DOI: 10.18267/j.pep.487

A STUDY OF INCOME STABILITY IN THE CZECH REPUBLIC BY FINITE MIXTURES

Jitka Bartošová, Nicholas T. Longford*

Abstract:

Income, expenditure and similar variables in monetary units tend to have distributions similar to log-normal. Description of such variables after logarithmic transformation by the normal model is often not accurate enough, especially for multivariate data. Deviations of their empirical distributions from the theoretical lognormal distribution often require more sophisticated analysis. Mixtures represent a very flexible way of reconstructing complex distributions with irregular features and are suitable for detailed modelling. Multivariate mixture models are applied to the Czech longitudinal survey of household income in the European Union Statistics on Income and Living Conditions (EU-SILC) in 2005–2008. The analysis identifies distinct patterns of progression of income, with a high percentage of households having steady annual increases over the four years (three transitions). Graphical presentation of the results is emphasised.

Keywords: Equivalised household income; EU-SILC; income distribution; longitudinal analysis; multivariate mixtures; stability.

JEL Classification: C33, C38, H31

1. Introduction

The Czech Republic is generally regarded as a success story of transition from a centrally managed economy until 1989 to a modern market-oriented economy, accomplished without any revolutionary social upheavals. This applies not only to changes in the political and economic areas, but also in the system of wage remuneration and social position of the population. Večerník (2009) examined changes in socio-economic policies and structures that occurred in the Czech Republic in the post-revolutionary period.

Study of household income and of its progression (dynamics) is important for monitoring the level of welfare and stability in the management of households' resources. In contrast to many undesirable features of the pre-revolutionary economic state of the Czech Republic, relatively small variation of income, across regions, occupations and experience (and therefore age) is a desirable feature of the structure of household income. It is therefore of interest to assess to what extent such homogeneity of income has survived the structural changes of the economy and of the composition of employment.

[*] University of Economics, Prague (bartosov@fm.vse.cz); SNTL and Universitat Pompeu Fabra, Barcelona (sntlnick@sntl.co.uk). The paper was supported by the Grant Projects GAČR 402/09/0515 from the Czech Science Foundation, IGF 6/3/2012 from University of Economics in Prague, VEGA 1/0127/11 from the Slovak Scientific Grant Agency and ECO 2011-28875 from the Spanish Ministry of Science and Technology.

DOI: 10.18267/j.pep.487

Analysis of income and its progression is useful for decisions in social policy and is crucial for predicting consumption. Information on how wage distribution is evolving in the Czech Republic in general and with regard to sex and age is found in Marek (2010) and Bílková (2012). Pacáková and Foltán (2011) analyse households with the highest wages in Slovakia, Benáček *et al.* (2010) compare the income of individuals and households in the Czech Republic with other countries of Central and Eastern Europe. Just as in other countries of the European Union, there is a considerable interest in monitoring income inequality, monetary poverty, material deprivation and regional disparities; see Labudová, Vojtková and Linda (2010), Stankovičová (2010), Želinský (2010), Marek (2011) and others.

A lot of attention is paid to the modelling of income distribution in the Czech and Slovak Republics. Pacáková, Sipková and Sodomová (2005) model cross-sectional data[1] about income distribution by quantile function, Bílková and Malá (2012) by lognormal distribution and Malá (2012) applies finite mixtures. Such a set of single-year (univariate) analyses does not inform about the stability and dynamics of income, that is, the pattern of changes. Our analysis of income in the Czech Republic is unusual in being longitudinal.[2] Only a few works use a dynamic approach to assess the financial situation of households in the Czech Republic; see *e.g*. Bartošová and Forbelská (2011).

We analyze the data from the survey that the Czech Republic contributes to the European Union Statistics on Income and Living Conditions, EU-SILC 2005–2008.[3] These surveys collect representative data about income, life style, quality and cost of housing, ownership of certain household appliances, access to utilities, services and the employment, health care and other conditions of the members of the household. The participating countries are contracted to provide annual survey data by the Directive of the European Commission No. 1177/2003. The survey has a rotating panel design, in which each participating household completes the same questionnaire in four consecutive years. This design enables the analysts to study not only the current state, but also the changes (development or progression) over time. The Czech Republic has been contributing to EU-SILC since 2005, when its sample contained 4,351 households. These households were followed up annually until 2008, when the sample contained a total of 11,294 households in various stages of the panel (in the first to fourth year of participation).

2. Studying Household Income

Income is studied for households, because a household pools the resources gained by its members. However, the total income of a household has to be adjusted for its composition. Larger households require more resources, but may have a greater capacity

[1] Cross-sectional data are a set of unrelated samples, one for each time point.

[2] Longitudinal data follow a sample of households over several time points.

[3] The datasets were obtained as part of the Project GAČR 402/09/0515.

DOI: 10.18267/j.pep.487

to gain higher income because they contain more adults. In contrast, being in a household tends to reduce the amount of resources required. For example, a typical household with two adults and two children might reasonably be expected to require fewer resources than two households, each comprising an adult and a child.

Conventionally, annual income is studied, to discount seasonal variation, but also to smooth out other short-term variation in income, which is rarely relevant. Part of the income may be saved, a household may fall into debt, may pay for obtained goods and services with delay and disparities between possession and expenditure may arise by other means (*e.g.* as deposits or hire purchase arrangements). It is convenient to inquire about annual income (of an individual or a household), because it is often used for reference, it is a key figure in a typical tax return and employment contract, and is sometimes used in a households' informal planning and accounting.

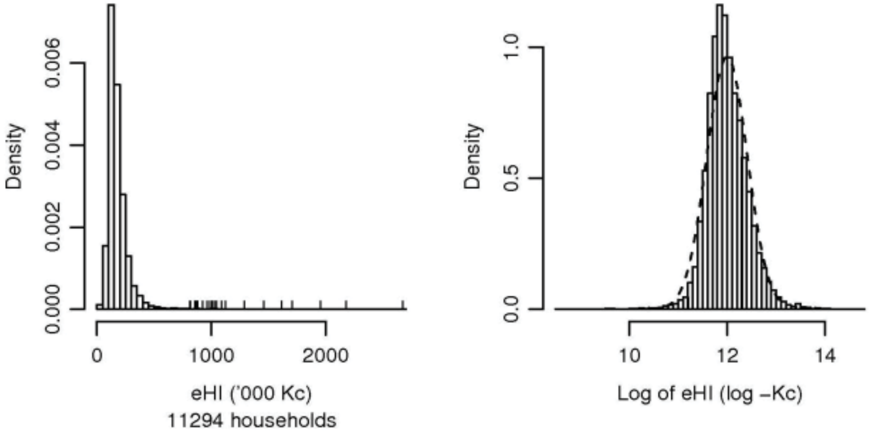
The economy of scale is reflected in the definitions of equivalised household size and equivalised household income. We adopt the definition promoted by Eurostat, according to which one adult member of a household (*e.g.* its head) counts as one unit, every other adult member as 0.5 units and every child (below 14 years of age) as 0.3 units. Formally, suppose a household comprises H_1 adults and H_2 children. Then its equivalised size is $0.5 + 0.5H_1 + 0.3H_2$, denoted by *eHs*. Suppose the total annual income of this household is I; then its equivalised income is I/eHs, denoted by *eHI*. See Longford and Nicodemo (2009) for a study of the sensitivity of these and related definitions.

