MODELLING DEPENDENCE INDICATORS OF LABOR MARKET USING ADVANCED STATISTICAL METHODS

A. Jindrová, H. Vostrá Vydrová

Abstract

JINDROVÁ, A., VOSTRÁ VYDROVÁ, H.: *Modelling dependence indicators of labor market using advanced statistical methods.* Acta univ. agric. et silvic. Mendel. Brun., 2012, LX, No. 4, pp. 165–172

Aim this paper is an analysis of disparities in the labor market in the Czech Republic. It is based on qualitative indicators. Unemployment is today increasingly perceived as a negative factor that affects the labor market and economy of the state. Due to the nature of the data, categorical data analysis and logistic regression was selected to reveal opportunities to change the role of unemployed persons to employed ones on the labor market. Cluster analysis on categorical data was used for division of regions into similar groups based on parameters affecting the market status of mentioned person in the labor market. Statistical calculations were performed in SPSS statistical software, version 18. The data from the Labor Force Survey was used for evaluation. Specifically, the fourth quarter of 2009, differentiated according to the NUTS 3 (regions of the Czech Republic). These factors were evaluated: age group, highest completed education, disability, participation in informal education, registration in employment office and family status. Based on the analysis we can conclude that the labor market in the CR is considerably regionally, educationally and physically structured. The workforce of individual does not always have the same weight for an employer. More likely to become unemployed, are people with lower education, disabled people and people who are divorced or widowed. Contrariwise, higher chances to be to employed have people who are registered in employment office and are further self-educating people.

labor market, the regions of the Czech Republic, categorical data, cluster analysis, logistic regression

Regional labor markets in the Czech Republic are very specific. Differences between supply and demand for labor is spatial and structural point of view. The main reasons are differentiated qualitative characteristics, such as demands for labor in different regions, etc. These lead to the differentiation of quantitative characteristics of labor markets (number of employed, unemployed, vacancies, etc.). Balanced regional development and reducing disparities between regions are the main objectives of regional policy not only in the Czech Republic, but also throughout the European Union.

This paper is extension of work of authors, which have been published Jindrová and Vostrá Vydrová (2011). This paper is an analysis of the labor market in the Czech Republic based on of qualitative indicators, which were obtained from the Labor Force Survey (LFS). The paper asked the following questions: "What chance do different groups of people have to become unemployed?" Whether the information about the chances to become unemployed are the same or different from the socio-demographic indicators?

Constantly pointed out problems of differences between genders in professional life and thus their success in the labor market showed in their work Buchmann *et al.* (2010), Prix (2009), Smyth (2003).

Unemployment is not only as a factor negatively affecting living standards, but also as a factor hampering social relationships, which creates a sense of limitation. It is particularly repeated and long-term unemployment, which is associated with social exclusion (Jahoda, 1982).

The paper is focused on indicators of quality of life, social and cultural determinants. Based on these, authors evaluated demographics, age and educational attainment, employment and unemployment in the labor market, health care and civic amenities. Method of regression and correlation analysis and cluster analysis were used for evaluation. Finally, it is stated that environmental, social and cultural resources and the region's economic performance affect the quality of life (Živělová, Jánský, 2008).

In the study "Psychosocial work environment and its association with socioeconomic status a comparison of Spain and Denmark, "the aim was to describe psychosocial work environment and wage inequality between Spain and Denmark. For statistical evaluation of obtained data the authors used multivariate methods (correspondence analysis) and ordinal logistic regression (Moncada *et al.*, 2010).

MATERIALS AND METHODS

Analyzed data are obtained from the Labor Force Survey (LFS) at the Czech Statistical Office in Prague. The subjects of investigation, all people usually residents in the studied house that remain or intend to stay in the Czech Republic for at least one year. Sampling households are based on the two-stage selection methods. The unit of first selection is the counting circuit. This is carried by coincidental systematic selection with probability of inclusion proportional to the number of permanently occupied houses in the district. For the second stage it is a flat. Selection is done by simple random selection.

