MODELING INCOME DISTRIBUTION OF HOUSEHOLDS IN THE REGIONS OF THE CZECH REPUBLIC

JITKA BARTOŠOVÁ, VLADISLAV BÍNA
University of Economics, Prague, Faculty of Management,
Department of Exact Methods,
Jarošovská 1117/II, Jindřichův Hradec, Czech Republic
e-mails: bartosov@fm.vse.cz, bina@fm.vse.cz

Abstract

Analysis of incomes of inhabitants is in focus in all developed countries mainly because of the assessment and comparison of lining standards of inhabitants. Knowledge of the income distribution and its comparison using different socio-economic, demographic and spatiotemporal perspectives is a prerequisite for the quantitative evaluation of living standard, level of social welfare and equality in redistribution of the goods created in society.

The presented paper focuses on an analysis and comparison of the income distribution shape in the households of all 14 Czech regions. According to the fact that regions of the Czech Republic mutually differ not only in the extent of job opportunities, but also in demographic structure of inhabitants (age, education, etc.), we can expect also differences in the shape of regional income distributions. The aim of the paper is to construct suitable models of income distribution and identify regional differences. For this purpose, both parametric and nonparametric models of frequency distributions will be used. Nonparametric approaches are represented by Gaussian kernel estimates. Among parametric methods we can employ some of simple probability distribution models – most frequently two- and three-parametric lognormal models – or chose more demanding but also more accurate methods, i.e. estimate a model of income distribution based on finite mixtures of densities. This method is usually used for the modeling of distributions of random variables in heterogeneous populations and therefore in our case it comprises a better alternative.

***Keywords**: income distribution, kernel density estimate, finite mixtures of densities, regions of CR*

***JEL Codes**: C14, I31, J17*

1. Introduction

Incomes of households reflect the socio-economic level of society and are, therefore, permanently in the focus of the interest of economists of all developed countries. The results of analyses of income distribution based on various socio-economic, demographic and spatiotemporal viewpoints constitute a prerequisite of a quantitative assessment of living condition, equality and standard of social security. They also comprise a basis for adjusting social and tax policy, drawing up a government budget and other different decision making problems on the level of states and regions. Moreover, the direct link between the incomes and purchasing power of inhabitants provides a possibility to identify contemporary and predict future level of consumption of products of long- and short-term consumption.

Estimation of basic characteristics of income distribution and construction of suitable models of incomes and wages in Czech households (or individuals) is studied in many papers. Characterization of distribution proper-ties, its development and future prediction is a focus of papers, e.g., by Bílková (2012), or Marek (2017). Other papers (e.g. Bílková, 2017) discuss possibilities of estimation of empirical distribution using simple lognormal model with two or three parameters. This simple parametric estimate can be augmented by a more sophisticated and precise type – an estimation using finite mixtures of normal distributions (see, e.g., papers Bartošová and Longford, 2014; Marek and Vrabec, 2013).

2. Methods

It is well known that the empirical distribution of frequencies of annual incomes in case of the Czech households is multimodal and strongly skewed to the right side. Thus, it shows a heterogeneous character and it is not possible to approximate it with sufficient accuracy using a simple parametric model. Such estimate of an empirical distribution does not provide sufficient flexibility even in case when we employ more flexible models with more parameters. In general, it is more reasonable to use finite mixtures of densities providing a possibility of arbitrary precise approximation (see e.g. Bartošová and Longford, 2014; Malá, 2013, 2014). Empirical distribution of incomes has approximately lognormal distribution and therefore it is advantageous to use instead of direct estimate using mixtures of lognormal components to use a logarithmic transformation and estimate the transformed distribution using a mixture of normal components. Depiction of household incomes on logarithmic scale also emphasizes heterogeneity of empirical distribution in graphical outputs.

