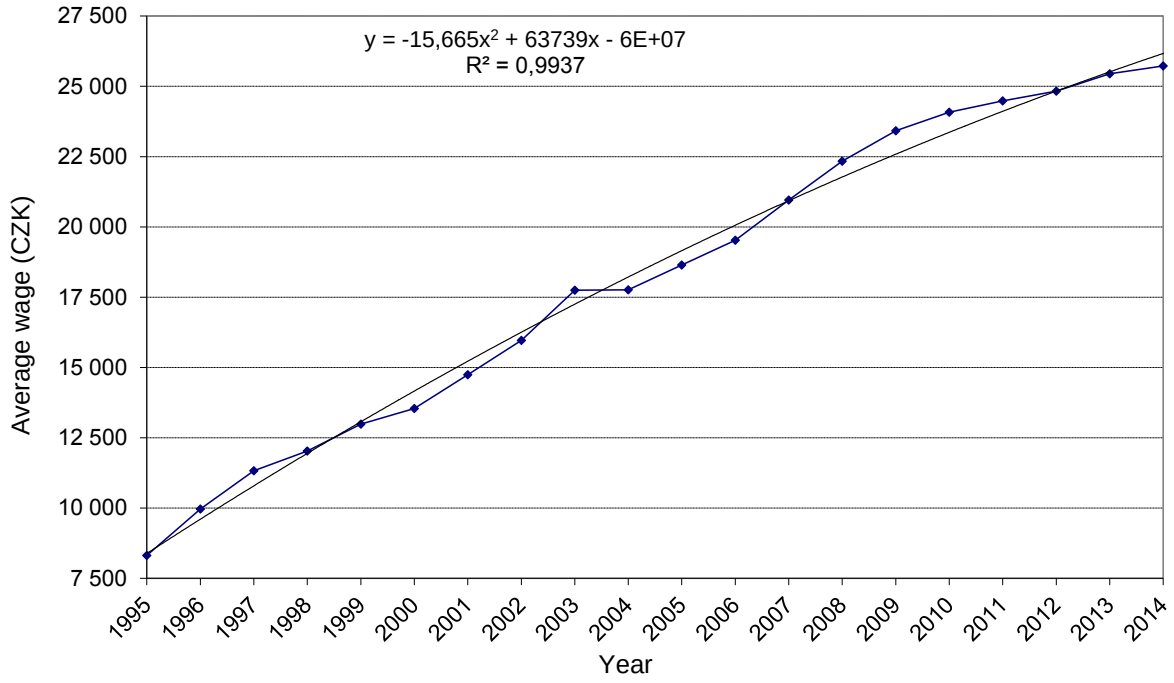
Table 2: Example of the descriptive characteristics available

Size of the sample	Number of firms	Average wage (CZK)	Standard deviation (CZK)	10 % quantile (CZK)	25 % quantile (CZK)	Median (CZK)	75 % quantile (CZK)	90 % quantile (CZK)	Fund of working days
2,098,854	17,600	26,369	19,903	12,978	17,290	22,658	29,566	40,162	157

Source: Trexima (www.Trexima.cz).

We used the characteristics above for the calculation of the distribution parameters. The next graph in Figure 2 shows the average wage in Czech Republic over past years. We have data for many similar graphs, but the scope of this article is limited. The graph also shows the fitted curve for the average wage (variable y) regressed on individual years (variable x). Also the corresponding R squared is displayed.

Figure 2: Average monthly wage between 1995 and 2014



Source: the authors.

3. Data Analysis

3.1. Kolmogorov-Smirnov Test

The quality of the model was tested using the Kolmogorov-Smirnov test. We tested the null hypothesis "H_0: the data follow the specified distribution" against the alternative hypothesis "H_1: the data do not follow the specified distribution".

The test is based on the empirical distribution function

$$F_n(x) = \frac{1}{n}\left[\text{Number of observations} \le x\right], \tag{1}$$

where $0 < x < \infty$ and n is the number of observations.