Observed data are weighted by a process where for each person is assigned a certain number of people of the Czech population, which the person presents. Weight is constructed by dividing the number of persons in the population with the number of investigated persons who have the same gender, age and county of residence.

The set of surveyed households (home) is changed regularly so-called rotation of Enumeration Districts. Each household remains under investigation 5 consecutive quarters, which means that the recovery set every quarter is 20%. To illustrate it in the sample survey in each quarter, there is approximately 25 000 flats, which is about 64 thousand people.

To explore the impact of unemployment various statistical methods can be used. To evaluate the impact of zoning on unemployment in the CR multivariate statistical method-cluster analysis was selected. For modeling dependencies between selected indicators of labor market and personal information of respondents was then used regression analysis logistic-regression.

The task of cluster analysis is to find the optimal grouping of data when an individual observation or objects of each cluster are each other similar, but different clusters are the most diverse. Clustering techniques are among the tools of indirect knowledge discovery. In the case of a few (2–3) dimensional clusters it can be recognize by own eyes, with growth of dimension increases the demand to visually identify clusters. The more dimensions, the growing importance of geometric analysis. The reason for the implementation of cluster analysis is the assumption that in the investigated data sets we find meaningful natural groupings of data (Berry, Linoff, 2004).

Common types of clustering methods include hierarchical clustering and decomposition method. A hierarchical clustering starts with *n* clusters, where each observation is a separate cluster, and ends with one cluster, which includes all observations. In each step the two closest observations or clusters of observations are combined into one new cluster. This method is referred to as agglomerative. Less used opposite approach is called divisive. Clustering procedure captures a special tree graph, called dendrogram. It illustrates the steps of hierarchical clustering, including the distance over which the individual clusters (or observation) were combined. This process is irreversible, neither of the two clusters, which were combined (split), cannot be later separated (merged), no earlier error, can be repaired. More information can be found eg. Hebák *et al.* (2007), Řezanková (2007).

Logistic regression is one type of regression model that takes into account the specifics of categorical dependent variable. Also it allows us a description of one-sided dependency of dependent variable using one or more explanatory variables. Logistic regression differs from the linear regression primarily in predicting the probability of an event that either happen (the value is equal to 1) or did not happen (value is equal to 0). To create the conditions of detention is used the logit transformation (Meloun, Militký, 2006).

Created regression model should describe not only the best selection data, but on its basis should be possible to generalize the whole population. This model is evaluated by the ability to determine whether the observed effect happens or not (the procedures appropriate for the non sorted data), then according to the expectations match – the actual and the expected frequency of observed effect happens (the procedures appropriate for the sorted data).

In the case of logistic regression step analysis is to evaluation of the significance of the model based on chi-square statistics. This test should be applied only with great caution and should be taken as a recommendation rather than a true test of significance (Meloun, Militký, 2006).

Statistics, which represents the degree of tightness of interspaced data logistic model, is called the -2 log likelihood (-2 LL), which is asymptotically chi-square distributed. This statistic takes positive values and higher values indicate worse prediction of the dependent variable (Řeháková, 2000).

Estimation of parameters of logistic regression function $(b_0, b_1, b_2, \ldots, b_k)$ is based on the maximum likelihood method, based on maximizing the likelihood function of sample data due to unknown parameters.

Parameters reflect a transformed mean of dependent variable - logit. Logit is the logarithm of the proportion $\pi / (1 - \pi)$ expressing the chance that the value Y (dependent variable) reaches value 1. Parameter β_0 specifies the size of logit at zero value of all explanatory variables. For $\beta_0 = 0$ the chance that $Y = 1$, is one to one, or $\pi = 0.5$. Positive values of this parameter indicate that the chance is greater than one, negative values, then that the chance is less than one (Pecáková, 2007).

