CLUSTER ANALYSIS OF EU'S REGIONS ACCORDING TO DEMOGRAPHIC CRITERIA

Ondřej Šimpach – Marie Pechrová

Abstract

Population ageing is a serious problem in all countries of the European Union which impact to social, labour and retirement policy. Therefore, the aim of the paper is to analyse and compare the demographic situation in NUTS 2 regions of 28 EU countries based on various demographic criteria. The data are firstly a subject of principal component analysis to reveal possible correlations among them. Only those indicators which explain substantial part of variation are used. Consequently, a cluster analysis is used to group the states with similar situation in particular year (2004, 2012). A hierarchical clustering using Ward's method with Squared Euclidean distances grouped the regions into 5 groups. This enable to distinguish those with potential for favourable development as their fertility rate and number of live births is high. There were also regions (in cluster 5) identified where the situation in 2004 was better than in 2012 as their potential worsened.

Key words: cluster analysis, demographic indicators, ageing, NUTS 2 regions

JEL Code: C38, J11

Introduction

European Union (EU) is facing serious problem of population ageing which impact on social, labour and retirement policy. It results from both longer life expectancy and declining fertility rates (Kurek and Rachwał, 2011 or Šimpach, 2015). All member states and their regions face these problems with different intensities. "The population of European Union grew by 13,7 million people in 2010 due to net migration (8,6 million people) and natural change (5,1 mil. people), but compared to previous years, both components (net migration and natural change) decreased" (Şerban, 2012). EU and its member states should act proactively and take actions in social and economic-financial sphere to be able to combat the negative consequences of population ageing (Długosz, 2011).

The process of policy-making must be covered by reliable information based on accessible data. Therefore, the EU has Eurostat office. "Eurostat does not collect data. This is

done in Member States by their statistical authorities. They verify and analyse national data and send them to Eurostat. Eurostat's role is to consolidate the data and ensure they are comparable, using harmonized methodology" (Eurostat, 2015). However, the Eurostat data faces number of difficulties, especially that too many indicators and data are available at very different periods (terms) (Mitruț and Simionescu (Bratu), 2014). Nevertheless, they are often used in researches related to regional policy. For example Palevičienė and Dumčiuvienė (2015) analysed the impact of the structural support from the EU on economic growth of regions. They performed multivariate statistical analysis for European Union states' NUTS2 level socio-economic data and created clusters of regions according to their development. Similar research using cluster analysis (Ward´s method with squared Euclidean distances) was done by Pechrová and Šimpach (2013), Löster (2014) or Hrubcová and Löster (2015).

For the purposes of regional planning and decision-making on public or private sector investments (see Nutt, 2006) and for simplification of administrative process is advantageous, when based on certain socio-economic factors we know, how similar are certain territorial units to each other. The presented study will follow authors Lv et al. (2011), who used similar demographic indicators for the creation of clusters of the selected population, but their study focused more on the urban population of adults. Authors Ozus et al. (2012) used the hierarchical cluster analysis for the development of multicenter and travel patterns. They used data on population development, employed and unemployed persons from 1970–2000 and travel statistics. The aim of our paper is to analyse and compare the demographic situation in NUTS 2 regions of 28 EU countries based on various demographic criteria, where the data are firstly a subject of principal component analysis (PCA) to reveal possible correlations among them and only those indicators which explain substantial part of variation will be used.

1 Data and methods

Data about life expectancy, number of deaths and live births, total fertility rate, crude rates of population change, and population on 1st January by five year age group for NUTS 2 regions in the EU were obtained from Eurostat (2016) for years 2004 and 2012. The data were available for 261 of 276 NUTS 2 regions. NUTS 2 are created to contain roughly similar number of inhabitants (from 800 000 to 3 millions) and hence relatively comparable. Variables were subject of PCA to de-correlate input variables and to reduce the volume of input variables with the less possible information loss. Original data were transformed from centred matrix $\mathbf{X}$ (with dimension $n \times d$ where n are columns and d rows) to output matrix $\mathbf{Y}$.

