HOW ARE THE CZECH REGIONS DIFFERENT AND MUTUALLY SIMILAR IN TERMS OF WAGES? CLUSTER ANALYSIS AND WAGE MODELS

Diana Bílková

Abstract

Income levels of the population have been constantly researched by economists in the developed countries mainly due to their connection with the living standards of the population. Knowledge of the wage distribution and its comparison from various socio-economic and time-spatial aspects is a precondition for the assessment of living standards, social security and equality in the division of material values produced by the society. Statistical analysis of the wage distribution also forms the basis for government social policy, taxation, budgetary and other decisions. Moreover, the direct connection between wages and the purchasing power of the population brings tracking the level, structure and development of the wage distribution to the foreground when identifying sales opportunities for the products of both long- and short-term consumption.

The present paper deals with a comparison of wage levels of fourteen regions in the Czech Republic. Similar wage-level clusters were created using the methods of cluster analysis. Three regions with the highest and lowest wage levels, respectively, were selected. For these six regions, the model wage distribution was presented to enable the comparison of wage development over the past seven years. Three-parameter lognormal curves represent the basis of the theoretical wage distribution.

Key words: cluster analysis, method of the furthest neighbour, Euclidean distance matric, wage models, Akaike and Bayesian information criterions

JEL Code: J31, D31, E24

Introduction

Standards of living cannot be exhaustively quantified since they are defined as the level of comforts that include the aggregate of all living conditions, both material and social. We therefore focus only on statistically measurable components of living standards. It is

necessary to capture the level and structure of wages comprehensively, proposing appropriate probability models of the wage distribution for particular social groups and the entire population, respectively, allowing for proper quantification of constituent wage-based elements of living standards of the population.

A number of authors deals with the issue of the labor market and living standards of the population of the Czech Republic, see for example (Pavelka & Löster, 2013) and (Pivoňka & Löster, 2014).

Probability models represent simple approximations of often complicated empirical distributions. Their parameters' developmental trends form the basis for future consumption estimates and the predictions of the consequences of various social and economic provisions.

Economic developments always follow political ones, thus certain inertia in the development of the wage distribution can be seen, its changes emerging gradually with the passing of time. An impact of the 2008 financial crisis, for instance, became fully evident in the wages of Czech employees as late as in 2011, which was a critical year in terms of the wage level development in the Czech Republic, the economic downturn slowing and eventually freezing the growth of wages.

1 Database

Data for the present study are collected from the official website of the Czech Statistical Office (CSO), the database containing the total wage distribution for the period 2009–2015 that covers all employees in the Czech Republic broken down by regions. Annual data are related to gross monthly nominal wages in the respective years, the average (median) wage, for instance, representing average (median) gross monthly wage over the year.

There are also data in the form of the interval frequency distribution with uneven and extreme open intervals. Neither more detailed nor individual data have been currently available. Since only nominal wage data are provided by the CSO, the obtained average and median nominal wages had to be converted to average and median real wages using the CSO-reported inflation rate data.

Only the data on nominal wages having been available, inflation rates had to be used for the conversion to a real wage that reflects purchasing power allowing for a comparison of the wage development without inflation effects in the research period. The rate of inflation is derived from the consumer price index (CPI), based on the Laspeyres price index. The real wage was calculated using the real wage index, the nominal wage index being divided by the

CPI (living cost index). The data were processed utilizing SAS and Statgraphics statistical programme packages and Microsoft Excel spreadsheets.

The late 1990s marked the culmination of Czech economy transformation to the market economy. Enterprises were being privatized, industries restructured and prices liberalized. These processes necessitated a change in the methodology of statistical analysis, the combination of exhaustive and sample surveys becoming more appropriate since the number of small businesses was growing considerably. A high rate of inflation was recorded, namely 8.5–10.7 per cent between 1996 and 1998. In 2003, on the other hand, the inflation rate fell as low as 0.1 per cent, rising to 6.3 per cent in 2008, when the recession triggered a sharp slowdown in real wage growth. Currently, the rate of inflation remains at a very low level (0.3 and 0.7 per cent in 2015 and 2016, respectively).

The research data include wages and salaries paid to employees for work performed in the private (business) and public (state budget, non-business) sectors, respectively. In terms of the data presented on the CSO website, "wages" cover remuneration for work done in both the sectors.

An impact of the 2008 financial crisis and subsequent economic downturn was clearly noticeable in the Czech Republic. Having recorded a decline of 4.8 per cent in 2009, the economy revived slightly over the following two years, the GDP growth reaching maximum of 2.3 per cent. However, since firms did not have enough time to recover and boost their investments, the Czech economy declined further by 0.8 and 0.7 per cent in 2012 and 2013, respectively. Despite the 2010 warnings of independent analysts, a fall into a protracted recession was not prevented. (This double recession is considered even more severe than the 1997–1998 slump. Having fallen by less than one percentage point of GDP, the Czech economy reported a GDP growth of 1.4 per cent in 1999.) In 2014 and 2015, however, we witnessed a GDP growth of 2 and 4.3 per cent, respectively. For due reasons, the present study spans just the seven-year period 2009–2015.

2 Theory and Methods

2.1 Cluster Analysis

Cluster analysis was used to divide the Czech regions into relatively homogeneous groups according to their respective gross monthly wage levels. Multivariate data analysis, which is often done to process economic data (see, e.g., (Malec, 2016)), may include other approaches

to statistical data analysis, namely that of canonical correlation, or (Malec & Malec, 2013), deal with application of two-set multivariate statistical methods to the Czech Republic arrival tourism data. (Particular aspects of cluster analysis are dealt with in (Longford & Bartošová, 2014) (Makhalova &Pecáková, 2015), (Řezanková & Löster, 2013) or (Šimpach & Pechrová, 2016).)

Multidimensional observations can be used when classifying a set of objects into several relatively homogeneous clusters. We have a data matrix $\boldsymbol{X}$ of n X p type, where n is the number of objects and p is the number of variables. Assuming various decompositions $S^{(k)}$ of the set of n objects into k clusters, we look for the most appropriate decompositions. The aim is to find the objects within certain clusters that are as similar as possible to those from other clusters. Only decompositions with disjunctive clusters and tasks with a specified number of classes are conceded.

2.1.1 Criteria for Assessing the Quality of Decomposition

The general task is to assess to what extent the cluster analysis aim has been achieved in a given situation, while applying a specific algorithm. Several criteria – decomposition functions – are proposed for this purpose. The most frequently used ones exhibit the following characteristics. They are the matrices of internal cluster variance

$$\boldsymbol{E} = \sum_{h=1}^{k}\sum_{i=1}^{n_h}(\boldsymbol{x}_{hi} - \bar{\boldsymbol{x}}_h)\cdot(\boldsymbol{x}_{hi} - \bar{\boldsymbol{x}}_h)^{/} \tag{1}$$

and between-cluster variance

$$\boldsymbol{B} = \sum_{h=1}^{k} n_h\cdot(\bar{\boldsymbol{x}}_h - \bar{\boldsymbol{x}}\ )\cdot(\bar{\boldsymbol{x}}_h - \bar{\boldsymbol{x}}\ )^{/}, \tag{2}$$

whose sum is the matrix of total variation

$$\boldsymbol{T} = \sum_{h=1}^{k}\sum_{i=1}^{n_h}(\boldsymbol{x}_{hi} - \bar{\boldsymbol{x}}\ )\cdot(\boldsymbol{x}_{hi} - \bar{\boldsymbol{x}}\ )^{/}. \tag{3}$$