Now we will assume that the observed population represented by the log-transformed data from our sample is composed from K particular subpopulations (components) described using distributions with unknown parameters μ and σ. The density of observed random variable X is in general a finite mixture of K densities if

$$f^{(K)}\left(X;\mathbf{\Psi}^{(K)}\right)=\sum_{k=1}^{K} p_k^{(K)} f_k^{(K)}\left(\mathbf{x};\mathbf{\Theta}_k^{(K)}\right), \tag{1}$$

where $p_k^{(K)} > 0,\ k = 1,\dots,K,$ and $\sum_{k=1}^{K} p_k^{(K)} = 1$ are weights (marginal probabilities) of components, $f_k^{(K)}\left(\mathbf{x};\boldsymbol{\theta}_k^{(K)}\right)$ stand for densities of particular components and $\mathbf{\Psi}^{(K)} = (p_1^{(K)},\dots,p_{K-1}^{(K)},\boldsymbol{\theta}_1^{(K)},\dots,\boldsymbol{\theta}_K^{(K)})$ is a vector of unknown mixture parameters. In our case we will approximate logarithms of incomes using mixtures of normal densities and thus the vector of parameters is $\mathbf{\Psi}^{(K)} = (p_1^{(K)},\dots,p_{K-1}^{(K)},(\mu_1^{(K)},\sigma_1^{(K)}),\dots,(\mu_K^{(K)},\sigma_K^{(K)}))$. For more information, see, e.g., monograph of McLachlan and Peel (2000).

For estimation of parameters of mixture, the well-known EM algorithm is used (Dempster *et al.*, 1977). EM-algorithm works in the principle of Bayesian clustering which ranks among the non-hierarchical (optimization) methods and it is therefore necessary to provide the number of clusters (components K of mixture) in advance. According to the fact the number of components is usually not known in advance, it is necessary to estimate it. This can be performed using various information criteria – usually it is Akaike's information criterion AIC (Akaike, 1973), or Bayesian information criterion BIC (Schwarz, 1978). The criteria bear information concerning the ability of model to depict the reality and is not used for testing of statistical significance of the model. The criteria thus can be used for comparison of pair of models and to assess the extent of information loss under assumption that we employ some model instead of another. Modeling using finite mixtures can lead to any chosen accuracy of approximation. But each additional component not only increases precision but also complicates the model. Trade-off between the accuracy and complexity of model in agreement with parsimony principle is realized using information criteria. Therefore, both criteria contain a penalty for the increase of count of parameters used in a model, i.e. for the growth of complexity. Thus, with their help we choose an appropriate trade-off between the accuracy of approximation and complexity leading to choice of an optimal model. Akaike's information criterion works with smaller penalization of number of model parameters than the Bayesian information criterion and allows to construct models fitting better to the empirical density.

The EM-algorithm is implemented in the *mclust* package (see Scrucca, *et al*., 2016) which is one of, the packages of R software (R Core Team, 2017) and is used for the estimation of parameters of the mixture distribution. The AIC criterion was used for the estimation of optimal number of components.

The agreement of empirical distribution of logarithmic household incomes with parametric model is presented using a graph. We chose a non-parametric kernel estimate with Gaussian type of kernel. The kernel estimate of density $\hat{f}_n(x),\ x \in R$, is given by the formula:

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{2}$$

where K is a kernel of the estimate and h_n is so called smoothing parameter (chosen according to Scott, 1992). Its choice has an impact on the shape of kernel and degree of concentration in point X_i on the neighborhood. The Gaussian (normal) kernel is given by the formula

$$K\left(\frac{x - X_i}{h_n}\right) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x - X_i}{h_n}\right)^2\right]. \tag{3}$$

3. Results

The majority of papers dedicated to the estimation of parametric models of income of wage distribution concern the distribution of all Czech households, i.e. for the Czech Republic as a whole. This paper chooses more detailed approach and aims to construct parametric models on the level of particular regions. Similar issues were analysed also in the paper of Bílková (2017). But this article she used for an approximation of empirical distribution of wages simple lognormal model which does not allow to model the heterogeneity of the data

The data base comprises data from Czech variant of sample survey of EU-SILC "Living conditions" from 2015. It is a large random sample survey performed each year in all EU member states and its results can be rather straightforwardly generalized. For recalculation of household counts the PKOEF weights were used which assure the representativeness of a sample and thus provides the possibility to correctly generalize the results of analyses.

Table 1: Structure of households according to typology defined and used in EU.