To test whether the regression coefficient β_k is statistically significantly different from zero a Wald test criterion is used. It tests the statistical significance of null hypothesis for the individual estimates of regression coefficients. Wald statistic has asymptoticaly chi-square distribution, and therefore this test is suitable only for sufficiently large samples. Decisions made on the basis of Wald characteristics need therefore be supported by evaluation of changes in the logistic model (-2 log likelihood values) for the inclusion or non-inclusion of variables into the model.

Evaluation of classification capabilities of the model is based on the classification table. This is a 4-field table on which basis is possible to specify how many objects were classified correctly.

To test the hypotheses and subsequent analysis significance level $\alpha = 0.05$ was chosen. Statistical calculations were performed in SPSS statistical software, version 18.

RESULTS AND DISCUSSION

Subject to the following analysis are the answers of respondents from the Labor Force Survey, which was carried out in the 4th quarter of 2009. After controlling and simple analysis of individual cases, data matrix was compiled. It included responses from 26 927 respondents surveyed in the age limit from 20 to 64 years. These are economically active persons, which were determined according to age Agenda Europe 2020. It is the growth strategy of the European Union in the next decade (2010–2020). The aim is to create an intelligent and sustainable economy that is supported by social inclusion. The most observed areas are the labor market information and education in individual member countries. Europe 2020 continued in 2010 on the Lisbon Strategy.

All included indicators in the following analysis had the character of categorical variables – nominal and ordinal variables. They were selected after a consultation with an expert on the labor market. The choice of these variables was confirmed by finding values of the descriptors (characteristics of variability). In this exploratory analysis, there were detected higher levels of variability characteristics,

which might reveal the disparities in different regions of the CR.

- Territorial division: Capital city Prague, Central Bohemia, South Bohemia, Pilsen Region, Karlovy Vary Region, Usti Region, Liberec Region, Hradec Kralove Region, Pardubice Region, Highlands Region, South Moravia Region, Olomouc region, Zlin Region, Moravian-Silesian Region.
- Employment status: employed, unemployed.
- Age groups: 20–34 years, 35–49 years, 50–64 years.
- Education: no education and primary education, skilled and high school, college.
- Family status: Single, married, widowed + divorced.
- Disabled: yes, no.
- Informal education: yes, no.
- Registered at the employment office: yes, no.

Assessment of the impact of zoning on employment / unemployment was performed by using cluster analysis. Measuring distance between variables in cluster analysis based on the calculation on Eucleidean distance. On the above-mentioned data was applied cluster analysis, which is dealt with under the procedure for categorical data (Pecáková, 2007).

Individual clusters were tested by multivariate analysis of variance. Among all the clusters were found statistically significant differences ($p < 0.000$).

Cluster analysis, based on categorical indicators of labor market, split region of the Czech Republic into three different clusters. The first cluster contains the county, which can be evaluated as the average labor market. The second cluster focuses on the Moravia Regions, which are characterized by high unemployment and high long-term unemployment. The last cluster is made up of regions with mild low unemployment and high employment of vulnerable groups of people such as women and people aged 50–64 years. A similar procedure for identifying regional differences in the labor market by using cluster analysis was chosen in the paper of Odehnal and Michálek (2009).

Using logistic regression performed modeling of dependencies between selected indicators for individual clusters. Explained variables – employment status was for the analysis purposes converted to the binary variable that takes values unemployed = 0, employed = 1.

Explanatory variables in the model were selected on the bases of a logical assessment of dependencies and with the help of simple dependencies arising from contingency tables. It was a multi categorical variable age group, highest educational received and family status. The logistic regression are thus entering three sets of indicator variables, where the reference category was always chosen as the last category of categorical variables. In the chosen method of calculation based on the algorithm – indicator each category creates a 0–1 variable. Coefficients of indicator variables, then, represent the effect of each category compared with the reference category. Other explanatory variables

are of binary nature (dichotomous variable), it is a disability, the registration at employment office and other forms of education. These variables were coded for logistic regression, as one and zero and these variables are used as conjoin (Hosmer, Lemeshow, 2000).