The Kolmogorov-Smirnov statistic (D) is based on the largest vertical difference between the theoretical and the empirical cumulative distribution function

$$D = \max_{1 \le i \le n} \left(F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i) \right). \tag{2}$$

More information can be found in Malá (2013) or in Marek and Vrabec (2013). Table 3 shows the first ten distributions for the 5 past years. The results are ordered by the Kolmogorov-Smirnov statistic. For each test we computed the P-value. The summary of P-values is given in Table 4. There are many viable distributions for our model.

Table 3: Rank of the fitted distributions according to the values of the Kolmogorov-Smirnov statistic (KSS)

Rank	Year 2010		Year 2011		Year 2012		Year 2013		Year 2014	
	Distribution	KSS	Distribution	KSS	Distribution	KSS	Distribution	KSS	Distribution	KSS
1	Log-Logistic (3P)	0.0168	Log-Logistic (3P)	0.0167	Log-Logistic (3P)	0.0220	Log-Logistic (3P)	0.022	Log-Logistic (3P)	0.0211
2	Burr	0.0228	Burr	0.0237	Log-Gamma	0.0380	Burr	0.027	Log-Gamma	0.0352
3	Frechet (3P)	0.0298	Frechet (3P)	0.0323	Frechet (3P)	0.0388	Log-Gamma	0.036	Frechet (3P)	0.0382
4	Pearson 6 (4P)	0.0302	Pearson 6 (4P)	0.0334	Pearson 5 (3P)	0.0403	Frechet (3P)	0.038	Lognormal (3P)	0.0388
5	Pearson 5 (3P)	0.0305	Pearson 5 (3P)	0.0334	Pearson 6 (4P)	0.0403	Lognormal (3P)	0.039	Lognormal	0.0395
6	Log-Gamma	0.0332	Pearson 6	0.0337	Gen. Gamma (4P)	0.0408	Lognormal	0.040	Pearson 5 (3P)	0.0397
7	Lognormal (3P)	0.0360	Log-Gamma	0.0367	Lognormal (3P)	0.0408	Pearson 5 (3P)	0.040	Pearson 6 (4P)	0.0397
8	Log-Pearson 3	0.0363	Log-Pearson 3	0.0389	Lognormal	0.0414	Pearson 6 (4P)	0.043	Inv. Gaussian (3P)	0.0437
9	Lognormal	0.0367	Lognormal (3P)	0.0393	Log-Pearson 3	0.0458	Inv. Gaussian (3P)	0.044	Log-Pearson 3	0.0439
10	Pearson 6	0.0384	Lognormal	0.0399	Inv. Gaussian (3P)	0.0469	Log-Pearson 3	0.045	Fatigue Life (3P)	0.0464

Source: the authors.

In Table 4 we show the significance of the Kolmogorov-Smirnov statistics for the year 2014 and for the log-logistic three-parameter distribution as an example.

Table 4: Kolmogorov-Smirnov for Log-Logistic (3P)

Sample size	200	Various critical levels of significance (α) considered				
Statistic	0.02117	α	0.2	0.1	0.05	0.02	0.01
P-value	0.99998	Critical value	0.07587	0.08648	0.09603	0.10734	0.11519
Rank	1	Reject?	No	No	No	No	No

Source: the authors.

From Table 3 we can see that the best result is achieved for the 3-parametric Log-Logistic distribution. Alternatively, once could use Burr, Log-Gamma, Frechet (3P), Pearson 5 (3P), Pearson 6 (4P), Lognormal (3P), Gen. Gamma (4P), Lognormal, Inv. Gaussian (3P) and Log-Pearson 3 distributions. Probability density function and cumulative distribution function of the Log-Logistic (3P) distribution have the form

$$f(x) = \frac{\alpha}{\beta} \left(\frac{x-\gamma}{\beta} \right)^{\alpha-1} \left(1 + \left(\frac{x-\gamma}{\beta} \right)^{\alpha} \right)^{-2}, \tag{3}$$