There are vectors of the observations for the i^{th} object and h^{th} cluster $\boldsymbol{x}_{hi}$, the averages for the h^{th} cluster $\bar{\boldsymbol{x}}_h$ and those for the total set $\bar{\boldsymbol{x}}_.$. There are p^{th}-membered vectors, $\boldsymbol{E}$, $\boldsymbol{B}$ and $\boldsymbol{T}$ being symmetric square matrices of the p^{th} order. The principal aim, consisting in the creation of

mutually distant compact clusters, is fulfilled by reaching the minimum of the total sum of the deviation squares of all values of corresponding cluster averages

$$C_1 = \mathrm{st}\,\boldsymbol{E} = \sum_{h=1}^{k}\sum_{i=1}^{n_h}\sum_{j=1}^{p}(x_{hij} - \bar{x}_{hj})^2 \ , \tag{4}$$

i.e. the Ward criterion. Since the st $\boldsymbol{T}$ is the same for all decompositions, the minimization of the st $\boldsymbol{E}$ means the same as that of the st $\boldsymbol{B}$. In order to become independent on the used units of measurement (or, more generally, the invariance to the linear transformations), it is recommended to minimize the determinant of the matrix of the internal cluster variance

$$C_2 = |\boldsymbol{E}|$$

or to maximize the trace criterion

$$C_3 = \mathrm{st}\,(\boldsymbol{B}\boldsymbol{E}^{-1}) \text{ or else } C_4 = \mathrm{st}\,(\boldsymbol{B}\boldsymbol{T}^{-1}).$$

The criteria mentioned above are employed not only retrospectively to assess the decomposition quality accomplished, changes in criterion values also guiding the creation of clusters. Since the criteria ultimately reach the limits (C_1 and C_2 the minimum, C_3 and C_4 the maximum) at $k = n$, it is necessary to find the extreme of the purpose function that properly includes the loss following from the growth in the number of clusters. The Ward criterion, for instance, is proposed to move towards the minimization of the quantity

$$Z_1 = C_1 + z \cdot k \ , \tag{5}$$

where constant z represents the loss resulting from an increase in the number of clusters by one.

2.1.2 Distance and Similarity of Objects

Having selected the variables characterizing the properties of the clustered objects and found their values, we decided on the method of the evaluation of distance or similarity of objects, the calculation of appropriate measures for all pairs of objects often being the initial stage of clustering algorithm implementation. The symmetric square matrix of n X n type has zeros or ones on the diagonal, depending on whether it is the matrix of distance $\boldsymbol{D}$ measures or that of similarity $\boldsymbol{A}$ measures, respectively.

Let us now focus on measuring the distance of the objects described by quantitative variables. The Hemming distance can be used when individual variables are roughly on the same level or at least expressed in the same units of measurement

$$D_{\mathrm{H}}(x_i, x_{i'}) = \sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right| . \tag{6}$$

The Euclidean distance can be applied in the same case

$$D_{\mathrm{E}}(x_i, x_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2} \tag{7}$$

as well as the Chebyshev distance

$$D_{\mathrm{C}}(x_i, x_{i'}) = \max_j \left| x_{ij} - x_{i'j} \right| . \tag{8}$$

All the above mentioned measurements have some common drawbacks – the dependence on the used measuring units that sometimes hinders the meaningful acquisition of any sum for different variables and the fact that if the variables are considered in sum with the same weights, the strongly correlated variables have a disproportionately large effect on the outcome. The starting point is the transformation of variables. The adverse effect of the measuring units can be removed by dividing all the values by the balancing factor, which can be presented with the corresponding average $\bar{x}_j$, standard deviation s_j or the range after deletion of extremes

$$\max_i x_{ij} - \min_i x_{ij} .$$

Particular variables can be also assigned more weight – having decided subjectively or on the basis of relevant information – their values then appearing in the formulas for the calculation of distance.

Other measurements of distance and similarity of objects for numerical, ordinal, nominal and alternative variables are described in the professional literature. When dealing with variables of a different type, the Lance-Williams distance is recommended

$$D_{\mathrm{LW}}(x_i, x_{i'}) = \frac{\sum_{j=1}^{p} \left| x_{ij} - x_{i'j} \right|}{\sum_{j=1}^{p} \left(x_{ij} + x_{i'j} \right)} . \tag{9}$$

2.1.3 Algorithm for the Creation of Hierarchical Sequence of Decompositions

The creation of a hierarchical sequence of decompositions belongs to the most widely used techniques applied in the cluster analysis, occurring sequentially in the following steps:

1) $\boldsymbol{D}$ matrix calculation of appropriate measurements of distances;
2) the start of the decomposition process $\boldsymbol{S}^{(n)}$ from n clusters, each of them containing one object;
3) the assessment of the symmetric matrix $\boldsymbol{D}$ (a lower or upper triangle), finding two clusters (the h^{th} and h'^{th} ones) whose distance $D_{hh'}$ is minimal;
4) the connection of the h^{th} and h'^{th} clusters into a new g^{th} cluster, the replacement of the h^{th} and h'^{th} row and column in the matrix $\boldsymbol{D}$ with those of the new cluster, the order of the matrix being reduced by one;
5) renumbering of the order of the cycle $l = 1, 2, \ldots, n - 1$, the identification of the connected objects h, h' and the level of the connection $d_l = D_{hh'}$;
6) returning to step (3) if the creation of decompositions has not been completed by connecting all objects into a single cluster $S^{(1)}$.

A divisive hierarchical procedure, contrary to the agglomerative hierarchical one, is less-used, starting from a single cluster $S^{(1)}$, splitting one of the clusters into two in each step and obtaining $S^{(n)}$ at the end of the process. The results of hierarchical cluster procedures can be effectively displayed in the form of a graphical tree dendrogram.

Given the choice of variables $x_1, x_2, \ldots, x_p$ and the matrix of distances $\boldsymbol{D}$, the results of applying the described algorithm vary according to the way the distance between clusters is evaluated.

Nearest Neighbour Method

Within the nearest neighbour method, both clusters, whose connection is considered, are represented by objects that are the closest to each other. The $D_{hh'}$ distance between the h^{th} and h'^{th} clusters therefore represents the minimum of all $q = n_h n_{h'}$ distances between their objects, the procedure of the third phase of the above algorithm thus being specified. In the fourth step, the h^{th} and h'^{th} rows and columns in the distance matrix are replaced with the new g^{th} cluster's row and column of distances. In the l^{th} cycle, total $n - l - 1$ distances determined by

$$D_{gg'} = \min\left(D_{g'h}, D_{g'h'}\right). \tag{10}$$

can be written.

If the way of evaluation of the proximity or similarity of clusters is given, which also determines the conversion of the distance matrix in each cycle, the above algorithm allows for the creation of a hierarchical sequence of decompositions and construction of the dendrogram.

When using this method, even considerably distant objects can get together in the same cluster if a large number of other objects create a kind of bridge between them. This typical chaining of objects is considered as a drawback, especially if there is a reason for the clusters to acquire the usual elliptical shape with a compact core. This method, however, possesses many positive features that outweigh the above disadvantage.

Furthest Neighbour Method

The method of the furthest neighbour is based on the opposite principle. The criterion for the connection of clusters is the maximum of q possible between-cluster distances of objects. When editing the matrix of distances, we proceed according to

$$D_{gg'} = \max(D_{g'h}, D_{g'h'}) . \tag{11}$$

An adverse chain effect does not occur in this case. On the contrary, there is a tendency towards the formation of compact clusters, not extraordinarily large, though.

Average Linkage Method (Sokal-Sneath Method)

As a criterion for the connection of clusters, this method applies an average of the q possible between-cluster distances of objects. When recalculating the distance matrix, we use

$$D_{gg'} = \frac{n_h \cdot D_{g'h} + n_{h'} \cdot D_{g'h'}}{n_h + n_{h'}} . \tag{12}$$

The method often leads to similar results as the furthest neighbour one.