In order to find the best model the forward step method was used to calculate logistic regression for all three selected clusters. In terms of scope of contribution, the detailed interpretation of the results will be made only for the first cluster. Interpretation of other clusters will be focus primarily on description of the output models.

Cluster 1

Cluster 1 contains the following regions: Region of Karlovy Vary, Usti Region, Liberec Region, Hradec Kralove Region, and Pardubice Region. The classification table of the input model shows, that the model we created correctly classifies 100% of the respondents. The model can therefore be used for prediction. It should be emphasized that this model includes into the calculation only parameter b_0, which gives us the size of logit for zero values of all explanatory variables. A positive value $b_0 = 2.363$ tells us that the chances of being employed is greater than one ($\pi > 0.5$). This means that the ratio of chance that the respondent is employed is 10.611 greater than not employed.

Statistical significance of model parameter is expressed by chi-square statistics, which tests the null hypothesis of zero value for all parameters in the model ($\beta_0 = \beta_1 = \beta_2 = \ldots = \beta_k = 0$). For all these models achieved significance level (p-value) is less than the selected significance level ($\alpha = 0.05$). In this case the null hypothesis is rejected. It can be said that information about the values of explanatory variables allows us to predict better the dependent variable values than it would have been possible without the knowledge of this information.

In the outputs of logistic regression procedures are the characteristics R^2 Cox and Snell with R^2 Nagelkerka that helps us to find the best models that reflect the dependence. Interpretation of the values of these coefficients is the same as in the case of the coefficient of determination in linear regression.

The highest values of these coefficients (0.342, respectively 0.770) were achieved in the 4th step of regression. The latest information can be deduced that 77% of the dependent variable is explained by independent variables that were included in step 4 and we will interpret them latter.

The quality of tightness of data distribution can also be demonstrated by Hosmer-Lemeshow test. This is a test of good conformity that can be used assuming a sufficiently large sample of input data provided and assuming that all expected frequencies are no smaller than 1 and most of them are greater than 5. This test is applicable in our case because the Contingent table, which is part of the SPSS output shows, that none of the expected frequency is less than 1 and only one is less than five. The high value of the chi-square statistics or calculated small p-value (Sig.) indicates that data distribution is not a good fit. From this point of view (Tab. I) the best model appears to be a model that was created in the second step of forward logistic regression in which there is the highest p-value, and the lowest value of the characteristics of the chi-square. In none of these steps p-value is not less than the selected level of significance and can thus be stated that in none of the steps no null hypothesis of tightness of data distribution is rejected.

I: *Hosmer and Lemeshow Test*

Step	Chi-square	df	Sig.
1	0.000	0	.
2	1.821	2	0.402
3	1.873	2	0.392
4	4.289	4	0.368

Source: own computation

Estimates of regression model parameters are listed in Tab. II. Based on the results of the Wald test criteria, it can be state that except for two variables, family status categories – single and education – secondary education category, for all other variables entering the model a significant impact was recorded.

II: *The resulting regression model 4th step of logistic regression*

	Var.	B	Wald	df	Sig.	Exp(B)
Family status			16.348	2	0.000	
Family status (single)	x_1	0.055	0.047	1	0.828	1.056
Family status (married)	x_2	0.772	10.656	1	0.001	2.165
Disability (yes)	x_3	–1.718	35.081	1	0.000	0.179
Education			37.338	2	0.000	
Education (without and basic)	x_4	–1.231	11.815	1	0.001	0.292
Education (high school)	x_5	0.147	0.212	1	0.645	1.159
Registration at the employment office (yes)	x_6	7.036	922.614	1	0.000	1 137.323
Constant		–3.113	55.866	1	0.000	0.044

Source: own computation

The reference category were defined for each variables, all these models as.