Type	Structure	Type	Structure
0	Individuals	5	2 adults, 1 child
1	2 adults, both under 65	6	2 adults, 2 children
2	2 adults, at least one 65+	7	2 adults, 3 and more children
3	other household without children	8	other households with children
4	1 adult, 1 or more children	9	other (not typical) households

Source: CZ SILC 2015 data.

2.1 Basic regional characteristics

The aim of the paper is to identify regional differences in income distribution of the households and to approximate empirical distribution with suitable parametric models. According to the fact that the analysis is focused on total annual incomes of households it can be expected that one of the important factors determining interregional differences is varying proportion of particular household types in the regions. The division of households into the groups is performed using typology defined in EU (see Table 1). The summary of counts of surveyed households of particular types in all 14 Czech regions is presented in the Table 2.

Table 2: Percentages of household types in particular Czech regions (CZ SILC 2015).

Region	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7	Type 8	Type 9	Type 0
Prague	20.89	20.69	14.20	17.34	8.32	6.90	5.68	0.81	2.54	2.64
Central Bohemian	11.88	15.87	16.43	17.87	14.32	5.22	8.21	1.22	2.89	6.10
South Bohemian	13.72	20.58	14.98	18.05	14.08	4.69	5.60	1.26	2.17	4.87
Plzeň	13.31	14.97	18.30	18.30	15.80	4.57	6.03	1.66	2.49	4.57
Karlovy Vary	21.25	18.75	15.00	15.63	11.88	5.63	6.25	0.63	0.63	4.38
Ústí nad Labem	15.73	18.38	19.70	17.38	7.95	5.46	5.63	0.99	3.97	4.80
Liberec	12.35	21.08	17.77	15.06	13.25	7.23	6.02	0.90	1.81	4.52
Hradec Králové	11.68	16.59	16.12	20.79	11.45	4.67	9.58	1.87	1.64	5.61
Pardubice	11.37	13.70	19.38	21.19	10.85	5.17	7.49	1.81	1.29	7.75
Vysočina	7.94	18.61	14.89	21.34	16.63	3.23	6.45	2.23	1.74	6.95
South Moravian	11.45	15.58	15.23	18.89	16.41	5.31	6.97	0.94	2.83	6.38
Olomouc	11.98	21.49	14.05	18.60	13.43	5.17	5.37	1.65	2.89	5.37
Zlín	14.80	16.59	16.14	17.94	13.45	5.38	5.16	1.12	2.91	6.50
Moravian-Silesian	14.98	15.65	18.31	17.09	13.32	5.77	5.77	1.11	2.33	5.66
Czech Republic	**13.96**	**17.61**	**16.41**	**18.21**	**12.86**	**5.41**	**6.44**	**1.25**	**2.49**	**5.34**

Source: Own computations based on the CZ SILC 2015 data.

Table 2 shows that the least frequent (and rather exceptional) households are – regardless of the region – households of type 8 (other households with children) and 9 (other households with children). The most frequent types of households (underlined values in Table 2) are connected with a region.

In the capital of Prague and in Karlovy Vary region are the most frequent households of type 1 (2 adults, both under 65), in the South Bohemian, Liberec and Olomouc regions households of type 2 (2 adults, at least one 65+) and in regions of Ústí nad Labem and Moravian-Silesian households of type 3 (other household without children). But most frequently the type 4 (1 adult, 1 or more children) dominates which is the most frequent in the Central Bohemian, Hradec Králové, Pardubice, Vysočina, South Moravian and Zlín regions. In the Plzeň region are the most frequent households of type 3 and 4.

Table 3 contains values of basic regional characteristics of logarithmic incomes of households. The first and the second column contain basic characteristic of location and variability, i.e. parameters μ and σ of simple lognormal models of regional income distributions. Two other columns complement these basic characteristics with the robust ones (median and interquartile range).

Regions are sorted decreasingly according to the means and therefore it is easy to observe that in some regions the mean incomes of households were above the mean of the whole country. Namely it occurred in Central Bohemian, Prague, Hradec Králové, Plzeň, South Moravian, and Liberec regions. The highest values were achieved in the Central Bohemian region (tightly followed by the capital of Prague region), the lowest values appeared apparently in the Moravian-Silesian region.