3. Modelling Household Income

The normal distribution, with a particular symmetric shape of its density, is the mainstay of statistical modelling. Possibly after some adjustment for covariates, normality is an appropriate assumption for modelling many phenomena. The normal distribution is often unsuitable in studies of income because it tends to have a skewed distribution; it cannot be negative, a small fraction of the units has very small (or zero) income, a large fraction has income in a relatively narrow range, and a small fraction has high income spread across a wide range. Such a distribution is distinctly asymmetric, and not normal. Empirical evidence and motivation by the multiplicative scale support the idea of analysing the logarithmic transformation of income.

DOI: 10.18267/j.pep.487

Figure 1
Histogram of the Equivalised Household Income (*eHI*) in EU-SILC for the Czech Republic in 2008



Note: Original scale (left-hand panel) and after log-transformation (right-hand panel), with the density of fitted normal distribution overlayed (dashes)

Source: EU-SILC, own calculation

Figure 1 presents an example with data from the cross-sectional component of EU-SILC for the Czech Republic in 2008. The left-hand panel shows the histogram of the original values, prior to the transformation, and the right-hand panel the log-transformed values. The latter is much closer to normality, although some departures from it can be discerned. The roughness of the histogram is due to insufficiently large sample (n = 11 294), but other departures from normality are systematic; they would be present even if a larger sample were collected.

In the right-hand panel, the density of the best-fitting normal distribution, with mean 11.998 and standard deviation 0.414, is drawn by dashes. This mean corresponds to 162,431 korunas, and the standard deviation to multiplying or dividing by exp(0.414) = 1.513, that is, increase by 51% or decrease by 100 – 100/1.513 = 34%. Note that mean *eHI* differs from the back-transformed mean of log(*eHI*), because the operations of taking logarithm and averaging cannot be interchanged. In fact, the sample mean of *eHI* is equal to 178,098 korunas; it is so much higher because the influence of the highest values of *eHI* is reduced by the log-transformation.

4. Mixtures of Lognormal Distributions

A variable is said to have a lognormal distribution, if its log-transformation is normally distributed. That is our original model and a starting point in more detailed modelling. Figure 1 indicates that lognormal distribution is better suited for modelling income (in the particular setting) than normal distribution, but further improvement on the fit would be highly desirable.

DOI: 10.18267/j.pep.487

We seek improvement by mixture modelling. A sample (or a population) of units is said to be a mixture if it comprises several subsamples (groups), each with a distinct distribution of the studied variables $y_1, \dots y_n$. The mixture density of variable $(i = 1, \dots, n)$ is expressed as

$$f\left(y_i;\boldsymbol{\Psi}\right)=\sum_{k=1}^{K}\pi_k f_k\left(y_i;\boldsymbol{\theta}_k\right) \tag{1}$$

where the mixing proportions $\pi_1, \dots, \pi_K$ sum up to one and the group-conditional density $f_k(y_i; \boldsymbol{\theta}_k)$ is specified up to a vector $\boldsymbol{\theta}_k$ of unknown parameters ($k = 1, \dots, K$). The vector of all unknown parameters is given by $\boldsymbol{\Psi} = (\pi_1, \dots, \pi_{K-1}, \boldsymbol{\theta}_{1,} \dots, \boldsymbol{\theta}_k)$. Using an estimate of $\boldsymbol{\Psi}$, this approach gives a probabilistic clustering of the data into k clusters in terms of estimates of the posterior probabilities of component membership

$$\omega_k\left(y_i\right)=\frac{\pi_k f_k\left(y_i;\boldsymbol{\theta}_k\right)}{f\left(y_i;\boldsymbol{\Psi}\right)} \tag{2}$$

where $k = 1, \dots, K$ denotes the component of the mixture. For more details see *e.g.* McLachlan and Peel (2000). In our context, *eHI* in the Czech Republic in 2008 (or another year) may comprise several groups, or components, each with a distinct distribution of *eHI*. The component to which a household belongs is not identified.

Fitting mixture models is an application of the EM algorithm (Dempster, Laird and Rubin, 1977; Fraley and Raftery, 2002). The algorithm assumes that if in addition to the observed data we had some additional data or information, referred to as the missing data, the problem at hand could be solved by a simple method. For fitting mixture models, this information is the assignment to the group, that is, which household belongs to which component. The observed data is called incomplete, and its union with the missing data is referred to as the complete data.

EM algorithm is an iterative procedure for maximum likelihood estimation. Each iteration comprises two steps, E (estimation) and M (maximisation). In the E-step, certain summaries of the missing data are estimated, which are needed in the following M-step. In the M-step, the (relatively simple) analysis intended for the complete data is applied, with the contributions that would be made by the missing data replaced by their estimates from the immediately preceding E-step. The iterations are terminated when two consecutive M-steps yield very similar results. We assume that the mixture components have (multivariate) normal distributions. For details specific to fitting mixture models to the European Community Household Panel (ECHP), the predecessor of EU-SILC, see Longford and Pittau (2006), to detection of convergence clubs of countries by their *per capita* income see Pittau, Zelli and Johnson (2010), and to study of mixture models with an improper component for panel data see Longford and D'Urso (2011).

A mixture model is described by the distribution and (marginal) probability of each component. The fit by a mixture of two components, each with a lognormal distribution, is bound to be better than by a single lognormal, because a mixture of two (or more)

DOI: 10.18267/j.pep.487

components affords much greater flexibility. The density of a mixture is usually not symmetric (after log-transformation) and may have as many modes as it has components. The incentive to obtain a better fit has to be tempered by the complexity of the posited model. All formal tests for the choice of the number of components are tainted by the preference for simple models for small samples and complex models for large samples. We prefer an informal approach in which we increase the number of components until one of them is very small.

