Agris on-line Papers in Economics and Informatics

Volume V Number 3, 2013

Hierarchical Cluster Analysis – Various Approaches to Data Preparation

Z. Pacáková, J. Poláčková

Faculty of Economics and Management, Czech University of Life Sciences in Prague, Czech Republic

Anotace

V rámci článku jsou řešeny dva různé přístupy přípravy dat, které předchází výskytu multikolinearity. Cílem tohoto článku je pomocí hierarchické shlukové analýzy nalézt podobnosti mezi úrovní e-komunikace ve státech EU. Původní datový soubor čtrnácti ukazatelů byl nejprve redukován na základě korelační analýzy. V případě ukazatelů s vysokou hodnotou korelačního ukazatele, byl do následné analýzy zahrnut pouze ukazatel s vyšší variabilitou. Druhý ze zvolených přístup využívá transformaci vstupních proměnných pomocí analýzy hlavních komponent, jelikož vzniklé hlavní komponenty jsou vzájemně ortogonální. Pro následující analýzu bylo vybráno pět hlavních komponent, které vysvětlují 92 % rozptylu vstupních proměnných. Hierarchická shluková analýza byla aplikována jak na redukovanou množinu proměnných, tak na komponentní skóre pěti hlavních komponent. Na základě Pseudo t^2 statistiky a Pseudo F statistiky byly zvoleny vždy tři výsledné shluky, jejichž složení se liší. Kvalita nalezených řešení byla posuzována také pomocí R-kvadrát indexu, který vykazoval zhruba o deset procent vyšší hodnotu pro řešení založené na komponentním skóre (57.8 % ve srovnání s 47 %). Lze proto konstatovat, že v případě využití komponentních skóre jako vstupních proměnných pro shlukování s dostatečně vysokým podílem vysvětlené variability (zhruba 92 % v provedené analýze), je ztráta informace nižší než u redukce dat na základě korelační analýzy.

Klíčová slova

Hierarchická shluková analýza, PCA, korelace, Pseudo t^2, Pseudo F statistika, e-komunikace, index spokojenosti s Internetem, index spokojenosti s mobilními službami.

Abstract

The article deals with two various approaches to data preparation to avoid multicollinearity. The aim of the article is to find similarities among the e-communication level of EU states using hierarchical cluster analysis. The original set of fourteen indicators was first reduced on the basis of correlation analysis while in case of high correlation indicator of higher variability was included in further analysis. Secondly the data were transformed using principal component analysis while the principal components are poorly correlated. For further analysis five principal components explaining about 92% of variance were selected. Hierarchical cluster analysis was performed both based on the reduced data set and the principal component scores. Both times three clusters were assumed following Pseudo t-Squared and Pseudo F Statistic, but the final clusters were not identical. An important characteristic to compare the two results found was to look at the proportion of variance accounted for by the clusters which was about ten percent higher for the principal component scores (57.8% compared to 47%). Therefore it can be stated, that in case of using principal component scores as an input variables for cluster analysis with explained proportion high enough (about 92% for in our analysis), the loss of information is lower compared to data reduction on the basis of correlation analysis.

Key words

Hierarchical clustering, PCA, correlation, Pseudo t^2, Pseudo F Statistic, e-communication, Internet satisfaction index, Mobile phone satisfaction index.

Introduction

Methods of exploratory analysis are often helpful in understanding the structure and nature of multivariate datasets. Part of the exploratory analysis is searching for the structure of natural grouping. The aim of the cluster analysis is to group the objects into classes in a way that two objects in one group are more similar than any pair of objects where each is from different group. "Groupings can provide an informal means for assessing dimensionality, identifying outliers, and suggesting interesting hypotheses concerning relationships." (Johnson and Wichern, 2007)

Although the cluster analysis can also be understand as a part of exploratory analysis it should not be of first steps. The data preparation should ensure that only the relevant indicators are included in the analysis. The data preparation should handle following problems:

1. missing values;
2. variables selection;
3. multicollinearity;
4. standardisation.

The article deals with various approaches to data preparation for the use of hierarchical clustering. For the purpose of hierarchical cluster analysis the variables should be selected in respect to the problem being solved and also from the statistical point of view. The statistical viewpoint is closely connected with multicollinearity. In case of collinear variables these variables have stronger weight for the cluster analysis. In such a case one should either reduce the number of indicators or use a measure which is not so sensitive to multicollineraity, e.g. mahalanobis distance (Meloun et alter, 2005). Another possibility to avoid from multicollinearity is to use principal component analysis (PCA) while principal components are weakly correlated.

