EX-POST VERIFICATION OF PREDICTION MODELS OF WAGE DISTRIBUTIONS

LUBOŠ MAREK, MICHAL VRABEC

University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill Sq. 4, Prague, Czech Republic
email: marek@vse.cz, vrabec@vse.cz

PETR BERKA

University of Economics, Faculty of Informatics and Statistics, Dept. of Information and Knowledge Engineering,
W. Churchill Sq. 4, Prague, Czech Republic
and
University of Finance and Administration, Dept. of Computer Science and Mathematics,
Estonska 500, Prague, Czech Republic
email: berka@vse.cz

Abstract

Our paper deals with the ex-post verification of models of wage distributions designed to predict wage distributions in the last three years. We will use the prediction results of Lognormal, Lognormal (3p), Johnson SB, Log-Logistic, Log-Logistic (3p) and Normal Mixture distributions and compare them with the empirical distribution from the period 2015-2017. The selection of the used distributions is based on the wage distribution models for the years 2000-2014. Our results show, that the best (and comparable) results can be obtained using three-parameter Log-logistic distribution and Normal Mixture distribution with two components. These results confirm our expectation that due to the fact, that empirical wage distribution becomes less smooth over time, a mixture model should be preferred for the future.

***Keywords**: wage distribution, prediction, model verification*

***JEL Codes**: C22, E24*

1. Introduction

Statistical analysis of the development of the wage and income distribution is a crucial precondition for economic modeling of the labor market processes. There is an ongoing debate how to measure the wage level. The mostly used average wage loses its expressiveness as the wage distribution becomes less smooth and exhibits higher variance over the years. There are proposals to replace the average by median, and/or to consider additional characteristics like variability or percentiles. In our opinion, it is necessary to work with the entire wage distribution.

Various probabilistic distributions can be used to model the empirical wage distribution. And a good model that is able to make good predictions of the future wage distributions is necessary for various socio-economic considerations. To assess the quality of different models we performed their ex-post verification, where models that have been created from the historical data starting in the year 1995 and applied to make predictions of wage distributions for the years 2015-2017 are confronted with the true empirical wage distributions in 2015-2017.

The rest of the paper is organized as follows: section 2 describes the used data, section 3 shows the distributions used for modelling, section 4 presents the models and discusses their quality and section 5 concludes the paper.