We do not assume that any particular model is valid (correct); the models merely approximate the underlying distribution. The fit of a mixture model is sometimes interpreted as the presence (or discovery) of distinct groups or clusters in the studied population. Such an interpretation is contingent on the assumption of lognormality for each cluster. That is, if we posit a model, in which (some) components have distributions different from lognormal, then the fit may comprise very different groups. In brief, we apply mixtures merely to reproduce or approximate the observed data by a distribution from a class much greater than any single class of some well-known distributions.

The two-component fit (solution) is given by the lognormal distributions $LN(11.963, 0.104)$ and $LN(12.175, 0.486)$. On the log scale, the means of the two components differ by 0.212, not a great deal in relation to the sizes of the standard deviations, $\sqrt{0.104} = 0.322$ and 0.697. The estimated probabilities of the two components are 0.838 and 0.162, for the less and the more dispersed component, respectively.

The three-component fit is given by the lognormal distributions $LN(11.862, 0.067)$, $LN(12.360, 0.083)$ and $LN(12.069, 0.711)$ with respective probabilities 0.677, 0.236 and 0.087. Thus, there is a small component with variance much greater than the two other components, both of them with much greater probabilities. The group membership of many households, those in the region of high density (11.5–12.3), cannot be inferred with any precision, because all three distributions have their highest densities in this range. Households with the largest and smallest values of *eHI* almost certainly belong to the third component, because these values are several standard deviations from the means of the other two components, but only a few from the mean of the third component. The household-specific probabilities of belonging to each component can be obtained from the concluding E-step of the EM algorithm. We explore them in detail in next section.

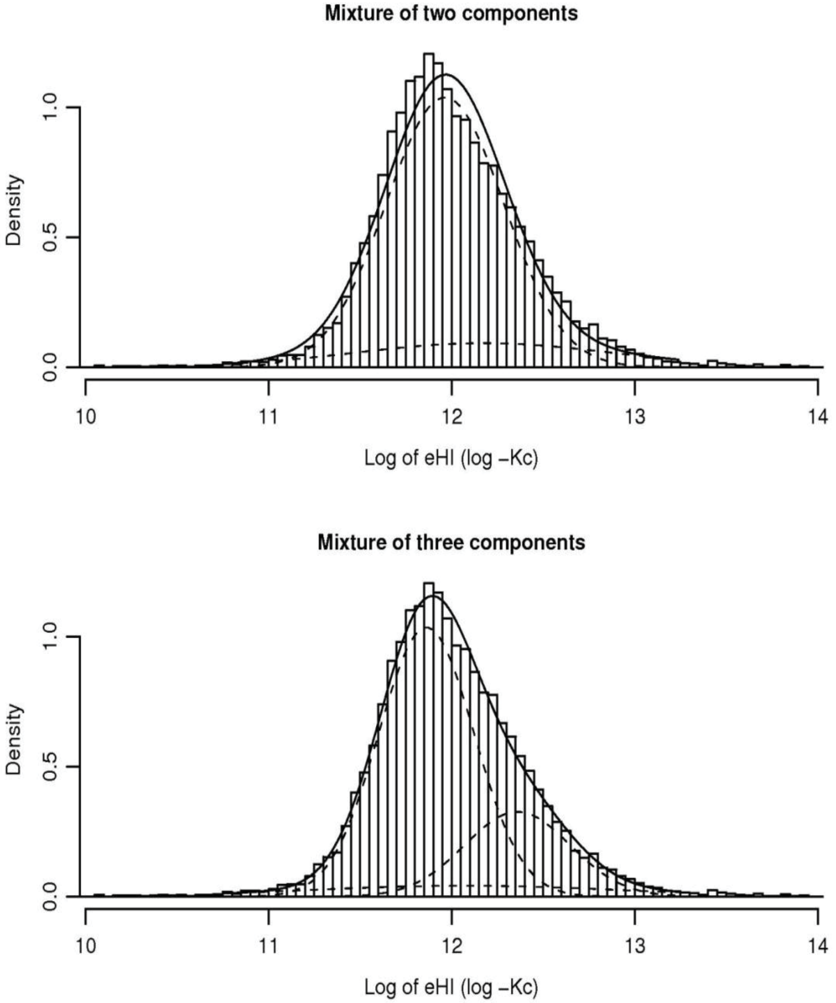
Figure 2 compares the fit of the mixture with two and three lognormal components to the values of eHI in the EU-SILC sample from the Czech Republic in 2008. The contributions of each component are marked by dashes, and the density of the mixture, equal to the total of these contributions, is drawn by a solid line in each panel. The two-component fit (the top panel) is poor in most of the range where the density is highest. Eighteen observations are off the scale in both panels; if they were accommodated in the figure, the resolution of the histograms would be much poorer.

The three-component solution (the bottom panel) fits much better, but it is not obvious whether an even better fit should be sought. The four-component solution is only marginally better. It comprises a component that accounts for only a minute fraction of the sample, and does not offer a better fit. Details are omitted.

DOI: 10.18267/j.pep.487

Figure 2

The Fits of the Two- and Three-Component Mixtures of Lognormal Distributions to the EU-SILC Data for the Czech Republic in 2008



Note: Ten observations with values below 10.0 and eight with values about 14.0 are off the scale.

Source: EU-SILC, own calculation

5. Multivariate Mixtures

The mixture models considered in the previous section are univariate; they are for a single variable, *eHI* in 2008. They can be applied to each of the years 2005–2008, for which the EU-SILC survey was conducted in the Czech Republic. With this approach we cannot study the evolution of *eHI* of the households, because they are not linked across the annual analyses and we fail to capture the longitudinal (dynamic) aspect of income.