Centroid method (Gower method)

Unlike the above methods, this one is not based on summarizing the information on between-cluster distances of objects, the criterion being the Euclidean distance of centroids

$$D_{\mathrm{E}}(\bar{x}_h, \bar{x}_{h'}) = \sum_{j=1}^{p} (\bar{x}_{hj} - \bar{x}_{h'j})^2 \ . \tag{13}$$

The recalculation of the distance matrix is done as follows

$$D_{gg'} = \frac{1}{n_h + n_{h'}} \cdot \left(n_h \cdot D_{g'h} + n_{h'} \cdot D_{g'h'} - \frac{n_h \cdot n_{h'}}{n_h + n_{h'}} \cdot D_{hh'} \right) . \tag{14}$$

Ward Method

The method uses a functional of the decomposition quality C_1 in formula (4). The criterion for the cluster connection is an increment to the total intra-group sum of the squares of observation deviations from the cluster average, thus

$$\Delta C_1 = \sum_{i=1}^{g} \sum_{j=1}^{p} (x_{gij} - \bar{x}_{gj})^2 - \sum_{i=1}^{h} \sum_{j=1}^{p} (x_{hij} - \bar{x}_{hj})^2 - \sum_{i=1}^{h'} \sum_{j=1}^{p} (x_{h'ij} - \bar{x}_{h'j})^2 \ . \tag{15}$$

The increment is expressed as a sum of squares in an emerging cluster which is reduced by the sums of squares in both vanishing clusters. Using arithmetic modifications, the expression can be simplified into the form

$$\Delta C_1 = \frac{n_h \cdot n_{h'}}{n_h + n_{h'}} \cdot \sum_{j=1}^{p} (\bar{x}_{hj} - \bar{x}_{h'j})^2 \ . \tag{16}$$

This equation is a product of the Euclidean distance between the centroids of clusters considered for the connection and a coefficient depending on the cluster size. The value of this coefficient grows with an increasing size of clusters, and for fixed $n_h + n_{h'}$ it represents the maximum in the case of same-size ($n_h = n_{h'}$) clusters. Since we create the connections to ensure the minimization of the criterion ΔC_1, the Ward method tends to eliminate small clusters, i.e. to form those of roughly the same size, which is often a desirable property. Starting from the matrix of Euclidean distances between objects in the process of its modification, we can use the formula

$$D_{gg'} = \frac{1}{n_h + n_{h'} + n_{g'}} \cdot [(n_h + n_{g'}) \cdot D_{hg'} + (n_{h'} + n_{g'}) \cdot D_{h'g'} - n_{g'} \cdot D_{hh'}] \ . \tag{17}$$

2.2 Lognormal Distribution

The importance of the lognormal distribution as a sample distribution model is beyond all discussion. It has many practical applications in various fields, e.g. economics, sociology,

technology or astronomy. The lognormal model allows for capturing differentiating features, such as random-to-systematic variance changes, sequential effects of interdependent factors or trends towards geometrically sequential development.

In the field of economics, wages and incomes of the population are among the many phenomena that the lognormal model enables to interpret. When choosing a curve for modelling frequency distribution, it is necessary to meet the following requirements. The curve is supposed to

- reflect the given shape of the frequency distribution, being fully compliant with the relevant distribution modelled according to its basic characteristics, i.e. location, variability, skewness and kurtosis;
- have a relatively simple shape so that it can be easily manipulated, depending on a small number of parameters estimated using a suitable method of point parameter estimation;
- show interpretable parameters allowing for the prediction of their values without using the methods of statistical time series analysis, especially in those cases when sufficiently long time series are not available.

Every option is always a compromise between the above requirements. The parameter functions of lognormal curves allow for an easy interpretation. In the case of a three-parameter lognormal curve, the parameter θ represents the minimum of the curve (the beginning of distribution, theoretical minimum), the expression $\exp(\mu)$ denote the distance of the median wage (income) from this theoretical minimum, parameters μ and σ^2 representing the expected value and variance of logarithms of wage (income) distances from the theoretical minimum θ.

An old notion that in the area of economy the logarithms of the distances of variables from the theoretical minimum θ are normally distributed stems from the fact that the effects of a large number of various stimuli, resulting in the value of a given quantity, are proportional to that quantity at the corresponding time.

A strong concordance of the model with the global wage or income distribution does not mean, however, that a lognormal distribution will suffice for all circumstances or extremely homogeneous subgroups of employees or households categorized in minute detail according to the selected demographic and socio-economic indicators. If the latter is not the case, it is possible to model the wage or income distribution accurately enough with the use of a lognormal curve, parameters of the lognormal distribution being appropriately estimated

from the sample. Alternatively, a shift of the curve can be made with either a subjectively determined wage/income minimum or another (shift) parameter, which is estimated from the sample. This solution led to positive results in the construction of wage and income models for a nationwide scale and large relatively homogeneous groups roughly categorized by some demographic and socio-economic indicators. The lognormal model, on the other hand, is not suitable for subsets formed from minutely classified employees or households. However, this is not the case of the present study that focuses on the nationwide wage distribution in the Czech and Slovak Republics.

2.2.1 Three-Parameter Lognormal Distribution

A continuous random variable X has a three-parameter lognormal distribution with parameters μ, σ^2 and θ, where $-\infty < \mu < \infty$, $\sigma^2 > 0$, $-\infty < \theta < \infty$, if its probability density function has the form

$$\begin{aligned} f(x;\, \mu, \sigma^2, \theta) &= \frac{1}{\sigma \cdot (x-\theta) \cdot \sqrt{2\pi}} \cdot \exp\left[-\frac{[\ln(x-\theta) - \mu]^2}{2\sigma^2} \right], \qquad x > \theta. \\ &= 0, \qquad \text{else.} \end{aligned} \tag{18}$$

The three-parameter lognormal distribution with parameters μ, σ^2 and θ is marked $LN(\mu, \sigma^2, \theta)$, where parameter θ is the beginning of the distribution (theoretical minimum). The probability density function of the three-parameter lognormal distribution is asymmetric, positively skewed. Figures 1 and 2 display the graphs of the probability density function of the three-parameter lognormal distribution depending on the values of parameters of this theoretical probability distribution.

The probability density function of the three-parameter lognormal distribution is sometime introduced in the form

$$\begin{aligned} f(x;\, \gamma, \delta, \theta) &= \frac{\delta}{(x-\theta) \cdot \sqrt{2\pi}} \cdot \exp\left\{ -\frac{1}{2} [\gamma + \delta \cdot \ln(x-\theta)]^2 \right\}, \qquad x > \theta, \\ &= 0, \qquad \text{else,} \end{aligned} \tag{19}$$

where it is valid between the expressions of probability density functions (18) and (19) that

$$\mu = -\frac{\gamma}{\delta} \quad \text{and} \quad \sigma = \frac{1}{\delta}.$$

The distribution function of the three-parameter lognormal distribution has a form

$$F(x) = \Phi\left[\frac{\ln(x-\theta)-\mu}{\sigma}\right], \quad x > \theta. \tag{20}$$

If the random variable X has a three-parameter lognormal distribution $LN(\mu, \sigma^2, \theta)$, then the random variable

$$Y = \ln(X - \theta) \tag{21}$$

has a normal distribution $N(\mu, \sigma^2)$ and the random variable

$$U = \frac{\ln(X-\theta)-\mu}{\sigma} = \gamma + \delta \cdot \ln(X - \theta) \tag{22}$$

has a standardized normal distribution N(0; 1).