The article introduces various approaches to data matrix preparation for the purpose of cluster analysis. The aim of the work is to compare various approaches used to avoid from multicollinearity and to propose a proper method of data preparation used for hierarchical clustering.

Materials and Methods

The data set consists of fourteen indicators characterizing e-communication in the European Union. The indicators were drawn from two different sources - Eurobarometer 75.1 survey and Eurostat database. The variables taken from Eurostat database are connected to 2011, the *Broadband penetration rate, E-government usage* and *Internet banking usage* were available for 2010 only. The Eurobarometer 75.1 was realized in 2011 (February - March). The survey was particularly focused on E-Communication in households: mobile phone, television and Internet. In all, Eurobarometer 75.1 interviewed 26.836 citizens in 27 countries of the European Union. All respondents were residents in the respective country, nationals and non-nationals but EU-citizens, and aged 15 and over.

The primary data set consists of fourteen variables as mentioned above. The variables are introduced in table 1.

The satisfaction indexes were taken from Eurobarometer survey. The mobile Internet satisfaction index was computed from the following questions: *mobile phone never cuts-off, it is always able to connect, user doesn't limit calls due to charges*, and user *doesn't limit mobile Internet due to charges*. The Internet satisfaction index was based on questions: *connection never breaks down, speed matches contract conditions, and the provider's support is useful*. The indicators are presented on a six point ordinal scale in the Eurobarometer survey. For the purpose of further analysis the responses of individual respondents were aggregated. The proportion of positive responses in each state was used in following computations. Also the proportions of positive responses of aggregated indicators from the Eurostat database were used.

The principal component analysis (PCA) was used for the reduction of dimensionality and multicollinearity in the model. The overall goal of principal component analysis is to reduce the dimensionality of a data set, while simultaneously retaining the information present in the data (see Lavine, 2000). By reducing a data set from a group of related variables into a smaller set of components, the PCA achieves parsimony by explaining the maximum amount of common variance using the smallest number of explanatory concepts (more in Field, 2005).

The original variables x_i, $i = 1, ..., m$, can be reduced to a smaller number of principal components y_j. The principal components are uncorrelated linear combinations of the original variables. All linear

Variable (expressed as percentage of population/households)	Data source
Having computer	Eurobarometer 75.1
Mobile Internet	
Phone calls over Internet	
Mobile phone satisfaction index	
Internet satisfaction index	
Broadband penetration rate	Eurostat database
E-government usage	
Ordering goods over Internet	
Never used the Internet	
Frequently using the Internet	
Using Internet banking	
High computer skills	
High Internet skills	
Households with Internet	

Source: own working

Table 1: Variables description and data sources.

combinations are related to other variables or to the data structure.

The principal components explaining the maximum amount of variance of the original variables (see Hebák et al., 2007, Meloun et al., 2001, or Rencher, 2002). The first principal component corresponds to the direction of maximum variance; the second principal component corresponds to the direction of maximizing the remaining variance, and so on. Each principal component corresponds to a certain amount of variance of the whole dataset.

The cluster technique was used to find the countries with similar e-communication level.

The automatic cluster detection is described as a tool for undirected knowledge discovery. The algorithms themselves are simply finding structure that exists in the data without regard to any particular target variable. The clustering algorithms search for groups of records composed of records similar to each other. The algorithms discover these similarities (see Berry and Linoff, 2004). The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters dissimilar to each other (Rencher, 2002).

We can search for clusters graphically by plotting the observations. If there are only two variables, we can do this in a scatter plot (Rencher, 2002). Even in three dimensions, picking out clusters by eye from a scatter plot cube is not too difficult. If all problems had so few dimensions, there would be no need for automatic cluster detection algorithms. As the number of dimensions (independent variables) increases, it becomes increasing difficult to visualize clusters. Our intuition about how close things are to each other also quickly breaks down with more dimensions (Berry and Linoff, 2004). For example for more dimensions it is possible to plot the data in two dimensions using principal components (Rencher, 2002).

In cluster analysis we generally wish to group the n rows into g clusters. Two common approaches to clustering the observation vectors are hierarchical clustering and partitioning. In hierarchical clustering we typically start with n observations. At each step, an observation or a cluster of observations is absorbed into another cluster (Rencher, 2002). This way is called agglomerative hierarchical approach. It is also possible to reverse this process. It is called divisive clustering and it starts with a single cluster containing all n observations and ends with n cluster of a single item each (Řezanková, 2007). In either type of hierarchical clustering, a decision must be made as to the optimal number of clusters. The results of a hierarchical clustering procedure can be displayed graphically using a tree diagram, also known as dendrogram, which shows all steps of the procedure, including distances at which clusters are merged.