Basically the input was rewritten to different ordinate system: $\mathbf{Y} = \mathbf{XP}$, where $\mathbf{P}$ is $d \times d$ matrix of own vectors of covariate matrix $\mathbf{C}_x$ which fulfil the relation $\mathbf{C}_x = \mathbf{P}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathbf{T}}$, where $\boldsymbol{\Lambda}$ is a matrix with C_x on diagonal, and matrix of the vectors $\mathbf{P}$ is ortonormal, i.e. $\mathbf{P}^{\mathbf{T}}\mathbf{P} = \mathbf{I}_{\mathbf{d}}$ ($\mathbf{I}_{\mathrm{d}}$ is an unit matrix). Vectors (columns of $\mathbf{P}$) formed new ordinate system.

Consequently a cluster analysis using hierarchical agglomerative approach was performed. The clustering procedure is forming hierarchical groups of mutually exclusive subsets, each of which has members that are maximally similar with respect to the chosen demographic indicators. "Given n sets, this procedure permits their reduction to $n-1$ mutually exclusive sets by considering the union of all possible $n(n-1)/2$ pairs and selecting a union having a maximal value for the functional relation, or objective function, that reflects the criterion chosen by the investigator" (Ward, 1963). This process repeats until only one group remains. Particularly Ward's method merges the clusters with minimal within-cluster sum of squared deviations from objects to centroids. Those distances of objects are usually measured by squared Euclidean distance. Euclidean distance (d) between two data points (X_i and Y_i) is calculated as the square root of the sum of the squares of the differences between corresponding values (1). (It is a special case of Minkowski distance with argument p = 2.) The Euclidean Squared distance metric uses the same calculation approach (1) without the square root. As a result, clustering with the Euclidean Squared distance metric is faster.

$$d = \sqrt{\sum_{i=1}^{n} \left(X_i - Y_i\right)^2} \tag{1}$$

Ward's method tends to create relatively small clusters because of the squared differences, but with similar numbers of observations that is desirable for our application. However, it is sensitive to outliers. Its disadvantage is also that the distance between clusters calculated at one step of clustering is dependent on the distance calculated in previous step. Results are displayed in dendrograms that illustrate the information in the amalgamation table in the form of a tree diagram. Calculations were done in Stata 11.2.

2 Results

First, a component analysis was done for selected indicators. The data were normalized. Originally, there were 26 variables. Therefore, the reduction was needed. In both years, 98.83% of variation was explained by the first six components (see Tab. 1). Those contributed to the variability the most and hence we pay attention only to them. Principal components (eigenvectors) show that the first component is not fed by any indicator from more than 0.3 in

both years. In 2004, second component is fed by life expectancy at birth, total fertility rate, and crude rates of population change.

Tab. 1: Results of principal components analysis (components)

	2004				2012			
	Eigen-value	**Differ.**	**Proportion**	**Cumulative**	**Eigen-value**	**Differ.**	**Proportion**	**Cumulative**
Comp1	21.6935	20.1649	0.8344	0.8344	21.6684	20.2598	0.8334	0.8334
Comp2	1.5286	0.5124	0.0588	0.8932	1.4086	0.4141	0.0542	0.8876
Comp3	1.0161	0.2039	0.0391	0.9322	0.9945	0.1411	0.0383	0.9258
Comp4	0.8122	0.3628	0.0312	0.9635	0.8534	0.2469	0.0328	0.9587
Comp5	0.4494	0.2531	0.0173	0.9808	0.6065	0.4431	0.0233	0.9820
Comp6	0.1964	0.0973	0.0076	0.9883	0.1634	0.0435	0.0063	0.9883

Source: own elaboration on data from Eurostat (2016)

The first and last mentioned indicators contribute also to the third component. For fourth component was the most important the number of people at age 85 years or over. To the fifth component with the highest magnitude contributed life expectancy at birth (positively) and total fertility rate (negatively). Last component was also fed by the number of death, live births, population aged less than 5 years and over 85 years. In 2012, the results are similar.