Table 3: Characteristics of logarithmic household incomes in regions (using PKOEF weights).

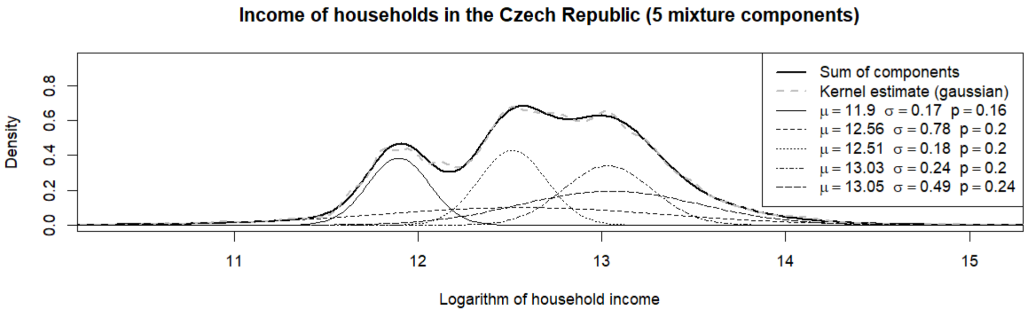
Region	Mean	Std. deviation	Median	Interq. range
Central Bohemian	12.783	0.607	12.816	0.777
Prague	12.777	0.646	12.772	0.869
Hradec Králové	12.702	0.584	12.706	0.903
Plzeň	12.677	0.573	12.734	0.831
South Moravian	12.677	0.590	12.707	0.844
Liberec	12.664	0.568	12.709	0.775
Czech Republic	**12.656**	**0.605**	**12.676**	**0.842**
Zlín	12.630	0.569	12.642	0.790
Vysočina	12.629	0.574	12.611	0.848
Pardubice	12.623	0.586	12.602	0.736
Olomouc	12.604	0.626	12.658	0.787
South Bohemian	12.601	0.587	12.609	0.887
Karlovy Vary	12.597	0.624	12.668	1.000
Ústí nad Labem	12.558	0.570	12.579	0.819
Moravian-Silesian	12.508	0.602	12.543	0.874

Source: Own computations based on the CZ SILC 2015 data.

2.2 Regional kernel estimates and mixture models of density

The following Figures 1 – 4 graphically present similarities and differences in mixture models of household incomes in the Czech Republic (Figure 1) and its regions (Figures 2 – 4). These parametric models (full line) are complemented by kernel estimates of empirical distributions (dashed line). K The number of components K in finite mixtures was optimized by choice of the best mixture according to the values of AIC. Graphs are augmented by maximum likelihood estimates of components in mixtures $p_1^{(K)}, \dots, p_K^{(K)}$ and values of parameters $(\mu_1^{(K)}, \sigma_1^{(K)}), \dots, (\mu_K^{(K)}, \sigma_K^{(K)})$.