To group the observations into clusters, many techniques begin with similarities between all pairs

of observations. In many cases the similarities are based on some measure of distance. A common distance function is the Euclidean distance between two vectors. Other cluster methods use a preliminary choice for cluster centers of a comparison of within - and between - cluster variability. The scale of measurement of the variables is important consideration when using the Euclidean distance measure. Changing the scale can affect the relative distances among the items. Each variable could be standardized in the usual way by subtracting the mean and dividing by the standard deviation of the variable (see Rencher, 2002, or Řezanková, 2007).

There are authors combining the principal component analysis with clustering to avoid high data-dimension and to reduce multicollinearity (e. g. Garcia-Cuesta et al., 2009; Sembiring et al., 2011 or Xu et al., 2010). There is also wide research on other alternative methods leading to dimension reduction for cluster analysis (e. g. Bharti and Singh, 2013; Shamsinejadbabki and Saraee, 2012).

Various methods for determining the number of clusters were introduced (see e. g. Collica, 2007). Apart from descriptive, graphical or exploratory methods, statistical significance test were introduced as well (for details see e. g. Bock, 1985). Milligan and Cooper (1985) and Cooper and Milligan (1988) compared thirty methods for estimating the number of population clusters using four hierarchical clustering methods. The three criteria that performed the best in these simulation studies with a high degree of error in the data were a **pseudo F statistic** developed by Calinski and Harabasz (1974), a statistic referred to as **Je(2)/Je(1)** by Duda and Hart (1973) that can be transformed into a pseudo t^2 statistic, and the **cubic clustering criterion** (CCC). The pseudo F statistic and the CCC are displayed by PROC FASTCLUS; these two statistics and the pseudo t^2 statistic, which can be applied only to hierarchical methods, are displayed by PROC CLUSTER. It may be advisable to look for consensus among the three statistics, that is, local peaks of the CCC and pseudo F statistic combined with a small value of the pseudo t^2 statistic and a larger pseudo t^2 for the next cluster fusion. It must be emphasized that these criteria are appropriate only for compact or slightly elongated clusters, preferably clusters that are roughly multivariate normal (for more information see e. g. SAS/STAT® 9.2, 2008).

Quality of clusters can also be evaluated using R Squared which informs about the proportion of variance accounted for by the clusters. The idea of computing R Squared is comparing the proportion of intercluster variability to the total variability (for details se e.g. Řezanková, 2007).

For the purpose of this analysis the SAS 9.3 software was used to construct the principal component and cluster analysis. The PRINCOMP Procedure was used to fit a principal component model. The CLUSTER Procedure was used to fit a cluster analysis.

Results and discussion

First application deals with data reduction on the basis of correlation coefficients. Pairs of variables with absolute value of correlation coefficient higher than 0.8 were further investigated. On the basis of coefficient of variation computed as $V = s/\bar{x}$, where s is the standard deviation and $\bar{x}$ is the arithmetic mean, variable of lower variation was excluded. For the purpose of this step, pairs of variables were sorted descending following the correlation coefficient. On the basis of correlation analysis six variables were excluded from further computations: **Households with Internet, Frequently using the Internet, E-government usage, Never used the Internet, Having computer, Broadband penetration rate.**

Therefore the following eight variables were selected for further analysis:

Mobile Internet
Phone calls over Internet
Mobile phone satisfaction index
Internet satisfaction index
Ordering goods over Internet
Using Internet banking
High computer skills
High Internet skills

The dendrogram of cluster analysis made upon the eight variables mentioned above is shown in the figure No. 1.