Tab. 2: Results of principal components analysis (Eigenvalues)

2004	**Component**					
Indicators	**1**	**2**	**3**	**4**	**5**	**6**
Life expectancy at birth	0.0511	*0.6235*	*-0.3464*	-0.1655	*0.6731*	0.0814
Deaths	0.2069	-0.0431	-0.1353	-0.0679	-0.1056	*0.3059*
Live births	0.2054	0.0015	0.0940	0.2229	0.1040	*-0.3332*
Total fertility rate	0.0320	*0.6758*	-0.1968	0.2288	*-0.6597*	-0.0888
Crude rates of pop. change	0.0239	*0.3686*	*0.8634*	0.0578	0.0800	0.2596
Pop. < 5 years	0.2071	-0.0008	0.0650	0.2162	0.0936	*-0.3159*
Pop. 70 - 74 years	0.2101	-0.0177	-0.0220	-0.1555	-0.0751	0.1932
Pop. ≥ 85 years	0.1991	0.0504	0.1289	*-0.3274*	-0.0128	*-0.3473*
2012	**Component**					
Indicators	**1**	**2**	**3**	**4**	**5**	**6**
Life expectancy at birth	0.0154	*0.4063*	*0.7987*	0.2516	*0.3441*	0.1014
Deaths	0.2061	-0.1410	0.0033	0.1205	-0.1004	0.2345
Live births	0.2031	0.0984	-0.0119	-0.2668	0.1285	-0.2756
Total fertility rate	-0.0204	*0.5378*	*-0.5981*	*0.3336*	*0.4636*	0.1103
Crude rates of pop. change	0.0153	*0.6923*	-0.0029	-0.2287	*-0.6716*	0.0827
Pop. < 5 years	0.2062	0.0781	-0.0069	-0.2376	0.1303	-0.2183
Pop. 70 - 74 years	0.2012	-0.0692	-0.0247	0.2591	-0.1889	*0.3097*
Pop. ≥ 85 years	0.2027	0.0585	-0.0023	0.2676	-0.1111	*-0.4190*

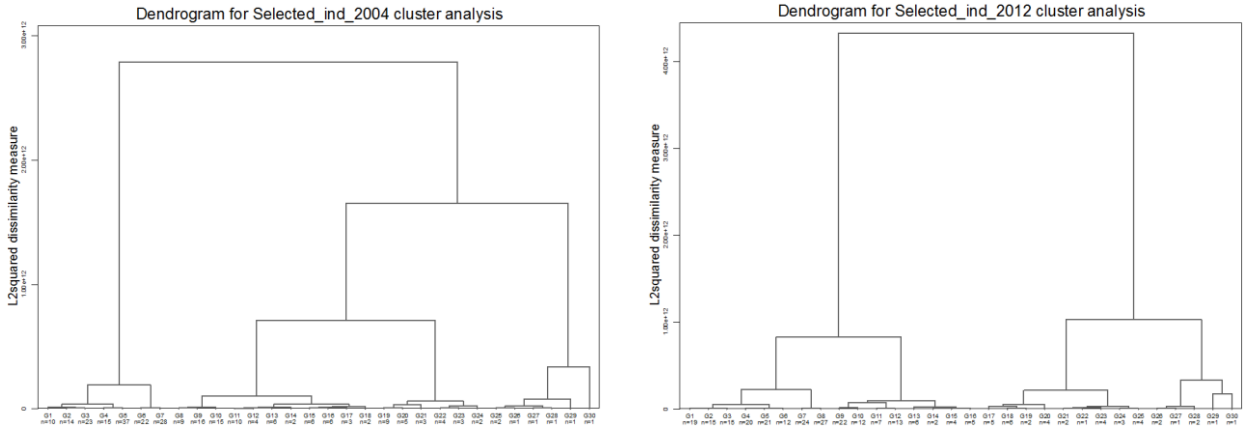
Source: own elaboration on data from Eurostat (2016)

Note: values higher than 0.3000 are marked in italics.

Life expectancy at birth feed second, third and fifth components. Total fertility rate is major in second, third, fourth and fifth component. Surprisingly crude rates of population

change are not important for the third component any more as same live births for the sixth component. Category of population aged less than five years does not contribute to the sixth component, over 85 years to fourth component. On the other hand, category from 70 to 74 years is important for sixth component. The results are displayed in Tab. 2 (only important indicators are included). Despite that some indicators did not feed any components in one of the years, they were included to both cluster analyses in order to keep the structure of indicators for clustering the same in both years. Ward linkage with squared distance enabled to create seven clusters (stopping rule was chosen based on the displayed dendrogram). Surprisingly, the CR regions were included into one cluster. Not all countries had the regions such homogenous. The results of clustering for both years are presented in dendrograms at Fig. 1. It is evident that the input data variability was greater in 2012, which is also confirmed by the calculated dissimilarity measures for individual clusters.