Age groups: 50–64 years, education: college, family status: divorced + widowed, disabled: well, informal education: well, registered at the employment office: no.

Important statistics that allow us to interpret the results of logistic regression is the value changes of chance Exp (B). Change in the chance greater than 1 indicates that with increasing value of explanatory variables increases the chances of realization of the output variable. On the other hand, if the value is less than one it indicates that with increasing explanatory variables decreases the chances of realization of the output variable. Dependent variable, which first enters significantly into the model, is family status – category married. In perceiving of opportunities it is evident that these people compared to divorcees or widowers have 2 165 times greater chance to be employed. Another explanatory variable characterizes the status of employed – yes / no is education. From the values of the coefficient changes it is clear that if the interviewee has a basic education or not even that, his chances of being employed are 0.292 times smaller than the university educated one. Health status is significantly influencing variable that has a significant impact on the status of employed / unemployed. In the case of disabled person is the chance to be hired is about 0.179 times smaller than when a person is healthy. Another variable that is statistically important in the model is a variable registration in the employment office. Here is a very high value of chances 1 137.323 and it means that if the respondent is registered at the employment office the chances of success in finding employment is many times higher than those who are not registered there. In following part of the analysis we will proceed to build a regression model. Logit model formulation can be written as follows:

$$\ln \frac{\pi}{1-\pi} = -3.113 + 0.055\, x_1 + 0.772\, x_2 -$$
$$- 1.718\, x_3 - 1.231\, x_4 + 0.147\, x_5 + 7.036\, x_6.$$

The success of the developed model can also be determined from the classification table, which gives us information on the percentage expression of correctly classified objects. In the model we created there are generally 98.2% cases correctly classified.

Cluster 2

Based on the answers of the respondents included in the cluster number 2 (Highlands Region, South Moravia Region, Olomouc Region, Zlin Region, Moravian-Silesian Region), model of logistic regression function was calculated. The Hosmer-Lemeshow test of good conformity sets out that the best model will be created in the third step of the logistic regression.

The calculation of chi-square statistic that has a value of 4 090.7 at 4 degrees of freedom it corresponds to the significance of 0.000 indicating that information about the values of independent variables allows better prediction of the dependent variables than it would be possible without this information. The model explains 73.5% of variation of (R^2) dependent variable.

Estimates of the parameters entering into the regression model are shown in the following table (Tab. III).

The tests of significance of each category of variables based on the calculation of the Wald test criterion shows that family status indicator-single is not statistically significant and thus has no significant influence on the prediction and explanation of the dependent variable. Based on the calculation of statistics -2LL calculated for each set of categorical variables, it is possible to say that variable as a whole has a statistical significance (Tab. IV).

III: *The resulting regression model 3rd step of logistic regression*

	Var.	B	Wald	df	Sig.	Exp(B)
Family status			28.441	2	0.000	
Family status (single)	x_1	0.269	1.806	1	0.179	1.309
Family status (married)	x_2	0.886	22.942	1	0.000	2.426
Disability (yes)	x_3	−1.657	50.620	1	0.000	0.191
Registration at the employment office (yes)	x_6	6.747	1 519.772	1	0.000	851.445
Constant		−3.114	206.864	1	0.000	0.044

Source: own computation

IV: *Model if Term Removed*

	Variable	Model Log Likelihood	Change in -2 Log Likelihood	df	Sig.
	Family status	−996.182	28.063	2	0.000
Step 3	Disability	−1 001.952	39.604	1	0.000
	employment office	−2 860.388	3 756.475	1	0.000

Source: own computation

The resulting model includes 3 statistically significant categories of variables whose interpretation shows that the single (married man / married woman) are 1.3 (2.4) times more likely to be employed than widowed and divorced. In the case that a person is registered at the employment office the chance to be employed is 851.45 times than when not registered. On contrary the chances of becoming employed for persons with disabilities is reducing by 0.191 times.