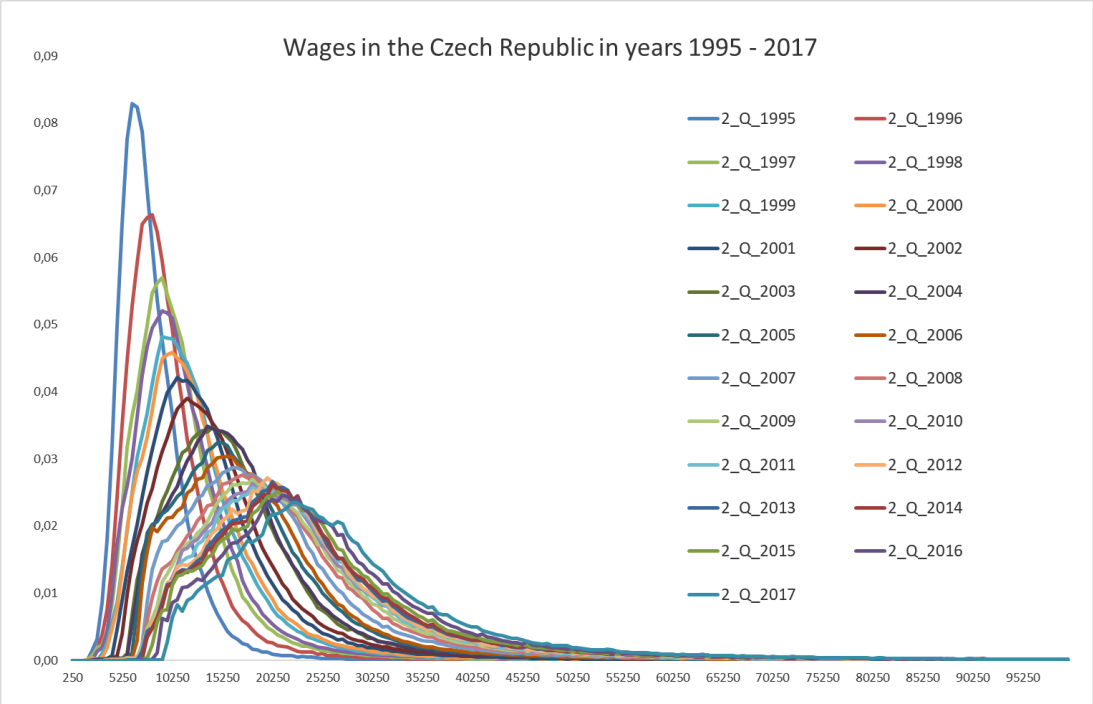
2. Wage Data

We work with time series of wages in Czech Republic covering the years 1995 - 2017. 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). The annual data are reported in quarterly units; our study observes the average wages in the second quarter of each year as we consider the months April-June to be the most stable period w.r.t wages of the year. The amount of the data gradually increases from the sample size of about 300 000 in 1995 to more than two million in 2017. This increase is due to the improved process of collecting the wage data by the Trexima company. The wage values are divided into intervals with widths of 500 CZK. Table 1 gives basic characteristics of the data and Fig. 1 visualizes the distribution of wages from these data. The curves shown in the graph are produced by connecting points of frequency for 500 CZK intervals, there is no method of empirical distribution smoothing applied. The figure clearly shows that the empirical wage distributions:

- are bounded by minimum wages (we also bound the empirical wage distributions by 100 000 CZK as there were very few employees with wages above this value in the data),
- are skewed, and
- change over time as the average value increases, the variability increases and the distributions become less smooth (see also Marek, 2010).

So modeling wage distribution of late 2010th is more difficult and more challenging than modeling wage distribution of late 1990th.