DOI: 10.18267/j.pep.487

In multivariate analysis, the basic data element is a set of observations on an unit, such as the values of *eHI*, or their log-transformation, in a sequence of years. The outcome of the analysis are the sequences of the annual means and variances, supplemented by the covariances (or correlations) for pairs of years. For example, the fit of the model with a single multivariate lognormal distribution is $LN4(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ where

$$\hat{\boldsymbol{\mu}} = (11.831, 11.879, 11.952, 12.029)$$

and

$$\hat{\Sigma} = \begin{pmatrix} 0.204 & 0.159 & 0.149 & 0.139 \\ 0.159 & 0.201 & 0.165 & 0.150 \\ 0.149 & 0.165 & 0.198 & 0.164 \\ 0.139 & 0.150 & 0.164 & 0.193 \end{pmatrix} \qquad (3)$$

are the vector of the mean profile and the variance matrix of log(*eHI*). It shows that the average *eHI* has been rising, with estimated annual increases of about 4.9, 7.5 and 8.0%. These figures are obtained by transforming the differences of the means, such as 11.879 – 11.831 = 0.048, which, after exponentiation, converts to exp(0.048) = 1.049, that is, an increase by 4.9%. (In general, $\exp(x) \approx 1 + x$ for x close to zero.) The fitted variances have been declining, from 0.204 in 2005 to 0.193 in 2008. The latter figure corresponds to the standard deviation of 0.439, that is, *eHI* greater by 55% or smaller by 36% than the fitted median income of exp(12.029) = 167,544 korunas.

The results quoted here differ from those in the previous section in two respects. First, the multivariate analysis is based only on the 3,320 households that participated in the survey in all four years 2005–2008. Second, each household is associated with a sampling weight, and the means and the variance matrix are estimated with them. In general, a household has a different sampling weight in a longitudinal sample (2005–2008) than in a cross-sectional sample.

The 4 × 4 variance matrix in (1) can be expressed in terms of the variances (its diagonal) or standard deviations (their square roots), and the correlation matrix

$$\hat{\boldsymbol{\rho}} = \begin{pmatrix} 1.000 & 0.784 & 0.741 & 0.698 \\ 0.784 & 1.000 & 0.830 & 0.761 \\ 0.741 & 0.830 & 1.000 & 0.839 \\ 0.698 & 0.761 & 0.839 & 1.000 \end{pmatrix}. \qquad (4)$$

The correlations decrease with distance in time (or distance from the diagonal). Correlation can be interpreted as a measure of stability; with high correlations, households tend to have values of *eHI* in one year similar to those in the next, after a linear adjustment. Those in prosperity (with high *eHI*) then remain in prosperity, and most of those in poverty (with low *eHI*) remain in poverty (in long-term or persistent poverty). Greater

DOI: 10.18267/j.pep.487

stability of income is associated with lower income mobility. In fact, an index of income mobility is defined as the complement of the correlation of *eHI*, or of another variable that characterises income, $1 - \rho_{h_1, h_2}$ for years h_1 and h_2; see Glewwe (2005). The inter-year index of mobility at the end of the studied period (2007–2008), 1 – 0.0839 = 0.161, was about a quarter lower than at beginning of the studied period (2005–2006), when it was 0.216. Thus, the position of Czech households on the income ladder tended to become more entrenched over this period.

The variation on the log scale, var{log(*eHI*)}, is an index of (year-specific) income inequality. The fitted distribution in (1) implies that the income inequality has not changed a great deal over the short term. The income mobility over several years is defined from the total income over the period. Although Gibson and Glewwe (2006) provide some approximations, we evaluate these indices directly from the data. The indices of income inequality for the pairs of consecutive years are 0.184, 0.183 and 0.178, for the two sets of three consecutive years are 0.175 and 0.171, and for the set of four years (2005–2008) 0.155. This indicates that the financial potential of Czech households stabilised somewhat in the period 2005–2008, while undergoing weak differentiation. By a mixture analysis, we explore whether households have started forming clubs (clusters), and if so, whether some convergence or divergence took place in them.

To avoid convoluted expressions, we refer to component *k* of the *K*-component model fit as *k*(*K*); for example, the symbols 1(2) and 2(2) denote the first and second component of the two-component mixture. We assume first that the sample comprises two (and later three, or four) subsamples 1(2) and 2(2), each associated with a different four-variate lognormal distribution. The fit of this model is given by the pair $\left\{\hat{P}_{1(2)}; \left(\hat{\boldsymbol{\mu}}_{1(2)}, \hat{\sum}_{1(2)}\right)\right\}$, and $\left\{\hat{P}_{2(2)}; \left(\hat{\boldsymbol{\mu}}_{2(2)}, \hat{\sum}_{2(2)}\right)\right\}$, where

$$\hat{P}_{1(2)} = 0.566$$

$$\hat{\boldsymbol{\mu}}_{1(2)} = (11.794, 11.839, 11.911, 11.982)$$

$$\hat{\sum}_{1(2)} = \begin{pmatrix} 0.109 & 0.104 & 0.101 & 0.099 \\ 0.104 & 0.109 & 0.105 & 0.102 \\ 0.101 & 0.105 & 0.110 & 0.107 \\ 0.099 & 0.102 & 0.107 & 0.112 \end{pmatrix} \quad (5)$$

and

$$\hat{P}_{2(2)} = 0.434$$

$$\hat{\boldsymbol{\mu}}_{2(2)} = (11.879, 11.931, 12.006, 12.091)$$

$$\hat{\sum}_{2(2)} = \begin{pmatrix} 0.324 & 0.226 & 0.207 & 0.185 \\ 0.226 & 0.315 & 0.240 & 0.206 \\ 0.207 & 0.240 & 0.307 & 0.233 \\ 0.185 & 0.206 & 0.233 & 0.292 \end{pmatrix} \quad (6)$$

DOI: 10.18267/j.pep.487

are the marginal probabilities, mean profiles and variance matrices of the components (on the log-scale). Thus, the means of both components have small annual increments, and the second component has a greater mean in all four years, with marginally greater increments. The first component has much smaller variances and much greater correlations,

$$\hat{\rho}_{1(2)} = \begin{pmatrix} 1.000 & 0.947 & 0.920 & 0.893 \\ 0.947 & 1.000 & 0.953 & 0.923 \\ 0.920 & 0.953 & 1.000 & 0.961 \\ 0.893 & 0.923 & 0.961 & 1.000 \end{pmatrix}, \qquad (7)$$

it represents a subpopulation with a high level of stability of income, and therefore low level of mobility, and low internal (within-component) differentiation. This finding is remarkable only in conjunction with the estimated size of this component, 56.6%. A component with much smaller probability would be much less remarkable, because some types of (intact) households, such as single-member households with pensioners relying on the state pension as a single source of income, are bound to have near-identical income (after annual adjustment for inflation) across the years. The fitted correlations for component 2(2), 0.603–0.778, are much smaller. We could evaluate the various indices for each fitted component. However, these components do not have any intrinsic existence; their principal role is to approximate the overall distribution of *eHI*.