Parameter μ is then the expected value of a random variable (21), parameter σ^2 being its variance. Parameter θ is the beginning of the distribution, i.e. theoretical minimum of the random variable X. For $\omega = \exp(\sigma^2)$, the r^{th} common and central moments of the three-parameter lognormal distribution have the forms

$$\mu_r^{\backslash} = E(X^r) = \theta + \exp\left(r \cdot \mu + \frac{r^2\sigma^2}{2}\right), \tag{23}$$

$$\mu_r = E[(X - \mu_1^{\backslash})^r] = \omega^{r/2} \cdot \left[\sum_{j=0}^{r} (-1)^j \cdot \binom{r}{j} \cdot \omega^{(r-j)\cdot(r-j-1)/2}\right] \cdot \exp(r \cdot \mu), \tag{24}$$

specifically,

$$\mu_3 = \omega^{3/2} \cdot (\omega - 1)^2 \cdot (\omega + 2) \cdot \exp(3 \cdot \mu), \tag{25}$$

$$\mu_4 = \omega^2 \cdot (\omega - 1)^2 \cdot (\omega^4 + 2\omega^3 + 3\omega^2 - 3) \cdot \exp(4 \cdot \mu). \tag{26}$$

We obtain the expressions for the expected value and the variance of the random variable X with a three-parameter lognormal distribution from the expressions (23) and (24)

$$E(X) = \theta + \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{27}$$

$$D(X) = \exp(2\mu + \sigma^2) \cdot [\exp(\sigma^2) - 1] = \exp(2\mu) \cdot \omega \cdot (\omega - 1). \tag{28}$$

The formula for median

$$Median(X) = \theta + \exp(\mu) \tag{29}$$

is derived from the relationship for the 100-percent-quantile of the distribution

$$x_P = \theta + \exp(\mu + \sigma \cdot u_P). \tag{30}$$

The three-parameter lognormal distribution is unimodal, having a single mode

$$Mode(X) = \theta + \exp(\mu - \sigma^2) = \theta + \frac{\exp(\mu)}{\omega}. \tag{31}$$

The relationship between the expected value, median and mode follows from the expressions (27), (29) and (31)

$$E(X) > Median(X) > Mode(X), \tag{32}$$

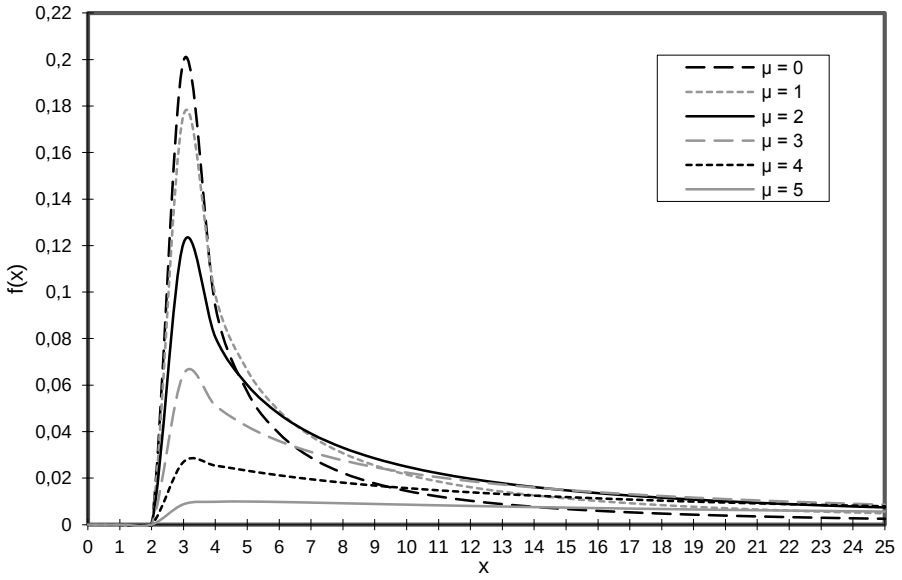
which is typical for a positively skewed frequency distribution.

The coefficient of variation of the three-parameter lognormal distribution is a function of all three parameters

$$V(X) = \frac{\exp\left(\mu + \frac{\sigma^2}{2}\right)\sqrt{\exp(\sigma^2) - 1}}{\theta + \exp\left(\mu + \frac{\sigma^2}{2}\right)} = \frac{\exp\left(\mu + \frac{\sigma^2}{2}\right)\sqrt{\omega - 1}}{\theta + \exp\left(\mu + \frac{\sigma^2}{2}\right)}. \tag{33}$$

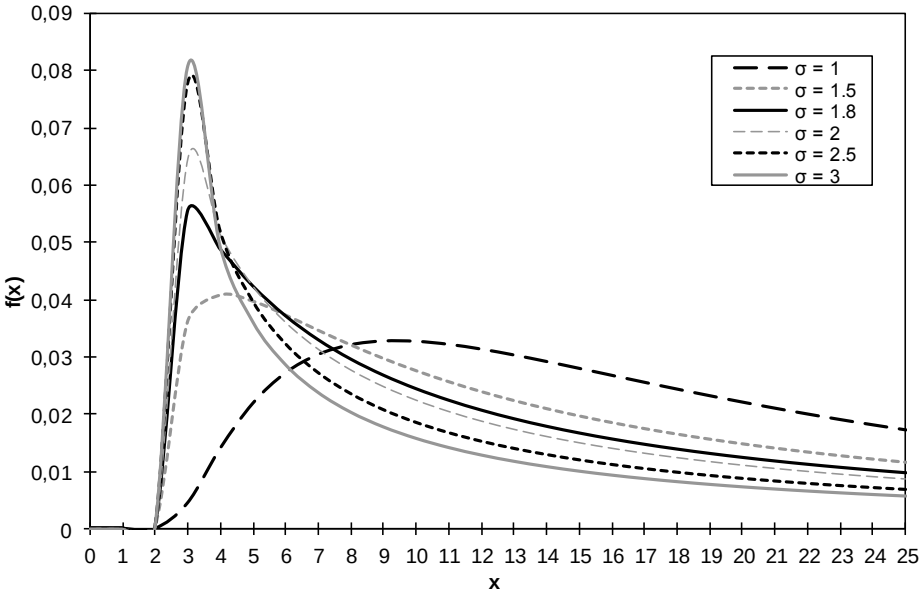
The Gini coefficient of differentiation also depends on the values of all three parameters μ, σ^2 and θ of the distribution

Fig. 1: Probability density function of lognormal distribution for parameter values σ = 2 ($\sigma^2 = 4$); θ = 2



Source: Own research

Fig. 2: Probability density function of lognormal distribution for parameter values μ = 3; θ = 2



Source: Own research

$$G = \frac{\exp\left(\mu + \frac{\sigma^2}{2}\right) \cdot \operatorname{erf}\left(\frac{\sigma}{2}\right)}{\theta + \exp\left(\mu + \frac{\sigma^2}{2}\right)}. \tag{34}$$

Moment measurements of skewness and kurtosis depend on a single parameter σ^2

$$\beta_1 = \sqrt{\exp(\sigma^2) - 1} \cdot [\exp(\sigma^2) + 2] = \sqrt{\omega - 1} \cdot (\omega + 2), \tag{35}$$

$$\beta_2 = [\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 3] = (\omega^4 + 2\omega^3 + 3\omega^2 - 3). \tag{36}$$

2.2.2 Maximum Likelihood of Point Parameter Estimation

The issue of point parameter estimation was studied by many authors in statistical literature, see for example (Malá, 2016), (Sládek, 2017) or (Šimpach, 2012).

Let a random sample of size *n* be taken from the three-parameter lognormal distribution with probability density function (18). The likelihood function then has the form

$$L(\boldsymbol{x}; \mu, \sigma^2, \theta) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2, \theta) =$$

$$= \frac{1}{(\sigma^2)^{n/2} \cdot (2\pi)^{n/2} \cdot \prod_{i=1}^{n} (x_i - \theta)} \cdot \exp\left\{ \sum_{i=1}^{n} -\frac{[\ln(x_i - \theta) - \mu]^2}{2\sigma^2} \right\}. \tag{37}$$

We determine the logarithm of the likelihood function

$$\ln L(\boldsymbol{x}; \mu, \sigma^2, \theta) =$$

$$= \sum_{i=1}^{n} -\frac{[\ln(x_i - \theta) - \mu]^2}{2\sigma^2} - \frac{n}{2} \cdot \ln \sigma^2 - \frac{n}{2} \cdot \ln(2\pi) - \sum_{i=1}^{n} \ln(x_i - \theta). \tag{38}$$

Then we set the first partial derivatives of the likelihood function logarithm equal to zero

$$\frac{\partial \ln L(\boldsymbol{x}; \mu, \sigma^2, \theta)}{\partial \mu} = \frac{\sum_{i=1}^{n} [\ln(x_i - \theta) - \mu]}{\sigma^2} = 0,$$

$$\frac{\partial \ln L(\boldsymbol{x}; \mu, \sigma^2, \theta)}{\partial \sigma^2} = \frac{\sum_{i=1}^{n} [\ln(x_i - \theta) - \mu]^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0.$$