$$F(x)=\left(1+\left(\frac{\beta}{x-\gamma}\right)^{\alpha}\right)^{-1}, \tag{4}$$

for $\gamma \le x < +\infty$, parameters α – continuous shape parameter ($\alpha > 0$), β – continuous scale parameter ($\beta > 0$) and γ – continuous location parameter ($\gamma \equiv 0$ yields the two-parameter Log-Logistic distribution). The last parameter gamma allows us to shift the distribution along the *x* axis. When $\gamma = 0$, we obtain a two-parameter Log-Logistic distribution with the probability density function

$$f(x)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{-2}, \tag{5}$$

and cumulative distribution function

$$F(x)=\left(1+\left(\frac{\beta}{x}\right)^{\alpha}\right)^{-1}. \tag{6}$$

3.2. Anderson-Darling Test

Program EasyFit offers the Anderson-Darling test for other testing. This test is very similar to the KS test and works with the difference of squares between theoretical and empirical distribution function

$$n\int_{-\infty}^{\infty}\left(F_n(x)-F(x)\right)^2 w(x)\,dF(x), \tag{7}$$

where $F(x)$ is theoretical distribution function (the hypothesized distribution), $F_n(x)$ is empirical (sample) cumulative distribution function, and $w(x)$ is weighting function. For $w(x)=1$ it is called Cramér-von Mises statistic. The Anderson-Darling test developed in 1954 is based on the distance

$$A=n\int_{-\infty}^{\infty}\frac{\left(F_n(x)-F(x)\right)^2}{F(x)\left(1-F(x)\right)}\,dF(x), \tag{8}$$

which is obtained when the weight function is

$$w(x)=\left[F(x)\left(1-F(x)\right)\right]^{-1}. \tag{9}$$

In the EasyFit program it is defined as

$$A^2=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(X_i)+\ln\left(1-F(X_{n-i+1})\right)\right]. \tag{10}$$

When we use this statistic, the most suitable distribution is again the Log-Logistic one, followed by the Power Function. The probability density function of the Power Function is

$$f(x)=\frac{\alpha(x-a)^{\alpha-1}}{(b-a)^{\alpha}}. \tag{11}$$

Cumulative distribution function is

$$F(x) = \left(\frac{x-a}{b-a} \right)^{\alpha}, \tag{12}$$

for $a \leq x \leq b$, α - continuous shape parameter ($\alpha > 0$), a, b - continuous boundary parameters ($a > b$).Given the results we prefer Log-Logistic (3P) distribution.

The estimated parameters for the Log-Logistic distribution are in Table 5 (we show only 12 last years). The rows with word "predict" are parameter forecasts. We obtain these forecasts as the results of the time series analysis. This method was successfully used in other papers (e.g. Marek, 2013).