The three-component fit is summarised in Table 1. It is difficult to relate to the two-component fit, but we can discern components with variances of distinct magnitudes. The second component has extremely high correlations (0.950–0.986), which corresponds to high income stability. The components are well ordered with respect to their annual means, although they differ by much less than the corresponding standard deviations.

Table 1
The Three-Component Mixture Model Fit

Year	2005	2006	2007	2008	
Component	**Mean profile**				**Probability**
1(3)	11.893	11.949	12.029	12.126	0.558
2(3)	11.718	11.761	11.821	11.883	0.358
3(3)	11.820	11.842	11.916	11.931	0.084
	Variance				**Correlation**
1(3)	0.173	0.173	0.156	0.145	0.710 – 0.862
2(3)	0.083	0.083	0.085	0.088	0.950 – 0.986
3(3)	0.680	0.637	0.680	0.652	0.560 – 0.734

Source: EU-SILC, own calculation

DOI: 10.18267/j.pep.487

The order, in which the components are listed, is not a feature of the model fit, so we display them in the order of size (marginal probability). Thus, component 2(3) seems to be 'distilled' from component 1(2) by selecting the households that correspond even more closely to the stereotype of small and even annual increases of *eHI* (small variances and high correlations). Component 3(3) collects the households that have least predictable income from one year to the next (large variances and small correlations). We explore these associations graphically in Section 6.

Table 2
The Four-Component Mixture Model Fit

Year	2005	2006	2007	2008	
Component	**Mean profile**				**Probability**
1(4)	11.899	11.952	12.032	12.127	0.554
2(4)	11.709	11.751	11.811	11.872	0.340
3(4)	11.757	11.873	11.928	12.012	0.097
4(4)	12.268	11.617	11.891	11.444	0.009
Component	**Variance**				**Correlation**
1(4)	0.166	0.165	0.152	0.142	0.728 – 0.871
2(4)	0.079	0.079	0.082	0.088	0.951 – 0.986
3(4)	0.514	0.587	0.524	0.411	0.568 – 0.770
4(4)	1.162	0.362	1.216	1.649	0.598 – 0.818

Source: EU-SILC, own calculation

Table 2 displays the fit of the model with four mixture components. Components 1(3) and 1(4) are very similar in all aspects, as are components 2(3) and 2(4). The fitted means of component 3(4) also correspond to a plausible pattern of *eHI*, although households in the component are widely dispersed and relatively weakly correlated. The remaining component 4(4) corresponds to an implausible pattern of the means (reduction of *eHI* by 48% in 2005–2006 followed by increase by 31% in the next year), with very large variances. It accounts for only about 1% of the households. We regard the presence of such a component in the fit, with a small marginal probability, as a signal that models with more components are not useful, because the additional components would merely identify further esoteric patterns in small subpopulations that are of rudimentary interest in the study of *eHI* in the country as a whole.

6. Graphical Summaries of the Fit

In this section, we present graphical summaries of the mixture model fits that are easier to comprehend and digest than Tables 1–3, and which complement them with effect. Figure 3 presents such a summary. Each panel corresponds to a model fit. For simplicity, we regard the fit by a single distribution as the one-component solution, 1(1).

DOI: 10.18267/j.pep.487

It is summarised in panel A by the solid line connecting the four estimated means $\hat{\boldsymbol{\mu}}_r, r = 2005, ..., 2008$, and thinner dashed lines connecting the values $\hat{\mu}_r \pm \hat{\sigma}_r/4$, where $\hat{\sigma}_r$ is the estimated standard deviation in year r. The range $\hat{\mu} \pm 2\hat{\sigma}$ is conventionally used to delineate values that are not exceptional. The ranges $\hat{\mu}_r \pm 2\hat{\sigma}_r$ are very wide in our case, so we choose a much smaller multiple of $\hat{\sigma}_r$ for representing the dispersion of the values around their means. The covariances or correlations for the pairs of successive years are represented by vertical bars drawn midway between the two years. Their lengths are proportional to the geometric average $\sqrt{\hat{\sigma}_r \hat{\sigma}_{r+1}}$, so for perfect correlation the bar would reach the dashes that connect the values $\hat{\mu}_r \pm \hat{\sigma}_r/4$ and $\hat{\mu}_{r+1} \pm \hat{\ }_{r+1}/4$.

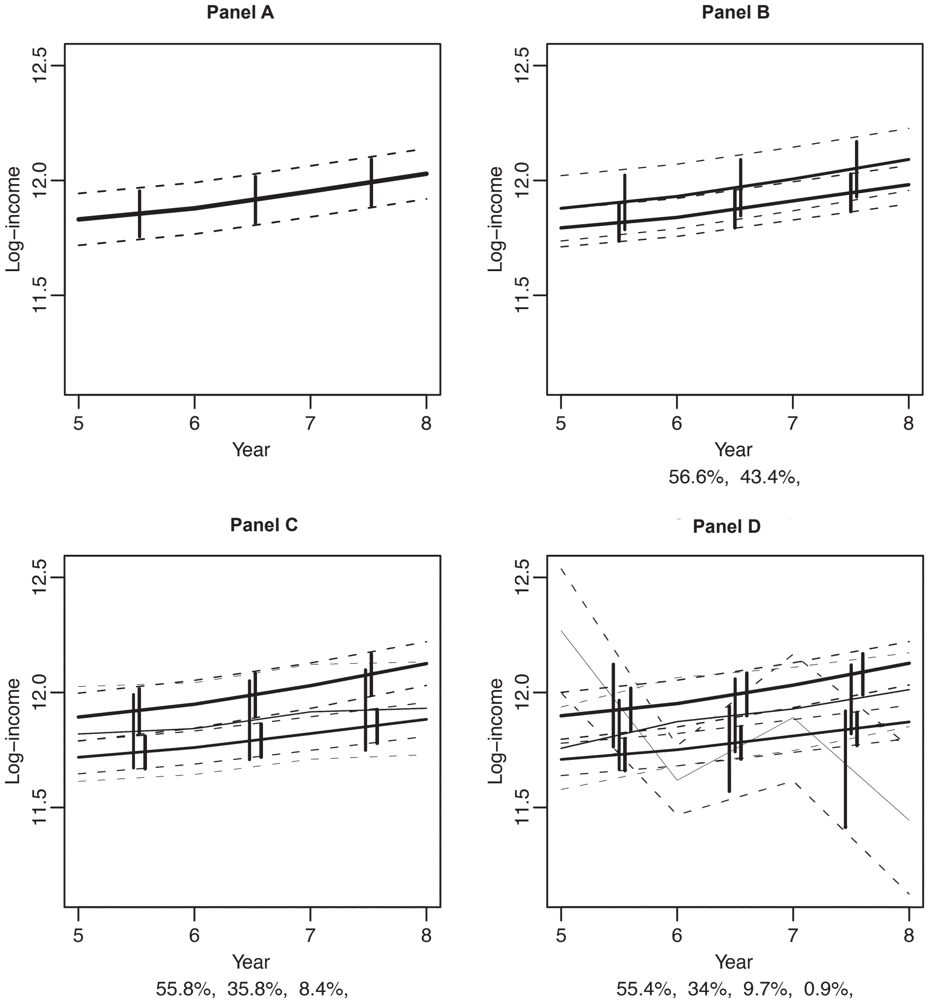
In respective panels B–D, for solutions with two to four components, each vector of means (mean profile) is represented by connected solid segments of thickness proportional to the probability of the component. Figure 3 shows that the mean profiles for components with large estimated probabilities are nearly parallel, components 3(3) and 3(4) differ from their direction only slightly, and component 4(4) represents an esoteric pattern with large variance in years (2005–2008). Imperfect quality of the data may well be its cause.