After adjustments, the maximum likelihood estimates of parameters μ and σ^2 for the given parameter θ are obtained as follows

$$\hat{\mu}(\theta) = \frac{\sum_{i=1}^{n} \ln(x_i - \theta)}{n}, \tag{39}$$

$$\hat{\sigma}^2(\theta) = \frac{\sum_{i=1}^{n} [\ln(x_i - \theta) - \hat{\mu}(\theta)]^2}{n}. \tag{40}$$

If the value of the parameter θ is known, we get the maximum likelihood estimates of the remaining two parameters of the three-parameter lognormal distribution using the expressions (39) and (40). However, if the value of the parameter θ is unknown, the situation is more complicated. It can be proved that if the parameter θ is close to $\min\{X_1, X_2, \ldots, X_n\}$, then the maximum likelihood approaches infinity. Moreover, the maximum likelihood method is often combined with the Cohen method, the smallest sample value being set equal to the 100-$(n + 1)^{-1}$ -percent quantile

$$x_{\min}^{V} = \hat{\theta} + \exp(\hat{\mu} + \hat{\sigma} \cdot u_{(n+1)^{-1}}). \tag{41}$$

Equation (41) is then combined with equations (39) and (40).

2.2.3 Akaike and Bayesian Information Criteria

Let us consider *L* as the maximum value of the likelihood function for an assumed model of data, *k* and *n* denoting the number of parameters estimated and the sample size, respectively. The Akaike information criterion (AIC) has the form

$$AIC = 2k - 2\ln L \tag{42}$$

and the Bayesian information criterion (BIC) is defined as

$$BIC = k\ln n - 2\ln L. \tag{43}$$

The model with minimal *AIC* or *BIC* values is preferred over other alternatives, AIC and BIC criteria also including a penalty which is an increasing function of the number of estimated parameters.

3 Results

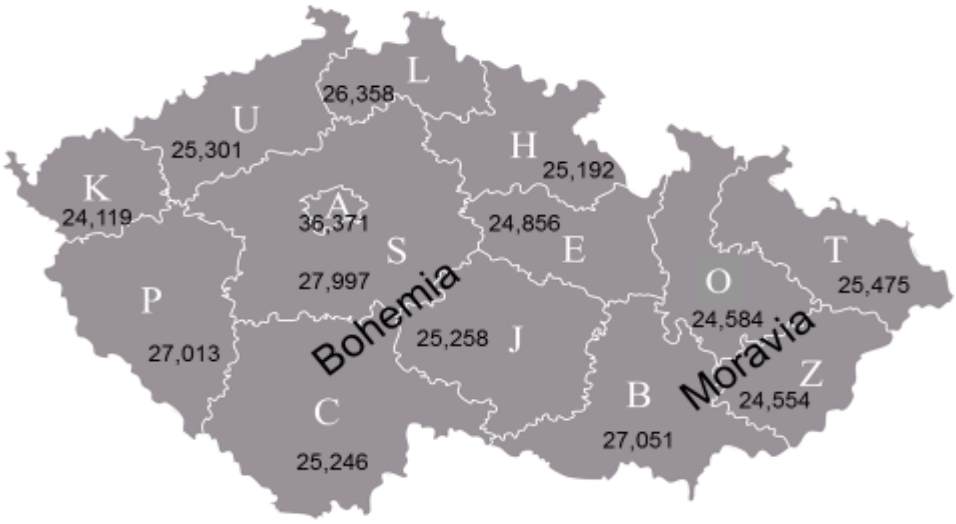
Figures 3 and 4 provide information on the geographical location of each region of the Czech Republic (their official names are presented in Table 1) and the respective level of the gross monthly wage. The figures clearly show a substantially higher wage level in the region of the capital Prague. A relatively high level of wages in Central Bohemian and Pilsen regions is noticeable, low levels, on the other hand, being reported in Karlovy Vary, Zlin and Olomouc regions.

Tab. 1: Official names[1)] of regions of the Czech Republic[2)]

Region	Code	Region	Code
Capital Prague Region	A	Hradec Kralove Region	H
Central Bohemian Region	S	Pardubice Region	E
South Bohemian Region	C	Vysocina Region	J
Pilsen Region	P	South Moravian Region	B
Karlovy Vary Region	K	Olomouc Region	O
Usti Region	U	Zlin Region	Z
Liberec Region	L	Moravian-Silesian Region	T

Source: www.mdcr.cz

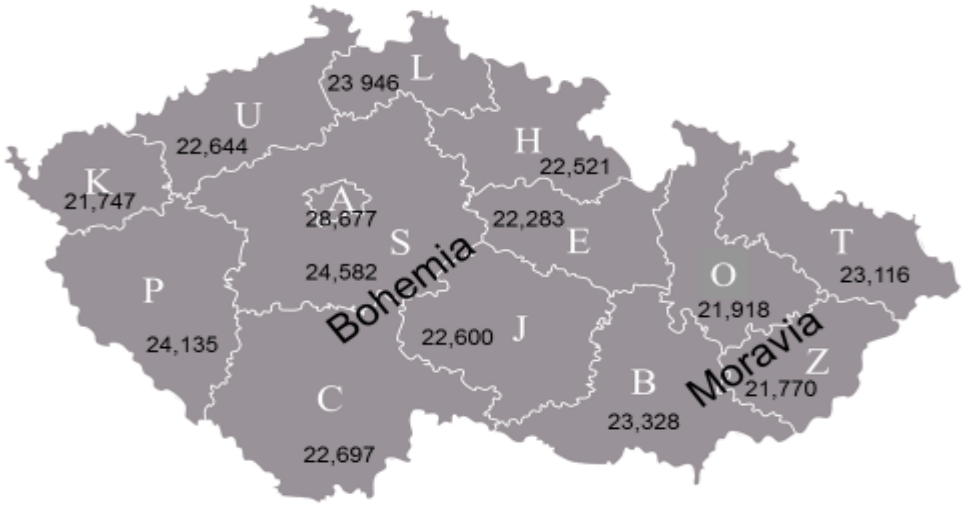
Fig. 3: Average gross monthly wages in respective regions of the Czech Republic in 2015



Source: www.czso.cz

[1)] The names of most regions match those of their respective capitals.
[2)] Different backgrounds distinguish the Bohemian regions (grey) from Moravian ones (white).

Fig. 4: Median gross monthly wages in respective regions of the Czech Republic in 2015



Source: www.czso.cz

Tab. 2: Average unemployment rates (in %) in regions of the Czech Republic in 2015

	Region													
	Central Prague Region	Central Bohemian Region	South Bohemian Region	Pilsen Region	Karlovy Vary Region	Usti Region	Liberec Region	Hradec Kralove Region	Pardubice Region	Vysocina Region	South Moravian Region	Olomouc Region	Zlin Region	Moravian-Silesian Region
Unemp. Rate	2.8	3.5	4.0	3.8	6.7	7.6	5.5	5.6	4.6	4.7	5.0	5.9	4.7	8.1

Source: www.czso.cz

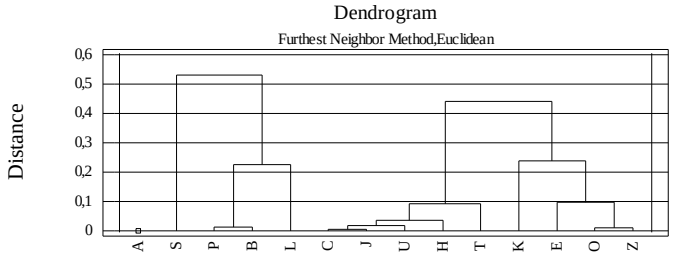
The region of the capital Prague is not food and energy self-sufficient. Its macroeconomic statistics, however, considerably exceed those of other regions, wages earned in Prague resembling those received in more developed countries. One of the reasons is the concentration of industries with higher labour productivity such as finance and informatics. Corporate policies also play an important role, since Prague-based firms often produce values outside the capital but divide the profits at a place where they are headquartered. As expected,

the three regions with the highest wage levels are identical to those with the lowest unemployment rates in the same order. However, the order of the regions at the bottom wage levels is not the same as that of the regions with top rates of unemployment, neither the Moravian-Silesian Region nor Usti Region belonging to the three lowest wage level areas; for details, see Table 2.