Figure 1: Wages in the Czech Republic in years 1995 - 2017



Source: The author's work

Table 1: Basic characteristics of the used data

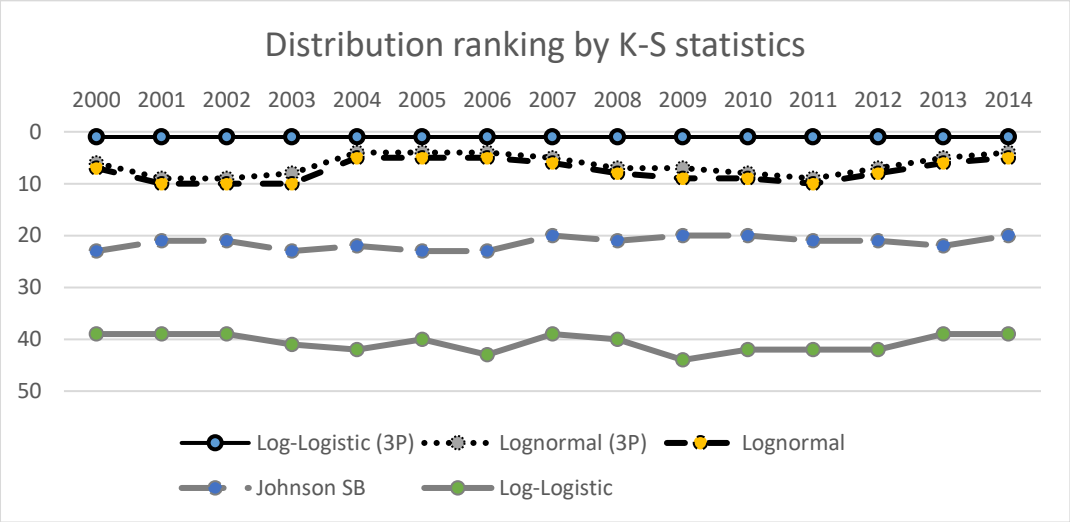
Year	Number of employees	Average	Std. dev.	Coeff. Off variation	D1	Q1	Median	Q3	D9	Mode
1995	321,277	8,311	4,133	0.50	4,879	5,963	7,500	9,691	12,314	6,920
1996	405,138	9,962	5,393	0.54	5,645	7,047	8,956	11,505	14,748	6,960
1997	622,505	11,322	6,490	0.57	6,178	7,910	10,171	13,083	16,774	8,750
1998	953,691	12,026	8,261	0.69	6,287	8,114	10,563	13,801	17,911	8,450
1999	1,024,898	12,982	8,262	0.64	6,894	8,859	11,506	14,911	19,499	6,760
2000	1,053,536	13,541	9,651	0.71	6,981	9,077	11,860	15,570	20,435	6,760
2001	1,075,875	14,743	10,372	0.70	7,693	9,870	12,901	16,794	22,234	4,740
2002	1,107,991	15,964	12,994	0.81	8,181	10,564	13,857	18,058	24,003	5,372
2003	1,230,282	17,748	13,504	0.76	9,143	11,829	15,519	20,070	26,271	6,520
2004	1,680,800	17,759	13,062	0.74	9,185	12,073	15,789	20,168	26,143	6,296
2005	1,818,369	18,640	13,796	0.74	9,371	12,403	16,432	21,376	27,754	6,715
2006	1,976,571	19,526	17,696	0.91	9,710	12,882	17,143	22,192	28,828	7,018
2007	2,059,416	20,953	18,055	0.86	10,381	13,659	18,185	23,602	31,257	7,552
2008	2,079,765	22,338	20,714	0.93	11,060	14,583	19,267	25,094	33,306	7,600
2009	1,933,772	23,418	19,014	0.81	11,681	15,339	20,138	26,241	35,093	7,552
2010	1,956,702	24,077	19,316	0.80	12,084	15,778	20,753	27,009	36,143	7,600
2011	1,973,468	24,484	24,802	1.00	12,199	15,996	21,020	27,225	36,677	7,600
2012	1,999,934	24,829	20,109	0.81	12,255	16,281	21,319	27,583	37,328	7,552
2013	2,015,903	25,448	20,564	0.81	12,416	16,595	21,779	28,322	38,598	7,600
2014	2,056,133	25,728	19,612	0.76	12,570	16,821	22,074	28,794	39,182	7,995
2015	2,098,854	26,369	19,903	0.75	12,978	17,290	22,658	29,566	40,162	8,635
2016	2,119,396	27,668	20,478	0.74	13,944	18,391	23,757	30,963	42,026	9,275
2017	2,185,573	29,166	20,749	0.71	14,982	19,547	25,135	32,610	44,334	10,296

Source: The author's work

3. Used Distributions

We used Log-normal, Log-normal (3p), Johnson SB, Log-Logistic, Log-Logistic (3p) and Normal Mixture distributions to model the wage distributions. This selection was based not only on the fact that these distributions are widely used to model wage distributions, but also on our modeling experiments of wage distributions for the period 2000-2014. Fig. 2 summarizes the results of these experiments. Here each curve shows the rank for the used distributions (except Normal Mixture) assigned according to the value of the Kolmogorov-Smirnov statistics to more than 50 probabilistic distributions available in the EasyFit system. The average rank for three-parameter Log-Logistic distribution was 1.0 (this distribution was always the best one), the average rank for three-parameter Log-Normal distribution was 6.4, the average rank for two-parameter Log-Normal distribution was 7.5, the average rank for Johnson(SB) distribution was 21.4 and the average rank for two-parameter Log-Logistic distribution was 40.7. Among other distributions reported in literature as suitable to model wage distributions, the Dagum distribution (Dagum, 2008), used e.g. by Matějka and Duspivová (2013) had the average rank 53.2 and therefore was not included in the prediction experiments.

Figure 2: Ranking of distributions based on Kolmogorov-Smirnov statistics 2000 - 2014



Source: The author's work

3.1 Log-normal distribution

Log-normal distribution (sometimes also called Galton distribution) is a continuous probability distribution of a random variable whose logarithm is normally distributed. The parameters of the distribution are:

σ - continuous parameter ($\sigma > 0$),

μ - continuous parameter,

γ - continuous location parameter ($\gamma = 0$ yields the two-parameter Lognormal distribution) and the domain is $\gamma < x < +\infty$.