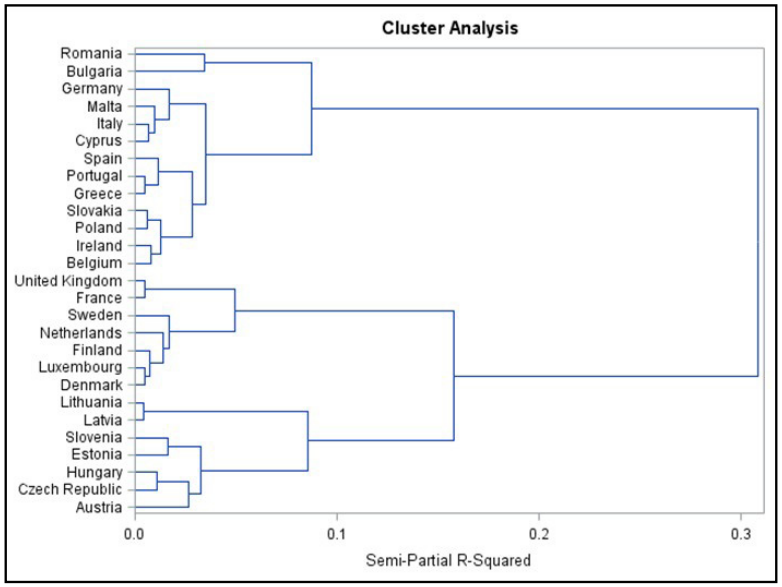
To determine the number of clusters more than one characteristic should be investigated. Figure No. 2 compares values of Pseudo t-Squared and Pseudo F Statistic. Higher values of Pseudo F Statistic provide evidence for the given number with clusters together with lower values of Pseudo t-Squared followed by larger pseudo t^2 for the next cluster fusion.

Following the Pseudo t-Squared the smallest value can be observed for four clusters. The F Statistic provides an evidence for determining three clusters

Correlation Coefficient	Variables	Coefficient of Variation (%)	Excluded Variables
-0.953	Never used the Internet	54.19	Households with Internet
	Households with Internet	18.86	
-0.946	Never used the Internet	54.19	Frequently using the Internet
	Frequently using the Internet	25.66	
0.944	Frequently using the Internet	25.66	NO
	Households with Internet	18.86	
0.928	E-government usage	48.37	NO
	Frequently using the Internet	25.66	
0.914	Frequently using the Internet	25.66	NO
	Using internet banking	58.25	
0.913	E-government usage	48.37	E-government usage
	Using internet banking	58.25	
0.911	Ordering goods over Internet	43.27	NO
	Households with Internet	18.86	
-0.897	Never used the Internet	54.19	Never used the Internet
	Using internet banking	58.25	
0.893	Having computer	17.79	Having computer
	Using internet banking	58.25	
0.88	Having computer	17.79	NO
	Broadband penetration rate	28.81	
0.88	Having computer	17.79	NO
	Households with Internet	18.86	
0.879	Using internet banking	58.25	NO
	Households with Internet	18.86	
-0.876	E-government usage	48.37	NO
	Never used the Internet	54.19	
0.876	Having computer	17.79	NO
	Frequently using the Internet	25.66	
0.874	E-government usage	48.37	NO
	Households with Internet	18.86	
0.871	Broadband penetration rate	28.81	NO
	Frequently using the Internet	25.66	
0.869	Broadband penetration rate	28.81	NO
	Households with Internet	18.86	
0.867	Having computer	17.79	NO
	E-government usage	48.37	
-0.847	Ordering goods over Internet	43.27	NO
	Never used the Internet	54.19	
0.842	Ordering goods over Internet	43.27	NO
	Frequently using the Internet	25.66	
0.84	Broadband penetration rate	28.81	Broadband penetration rate
	Ordering goods over Internet	43.27	
0.82	Broadband penetration rate	28.81	NO
	Using internet banking	58.25	
0.819	Broadband penetration rate	28.81	NO
	E-government usage	48.37	
-0.808	Broadband penetration rate	28.81	NO
	Never used the Internet	54.19	
-0.801	Having computer	17.79	NO
	Never used the Internet	54.19	

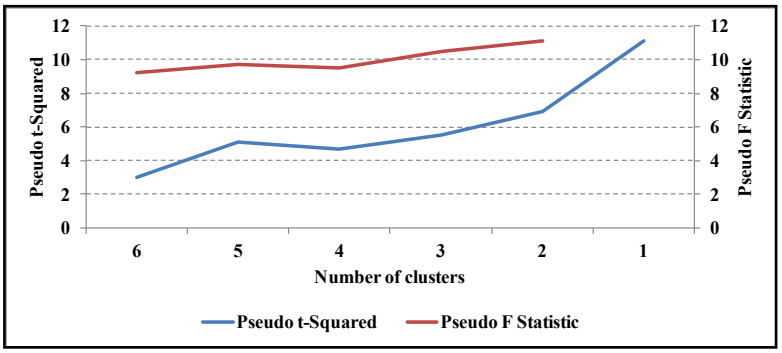
Source: own working

Table 2: Correlation coefficients and reduction of variables



Source: own working

Figure 1: Dendrogram for hierarchical cluster analysis on the basis of eight variables.