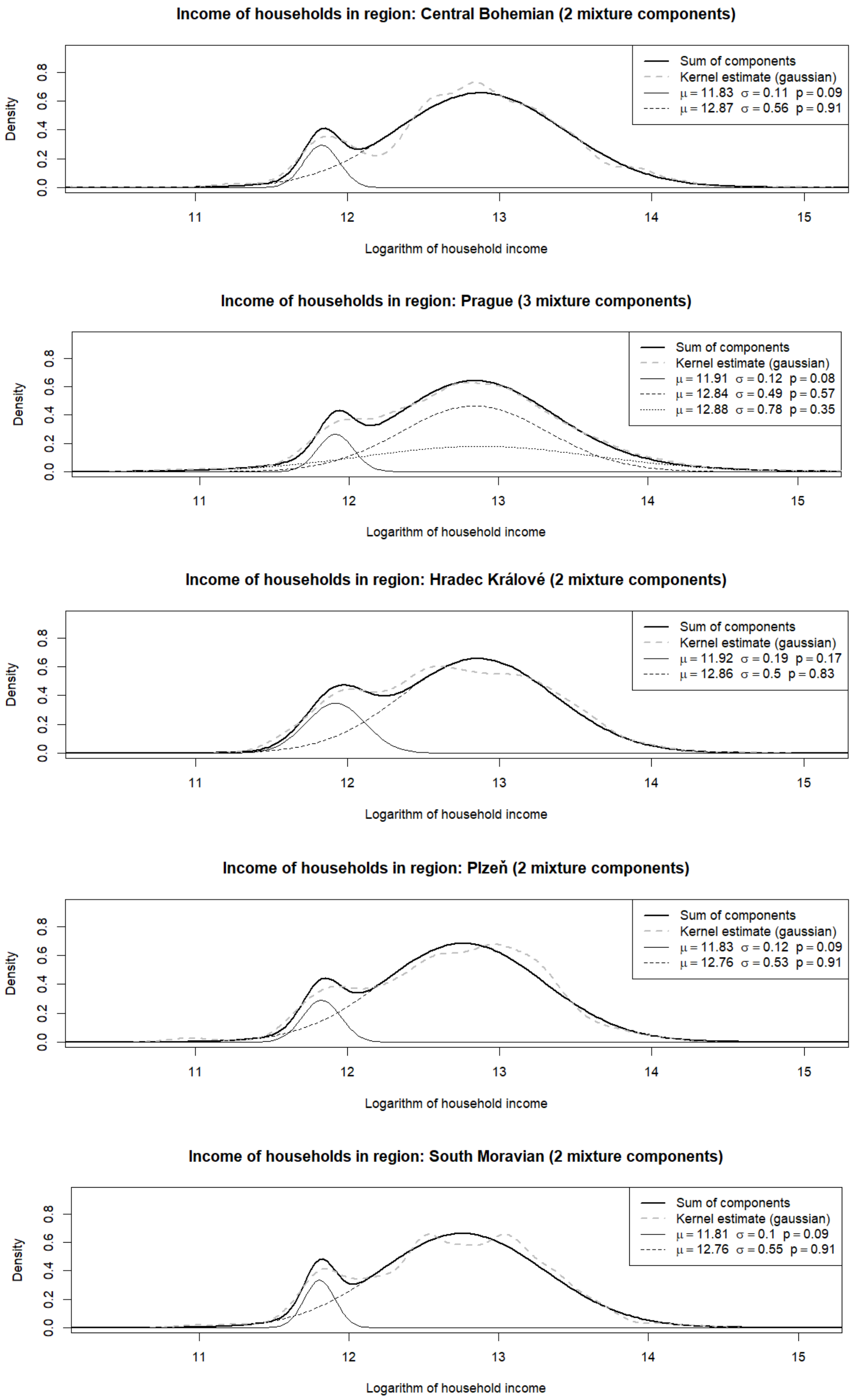
Figure 1: Finite mixtures of Gaussian densities for household incomes in the Czech Republic.



Source: Figures of mixtures estimated form the CZ SILC 2015 data.

Figures 1 – 4 show that the heterogeneity of empirical distribution varies in different regions. AIC found optimal number of components K between 2 and 6. Two components were chosen as the best fit in eight regions and three were chosen in case of three other regions. The highest heterogeneity appeared in case of Vysočina region (6 components), Liberec region (5 components) and Karlovy Vary region (4 components). Five components were optimal also in the case of incomes in the Czech Republic as a whole.

Figure 2: Finite mixtures of Gaussian densities for incomes of the Czech households in regions.



Income of households in region: Central Bohemian (2 mixture components)

Income of households in region: Prague (3 mixture components)