Logit model form is:

$$\ln \frac{\pi}{1-\pi} = -3.114 + 0.269\,x_1 + 0.886\,x_2 - 1.657\,x_3 + 6.747\,x_6.$$

The final classification table of resulting model shows that this model correctly classifies 99.5% of respondents who are employed and 81.1% of respondents who are unemployed. Overall 98% of respondents is classified correctly.

Cluster 3

To describe the dependence of indicators included in the third cluster (Prague, Central Bohemia, South Bohemia, Pilsen Region) 4th step of forward logistic regression was selected. According to all available statistics, it corresponds with observed data and explains 67.4% of (R^2) dependent variable.

The above table shows that two of the coefficients are statistically not significant. These are the categories of variables, family status - single, and education – high school. The results of statistics -2LL indicates that these variables as a whole have a statistically significant effect on the prediction of the dependent variable.

$$\ln \frac{\pi}{1-\pi} = -3.064 - 0.234\,x_1 + 0.654\,x_2 -$$

$$- 1.833\,x_3 - 0.971\,x_4 + 0.037\,x_5 + 7.255\,x_6.$$

The interpretation of identification of opportunities ratio and logistic model is similar to cluster 1 and 2.

Tab. VI summarizes the results of the number of opportunities ratios in individual clusters of variables included.

The largest inter-regional differences in the labor market have been reported in the variable family status. In the first two clusters are single people more likely to have a job than those who are divorced or widowed. This is specifically about South Moravia and Zlín Region, where there is large percentage of single people. The opposite is the third cluster (Prague, Central Bohemia, Pilsen Region and South Bohemia). Here we see 0.791 times less likely to become employed for single persons. The preferences of employers in these four regions, which tend rather to people who have family support and confidence than divorced and widowed. Which shows the following row in

V: *The resulting regression model 4th step of logistic regression*

	Var.	B	Wald	df	Sig.	Exp(B)
Family status			21.801	2	0.000	
Family status (single)	x_1	–0.234	0.863	1	0.353	0.791
Family status (married)	x_2	0.654	7.111	1	0.008	1.923
Disability (yes)	x_3	–1.833	32.888	1	0.000	0.160
Education			13.951	2	0.001	
Education (without and basic)	x_4	–0.971	8.229	1	0.004	0.379
Education (high school)	x_5	0.037	0.022	1	0.881	1.038
Registration at the employment office (yes)	x_6	7.255	738.49	1	0.000	1415.42
Constant		–3.064	58.772	1	0.000	0.047

Source: own computation

VI: *Opportunities ratios in the various clusters of included variables*

Variables	Cluster 1	Cluster 2	Cluster 3
Family status			
Family status (single)	1.056	1.309	0.791
Family status (married)	2.165	2.426	1.923
Disability (yes)	0.179	0.191	0.160
Education			
Education (without and basic)	0.292		0.379
Education (high school)	1.159		1.038
Registration at the employment office (yes)	1 137.323	851.445	1 415.428
Constant	0.044	0.044	0.047

Source: own computation

table number VI, where you can see up to 2426 times higher chance to become employed when married. Regions with the highest rate of married workers are the capital city of Prague, Central Bohemia, South Bohemia and the Pilsen Region. A disabled person has in all clusters equally low chances of becoming employed. There are no disparities between regions or clusters. The results point to a very disadvantaged position of disabled people in the Czech Republic. Based on the obtained results, the government should help these handicapped persons and give more support to various charity and non-governmental organizations. The same results are in the results of opportunities in the variable of education. Different level of opportunities of becoming employed were recorded in the case that the person is registered at the employment office. The highest chance of becoming employed is in the first cluster. This is due to higher supply of jobs, which is published by the employment offices, especially in the capital city where there is greatest level of employment supply. Number of offices in each region plays an important role due to the availability and speed of information about job opportunities for unemployed. Depending on the willingness of persons seeking employment to work with the following information:

1 cluster has 61 offices (Karlovy Vary Region-10, Hradec Kralove Region-15, Pardubice Region-15, Usti Region-21), 2 cluster has 92 (South Moravian-24, Olomouc Region-19, Highlands-19, Zlin Region-13, Moravian-Silesian Region-27), 3 cluster has 71 (South Bohemia-20, Pilsen Region-15, Prague-10, Central Bohemia-26) (information available on the Public Administration server). The lowest value to become employed is in the second cluster. In this group there are regions of the Czech Republic that have problems with high unemployment (unemployment rate in the Moravian-Silesian Region reached 10.2%), with high long-term unemployment and low labor productivity, etc. Therefore, the opportunity to become employed is the lowest compared with those clusters of two clusters.

CONCLUSION

The goal of any regression model is to find the best and meaningful model that describes the relationship between the dependent variable and a group of explanatory variables. The aim of authors based on logistic regression model was to analyze qualitative data on employment or unemployment and thus to point out that analysis of this issue based on data from surveys can be useful and serve a deeper understanding of the context of the labor market.

Based on the analysis we can conclude that the labor market in the CR is considerably regionally, educationally and physically structured. An individual does not represent the workforce for an employer always at the same level. More likely to become unemployed are people with lower education, disabled people and people who are divorced or widowed. Contrariwise, higher chances of being hired have people who are registered in the employment office and further self-educating people.

The results achieved by construction the logistic models included in this work are comparable with the results published in the work of authors Katrňák and Mareš (2007). These authors emphasize the fact that a group of unemployed and employed in the Czech Republic is highly differentiated, and so should be the different measures of state social policy for different regions (Katrňák, Mareš, 2007).

Prix using logistic regression proved the impact of educational attainment on different status of persons in the labor market. It states that the higher the education the higher the chances of obtaining job. It further states that the basic idea of the labor market is that education is the most important asset in the labor market (Prix, 2009).

Statistical evaluation of regional differences on the basis of labor market indicators using multivariate methods: correspondence analysis and ordinal logistic regression was also applied when comparing Spain and Denmark. The examination results shows that among the surveyed regions there are many disparities in the labor market and the methods used are suitable for inter-regional comparisons (Moncada *et al.*, 2010).

SUMMARY

Aim this paper is an analysis of disparities in the labor market in the Czech Republic. It is based on qualitative indicators. Unemployment is today increasingly perceived as a negative factor that affects the labor market and economy of the state. Due to the nature of the data, categorical data analysis and logistic regression was selected to reveal opportunities to change the role of unemployed persons to employed ones on the labor market. Cluster analysis on categorical data was used for division of regions into similar groups based on parameters affecting the market status of mentioned person in the labor market. Statistical calculations were performed in SPSS statistical software, version 18. The data from the Labor Force Survey was used for evaluation. Specifically, the fourth quarter of 2009, differentiated according to the NUTS 3 (regions of the Czech Republic). These factors were evaluated: age group, highest completed education, disability, participation in informal education, registration in employment office and family status.

Based on the analysis we can conclude that the labor market in the CR is considerably regionally, educationally and physically structured. The workforce of individual does not always have the same weight for an employer. More likely to become unemployed, are people with lower education, disabled people and people who are divorced or widowed. Contrariwise, higher chances to be to employed have people who are registered in employment office and are further self-educating people.

Based on the analysis we can conclude that the labor market in the CR is considerably regionally, educationally and physically structured. An individual does not represent the workforce for an employer always at the same level. More likely to become unemployed are people with lower education, disabled people and people who are divorced or widowed. Contrariwise, higher chances of being hired have people who are registered in the employment office and further self-educating people.

Acknowledgement

REFERENCES

BERRY, M., LINOFF, G., 2004: *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. Second Edition. Indianapolis: Wiley Publishing, 821 p. ISBN 0-471-47064-3.