Table 5: Parameter estimates

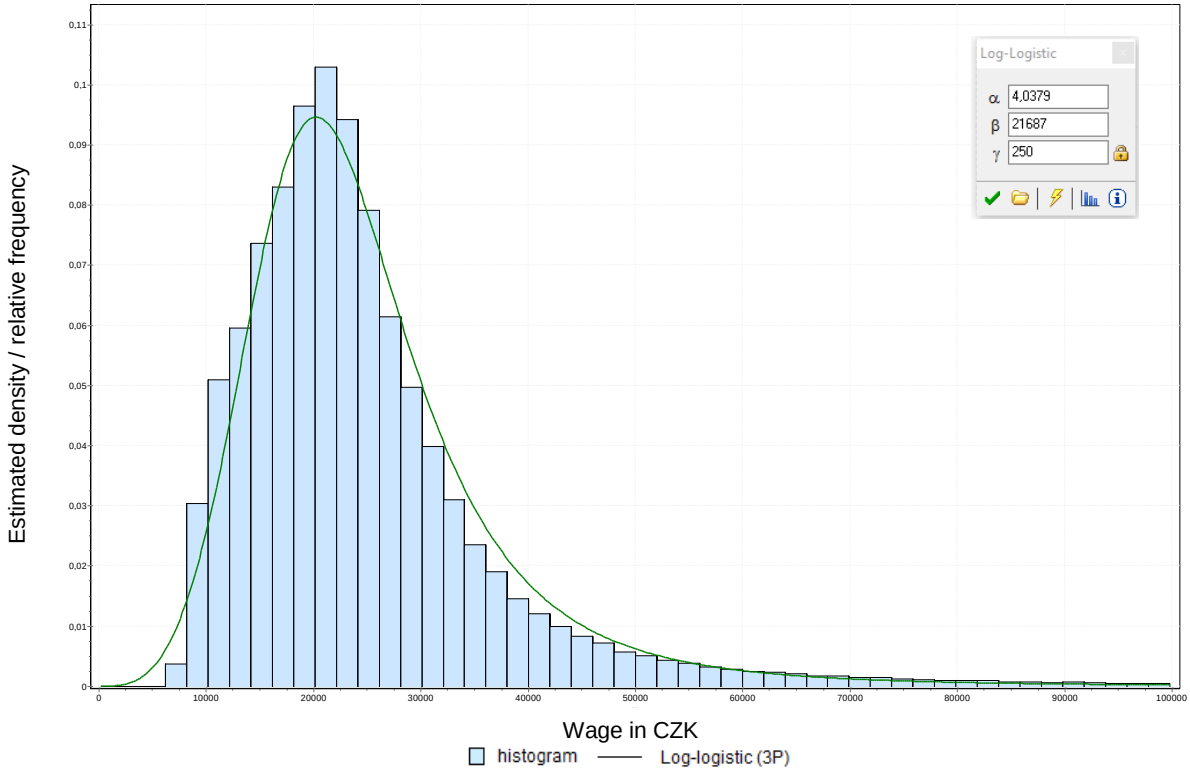
Year	a	b	α
2004	4.2399	15,402	250
2005	4.0996	16,064	250
2006	4.0961	16,711	250
2007	4.1279	17,980	250
2008	4.1190	19,115	250
2009	4.0946	19,781	250
2010	4.1208	20,372	250
2011	4.1114	20,611	250
2012	4.0893	20,930	250
2013	4.0439	21,380	250
2014 prediction	4.0541	22,675	250
2014	4.0379	21,687	250
2015 prediction	4.0373	23,014	250

Source: the authors.

