Two-Step Classification of Unemployed People in the Czech Republic

Zdeněk Šulc[1] | *University of Economics, Prague, Czech Republic*
Marina Stecenková | *University of Economics, Prague, Czech Republic*
Jiří Vild | *University of Economics, Prague, Czech Republic*

Abstract

The paper analyzes structure and behavior of unemployed people in the Czech Republic by means of latent class analysis (LCA) and CHAID analysis where the output of LCA serves as the input for CHAID. The unemployed are classified in two steps; for each step different characteristics are used. In the first step, respondents are split into latent classes according to their answers to questions concerning ways of searching for a new job. In the second step, CHAID analysis is performed with results obtained from LCA as a dependent variable. In the paper, data from periodical Labor Force Survey conducted in Czech Republic in spring 2011 are used. The results indicate that unemployed people in the Czech Republic can be divided into four segments: Active, Passive, Typical and Specific. A special attention is paid to extreme segments Active and Passive.

Keywords	**JEL code**
Double-step classification, unemployment, latent class analysis, CHAID algorithm	*E24, C38*

INTRODUCTION

Unemployment is thoroughly observed socio-economic phenomenon. High unemployment shows negative effects on economic, social and psychic situation of individuals, households, and generally, on a whole society. It is connected with many adverse effects, such as poverty, criminality, increased social (and other) expenditures of a state, problems caused by long-term unemployment and other, see Katrňák, Mareš (2007). Therefore, it is necessary to fully understand all aspects of unemployment and to make all possible arrangements to minimize its negative effects. In this paper, we focus on explaining the structure of unemployed people in the Czech Republic. This knowledge can help to improve the unemployment strategy of the Czech Republic and this procedure can also be applied in other countries.

The unemployment rate in the Czech Republic was influenced by economic crisis with a slight delay. The first signs of rising unemployment became obvious in 2009. Since then the unemployment rate holds the constant level around 7%,[2] in 2014 it dropped even below 6%. In comparison with other EU

[1] Faculty of Informatics and Statistics, Department of Statistics and Probability, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. Corresponding author: zdenek.sulc@vse.cz.
[2] Czech Statistical Office <www.czso.cz>.

countries, the Czech unemployment rate was among the lowest. In September 2014, it was lower only in Malta, Austria and Germany.[3] From among the Czech regions, the steadily lowest unemployment rate is in Prague and surrounding Středočeský region. It is just because many companies have their headquarters in these regions. On the other hand, the highest unemployment rate occurs in the Ústecký and Moravskoslezský regions because of high concentration of heavy industry from the past, which was liquidated or significantly reduced in the beginning of the 90's. The situation there improves very slowly. The unemployment rates in the rest of regions can be found in Statistical yearbook of the Czech Republic (2013).

An efficient hybrid methodology, which was introduced by Magidson and Vermunt (2004a), was applied to the data. This methodology combines features of the CHAID algorithm and latent class modelling. Using this methodology, the paper aims to achieve a better understanding of structure and behavior of unemployed people in the Czech Republic. The approach is performed in two steps. In the first step, the unemployed people are divided into several segments (latent classes) according to the way of seeking for a new job. In the second step, CHAID analysis is used to estimate the membership of each object in one of the latent classes, which were created in the first step. The second step is performed in two ways. First, there are used the same indicators, which served for construction of the latent classes. Second, CHAID analysis with five socio-demographic variables is performed with the aim to describe the created latent classes from a different perspective. Thus, we are able to reveal the background of people's attitudes in terms of other social characteristics. There are other interesting approaches to achieve the same goal. For example, Shaunna and Muthen (2009) introduced an approach based on posterior probabilities from LCA. The rest of the paper is organized as follows: Section 1 describes how the unemployment is measured. Section 2 introduces the principles of LCA and in Section 3 the principles of CHAID analysis. Section 4 offers practical application which is performed on data from periodical Labor Force Survey conducted by Eurostat in the Czech Republic in 2011. The results are summarized in Conclusion.