Source: own working

Figure 2: Pseudo t-Squared and Pseudo F Statistic for the first cluster analysis.

only, so there is no full agreement following these two statistics. On the other hand, the small value of Pseudo t-Squared should be followed by rapidly increasing value of t-Squared. This can also be observed for three clusters - Pseudo t-Squared is rapidly increasing for two clusters. That is why three clusters were determined as an output from the cluster analysis. Furthermore, in case of dividing the states into four clusters, one would be of two states only.

The three clusters are of seven (two times) and thirteen observations. The second cluster (following the dendrogram) consists of northern states, Netherlands, Luxembourg, France and United Kingdom is obviously of much better e-communication level. Most of the households are equipped with computer (about 82% in average) and covered by Internet – about 64% of households overall. People are frequently using Internet (74% of population in average) and they have very good computer and Internet skills. On the other hand people are the least satisfied with mobile and Internet services.

The second cluster of seven states including Czech Republic, Austria, Estonia, Hungary, Latvia, Lithuania and Slovenia is somewhat in the middle. Although the prevalence of computers and Internet is not much higher in comparison to the third clusters of thirteen states, people are of higher ability to use the Internet. Percentage of those who use e-government services or those who use Internet for ordering goods, Internet banking varies between 30 and 40%. Following the Internet and mobile phone satisfaction indices people from this

group of states are the most satisfied with services provided.

The biggest group of thirteen states covers mostly southern European states together with a group of middle-western European states such as Germany, Poland or Slovakia. These states are characterized by the lowest prevalence of both computers and Internet which is about 60, resp. 64% in average. The percentage of people with high Internet skills ranges between 5 and 13 percent only. That is why the overall Internet usage is at lower level in comparison to other clusters (except ordering goods over the Internet). On the other hand people are more or less satisfied with Internet and mobile phone services, about 70 to 75% of inhabitants are satisfied or very satisfied.

Using the principal components

Second application of cluster analysis was based on the results of principal component analysis (PCA). In PCA, we seek to maximize the variance of a linear combination of the input variables. The eigenvalues indicate that three components could provide a good summary of the data. Five components were selected for the purpose of complex analysis. These components account for almost 93% of variance of the whole dataset.

The first principal component is the linear combination with maximal variance. It explains almost 60% of the total dataset. It largely represents 10 input variables, which are logically related. The corresponding eigenvector expresses an association of input variables with the first principal component. The first principal component has high negative loadings on variables *Never used the Internet* and high positive loadings on 9 input variables related to equipment and Internet use. Therefore it is obvious that the higher component score of this component means a higher level of e-communication in the country.

The second principal component accounts for 17% of variance and it has high positive loadings on four indicators. It is correlated with indicators of the quality of services (mobile phone and Internet satisfaction index), and also with variables *Phone calls over Internet* and *High Internet skills*. It refers to the relationship between the level of the quality of services and the proportion of advanced Internet users.

The eigenvalue of the third component is 1.18 and it accounts for 8% of the total variance. It positively corresponds with *Internet satisfaction index* and negatively with *High Internet skills*.

Forth component accounts for 4% of the total variance. It positively corresponds with *Phone calls over Internet* and negatively with *High computer skills*.

Fifth component accounts for 3.6% of the total variance. It positively corresponds with High Internet skills and negatively with *Phone calls over Internet* and *Mobile Internet*.

Subsequent components contribute less than 3% of the total variance each and these will not enter into following computations.

While the first five components explain more than 90% of overall variance, components scores for the first to the fifth component were used as input variables for the cluster analysis. The use of principal components instead of original data, ensure very low correlation among the inputs.