3.3 Visualization

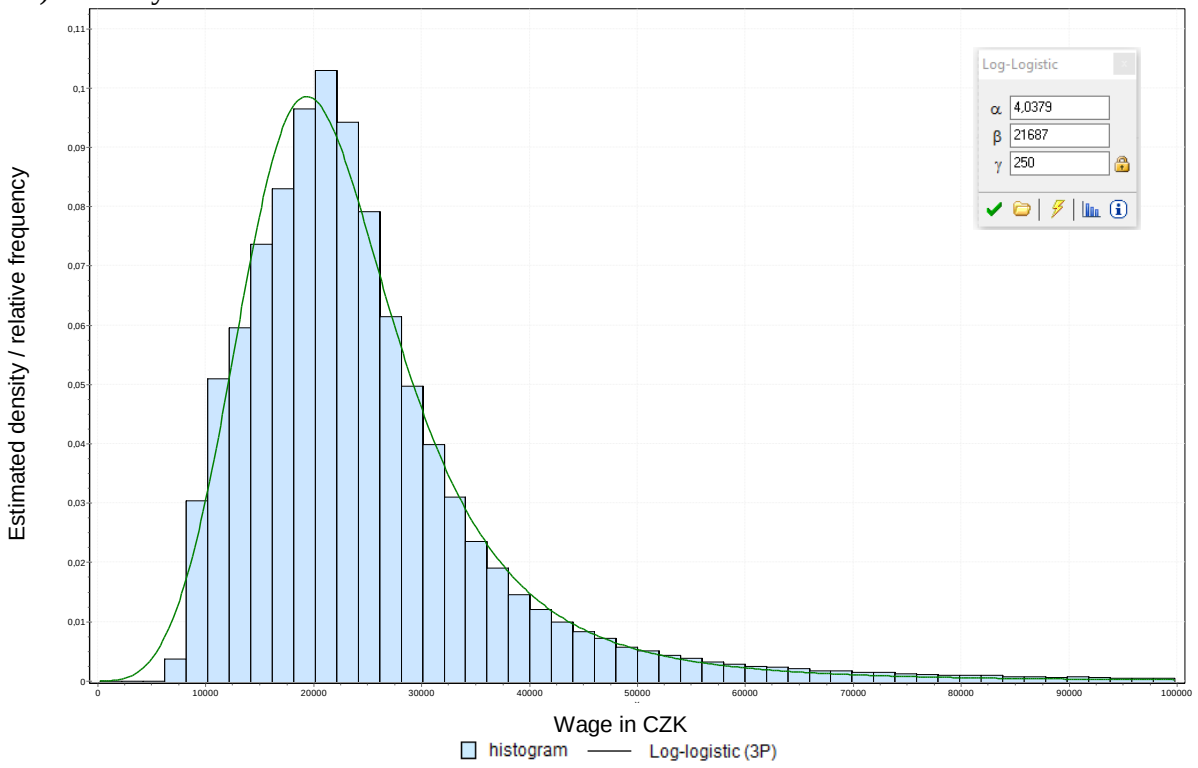
Comparison of the empirical data and theoretical density function for year 2014 is in Figure 3. The important fact is that this model is built from data over years 1995 – 2013. Therefore, the model for 2014 is the forecast of the density function and we compare this model with the reality. P-value for the KS test is 0.0481. A very similar graph can be seen in Figure 4. However, the theoretical density function is computed from the 2014 data only. P-value for the KS test is 0.0478 – we can confirm that both models are comparable, because the parameter estimation and P-value are very similar. The last Figure 5 shows the estimate of the density function for year 2015. This model is computed from data over years 1995 – 2014. This way allows us to compute the estimates of the density function for next years.

Figure 3: Comparison of empirical data and theoretical density function (model from the 1995 – 2013 data) for the year 2014



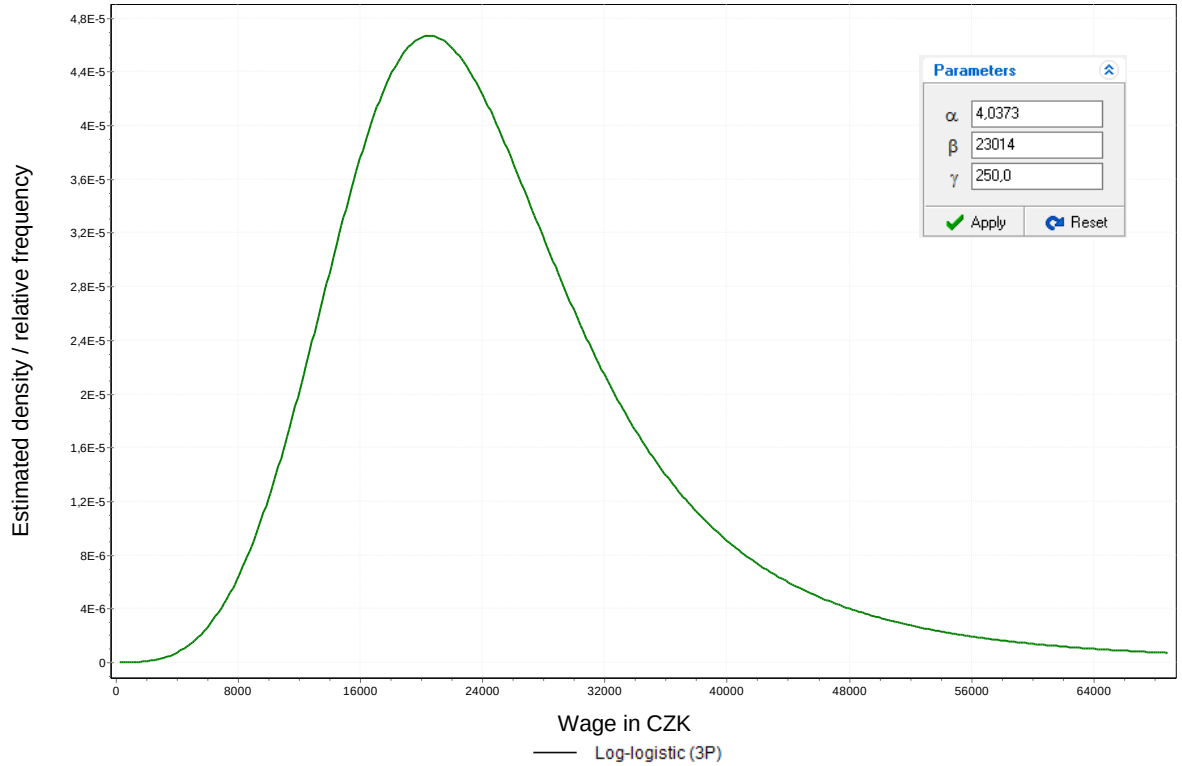
Source: the authors.