1 UNEMPLOYMENT AND THE LABOR FORCE SURVEY

Indicator of unemployment used in the research comes from Labor Force Survey (LFS). LFS is organized by Eurostat and it provides results which are comparable in all countries of EU and some other countries. It is conducted on a random sample of private households in which all persons are surveyed according to their labor status (employed, unemployed or inactive). The labor status is determined by ILO conditions. According to them, the unemployed are at least 15 years old and have to meet the following conditions during a reference week to be involved in the survey. First, they do not work as paid employees during a reference week. Second, they have been actively looking for a job within a four-week period ending with a reference week. Third, they are available for paid employment within two weeks since the end of a reference week. The indicator of unemployment is computed as the ratio of unemployed people according to ILO conditions to the total labor force.

The data from Eurostat can be broken down by many additional criteria like age, nationality, full-time/part-time employment etc. Thus, they allow for a detailed view at unemployment issues, which suits well for the research performed in this paper.

2 LATENT CLASS ANALYSIS

Latent class analysis (LCA) is a latent variable model in which a categorical latent variable is constructed comprising set of discrete, mutually exclusive latent classes. It is described in detail for example from Haberman (1979) or Vermunt and Magidson (2004). The latent variable is not measured directly but indirectly by means of two or more categorical observed variables. LCA is used to divide objects (respondents, individuals) into homogeneous groups (clusters) according to their characteristics, see Collins

[3] Eurostat <ec.europa.eu/eurostat>.

and Lanza (2010). It is often used in questionnaire surveys, where it helps to identify groups of similar respondents. For each object, probability of belonging (prevalence) to each latent class is computed. Usually, the object is assigned to the latent class with the highest prevalence. When performing LCA, there is a constriction that all input variables have to be mutually independent.

In LCA, two sets of parameters are estimated – the latent class membership probabilities (prevalence), which represent the proportion of the researched population in particular latent class, and the item-response probabilities, which express probability of a particular response to an observed variable, conditional on latent class membership.

2.1 Model fit and model selection

The goodness-of-fit of an estimated latent class model is usually tested by the likelihood-ratio chi-squared statistic which is compared to a critical value of chi-square distribution. To approximate the G2 statistics with the chi-square distribution, it is necessary that each cell of a contingency table has a sufficient number of cases. This situation may not been fulfilled when there are too many observed variables or the observed variables have too many categories compared to the total sample size. One possibility to solve such a problem is to estimate p-values with a bootstrapping technique.

Other methods of evaluating model fit are based on information criteria which penalize models with a higher number of parameters. The most common information criteria are Akaike information criterion (AIC) and Bayesian information criterion (BIC), see Akaike (1973), Schwartz (1978) or Bozdogan (1987). A model with minimum value of AIC or BIC is then selected.

2.2 Classification

When a latent class model is estimated, it can be used for assigning objects to latent classes. The classification is based on their response pattern and posterior probability of membership in each of the latent classes. The classification probabilities are obtained using Bayes rule with estimates of prevalence and item response probabilities. The most common classification rule is modal assignment, which assigns each individual to the latent class with the highest posterior probability. Correctness of the classification can be measured by the entropy, which is measured on a zero to one scale with a value one indicating that the individuals are perfectly classified into latent classes. Generally, higher values indicate better classification of objects.

3 CHAID ANALYSIS

The Chi Square Automatic Interaction Detection Analysis (CHAID), which was originally introduced by Kass (1980), belongs to group of classification methods using unsupervised learning. It provides much better results in comparison with not so appropriate methods for response modelling, such as discriminant analysis or multiple regression, see Madgison (2006). The main aim of the analysis is to find a combination of variables, which best explains the outcome of a given dependent variable. The main output displays relationships among variables in a hierarchical form. CHAID offers two main fields of use. The first one is to determine relationships among variables; the second one to classify objects into classes of dependent variable. It can also be used to find interactions among observed predictors; thus, it can serve to improve results of other data analyses (e. g. classification by neural networks). More detailed view into the CHAID analysis is well described e.g. in Tufféry (2011) or in Madgison (1994). CHAID analysis has several advantages, especially, it does not impose almost any conditions on data. The method is determined for categorical data primarily; however, continuous variables can be categorized into a suitable number of intervals. CHAID visualizes results easily in form of a classification tree. The CHAID algorithm builds multiple classification trees, in which each node can be further divided into two and more nodes. Tree structure allows to reveal interesting interactions among variables.