BUCHMANN, MC., KRIESI, I. and SACCHI, S., 2010: Labour market structures and women's employment levels. *Work employment and society*. 24, 2: 279–299. ISSN 0950-0170.

HEBÁK, P., HUSTOPECKÝ, J., PECÁKOVÁ I., PRŮŠA, M., ŘEZANKOVÁ, H., SVOBODOVÁ, A., VLACH, P., 2007: *Vícerozměrné statistické metody [3]*. Praha: Informatorium, 256 p. ISBN 80-7333-039-3.

HOSMER, D. W. Jr., LEMESHOW, S., 2000: *Application logistic regression. II. Edition.* New York: Wiley-Interscience. ISBN 0471356328.

JAHODA, M., 1982: *Employment and Unemployment: A social psychological analysis*. Cambridge: Cambridge University Press. p 111.

JINDROVÁ, A. and VOSTRÁ VYDROVÁ H., 2011: Evaluation of indicators of labour market using logistic regression. *Sborník příspěvků z konference Agrarien perspectives XX*, 447–454. ISBN: 978-80-213-2196-0.

KATRŇÁK, T., MAREŠ, P., 2007: Segmenty zaměstnaných a nezaměstnaných v České republice v letech 1998 až 2004. *Sociologický časopis Czech Sociological Review.* Praha: Sociologický ústav AV ČR. 43, 2. ISSN 0038-0288.

MELOUN, M., MILITKÝ, J., 2006: *Kompendium statistického zpracování dat.* Praha: Academia, 984 p. ISBN 80-200-1396-2.

MONCADA, S., PEJTERSEN, J., NAVARRO, A., LLORENS, C., BURR, H., HASLE, P. and BJORNER, J., 2010: Psychosocial work environment and its association with socioeconomic status. A comparison of Spain and Denmark. *Scandinavian journal of public health*. England, 38, 3: 137–148. ISSN 1403-4948.

ODEHNAL, J., MICHÁLEK J., 2009: Assessment of Competitiveness of Chosen European Union Regions. *Ekonomický časopis*. Brno, 57, 2: 113–131. ISSN 0013-3035.

PECÁKOVÁ, I., 2007: *Logistická regrese s vícekategoriální vysvětlovanou proměnnou*. Praha: Acta Oeconomica Pragensia, 15, 1: 86–96. ISSN 0572-3043.

PRIX, I., 2009: Gender and recent graduates' occupational stratification: the interactive role of the educational and employment sectors in four countries. *Comparative education*, 45, 4: 545–568. ISSN 0305-0068.

ŘEHÁKOVÁ, B., 2000: Nebojte se logistické regrese. *Sociologický časopis*. Sociologický ústav AV ČR, 36, 4: 475–492. ISSN 0038-0288.

ŘEZANKOVÁ, H., 2007: *Analýza dat z dotazníkových šetrení /1*. Vydání, Praha: Professional Publishing. ISBN 978-80-86946-49-8.

SMYTH, E., 2003: online: http://www.socsci.ru.nl/~maartenw/school_to_work_2ed%20MZES%20publication.pdf#page=60, Gender differentiation an early labour amrket integration Gross Europe.

ŽIVĚLOVÁ, I., JÁNSKÝ, J., 2008: Analýza kvality života z pohledu sociálních a kulturních zdrojů v krajích České republiky. [Online], Available: http://www.mmrvyzkum.cz/infobanka/DownloadFile/38628.aspx [20 Sep 2011].

Address

Ing. Andrea Jindrová, Ing. Hana Vostrá Vydrová, Katedra statistiky, Provozně ekonomická fakulta, Česká zemědělská univerzita, Kamýcká 129, 165 21 Praha 6 – Suchdol, Česká republika, e-mail: jindrova@pef.czu.cz, vydrova@pef.czu.cz