Fig. 1: Dendrograms for cluster in year 2004 (left) and 2012 (right)



Source: own elaboration on data from Eurostat (2016)

2.1 Clusters in year 2004

First, the regions were clustered based on the data from year 2004. Together there are especially southern regions which are characterized by their higher life expectancy at birth and lower total fertility rate. On the contrary, regions in the north have more live births and higher level of total fertility rate. Regions lying to the west are represented by a higher proportion of the elderly population and by regressive demographic tree. The results of clustering are presented in Tab. 3.

Majority of regions was included in **first** cluster (99). There were all regions of the Czech Republic, 14 from UK, 13 from Germany, Estonia (NUTS 2 region is identical to the NUTS 1), and others. The values for all demographic criteria in this group are not extreme

and relatively close to average values. **Second** cluster included for example Cyprus, Lithuania and Malta – all countries have NUTS 1 and NUTS 2 classification identical. Nine regions from Greece, five from Spain, Italy, Netherlands, and Austria were also present in this cluster.

Tab. 3: Results of cluster analysis for year 2004

Indicators	Mean	2004 (clusters 2004)						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Life expectancy at birth	78,40	77,83	**79,43**	78,72	*77,17*	79,38	*80,50*	81,70
Deaths	17 680	12 767	*3 947*	17 922	27 470	**38 907**	55 620	68 457
Live births	19 717	12 105	*4 280*	20 084	27 238	**42 926**	*78 358*	*175 282*
Total fertility rate	1,52	1,49	1,54	**1,56**	*1,46*	1,48	1,49	1,95
Crude rates of pop. change	3,97	2,18	**5,82**	4,70	*1,73*	5,21	14,12	8,10
Pop. < 5 years	94 623	59 895	*21 863*	100 697	135 467	**211 836**	367 333	793 138
Pop. 70 - 74 years	77 753	52 852	*18 188*	77 157	112 333	**180 984**	291 613	346 111
Pop. ≥ 85 years	28 357	18 624	*7 353*	29 588	37 817	**68 339**	105 222	158 832

Source: own elaboration on data from Eurostat (2016)

Note: maximal values of indicators are marked in bold, minimal in italics (excluding 6th and 7th cluster)

All regions are relatively small in terms of the number of inhabitants and therefore the values of all indicators are minimal. This group is characterized by high life expectancy at one side, but the smaller average number of live births on the other. Also the average number of deaths was the lowest. Average number of people in each observed categories was the lowest too. Eleven clusters from UK, seven from Denmark, and six from France were grouped to **third** cluster. Its typical feature is the highest average total fertility and relatively high crude rate of population change. **Fourth** cluster included regions mainly from Germany (8), UK (7), and Romania (5). There are 3.58 times more children under 5 than old people above 85 years which is the highest relation from all categories. The values for total fertility and crude rates of population change as same as the life expectancy were the lowest from all. Seven Italian, six German, and five French regions are included in **fifth** cluster. They are relatively populated; hence the values of deaths, live births, population in observed age categories were the highest. Sixth cluster contains only six regions and seventh only one (Île de France).

2.2 Clusters in year 2012

If the regions stayed in the same clusters in 2012 as in 2004 there would not be many changes happening in their characteristics. Only the minimal crude rate of population change would not be in second cluster, but in the third (Compare Tab.3 with Tab. 4).

We created new clusters based on the data from year 2012 (see lower part of Tab. 4). In this case, the number of regions in **first**, **second** and **third** cluster increased. On the other hand, other clusters were smaller. Cluster seven included besides this time besides Île de

France also Italian region Lombardia. For both is typical high number of live birth. **Forth** cluster is characterized by above average values of all variables. Only the total fertility rate and life expectancy at birth was maximal in second cluster. This contained 10 regions from Greece and 5 from Italy, Netherland, Austria, and from Spain. The development potential of this region is high. The total fertility rate is very important for EU not only from demographic, but also from economic point of view. As Hondroyiannis and Papapetrou (2004) found out "the sample of the eight European countries, an increase in fertility will be associated with higher real per capita output". However, as stated e.g. Bílková (2012), it is important to note that this increase in the total fertility rate should be associated with the change of the quality and structure of education.