3.1 Algorithm

The algorithm is performed in three steps. In the first one, each predictor is tested whether all its categories are significantly different in terms of a dependent variable. If not, these categories are merged into so called reduced categories. The algorithm continues iteratively until all pairs of categories are treated as statistically different and the initial contingency table turns into reduced contingency table. In the second step, the algorithm searches for a predictor, which best differentiates values of dependent variable. Criterion for the selection is an adjusted p-value of chi-square test of independence in a reduced contingency table. The adjusted p-value takes into account a number of categories of the newly reduced predictor (Bonferroni adjustment). The predictor with the lowest adjusted p-value is then chosen as a branching variable and each of its categories builds one node of a classification tree. In the first two steps, Pearson's Chi-square test of independence in a contingency table is used. In the third step, the first and the second step are recurring until any of the rules for stopping the algorithm is satisfied, see Kass (1980). The rules for stopping are following: first, there is no other predictor, for which an adjusted p-value would be lower than a given significance level; second, the maximal number of levels of the classification tree was achieved; third, the minimal number of observations cannot be reached in any new node. These criteria must be set before the analysis. There are no strict rules for setting these parameters. Usually, such a combination of parameters securing sufficient interpretation of a classification tree is selected.

3.2 Evaluation of Classification Results

Classification quality indicators are calculated from a confusion matrix. Rows of this matrix represent instances in actual categories, whereas columns correspond to estimated categories by the model. In this article, the average probability of correct classification is denoted as p. Average probability is calculated as the proportion of well-classified objects to all objects.

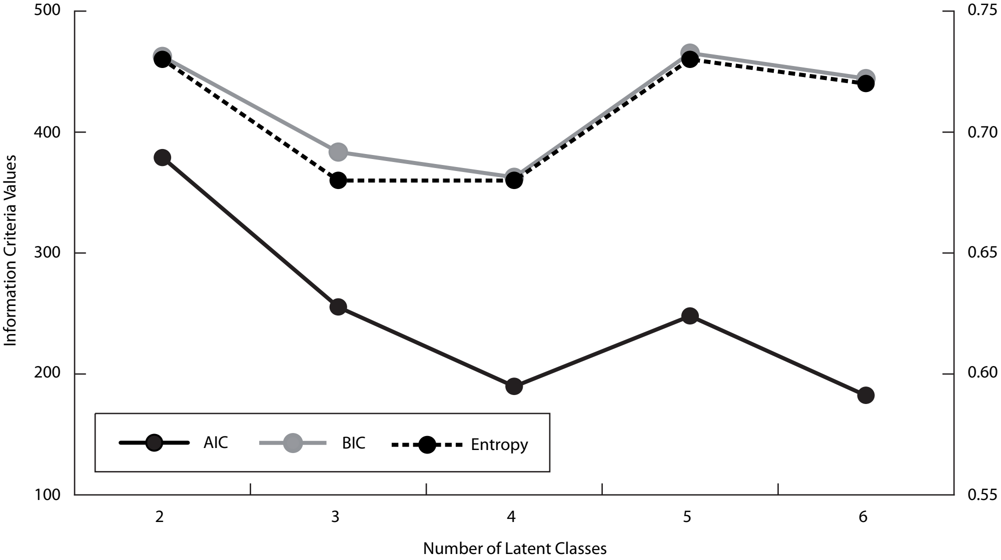
4 RESULTS

4.1 First step – latent class analysis on survey questions

LCA was performed on seven bivariate questions concerning ways of search for a new job. People specified whether they search for a new job being registered at the Labor Bureau, through a personal agency, by directly contacting a potential employer, getting contacts from acquaintances, answering to advertisements or just browsing through them or by other way. According to information criteria, the most suitable model appears to be the model with four latent classes, even though the entropy favors a higher number of latent classes (Figure 1).