Figures 5–8 provide an overview of the results of regional cluster analysis according to the wage level employing the method of the furthest neighbour and Euclidean distance metric.

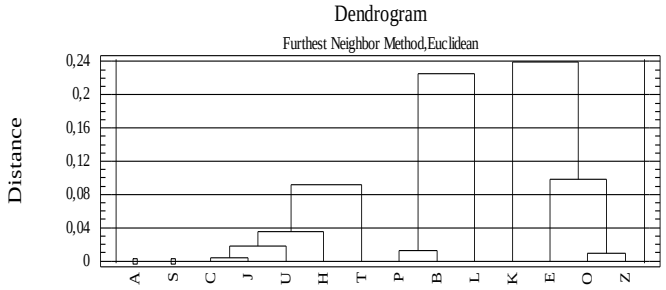
The first cluster always contains only one element – the Capital Prague Region – in the case of both the average and median monthly wage (i.e. three- and five-cluster analysis, respectively), due to markedly higher wage levels in this respective region.

Fig. 5: Cluster analysis using three clusters, furthest neighbour method and Euclidean distance metric; 2015 average wage



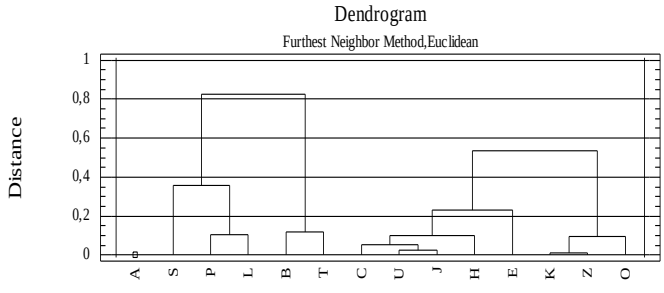
Source: Own research

Fig. 6: Cluster analysis using five clusters, furthest neighbour method and Euclidean distance metric; 2015 average wage



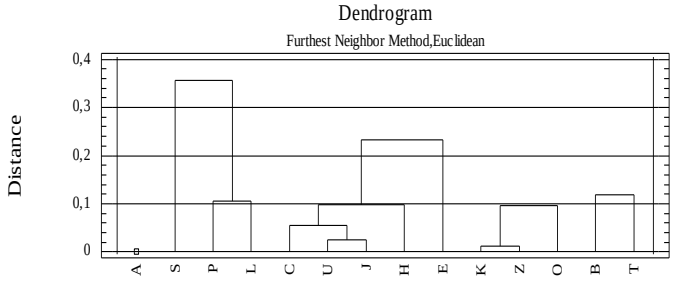
Source: Own research

Fig. 7: Cluster analysis using three clusters, furthest neighbour method and Euclidean distance metric; 2015 median wage



Source: Own research

Fig. 8: Cluster analysis using five clusters, furthest neighbour method and Euclidean distance metric; 2015 median wage



Source: Own research

Within the division of the regions into three clusters by the average wage, the second cluster includes four elements – Central Bohemian, Pilsen, Liberec and South-Moravian regions. According to the median wage division, however, the second cluster has five elements; along with those four mentioned above, this cluster contains the Moravian-Silesian Region, which seems to be rather surprising since this region's general unemployment rate reaches the highest value of the whole Czech Republic. The remaining regions form the third clusters.

Within the division of the regions into five clusters by the average wage, the second cluster contains only one element, namely the Central Bohemian Region. However, according to the median wage, the second cluster consists of three elements – Central Bohemian, Pilsen and Liberec regions. The third cluster always comprises five elements – South Bohemian, Usti, Hradec Kralove, Vysocina and Moravian-Silesian regions by the average wage and South Bohemian, Usti, Hradec Kralove, Pardubice and Vysocina regions, respectively,

according to the median wage. The fourth cluster has only three elements in both cases – Pilsen, Liberec and South Moravian regions according to the average wage, and Karlovy Vary, Olomouc and Zlin regions by the median wage, the latter being those with the lowest wage levels. The fifth clusters formed by the average and median wage contain the four and two remaining regions, respectively; for details, see Figures 5–8).

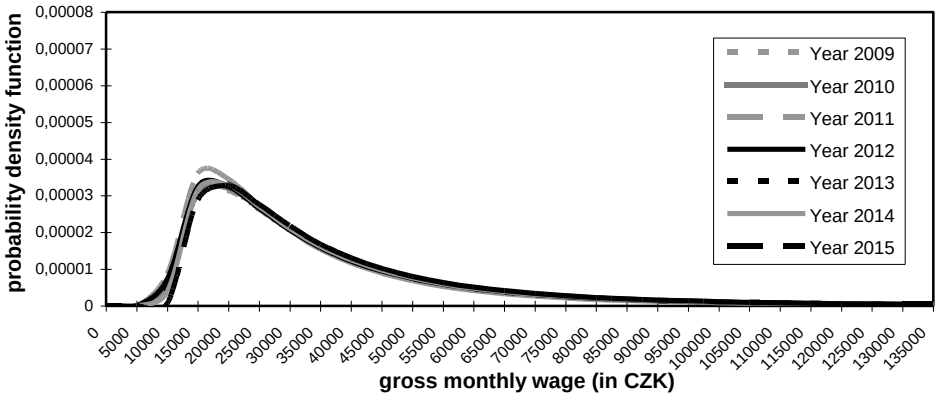
Theoretical models for the wage distribution of each region from 2009 onwards have been constructed. They are based on the use of the probability density function of three-parameter lognormal curves, whose parameters were estimated using the maximum likelihood method. The beginning of these curves is represented by the value of the minimum wage in respective years; see Table 3. The accuracy of the models obtained was compared applying the Akaike and Bayesian information criteria, both of which take a number of the corresponding wage model parameters into account.

Tab. 3: Minimum wage development (in CZK) since 2009

Year	2009	2010	2011	2012	2013[3)]	2014	2015	2016	2017
Minimum wage	8,000	8,000	8,000	8,000	8,000[4)] 8,500[5)]	8,500	9,200	9,900	11,000

Source: www.mpsv.cz

Fig. 9: Development of model wage distributions − Capital Prague Region



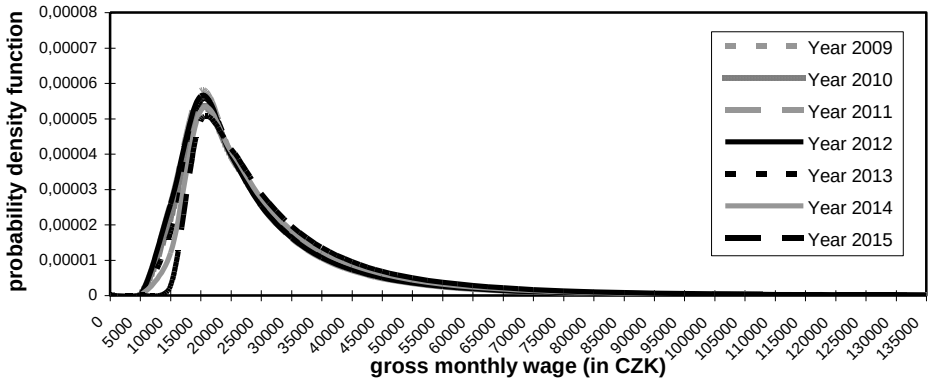
Source: Own research

[3)] In 2013, the beginning of lognormal curves was determined proportionally, i.e

$$8\,000 + 7 * \frac{8\,500 - 8\,000}{12} \doteq 8\,292\,.$$

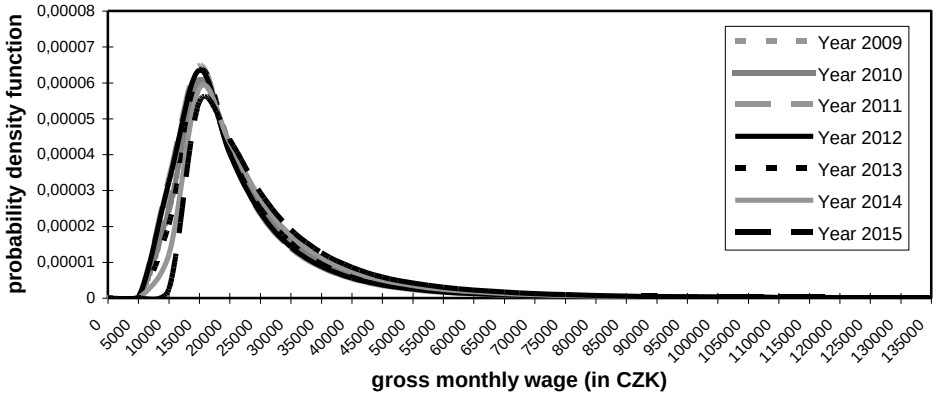
[4)] From 1st January 2013 to 31st July 2013.
[5)] From 1st August 2013 to 31st December 2013.