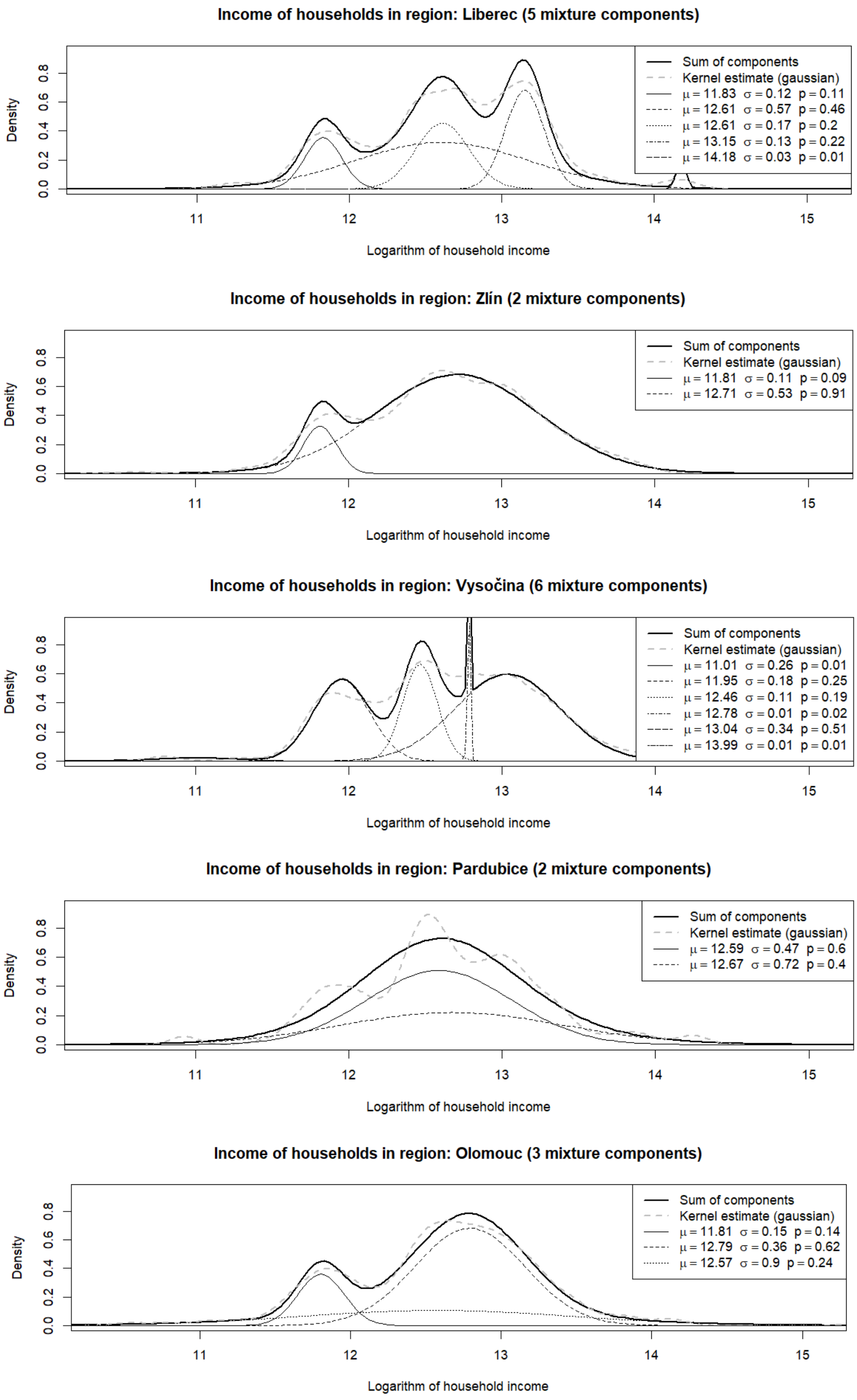
Income of households in region: Hradec Králové (2 mixture components)

Income of households in region: Plzeň (2 mixture components)

Income of households in region: South Moravian (2 mixture components)

Source: Figures of mixtures estimated form the CZ SILC 2015 data.

Figure 3: Finite mixtures of Gaussian densities for incomes of the Czech households in regions.