The three-parameter Log-normal distribution has probability density function

$$f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^2\right)}{(x-\gamma)\sigma\sqrt{2\pi}} \tag{1}$$

and cumulative distribution function

$$F(x) = \Phi\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right) \tag{2}$$

The two-parameter Log-normal distribution has probability density function

$$f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x-\mu)}{\sigma}\right)^2\right)}{x\sigma\sqrt{2\pi}} \tag{3}$$

and cumulative distribution function

$$F(x) = \Phi\left(\frac{\ln(x-\mu)}{\sigma}\right) \tag{4}$$

where Φ is the Laplace Integral.

3.2 Johnson SB distribution

Johnson distributions (Johnson, 1949) are based on a transformation of the standard normal variable. Given a continuous random variable X whose distribution is unknown and is to be approximated, Johnson proposed three normalizing transformations having the general form:

$$Z = \gamma + \delta f\left(\frac{X-\xi}{\lambda}\right), \tag{5}$$

where f (.) denotes the transformation function, Z is a standard normal random variable, γ and δ are shape parameters (δ > 0), λ is a scale parameter (λ > 0) and ξ is a location parameter.

We will consider the Johnson SB distribution where

$$Z = \gamma + \delta \ln\left(\frac{X-\xi}{\xi+\lambda-X}\right). \tag{6}$$

The domain of this distribution is 0 < y < 1, the density function is

$$f(y) = \frac{\delta}{\sqrt{2\pi}}\frac{1-y}{y}\exp\left(-\frac{1}{2}\left(\gamma + \delta \ln\left(\frac{y}{1-y}\right)\right)^2\right), \tag{7}$$

and the cumulative distribution function is

$$F(y) = \Phi\left(\gamma + \delta\, ln\left(\frac{y}{1-y}\right)\right), \tag{8}$$

where $y = \frac{x-\xi}{\lambda}$, and Φ is the Laplace integral.

3.3 Log-Logistic distribution

Log-logistic distribution is the probability distribution of a random variable whose logarithm has a logistic distribution. The parameters of the distribution are

α - continuous shape parameter ($\alpha > 0$),

β - continuous scale parameter ($\beta > 0$),

γ - continuous location parameter ($\gamma = 0$ yields the two-parameter Log-Logistic distribution)

and the domain $\gamma \le x < +\infty$.

The three-parameter Log-logistic distribution has probability density function

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2} \tag{9}$$

and cumulative distribution function

$$F(x) = \left(1+\left(\frac{\beta}{x-\gamma}\right)^{\alpha}\right)^{-1}. \tag{10}$$

The two-parameter log-logistic distribution has probability density function

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{-2} \tag{11}$$

and cumulative distribution function

$$F(x) = \left(1+\left(\frac{\beta}{x}\right)^{\alpha}\right)^{-1}. \tag{12}$$

3.4 Normal Mixture distribution

The probability density for a general model of a normal mixture can be written as

$$f(x) = \sum_{i=1}^{n} p_i g_i(x), \tag{13}$$

where $g_i(x)$ is the probability density of normal distribution

$$g_i(x) = \frac{1}{\lambda_i\sqrt{2\pi}}\exp\left(-\frac{(x-\theta_i)^2}{2\lambda_i^2}\right), \tag{14}$$

n is the number of components in the mixture and p is the vector of weights, for which

$$0 < p_i < 1, \ \forall i, \ \sum_{i=1}^{n} p_i = 1. \tag{15}$$

4. Ex-post verification of wage distribution models

We used the distributions described in section 3 to model the wage distributions. Models based on all these six distributions have then been used to predict the empirical wage distributions for the years 2015-2017. To perform the ex-post verification of models of wage distributions we used following setting of our experiments:

- wage data for the period 1995-2016 have been used to predict the parameters of the distributions for the year 2017 (we will denote this as Prediction1),
- wage data for the period 1995-2015 have been used to predict the parameters of the distributions for the years 2016 and 2017 (we will denote this as Prediction2),

- wage data for the period 1995-2014 have been used to predict the parameters of the distributions for the years 2015, 2016 and 2017 (we will denote this as Prediction3),
- distributions based on predicted parameters have been compared with the empirical wage distribution in year 2017; we performed the Kolmogorov-Smirnov test testing the null hypothesis "H0: the data follow the specified distribution created using the predicted parameters" against the alternative hypothesis "H1: the data do not follow the specified distribution created using the predicted parameters".