Fig. 10: Development of model wage distributions − Central Bohemian Region



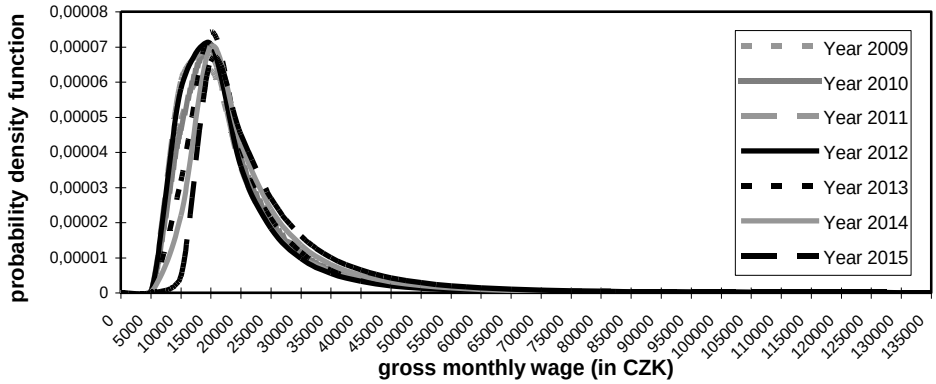
Source: Own research

Fig. 11: Development of model wage distributions − Pilsen Region



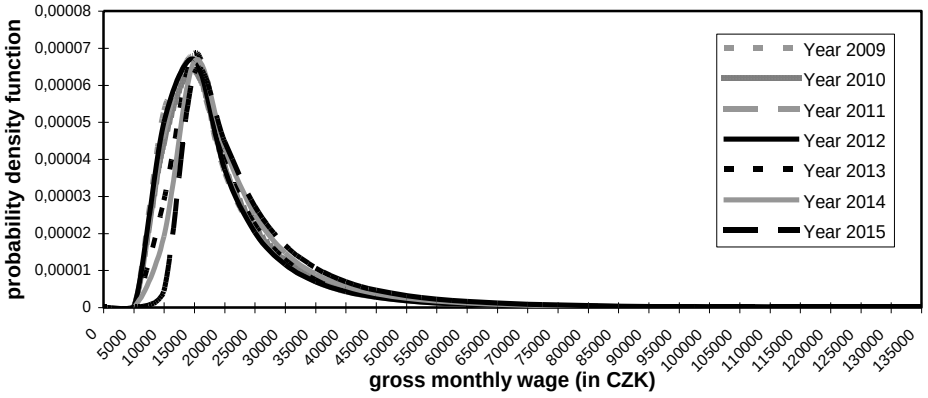
Source: Own research

Fig. 12: Development of model wage distributions − Karlovy Vary Region



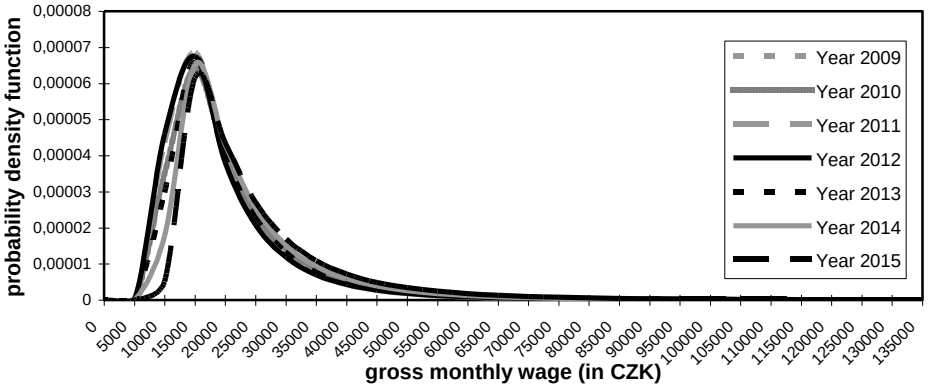
Source: Own research

Fig. 13: Development of model wage distributions − Zlin Region



Source: Own research

Fig. 14: Development of model wage distributions − Olomouc Region



Source: Own research

Theoretical wage models using three-parameter lognormal curves and the maximum likelihood method of parameter estimation were created. Table 4 indicates the values of parameters estimated, Table 5 presenting the values of the Akaike and Bayesian information criteria, which enable to assess the estimation accuracy. Using Figures 9–14, the theoretical wage models capture the three regions with the highest (Prague, Central Bohemian and Pilsen regions) and the other three with the lowest wage levels (Karlovy Vary, Zlin and Olomouc regions).

The above figures allow for comparison of the wage distribution development of the regions with the highest and lowest wages over the last seven years. As observed in the figures, distributions with a high wage level are characterized by higher variability than those

with a low wage level. Moreover, distributions with a low level of wages are more skewed and have higher kurtosis than those with high wage levels.

Tab. 4: Parameter estimates of three-parameter lognormal distribution using maximum likelihood method (parameter θ equalling respective annual minimum wage)

Region	Est.	Year						
		2009	2010	2011	2012	2013	2014	2015
Capital Prague Region	μ	9,937	9,911	9,803	9,885	9,875	9,885	9,907
	σ^2	0,765	0,746	0,740	0,761	0,752	0,765	0,777
Central Bohemian Region	μ	9,426	9,443	9,342	9,366	9,412	9,467	9,512
	σ^2	0,732	0,707	0,683	0,702	0,699	0,711	0,742
South Bohemian Region	μ	9,112	9,156	9,124	9,092	9,176	9,250	9,312
	σ^2	0,701	0,688	0,633	0,633	0,627	0,632	0,668
Pilsen Region	μ	9,298	9,316	9,202	9,226	9,275	9,371	9,426
	σ^2	0,653	0,634	0,639	0,649	0,630	0,639	0,682
Karlovy Vary Region	μ	9,113	9,054	8,978	8,963	9,060	9,149	9,239
	σ^2	0,737	0,627	0,686	0,635	0,577	0,616	0,670
Usti Region	μ	9,264	9,280	9,130	9,169	9,216	9,266	9,337
	σ^2	0,711	0,674	0,671	0,664	0,637	0,654	0,707
Liberec Region	μ	9,326	9,292	9,146	9,182	9,247	9,308	9,392
	σ^2	0,950	0,656	0,636	0,625	0,618	0,626	0,689
Hradec Kralove Region	μ	9,144	9,201	9,087	9,139	9,190	9,259	9,304
	σ^2	0,640	0,671	0,623	0,625	0,620	0,636	0,661
Pardubice Region	μ	9,225	9,168	9,105	9,118	9,160	9,231	9,306
	σ^2	0,920	0,697	0,642	0,660	0,642	0,650	0,704
Vysocina Region	μ	9,195	9,204	9,093	9,133	9,193	9,255	9,322
	σ^2	0,754	0,691	0,616	0,640	0,605	0,627	0,685
South Moravian Region	μ	9,358	9,394	9,268	9,311	9,361	9,404	9,460
	σ^2	1,028	1,025	1,028	1,027	1,018	1,014	1,009
Olomouc Region	μ	9,204	9,203	9,088	9,086	9,164	9,239	9,296
	σ^2	0,661	0,659	0,632	0,657	0,651	0,633	0,719
Zlin Region	μ	9,077	9,139	9,075	9,065	9,149	9,215	9,277
	σ^2	0,739	0,704	0,656	0,680	0,626	0,633	0,686
Moravian-Silesian Region	μ	9,196	9,251	9,220	9,233	9,262	9,290	9,345
	σ^2	0,681	0,664	0,665	0,662	0,656	0,654	0,702