Table 1 shows parameter estimates, both prevalence and item-response probabilities. The latent class *Typical* has the highest prevalence, 0.58. That means there is a 58% probability that a given object is going to be classified into this class. Because of this high probability, we assume the unemployed people in this cluster represent the typical behavior when searching for a job. The remaining three classes, which have prevalence under 0.20, are to be profiled deeper according to item-response probabilities. In this case, the item-response probabilities express conditional probability of *Yes* answer to each of the surveyed questions under the condition of being in a particular latent class. As the fourth latent class (*Active*) reaches maximal values across all questions, it clearly refers to those who search for a job by all possible ways. We can describe the behavior of the unemployed in this latent class as *Active*. On the other side, the third latent class comprises people who only registered at the Labor Bureau and some of them ask their acquaintances. The unemployed in this cluster can be called *Passive*. The unemployed people in the last latent class are very similar to those in the *Typical* class; they are very likely to ask their acquaintances or contact the potential employer directly, but unlikely to the Typical class, they browse less through advertisements and nearly do not respond to them. We will refer to this cluster of unemployed as to *Specific*. Distribution of respondents into latent classes is following: *Typical* 69%, *Active* 12%, *Specific* 12%, *Passive* 7%.

Figure 1 Information criteria and entropy of the various LCA models



Source: Authors' computation

Table 1 Share of positive answers to job search questions and item-response probabilities

Way of a job search	Share of unemployed	Latent class			
		Typical	Specific	Passive	Active
Labor Bureau	0.90	0.92	0.82	0.85	0.96
Personal Agency	0.17	0.06	0.19	0.03	0.60
Employer	0.81	0.81	0.98	0.02	0.96
Acquaintances	0.91	0.98	0.80	0.40	0.98
Ads – Active	0.35	0.35	0.02	0.01	0.80
Ads – Passive	0.83	1.00	0.42	0.11	0.99
Other	0.37	0.33	0.26	0.14	0.70
Prevalence		0.58	0.16	0.08	0.18

Source: Authors' computation

4.2 Second step – CHAID on latent classes

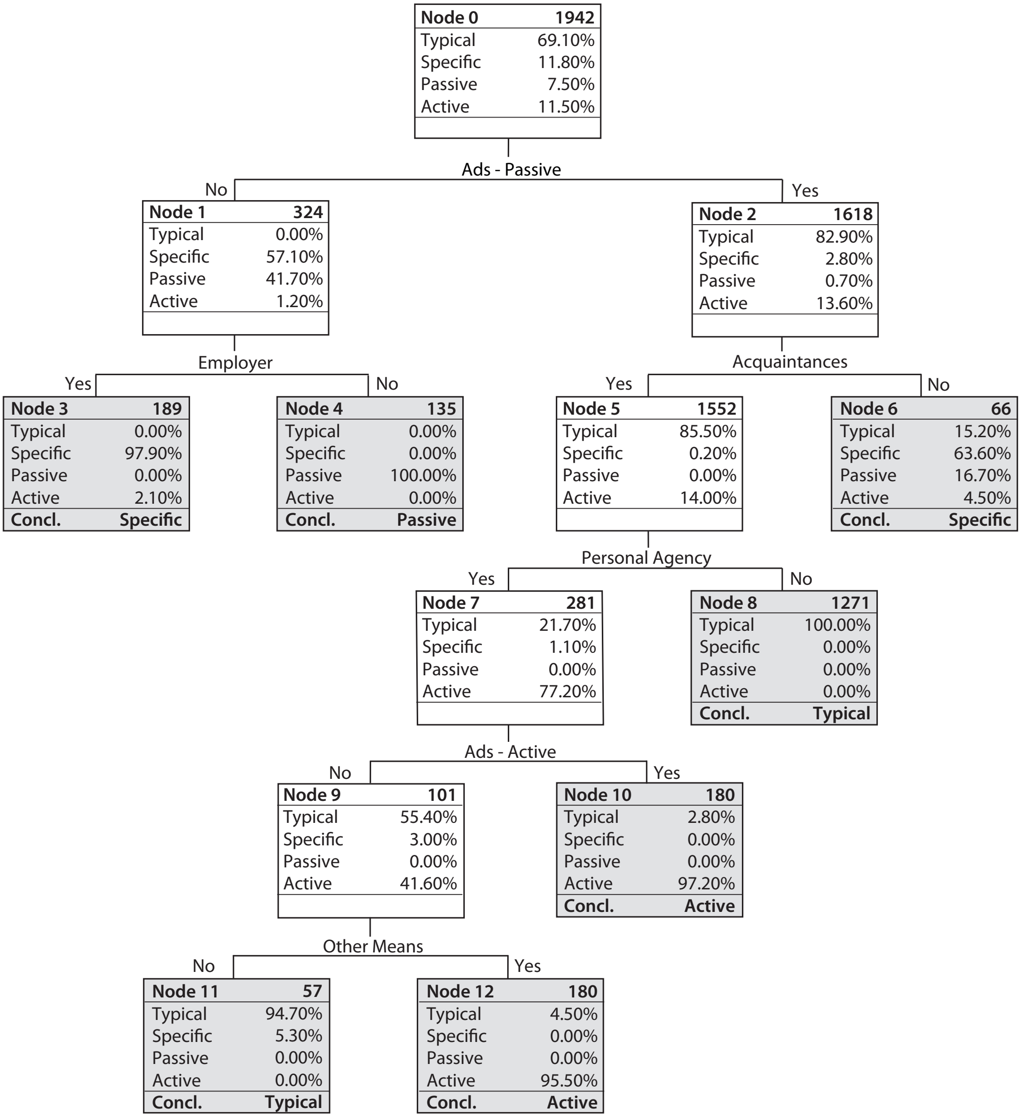
The four latent classes, revealed in LCA, are further examined by CHAID analysis, which is performed in two ways. First, it is performed on the same variables which were used for construction of latent classes, i.e. according to a method of a job search. Second, as its input serve socio-demographic variables, which are described in Table 2.