Income of households in region: Liberec (5 mixture components)

Income of households in region: Zlín (2 mixture components)

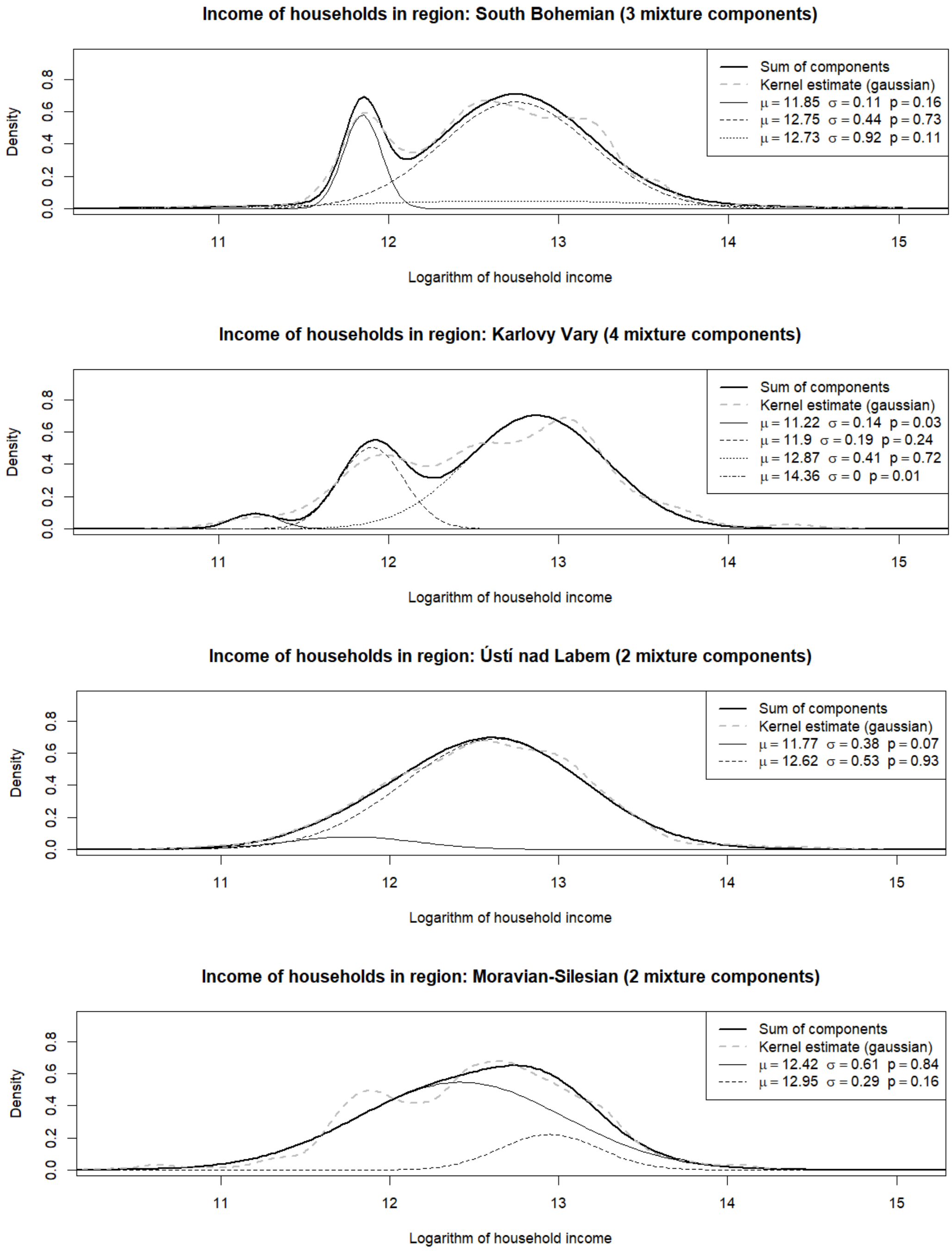
Income of households in region: Vysočina (6 mixture components)

Income of households in region: Pardubice (2 mixture components)

Income of households in region: Olomouc (3 mixture components)

Source: Figures of mixtures estimated form the CZ SILC 2015 data.

Figure 4: Finite mixtures of Gaussian densities for incomes of the Czech households in regions.



Income of households in region: South Bohemian (3 mixture components)

Income of households in region: Karlovy Vary (4 mixture components)

Income of households in region: Ústí nad Labem (2 mixture components)

Income of households in region: Moravian-Silesian (2 mixture components)

Source: Figures of mixtures estimated form the CZ SILC 2015 data.

4. Conclusion

The aim of the paper was to characterize the income distribution of the Czech households on the regional level and to provide a suitable approximation using appropriate parametric models.