In all these predictions, the parameters were predicted using linear trend.

When working with a single distribution, we created one model for each prediction, when working with a mixture, we created a mixture model with two components reflecting gender (male, female). Tables Tab.2 – Tab. 7 present the estimated parameters of the created models. Table 8 shows the quality of prediction for 2017 in terms of the Kolmogorov-Smirnov statistics and the rank of the model. A common expectation is that the more ahead a prediction is made, the less reliable it will be. So in our experiments we expected that the Prediction1 experiment will give the best results and the Prediction3 experiment will give the worst results. But this expectation was not confirmed by the values of the Kolmogorov-Smirnov statistics. Fig. 3 illustrates the fit of the respective model for Prediction1 (i.e. model created from the years 1995-2016 predicts for the year 2017). We used the SAS system, JMP and EasyFit programs for the computations.

Table 2: Parameters for three parameters Log-normal model

Prediction experiment		σ	μ	γ
Prediction 1	2017	0.442353	10.23054	250
Prediction 2	2016	0.444303	10.20235	250
	2017	0.445250	10.25424	250
Prediction 3	2015	0.444421	10.17861	250
	2016	0.445446	10.23256	250
	2017	0.446471	10.28651	250

Source: The author's work

Table 3: Parameters for two parameters Log-normal model

Prediction experiment		σ	μ
Prediction 1	2017	0.439073	10.23794
Prediction 2	2016	0.440871	10.21017
	2017	0.442182	10.26107
Prediction 3	2015	0.440880	10.18674
	2016	0.442288	10.23965
	2017	0.443696	10.29255

Source: The author's work

Table 4: Parameters for Johnson SB model

Prediction experiment		γ	δ	λ	ξ
Prediction 1	2017	2.722681	1.215317	51757.79	10794.11
Prediction 2	2016	2.793288	1.213620	53333.00	10577.51
	2017	2.667387	1.221829	28387.43	10815.66
Prediction 3	2015	2.860145	1.211931	54420.58	10451.72
	2016	2.729925	1.220617	27729.30	10698.08
	2017	2.599704	1.229303	1038.023	10944.45

Source: The author's work

Table 5: Parameters for three parameters Log-logistic model

Prediction experiment		α	β	γ
Prediction 1	2017	4.019687	24818.68	249.9934
Prediction 2	2016	4.002285	24194.83	249.9920
	2017	3.989048	24953.02	249.9919
Prediction 3	2015	4.003335	23685.16	249.9904
	2016	3.989206	24461.53	249.9902
	2017	3.975077	25237.90	249.9900

Source: The author's work

Table 6: Parameters for two parameters Log-logistic model

Prediction experiment		α	β
Prediction 1	2017	2.950575	19339.18
Prediction 2	2016	2.922208	18586.06
	2017	2.914340	19202.46
Prediction 3	2015	2.903506	17968.72
	2016	2.893694	18585.06
	2017	2.883883	19201.39