Source: Own research

Tab. 5: Akaike and Bayesian information criteria values

Region	Crit.	Year						
		2009	2010	2011	2012	2013	2014	2015
Capital Prague Region	*AIC*	1,715,847	1,605,510	1,514,319	1,456,900	1,362,167	1,291,967	1,219,753
	BIC	1,715,870	1,605,533	1,514,342	1,456,923	1,362,191	1,291,990	1,219,776
Central Bohemian Region	*AIC*	544 ,441	536,317	528,511	543,060	546,067	556,645	577,484
	BIC	544,462	536,338	528,533	543,082	546,089	556,666	577,505
South Bohemian Region	*AIC*	297,991	296,670	282,364	284,622	285,256	289,311	303,276
	BIC	298,011	296,691	282,384	284,643	285,277	289,331	303,296
Pilsen Region	*AIC*	269,898	267,230	272,194	278,447	275,599	281,651	298,363
	BIC	269,919	267,250	272,215	278,468	275,619	281,671	298,383
Karlovy Vary Region	*AIC*	127,711	115,055	123,387	117,626	110,162	116,655	124,583
	BIC	127,730	115,074	123,406	117,644	110,180	116,674	124,602
Usti Region	*AIC*	348,582	336,708	336,174	334,146	324,956	331,734	350,898
	BIC	348,603	336,728	336,195	334,167	324,977	331,755	350,919
Liberec Region	*AIC*	231,143	185,448	183,454	183,282	183,845	187,689	203,032
	BIC	231,162	185,468	183,474	183,302	183,864	187,709	203,052
Hradec Kralove Region	*AIC*	240,158	250,588	239,702	242,212	242,731	249,249	258,341
	BIC	240,178	250,609	239,722	242,233	242,751	249,270	258,361
Pardubice Region	*AIC*	275,403	235,724	226,401	234,568	233,737	239,552	257,057
	BIC	275,423	235,744	226,421	234,588	233,757	239,572	257,077
Vysocina Region	*AIC*	238,726	228,425	213,311	222,088	215,872	224,336	241,838
	BIC	238,746	228,445	213,331	222,108	215,892	224,356	241,858
South Moravian Region	*AIC*	786,921	788,918	793,299	796,116	795,440	796,875	797,641
	BIC	786,943	788,940	793,321	796,138	795,462	796,897	797,663
Olomouc Region	*AIC*	255,008	261,376	260,425	274,584	279,522	280,684	313,825
	BIC	255,028	261,396	260,445	274,604	279,543	280,705	313,846
Zlin Region	*AIC*	289,572	282,403	271,177	279,996	265,977	270,105	288,029
	BIC	289,593	282,423	271,197	280,017	265,997	270,125	288,050
Moravian-Silesian Region	*AIC*	581,796	574,637	578,439	579,575	578,706	580,905	613,021
	BIC	581,817	574,659	578,461	579,597	578,728	580,927	613,042

Source: Own research

A probability model, usually representing a simple approximation of a rather complex empirical distribution and the knowledge of the development trend of its parameters allow for the estimation of the whole wage distribution for future research purposes.

As we can see from Table 5, wage models for the capital Prague region show the lowest accuracy, while Karlovy Vary region with its lowest wage level indicates the best accuracy of wage models, the number of model parameters being implicated in both information criteria (AIC and BIC). The relationship between the model accuracy and the wage level in the respective region is obvious. It holds in principle that low model accuracy corresponds to a high wage level and vice versa.

Conclusion

The highest and lowest wages, respectively, are reported in Prague and Karlovy Vary regions, the average gross monthly wage amounting to 36,371 CZK in the former, compared to only 24,119 CZK in the latter region in 2015. Residents of Central Bohemian, Pilsen and South Moravian regions receive relatively high wages, averaging 27,997, 27,013 and 27,051 CZK, respectively, in the same year. High-income regions, however, are also characterized by relatively wide gender wage gaps.

The Czech Republic lagging behind economically, the purchasing power of its population currently reaches less than 60 per cent of the European average. Despite a 5 percent annual increase, the dividing line between Western and Eastern European countries still persists and will, unfortunately, likely to remain so for some time to come.

Acknowledgment

This paper was subsidized by the funds of institutional support of a long-term conceptual advancement of science and research number IP400040 at the Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

References

Longford, N. T., & Bartošová, J. (2014). A Confusion Index for Measuring Separation and Clustering. *Statistical Modelling, 14*(3), 229–255.

Makhalova, E., & Pecáková, I. (2015). The Fuzzy Clustering Problems and Possible Solutions. *In: The 9th International Days of Statistics and Economics, Conference Proceedings [online]*, Prague, 10.09.2015–12.09.2015, 1052–1061.

Malá, I. (2016). Properties of moment method estimates based on L moments from right censored data. *In: The 10th International Days of Statistics and Economics, Conference Proceedings [online]*, Prague, 08.09.2016–10.09.2016, 1159–1169.

Malec, L. (2016). Some Remarks on the Functional Relation between Canonical Correlation Analysis and Partial Least Squares. *Journal of Statistical Computation and Simulation, 86*(12), 2379–2391.

Malec, L., & Malec, M. (2013). Application of Two-set Multivariate Statistical Methods to the Czech Republic Arrival Tourism Data. *In: The 7th International Days of Statistics and Economics, Conference Proceedings [online]*, Prague, 19.09.2013 – 21.09.2013, 937-946.

Pavelka, T., & Löster, T. (2013). Flexibility of the Czech Labour Market from a Perspective of the Employment Protection Index. *In: The 7th International Days of Statistics and Economics, Conference Proceedings [online]*, Prague, 19.09.2013 – 21.09.2013, 1090–1099.

Pivoňka, T., & Löster, T. (2014). Clustering of Regions of the European Union by the Labour Market Structure. *In: The 8th International Days of Statistics and Economics, Conference Proceedings [online]*, Prague, 11.09.2014 – 13.09.2014, 1187–1196.

Řezanková, H., & Löster, T. (2013). Shluková analýza domácností charakterizovaných kategoriálními ukazateli (Cluster Analysis of Households Characterized by Categorical Indicators). *E+M Ekonomie a Management, 16*(3), 139–147.

Sládek, V. (2017). Comparison of Robust Estimates. *In: The 16th Conference on Applied Mathematics (APLIMAT 2017), Conference Proceedings [flash disk]*, Bratislava, 31.01.2017–02.02.2017, 1427–1438.

Šimpach, O. (2012). Faster Convergence for Estimates of Parameters of Gompertz-Makeham Function Using Available Methods in Solver MS Excel 2010. *In: Mathematical Methods in Economics 2012, Conference Proceeding [CD-ROM]*, Karviná, 11.09.2012–13.09.2012, 870–874.

Šimpach, O., & Pechrová, M. (2016). Searching for Suitable Method for Clustering the EU Regions according to Their Agricultural Characteristics. *In: Mathematical Methods in Economics 2016,* Conference Proceeding *[CD-ROM]*, Liberec, 06.09.2016–09.09.2016, 821–826.

Contact

Diana Bílková
University of Economics, Prague
Faculty of Information and Statistics
Department of Statistics and Probability
Sq. W. Churchill 1938/4
130 67 Prague 3
Czech Republic
Mail: bilkova@vse.cz