CHAID analysis based on the first way produces a classification tree which correctly classifies 98% of unemployed persons into one of four defined latent classes.[4] Thus, the four latent classes can be profiled

[4] The model has following setting. Significance level for splitting nodes and merging categories is 0.05, maximal tree depth is 5 levels, a minimal number of cases in parent nodes is 80, and a number of cases in child nodes is 40.

by considering both the information from the classification tree (Figure 2) and item-response probabilities in Table 1. *Typical* individuals do not use services of personal agencies very much. They usually check advertisements and ask people in their surroundings. These statements are true for 95% persons in this group (path to node 10). The *Specific* individuals do not answer to advertisements and their most common way of searching for a job is to contact the potential employer directly (path to node 3). Such behavior is characteristic for 80% of persons in this group.

Figure 2 Classification tree for CHAID model with seven explanatory variables related to the way of a job search for classification into four classes



Source: Authors' computation

An important question is why the *Specific*, unlike *Typical*, do not browse through advertisements and do not answer them. Unfortunately, we are unable to find it out from the data we have available. We can assume two basic hypotheses – they do not want to or they cannot. More probable is the second one. These people could have such education and skills which can be used only in very specific and specialized fields (pilots, craft workers …) where it is not usual to use advertisement. It is also possible that they live in small towns where searching for a job is based more on direct personal contact than on any mediators.

The second approach to CHAID analysis is based on construction of a classification tree with socio-demographic variables. It leads to more accurate profiles of unemployed people in each of latent classes. Due to the fact that the class *Typical* contains 70% of objects, the classification process of four classes is very difficult, because a majority of the unemployed falls very easily into this class. Therefore, only extreme classes *Active* and *Passive*, which are easier to differentiate, will be taken into account for further analysis.

Following variables were chosen as an input for the second approach of CHAID analysis: Actual economic status (labelled *Status* in the analysis), education, usual economic status (*Usual_status*), a number of persons in the responder's household (*Num_of_Person*) and responder's age interval (*Age_interval*). Values of all the variables are in Table 2.