The analysis carried out showed that there exists a regional differentiation caused by conditions which particular regions provide to its inhabitants. As a consequence of interregional differences the regions mutually differ in the composition of types of household residing there (see Table 2). This regional structure of household types is tightly connected with the demographic structure of population (age, education, etc.) and is conditioned by characteristics of regions like number and type of job opportunities (including opportunities in neighbouring regions of Germany and Austria), educational institutions, etc.

The differences of particular household types are then tightly connected with the level and differentiation of total household incomes as shown in Table 3. The next part of paper presented a graphical analysis of distributions of total annual household incomes which showed that the character of income distribution is connected to the region. The common regional property appearing generally in finite mixtures of regions was the low-income component appearing in similar locations in all regional distributions and then 1 to 5 higher income components of households with medium or higher incomes (in dependence on heterogeneity of particular regional distribution).

From the Figures 1 – 4 we can also infer that for the approximation of empirical distribution the simple lognormal model is not suitable not only on the level of the Czech Republic as a whole but also on the level of the regions. The presented graphs showed different level of heterogeneity in particular regions and in some cases even higher heterogeneity than in the case of whole Czech Republic. On contrary, the finite mixture models showed rather good performance in approximation flexibility and it can be stated that they comprise a good possibility of approximation of income distribution on both the statewide and regional level.

The components of finite mixture are not easy to interpret. We know its parameters and sometimes can observe higher frequencies of some household types. E.g., in case of Liberec and Vysočina regions we can observe two components with lower incomes which in a way correspond to households with at least on 65+ member which are quite frequent in these regions. The further research may aim at the detailed view of household types in particular components and characteristics of household members (particularly of head of household). But there are so many effects which overlap and interfere that the connection of income distribution and proportion of particular household types is not clear. But the interpretation of mixture model is in general not straightforward.

Acknowledgements

Supported by Institutional support of the long-term development (Faculty of Management in Jindřichův Hradec, University of Economics, Prague) under grant number IP 600040.

References

[1] Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, Csaki (eds.) Proceedings International Symposium in Information Theory, Budapest: Akademiai Kiado, pp. 267–281.

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

[3] Bílková, D., 2017. How Are the Czech Regions Different and Mutually Similar in Terms of Wages? Cluster Analysis and Wage Models. In: Löster, T., Pavelka, T. (eds.) 11th International Days of Statistics and Economics. Slaný: Melandrium. ISBN 978-80-87990-12-4.

[4] Bílková, D., 2012. Recent Development of the Wage and Income Distribution in the Czech Republic. Prague Economic Papers, vol. 21, iss. 2, pp. 233-250.

[5] Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, vol. 39, iss. 1, pp. 1-38.

[6] Malá, I., 2013. Použití konečných směsí logaritmicko-normálních rozdělení pro modelování příjmů českých domácností. Politická ekonomie, vol. 61, iss. 3, pp. 356-372.

[7] Malá, I., 2014. A multivariate mixture model for incomes of the Czech households in 2006-2010. In: Löster, T., Pavelka, T. (eds.) 8th International Days of Statistics and Economics. Slaný: Melandrium. ISBN 978-80-87990-02-5.

[8] Marek, L., 2017. The effect of extreme wages on average wage values. In: Löster, T., Pavelka, T. (eds.) 11th International Days of Statistics and Economics. Slaný: Melandrium. ISBN 978-80-87990-12-4.

[9] Marek, L., Vrabec, M., 2013. Probability models for wage distributions. In: Vojáčková, H. (ed.) 31st International Conference on Mathematical Methods in Economics. Jihlava: College of Polytechnics. ISBN 978-80-87035-76-4.

[10] McLachlan, G., Peel, D., 2000. Finite Mixture Models. New York: Wiley. ISBN 978-0-471-00626-8.

[11] R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[12] Scott, D. W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. New York: Wiley. ISBN: 978-0-471-54770-9.

[13] Scrucca L., Fop M., Murphy T. B. and Raftery A. E. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 205-233

[14] Schwarz, G. E., 1978. Estimating the dimension of a model. The Annals of Statistics, vol. 6, iss. 2, pp. 461-464.