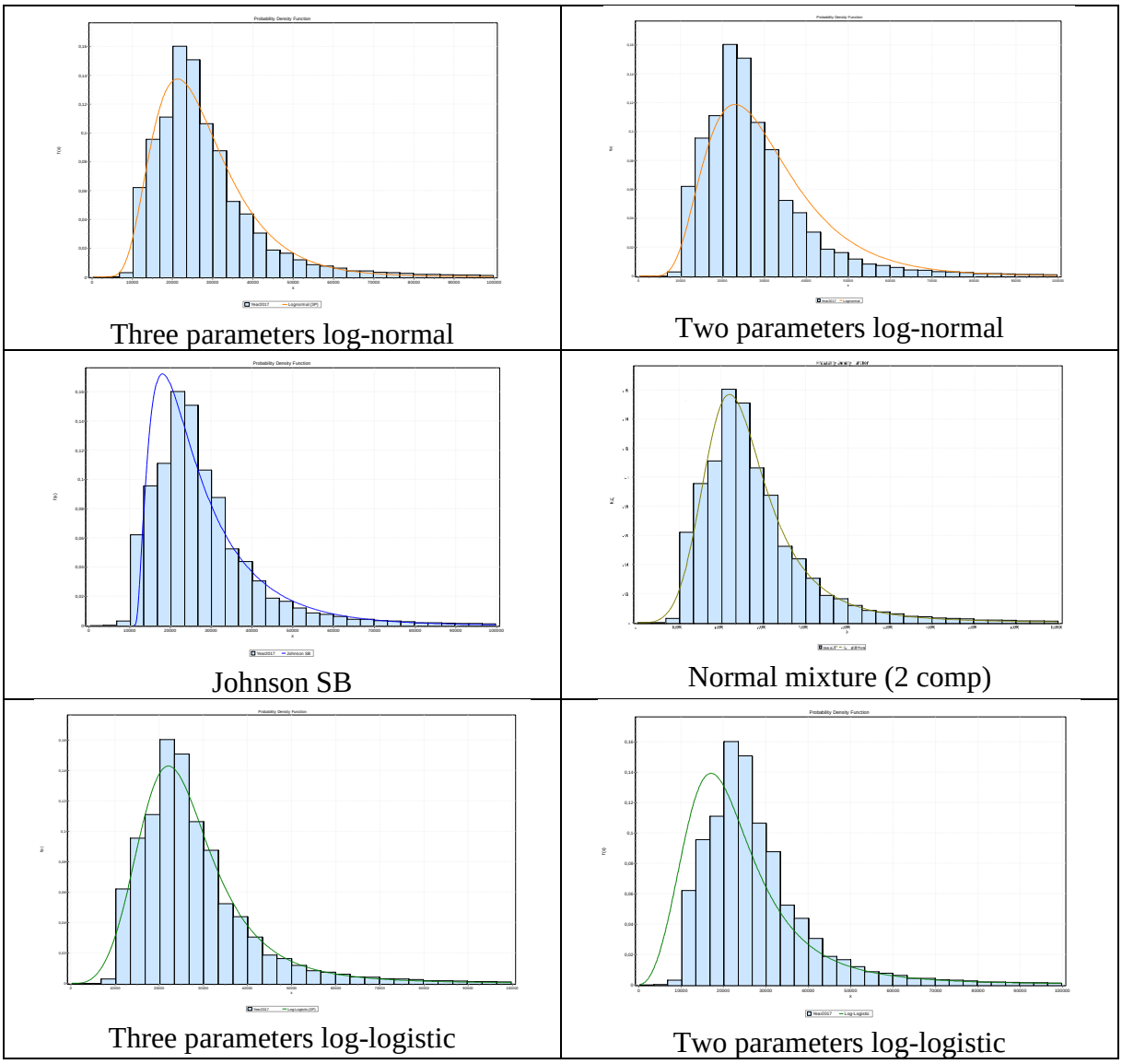
Source: The author's work

Table 7: Parameters for 2 components mixture model

parameter	2017	2016	2015
θ_1	23820.072	22657.212	21620.698
θ_2	46179.085	45101.877	43928.511
λ_1	7215.6347	7070.7166	6944.0011
λ_2	17454.832	17555.078	17558.683
p_1	0.8245363	0.8372576	0.8438542
p_2	0.1754637	0.1627424	0.1561458

Source: The author's work

Figure 3: Predicted wage distribution for 2017 based on models created from years 1995 – 2016



Source: The author's work

Table 8: Results of the Kolmogorov-Smirnov test

model	Prediction1		Prediction2		Prediction3	
	statistic	rank	statistic	rank	statistic	rank
Log-normal (3p)	0.03886	3	0.03809	3	0.03739	3
Log-normal	0.13290	5	0.15408	5	0.18248	5
Johnson SB	0.06210	4	0.06605	4	0.06982	4
Log-logistic (3p)	0.01900	1-2	0.02872	2	0.01830	1
Log-logistic	0.20191	6	0.21899	6	0.23095	6
Mixture (2comp)	0.01900	1-2	0.01961	1	0.02223	2