Next we explore how the membership of a component of one fit is related to the membership of another. The assignment of a household to a component is in general not known, and is therefore regarded as a random variable. In the E-step of the concluding iteration of model fitting, the conditional probabilities of belonging to component 1, ..., K, given the parameter estimates, are evaluated. These probabilities, specific to each household, are naturally interpreted as the likelihood of belonging to the components. For each household, they add up to unity.

DOI: 10.18267/j.pep.487

Figure 3

The Summaries of the Fits of the Mixture Models with up to Four Components



Note: panels A (one component), B (2 components), C (3 components) and D (4 components).

Source: EU-SILC, own calculation

For the two-component model fit, these probabilities are presented concisely after rounding in Table 3. The table classifies the households by the first decimal digit of the (conditional) probability of belonging to component 1(2). The within-category counts of households and their percentages do not agree, because the latter are evaluated with the sampling weights. The households in the first and last category (0.0–0.1 and 0.9–1.0, respectively) are very likely to belong to the respective components 2(2) and 1(2);

DOI: 10.18267/j.pep.487

they constitute about 75% of the households. Less than 10% of the households have probabilities in the range 0.3–0.7; each of these households may almost equally well belong to either component. A histogram of the probabilities is an alternative to Table 3.

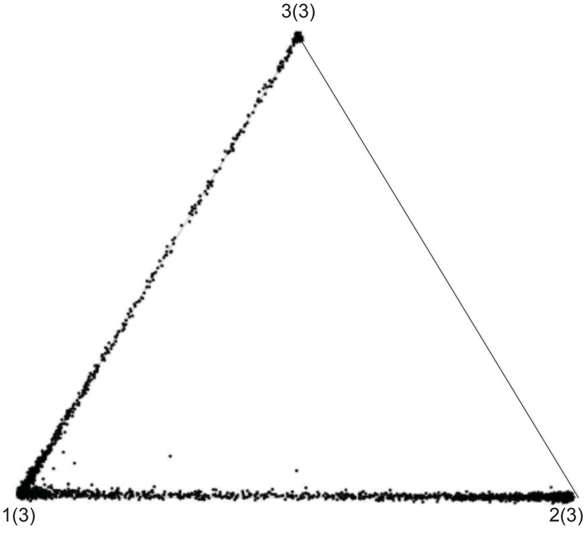
Table 3
The Probabilities of Belonging to Component 1(2)

Number (Percentage)	**Probability range**					
	0.0 – 0.1	**0.1 – 0.2**	**0.2 – 0.3**	**0.3 – 0.4**	**0.4 – 0.5**	-
Households	917	96	70	59	53	-
(Weighted %)	32.1	3.1	2.1	1.8	1.6	-
	0.5 – 0.6	**0.6 – 0.7**	**0.7 – 0.8**	**0.8 – 0.9**	**0.9 – 1.0**	**Total**
Households	79	102	124	202	1618	3320
(Weighted %)	2.6	3.2	3.6	6.2	43.6	100.0

Note: See Table 1

Source: EU-SILC, own calculation

Figure 4
Ternary Plot of the Fitted Probabilities of Belonging to the Components of the Three-Component Mixture Fit



Source: EU-SILC, own calculation

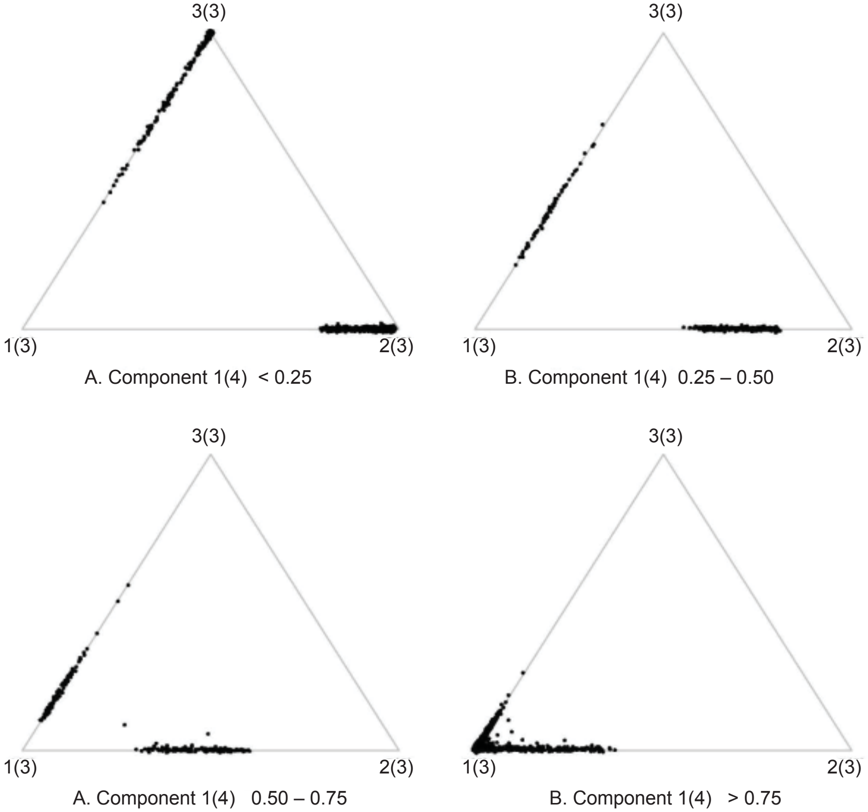
The probabilities of belonging to the components of a three-component mixture model are presented by a ternary graph in Figure 4; see Aitchison (1986). Each point in the equilateral triangle represents a household. The points for households with near-full