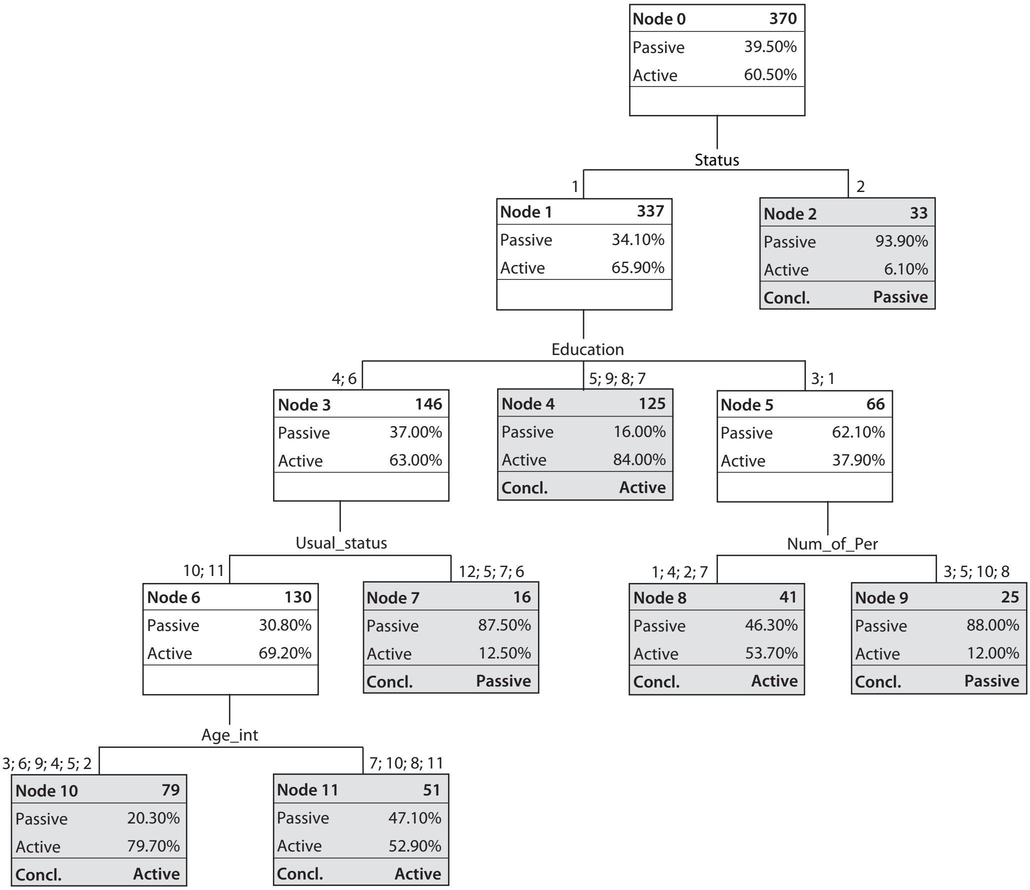
Table 2 Social-demographic variables and their categories

Variables	Categories of variables
Status	**1** – Is seeking any paid job; **2** – S/he has already found job which will start later.
Education	**1** – No education; **2** – Elementary school; **3** – Secondary school; **4** – High school without leaving exam; **5** – High school with leaving exam; **6** – Postsecondary school; **7** – Vocational school of tertiary education and conservatory; **8** – University with bachelor degree; **9** – University with master degree; **10** – University with doctor degree.
Usual_status – main status	**1** – On maternity leave; **2** – In education; **3** – On parental leave; **4** – In early retirement; **5** – In regular retirement; **6** – In retirement due to full disability; **7** – In retirement due to partial disability; **8** – Permanently disabled from healthy reasons; **9** – Works; **10** – Is unemployed; **11** – In the household; **12** – Other.
Num_of_Person	Number of persons (0 – 15).
Age_interval	**1** – "0-14"; **2** – "15-19"; **3** – "20-24"; **4** – "25-29"; **5** – "30-34"; **6** – "35-39"; **7** – "40-44"; **8** – "45-49"; **9** – "50-54"; **10** – "55-59"; **11** – "60-64"; **12** – "65+".

Source: Authors' computation

On the first branching level, persons are divided according to the fact whether they have already found a job, or they still have not found it, see Figure 3. It explains, why 21% of *Passive* have negative attitude towards searching for a job – they have already found it. On the second branching level, the unemployed are divided into three groups according to a level of education. The first group consists of persons with lower or no education, the second group contains persons with secondary education and the third group brings together persons with higher and tertiary education. It is obvious that the higher level of education the more active are the unemployed people in searching for a job. On the third level, the group of low-educated people is further divided into smaller and bigger households. The borders between these groups are not accurate; there are visible rather general tendencies. In smaller households, ratios of *Active* and *Passive* persons are nearly equal, whereas bigger households contain almost 90% of *Passive* persons. This is very typical for gipsy families which have a lot of members and that there is a high unemployment rate in this ethnic group. The group of people with the secondary education is divided according to a usual status of a person. One group is made of people who are usually unemployed or stay in a household. The ratio of *Active* persons is 69% in this group. The second branch mostly consists of people who are in a regular retirement, in a retirement due to partial disability or in retirement due to full disability. This group is dominated by the passive unemployed.