Figure 4: Comparison of empirical data and theoretical density function (model from the 2014 data) for the year 2014.



Source: the authors.

Figure 5: Density estimate for the year 2015.



Source: the authors.

4. Conclusions

Based on the results we can make the following conclusions: Commonly used log-normal distribution gives good results (not included in article), but it is possible to find densities with better results. The best model is the Log-Logistic distribution with 3 parameters. This distribution is best for each year. Therefore we recommend it as a probability model for the distribution of wages in the Czech Republic. The quality of all the models is very good. We tested them at level of $\alpha = 0.05$, we applied Kolmogorov-Smirnov test and Anderson-Darling test. All tests confirmed goodness between empirical and theoretical values. We solved the problem of estimating the future values of parameters using the time series analysis. We find the appropriate trend function and build the forecasts. This way allows us to build the model not only for past or current wage data but also for the next years. It would not have been possible with the conventional approach. The model for the next year allows us to compute confidence intervals, provide hypothesis tests, other characteristics etc.

Acknowledgements

This paper was written with the support of the Czech Science Foundation project No. P402/12/G097 „DYME – Dynamic Models in Economics“.

References

[1] BARTOŠOVÁ, J. LONGFORD, N. T. 2014. A study of income stability in the Czech Republic by finite mixtures. In Prague Economic Papers, 2014, vol. 23, iss. 3, pp. 330-348.

[2] MALÁ, I. 2015. Vícerozměrný pravděpodobnostní model rozdělení příjmů českých domácností. In Politická ekonomie, 2015, vol. 63, iss. 7, pp. 895-908.

[3] MALÁ, I. 2012. Modelling of conditional distributions of wages in the Czech Republic. Research In Journal of Economics, Business and ICT, 2012, vol. 6, pp. 1-5.

[4] MALÁ, I. 2013. Statistické úsudky. Prague : Professional Publishing. ISBN 978-80-7431-127-7.

[5] MAREK, L. 2010. Analýza vývoje mězd v ČR v letech 1995 – 2008. In Politická ekonomie, 2010, vol. 58, iss. 2, pp. 186-206.

[6] MAREK, L. 2013. Some aspects of average wage evolution in the Czech Republic. In The 7th International Days of Statistics and Economics Conference Proceedings, Prague, 19 – 21 Sep 2013. Slaný : Melandrium. ISBN 978-80-86175-87-4, pp. 947-958.

[7] MAREK, L., VRABEC, M. 2013 Probability models for wage distributions. In Mathematical Methods in Economics 2013. Jihlava, 11 – 13 Sep 2013. College of Polytechnics, 2013. ISBN 978-80-87035-76-4, pp. 575-581.