DOI: 10.18267/j.pep.487

probabilities of belonging to component $k(3)$, $k = 1$, 2 or 3, are placed at the vertex $k(3)$. The point that represents a household is placed so that its distances from the three sides of the triangle are proportional to the probabilities of belonging to the opposing vertices. For example, the point for a household with probabilities 0.8, 0.2, 0.0 for the respective components 1, 2 and 3(3) is placed on the side connecting vertices 1(3) and 2(3), one-fifth of the way from 1(3) toward 2(3). The points on or near a side of the triangle correspond to households that are unlikely to belong to the component indicated at the opposing vertex. Points (deep) inside the triangle represent households that have appreciable probabilities of belonging to either of the three components. There are only a handful of such points in Figure 4. To avoid overprinting points at the same location in the figure and to make the density of points transparent, a small amount of random noise is added to each point in both horizontal and vertical directions.

Figure 5

Ternary Plots of the Probabilities of Belonging to the Components of the Three- and Four-Component Mixture Fits



A. Component 1(4) < 0.25

B. Component 1(4) 0.25 – 0.50

A. Component 1(4) 0.50 – 0.75

B. Component 1(4) > 0.75

Note: In each panel only households with fitted probabilities that satisfy the condition stated in the subtitle are marked; panels A (component 1(4)<0.25) – D (component 1(4)>0.75).

Source: EU-SILC, own calculation

DOI: 10.18267/j.pep.487

Absence of any points on the side 2(3) – 3(3) suggests the ordering 2(3) – 1(3) – 3(3), which corresponds to the ordering of the components by their variances and correlations; see Table 2. The temporal dependence structure seems to be a more important feature of the data than the annual means of the components.

The association of the assignment based on the three-component mixture model and the component 2(4) can be explored similarly. To save space, we omit the graph and summarise it as follows. The household-specific probabilities of belonging to component 2(4) differ very little from the probabilities for component 2(3). In fact, the sample correlation of these two pairs of probabilities is 0.997, and the marginal probabilities, 0.340 and 0.358, differ only slightly. The correlation for the components 1(4) and 1(3) is 0.988.

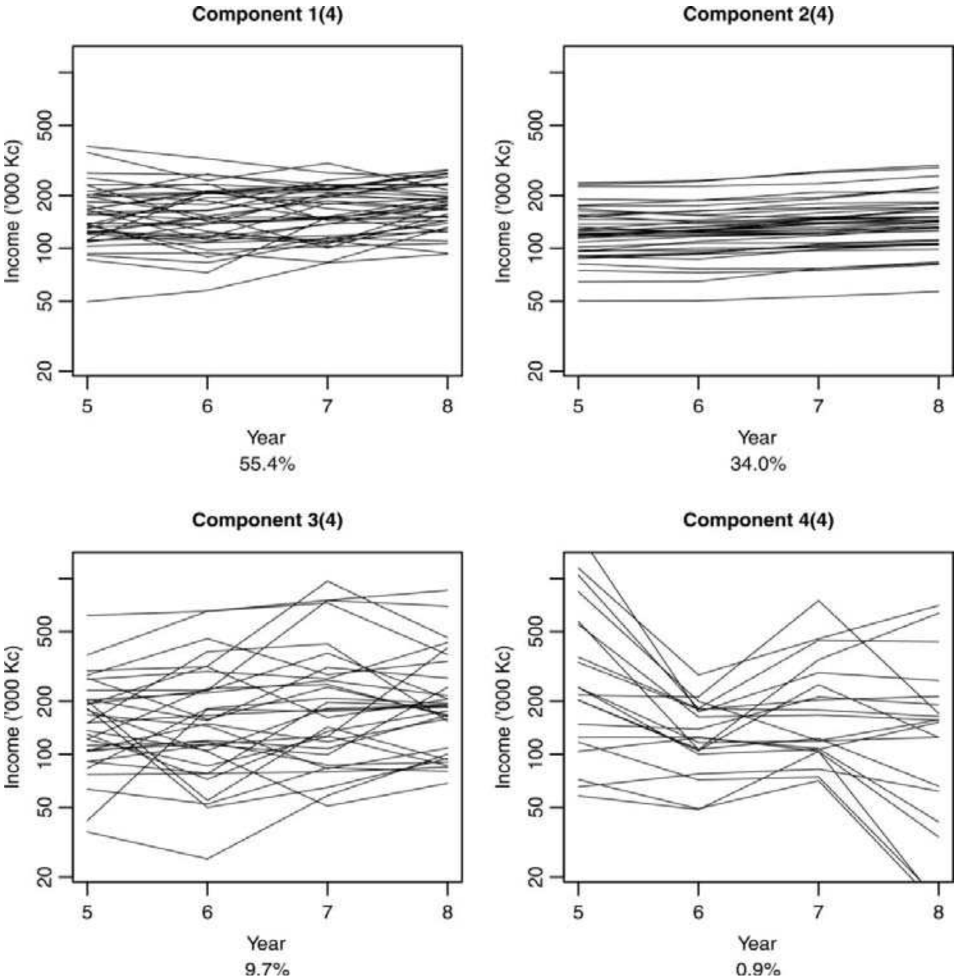
Figure 5 presents the ternary plots, which relate the probabilities of assignment in the three-component model to the probabilities of belonging to component 1(4). The graph confirms that the probabilities of belonging to component 1(4) are similar to the probabilities of 1(3). The probabilities of belonging to component 1(3) tend to be greater than for 1(4) when component 3(3) is a plausible alternative to 1(3), and tend to be smaller when 2(3) is a plausible alternative.

We conclude with Figure 6 in which the annual values of *eHI* are drawn for a small random sample from each component of the four-component model fit. A household is drawn into the sample represented in the panel for component *k*(*K*) with probability proportional to the product of the sampling weight and the conditional probability of belonging to the component. Forty profiles are drawn in the first two panels, 32 in the panel for component 3(4) and 20 in the panel for 4(4). These choices are made to avoid excessive clutter of lines for components with greater dispersion and to reflect the smaller marginal probabilities of components 3(4) and 4(4).