Figure 3 Classification tree for a CHAID model with 5 socio-demographic explanatory variables for classification into two classes



Source: Authors' computation

DISCUSSION AND CONCLUSION

In the paper, attitudes of Czech unemployed persons towards a job search was analyzed. Latent class analysis identified four main attitudes. Besides the *Typical* attitude (prevalence of 58%), which is characteristic by browsing through advertisements, asking acquaintances and directly contacting potential employers, there are groups of *Active* (18%) and *Passive* (8%) unemployed. *Specific* class (16%) consists of people who take nearly typical attitude but they do not use advertisements. After classifying the responders into latent classes, we analyzed them further by CHAID analysis where the constructed latent variables served as an input. First, we built a classification tree with the same variables which were used for the construction of the latent classes. This helped us to characterize the attitudes more precisely. We found out that 95% of people with *Typical* attitude browsed through advertisements and asked people in their surroundings but did not use services of personal agencies, whereas 80% of *Specific* unemployed did not read advertisements but contacted the potential employer directly. Second, we built a classification tree in which the attitude towards a job search was analyzed by the means of socio-demographic variables. Because of high prevalence of the *Typical* group, which would cause low classification performance of a classification tree, we decided to look for socio-demographic differences between the two extreme attitudes, i.e. Active and Passive unemployed. The results showed that *Passive* attitude towards a job search can be found mainly among people with lower education, people living in bigger households and retired people.

Further questions arose during performing the analysis in the paper. First, what is the differentiating factor between the *Specific* and the *Typical class*. It would be very useful to determine why the unemployed people from the Specific group do not use advertisements. Another question is stability of these attitudes in time. Next time, we would like to use longitudinal data to perform longitudinal LCA which would follow development of latent classes in time. Last but not least is the question how to adjust the data where one of the latent classes is prevailing (in our case the *Typical*) so that the classification performance would not be worse when incorporating this latent class into the analysis. All these questions are subjects of our next research.

ACKNOWLEDGMENTS

This work was supported by the University of Economics, Prague under Grant IGA F4/104/2014.

References

AKAIKE, H. *Information theory as an extension of the maximum likelihood principle, Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 1973.

BOZDOGAN, H. Model selection and Akaike´s information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 1987, 52 (3), pp. 345–370.

COLLINS, L. M., LANZA, S. T. *Latent class and latent transition analysis for the aocial, behavioral, and health sciences.* New York: Wiley, 2010.

CZSO. *Statistical yearbook of the Czech Republic 2013* [online]. Prague: Czech Statistical Office, 2013. [cit.20.9.2014]. <http://www.czso.cz/csu/2013edicniplan.nsf/engt/0E002418FB/$File/000113.pdf>.

HABERMAN, S. J. *Analysis of qualitative data.* New York: Academic Press, 1979.

KASS, G. V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 1980, 29, pp. 119–127.

KATRŇÁK, T., MAREŠ, P. Thee employed and the unemployed in the Czech labour market between 1998 and 2004. *Czech Sociological Review,* 2007, 43 (2), pp. 281–304.

MAGIDSON, J. Improved statistical techniques for response modelling. Progression beyond regression. *J. of Direct Marketing,* 1988, 2, pp. 6–18.

MAGIDSON, J. The CHAID approach to segmentation modelling: chi-squared automatic interaction detection. In: BAGOZZI, RICHARD P., eds. *Advanced Methods of Marketing Research.* Oxford: Blackwell, 1994.

MAGIDSON, J., VERMUNT, K. J. *An extension of the CHAID tree-based segmentation algorithm to multiple dependent variables.* In: Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Dortmund, March 9–11, 2004, pp 176–183.

SCHWARTZ, G. Estimating the dimension of a model. *The Annals of Statistics,* 1978, 6 (2), pp. 461–464.

SHAUNNA, C., MUTHEN, B. *Relating latent class analysis results to variables not included in the analysis,* 2009 [online]. [cit.10.5.2014]. <http://statmodel2.com/download/relatinglca.pdf>.

TUFFÉRY, S. *Data mining and statistics for decision making.* United Kingdom: Wiley, 2011.

VERMUNT, K. J., MAGIDSON, J. Latent class analysis. *The Sage Encyclopedia of Social Science Research Methods.* NewBury Park: Sage Publ., Inc., 2004.