Tab. 4: Results of cluster analysis for year 2012 (clusters 2004, clusters 2012)

Indicators	Mean	2012 (clusters 2004)						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Life expectancy at birth	80,26	79,66	**81,14**	80,58	*77,23*	81,07	*82,48*	83,80
Deaths	19 048	13 190	*4 192*	18 790	28 063	**42 126**	61 621	72 529
Live births	20 202	12 353	*4 273*	20 882	27 915	**42 695**	*77 119*	*181 229*
Total fertility rate	1,60	1,57	1,57	**1,59**	*1,37*	1,48	1,52	2,02
Crude rates of pop. change	1,76	0,78	2,36	*3,06*	*0,53*	2,84	2,43	4,50
Pop. < 5 years	100 743	63 718	*22 298*	107 202	141 556	**216 311**	407 381	835 215
Pop. 70 - 74 years	84 709	56 473	*19 262*	82 578	121 239	**200 196**	291 567	332 204
Pop. ≥ 85 years	43 385	27 955	*10 296*	43 586	56 581	**104 923**	168 182	235 546
Indicators	**Mean**	**2012 (clusters 2012)**						
		1	**2**	**3**	**4**	**5**	**6**	**7**
Life expectancy at birth	80,26	*79,70*	**82,72**	80,51	81,54	80,29	82,72	83,30
Deaths	19 048	12 702	*4 217*	22 139	**46 114**	30 921	51 260	83 143
Live births	20 202	12 360	*4 319*	21 772	**42 546**	42 013	79 261	*136 514*
Total fertility rate	1,60	1,60	**1,65**	1,58	*1,44*	1,71	1,64	1,77
Crude rates of pop. change	1,76	*1,09*	2,50	1,70	**4,46**	3,21	3,16	7,05
Pop. < 5 years	100 743	64 046	*22 632*	111 647	**215 814**	210 483	410 598	655 307
Pop. 70 - 74 years	84 709	52 849	*19 083*	101 402	**235 130**	126 867	226 496	429 011
Pop. ≥ 85 years	43 385	17 203	*10 332*	51 130	**117 511**	70 342	145 297	247 789

Source: own elaboration on data from Eurostat (2016)

Note: maximal values of indicators are marked in bold, minimal in italics (excluding 6th and 7th cluster)

The crude rates of population change were the highest in cluster number 4 which might be positive for its development. Mazilescu (2012) argue that "even if the natural balance and the migratory balance are apparently separate sources of the total population change, on long term they become strongly related". The migration to the regions was not evaluated in our article, but we may suppose that migrants are young and therefore negative migration balance could represent negative impact on the future natural balance. The cluster number 4 had also the most population in the youngest category and the highest number of live births. This is very important indicator of positive development. As Leridon (2005) is

warning in his research there is an increasing threat that for medical reasons the increasing number of couples might fail to have all the children they would have liked to have. "Births are postponed more often and the age at first birth rose by 3–4 years in 20 years in most European countries" (Leridon, 2005). The regions if fourth cluster have the second highest life expectancy at birth. Parallel increase in life expectancy at birth together with low fertility levels (it is lower than average in these regions) can have serious consequences. According to research of Cuaresma et al. (2016) it implies that the region will go through an unprecedented process of population ageing, leading to sizeable changes in the age structure of society.

Conclusions

Population ageing is a serious problem in all countries of the European Union which impact to social, labour and retirement policy. Therefore, the aim of the paper is to analyse and compare the demographic situation in NUTS 2 regions of 28 EU countries based on various demographic criteria. Particularly, after the principal component analysis following variables was chosen: life expectancy at birth, number of deaths and live births, total fertility rate, crude rates of population change, number of population in category less than 5 years, 70 to 74 years and 85 years or over. Seven clusters were created. Life expectancy at birth increased in all groups between 2004 and 2012. However, the most favourable was always in region 2004. The most favourable situation in terms of the number of live births and young population was in cluster 4 in both years. Interesting findings are supported by results from other cited studies. Regions are a lot of similar by geography, which means that Western Europe is relatively elderly, with higher life expectancy at birth and the average total fertility rate. East Europe is much younger but with a lower life expectancy at birth. Northern regions are represented by a higher number of live births, a higher total fertility rate and the average life expectancy at birth. Finally, the southern regions have the lowest total fertility rate in all regions, low numbers of live births and the average life expectancy at birth. The challenge for future research is to connect the demographic situation with economic criteria and to explore the relations.