DOI: 10.18267/j.pep.487

Figure 6

Random Samples from Each Component of the Four-Component Mixture Model Fit; four panels for component 1(4) – 4(4)



Source: EU-SILC, own calculation

7. Conclusion

The Czech Republic, as part of the Czechoslovak Socialist Republic until 1989, had a centrally planned economy. The 'Velvet Revolution' in 1989 initiated the period of transformation to democracy with a market oriented economy. The totalitarian regime controlled the wages by various devices, keeping their differentiation to minimum, and largely avoided unemployment and poverty (Mareš, 2000). After 1989, these controls were removed and the economic reforms that followed brought about more inequality,

DOI: 10.18267/j.pep.487

higher unemployment and more poverty (Večerník, 1991). The continuity and stability of income of many households was interrupted and the distribution of wages gradually changed. Its dispersion increased with more outliers at either extreme. Whilst a single distribution, such as lognormal, Weibull or Pareto, describes the pre-1989 situation adequately, mixture models are essential for the income distribution after 1989.

We presented a comprehensive analysis of the household income in the Czech Republic, using mixtures of lognormal distributions. The high level of stability of income, inferred from a relatively large component with extremely high correlations across the years, is an important feature of the results.

The results, if presented mechanically, are extensive and difficult to digest, but with suitable reduction and formatting, they offer clear insights into the structure and evolution (dynamics) of household income. Graphical displays are an important element of this, both for representing the estimated means, variances and correlations of the components (the estimates of the model parameters) and for studying the assignment of the households into the components. The results of our analysis indicate that stability (predictability) is the main feature of the annual (equivalised) household income, not their level. This is an important conclusion, showing an expectation of long-term poverty or prosperity of a household, depending on its current financial potential. During the current economic crisis, this means that households with income below the poverty line or just above it, are in danger of long-term poverty.

The high stability of income in the Czech Republic is largely in agreement with the (multivariate) distribution of income in several other countries of the European Union (Longford and Pittau, 2006). An open question remains to what extent this stability will be maintained if the current economic crisis persists. The methods presented are relevant to these issues and should be applied when more recent data (*e.g.* for 2008–2012) become available.

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Bartošová, J., Forbelská, M. (2011), “Differentiation and Dynamics of Household Incomes in the Czech EU-SILC Survey in the Years 2005–2008.” *APLIMAT – Journal of Applied Mathematics*, Vol. 4, No. 3, pp. 199–208.

Benáček, V., Michalíková, E., Mysíková, M., Nešporová, O., Nešpor, R., Večerník, J. (2010), *Individuals and Households in the Czech Republic and CEE countries*. Prague: Institute of Sociology, AS CR.

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.

Bílková, D., Malá, I. (2012), “Modelling the Income Distributions in the Czech Republic since 1992.” *Austrian Journal of Statistics* [online], Vol. 41, No. 2, pp. 133–152.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society*, Vol. 39, No. 1, pp. 1–38.

Fraley, C., Raftery, A. E. (2002), “Enhanced Software for Model-Based Clustering, Discriminant Analysis and Density Estimation.” *Journal of the American Statistical Association*, Vol. 97, pp. 611–631.

DOI: 10.18267/j.pep.487

Gibson, J., Glewwe, P. (2006), "Analysis of Poverty Dynamics." New York: United Nations Statistics Division. *http://unstats.un.org/unsd/methods/ poverty/pdf/Chapter-8.pdf.*

Glewwe, P. (2005), *How much of Observed Economic Mobility Is Measurement Error? A Method to Reduce Measurement Error, with an Application to Vietnam*. St. Paul, MN: Department of Applied Economics, University of Minnesota.

Labudová, V., Vojtková, M., Linda, B. (2010), "Application of Multidimensional Methods to Measure Poverty." *E+M Ekonomie a management*, Vol. 13, No. 1, pp. 6–21.

Longford, N. T., D'Urso, P. (2011), "Mixture Models with an Improper Component for Panel Data." *Journal of Applied Statistics*, Vol. 38, No. 11, pp. 2511–2521.

Longford, N. T., Nicodemo, C. (2009), "A Sensitivity Analysis of Poverty Definitions." *IRISS Working Paper Series 2009–15.* Differdange, Luxembourg: CEPS/ INSTEAD.

Longford, N. T., Pittau, M. G. (2006), "Stability of Household Income in European Countries in the 1990s." *Computational Statistics and Data Analysis,* Vol. 51, No. 2, pp. 1346–1384.

Malá, I. (2012), "Použití konečných směsí pro modelování příjmových rozdělení." *Acta Oeconomica Pragensia*, Vol. 20, No. 4, pp. 26–39.

Marek, L. (2010), "Analýza vývoje mezd v ČR v letech 1995–2008." *Politická ekonomie*, Vol. 58, No. 2, pp. 186–206.

Marek, L. (2011), "Gini Index in Czech Republic in 1995–2010." *Statistika*, Vol. 48, No. 2, pp. 42–48.

Mareš, **P.** (2000), "Poverty, Marginalisation, Social Exclusion." *Czech Sociological Review,* Vol. 36, No. 3, pp. 285–297.

McLachlan, G., Peel, D. (2000), *Finite Mixture Models.* New York: Wiley.

Pacáková, V., Foltán, F. (2011), "Analysis of the Highest Wages in the Slovak Republic." *Scientific Papers of the University of Pardubice*, *Series D*, Vol. 19, No. 1, pp. 172–180.

Pacáková, V., Sipková, Ľ., Sodomová, E. (2005), "Štatistické modelovanie príjmov domácností v Slovenskej republike." *Ekonomický časopis,* Vol. 53, No. 4, pp. 427–439.

Pittau, M. G., Zelli, R., Johnson, P. A. (2010), "Mixture Models, Convergence Clubs, and Polarization." *Review of Income and Wealth,* Vol. 56, No. 1, pp. 102–122.

Stankovičová, I. (2010) "Regionálne aspekty monetárnej chudoby na Slovensku." Herľany 13.10.2010, In *Sociálny kapitál, ľudský kapitál a chudoba v regiónoch Slovenska*, Košice: Technická Univerzita Košice, pp. 67–75.

Večerník, J. (1991), Introduction into Study of Poverty in Czechoslovakia. *Czech Sociological Review,* Vol. 27, No. 5, pp. 577–602.

Večerník, J. (2009), *Czech Society in the 2000s: A Report on Socio-economic Policies and Structures.* Prague: Academia.

Želinský, T. (2010), "Analýza chudoby na Slovensku založená na koncepte relatívnej deprivácie." *Politická ekonomie*, Vol. 58, No. 4, pp. 542–565.