Source: The author's work

5. Conclusion

The paper presents a comparison of wage distribution predictions based on several probabilistic distributions. Although some previous work (Marek, Vrabec, 2013, Malá, 2013) has shown that using a single distribution to model wages need not to be optimal and that mixture models can achieve better results, our experiments show that Log-logistic distribution with three parameters and normal mixture model with two components are still comparable (see Table 8). The experiments also confirm the conclusions of Matějka and Duspivová (2013) that log-normal distribution gives bad results. But unlike their results, our initial experiments with modelling the wage distributions for the years 2000-2014 show poor performance of the Dagum distribution. The initial experiments also show significant difference in performance between Log-logistic distribution with three parameters and Log-logistic distribution with two parameters. While three-parameter Log-logistic distribution was found to be the best one (see also (Vrabec, Marek, 2016) for similar results), the two-parameter Log-logistic distribution was worse than e.g. three-parameter Log-normal distribution. The reason is that for wage distribution that is bounded by minimal (non-zero) wage, a third parameter is necessary to get a suitable model.

Our prediction models were created using the most simple way, by linear trend. More advanced methods like nonlinear trend or Holt exponential smoothing can be considered as well (and this can be a possible direction of our future work) but even the linear trend gave the values of R^2 varying from 0.9704 to 0.9937. When comparing the results of prediction experiments for any of the used model, we do not see any great difference in goodness of prediction for the year 2017 based on the data from the period 1995-2016 (Prediction1), based on the data from the period 1995-2015 (Prediction2) and based on the data from the period 1995-2014 (Prediction3). The reason could be the stable economic environment in the Czech Republic in the last years in which linear trend well fits the parameters of the wage distribution.

When working with a normal mixture model, we considered only two components (males, females) because the categorization by gender has high impact on wage distribution (see e.g. Bílková, 2012). But other natural components can be considered as well. Another examples of interpretable normal mixture models can be mixture model with three components using the age categories “below 30”, “30 to 50”, “above 50” or a mixture model with four component considering the education categories “basic”, “secondary”, “university”, “PhD”. Some initial experiments in this direction are reported in Marek, Vrabec (2013). The above mentioned categories can be used not only separately, but also simultaneously thus resulting in a mixture model with 2x3x4 components. Such a model will be of course computationally very complex and will require to process data on very detailed level but has a potential to fit well the empirical wage distribution using an interpretable mixture model. This will be our future research direction. We will also work with mixtures of other probabilistic distributions than a normal mixture model as presented in this paper.

Acknowledgements

This paper was written with the support of the Czech Science Foundation project No. P402/12/G097 „DYME – Dynamic Models in Economics“ and was processed with contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

[1] Bílková, D. 2012. Recent Development of the Wage and Income Distribution in the Czech Republic. Prague Economic Papers. vol. 21, no. 2, pp. 233–250

[2] Dagum, C. A. 2008. New Model of Personal Income Distribution: Specification and Estimation, In Modeling Income Distributions and Lorenz Curves, Economic Studies in Equality, Social Exclusion and Well-Being, Vol. 5, pp. 3–25.

[3] Johnson, N. J. 1949. Systems of frequency curves generated by methods of translation. Biometrika, 36(3/4), pp. 297-304.

[4] Malá, I. 2013. Použití konečných směsí logaritmicko-normálních rozdělení pro modelování příjmů českých domácností. Politická ekonomie. vol. 61, no. 3, pp. 356–372.

[5] Marek, L. 2010. Analýza vývoje mezd v ČR v letech 1995-2008. Politická ekonomie, Vol. 58, Issue 2, pp. 186–206.

[6] Marek, L., Vrabec, M. 2013. Model wage distribution - mixture density functions. Int. Journal of Economics and Statistics, Vol. 1, Issue 3, pp. 113-121.

[7] Matějka, M., Duspivová, K. 2013. The Czech wage distribution and the minimum wage impacts: an empirical analysis. Statistika, 93(2), pp. 61-75.

[8] Vrabec, M., Marek, L. 2016. Model for distribution of wages. In Proc. of the Applications of Mathematics and Statistics in Economics AMSE 2016, pp. 378-386.