Acknowledgment

The research was supported by the Czech Science Foundation project no. P402/12/G097 DYME – "Dynamic Models in Economics" and also by Thematic task no. 4107/2016 of the Institute of Agricultural Economics and Information.

References

Bílková, D. (2012). Development of wage Distribution of the Czech Republic in Recent Years by Highest Education Attaintment and Forecasts for 2011 and 2012. *6th International Days of Statistics and Economics*. Prague: University of Economics Prague, 162–182.

Cuaresma, J. C., Loichinger, E. & Vincelette, G.A. (2016). Aging and income convergence in Europe: A survey of the literature and insights from a demographic projection exercise. *Economic Systems*, *40*, 4–17.

Długosz, Z. (2011). Population ageing in Europe. *Procedia Social and Behavioral Sciences, 19*, 47–55.

Eurostat. (2015). What we do [on-line] Retreived April 5, 2016, from http://ec.europa.eu/eurostat/about/overview/what-we-do

Eurostat. (2016). Demography and migration [on-line]. Retrieved February 25, 2016, from http://ec.europa.eu/eurostat/data/database

Hondroyiannis, G. & Papapetrou, E. (2004). Fertility and output in Europe: new evidence from panel cointegration analysis. *Journal of Policy Modeling*, *27*, 143–156.

Hrubcová, G. & Löster, T. (2015). Clustering of the Least Developed Countries by the Tourism Economic Impact Analysis. *9th International Days of Statistics and Economics*. Slaný: Melandrium, 587–596.

Kurek, S., & Rachwał, T. (2011). Development of entrepreneurship in ageing populations of The European Union. *Procedia Social and Behavioral Sciences, 19*, 397–405.

Leridon, H. (2005). Reproduction and demography in Europe. *International Congress Series, 1279*, 68–74.

Löster, T. (2014). The Evaluation of CHF coefficient in determining the number of clusters using Euclidean distance measure. *8th International Days of Statistics and Economics*. Slaný: Melandrium, 858–869.

Lv, J., Liu, QM., Ren, YJ., Gong, T., Wang, SF. & Li, LM. (2011). Socio-demographic association of multiple modifiable lifestyle risk factors and their clustering in a representative urban population of adults: a cross-sectional study in Hangzhou, China. *International Journal of Behavioral Nutrition and Physical Activity*, *8*: 40.

Mazilescu, R. (2012). Patterns of demographic development. *Procedia Economics and Finance*, *3*, 1075–1080.

Mitruţ, C. & Simionescu (Bratu) M. (2014). A Procedure for Selecting the Best Proxy Variable Used in Predicting the Consumer Prices Index in Romania. *Procedia Economics and Finance*, *10*, 178–184.

Nutt, P. C. (2006). Comparing Public and Private Sector Decision-Making Practices. *Journal of Public Administration Research and Theory*, 16 (2): 289–318.

Ozus, E., Akın, D. & Çiftçi, M. (2012). Hierarchical Cluster Analysis of Multicenter Development and Travel Patterns in Istanbul. *Journal of Urban Planning and Development*, 138(4): 303–318.

Palevičienė, A. & Dumčiuvienė, D. (2015). Socio-Economic Diversity of European Regions: Finding the Impact for Regional Performance. *Procedia Economics and Finance*, *23*, 1096–1101.

Pechrová, M. & Šimpach, O. (2013) The Development Potential of the Regions of the EU. *Proceedings of International Scientific Conference Region in the Development of Society*. Mendel University in Brno, Brno, 322–335.

Şerban, A. C. (2012). A better employability through labour market flexibility. The case of Romania. *Procedia-Social and Behavioral Science Journal*, 46(805): 4539–4543.

Šimpach, O. (2015). Analysis of Changes in Trends of Fertility Time Series in the Czech Republic: Normality and Principal Component Approach. *9th International Days of Statistics and Economics*. Slaný: Melandrium, 1548–1557.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Contact

Ing. Ondřej Šimpach
University of Economics Prague, Faculty of Informatics and Statistics
W. Churchill sq. 4, 130 67 Prague 3, Czech Republic
ondrej.simpach@vse.cz

Ing. Marie Pechrová, Ph.D.
Institute of Agricultural Economics and Information
Mánesova 1453/75, 120 00 Prague 2, Czech Republic
pechrova.marie@uzei.cz