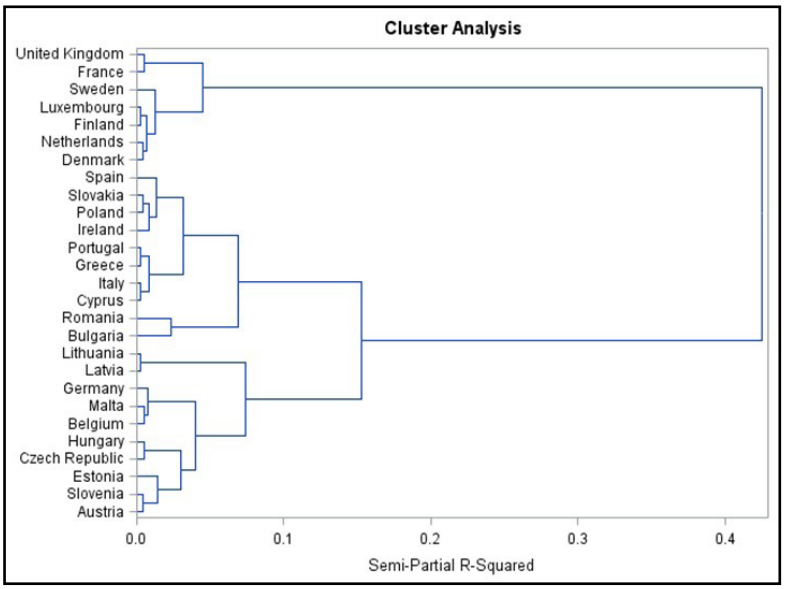
Three dominant clusters can be assumed as it is shown in figure No. 3. Distance for dividing the states into three clusters is denoted by the dashed line. The states were divided into three clusters of ten (two-times) and seven states.

Graph No. 4 shows the relation between the Pseudo t-Squared, Pseudo F Statistic and number of clusters.

Eigenvalues of the Correlation Matrix				
No.	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
1	8.3589	5.9428	0.5971	0.5971
2	2.4160	1.2362	0.1726	0.7696
3	1.1798	0.6326	0.0843	0.8539
4	0.5472	0.0421	0.0391	0.8930
5	0.5051	0.1656	0.0361	0.9291

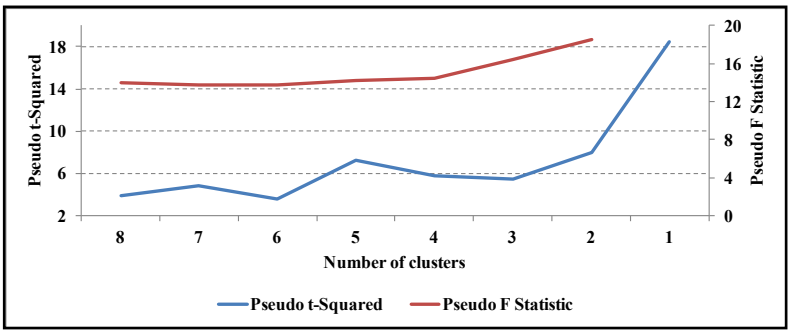
Source: own working

Table 3: First five principal components.



Source: own working

Figure 3: Dendrogram for hierarchical cluster analysis on the basis of principal components .



Source: own working

Figure 4: Pseudo t-Squared and Pseudo F Statistic for the second cluster analysis.

The minimum values of Pseudo t-Squared provide an evidence for determining six or three clusters. Following the F Statistic, there is no clear local peak, but the values of F Statistics are the highest for three and two clusters. Finally three clusters were chosen as well as in case of the previous cluster analysis.

The smallest cluster consists of northern states (Denmark, Sweden, Finland) together with Luxembourg, Netherlands, France and the UK. These states are of higher level of all indicators characterizing both the availability (Households with computer or Internet) and use (phone calls over the Internet. E-government usage, Ordering goods over the Internet, …) of e-communication services. It is obvious that Internet is commonly used in work, in everyday life and also in relation to the government. There is lower percentage of those who never used the Internet (less than 10% in average) in comparison to the other groups with average value above 20, resp. 35%. Also the percentage of those frequently using Internet is above 70% in average (74.43%), while the other groups are of averages about 40%, resp. 55%. Both the computer skills as well as the Internet skills are much better in this states and the Internet is much often used for various purposes including phone calls, ordering goods or Internet banking. The states are also more homogenous in many aspects.

On the other hand, which is maybe surprising, people are less satisfied both with the Internet and mobile phone services. Although the percentage of people who are satisfied with mobile phone

services is pretty high, it ranges between 66% and 78%, the average value is 72,7% which is more than seven percent below the average value of the first cluster. The average Internet satisfaction is about 70%, while in the other clusters it is 76, respectively 82%.

The remaining two clusters consist of ten states each. There is better situation from the view of characteristics being evaluated in the third cluster (Lithuania, Latvia, Germany, Malta, Belgium, Hungary, Czech Republic, Estonia, Slovenia and Austria). These states indicate higher prevalence of computers and Internet in households as well as higher ability to use it. People in these states are the most satisfied with Internet and mobile phone services, the average satisfaction is almost 76, resp. 80%.

The remaining cluster covers southern states such as Spain, Portugal, Greece and Italy, together with Bulgaria, Romania, Slovakia, Poland, Ireland and Cyprus. This group of states shows the lowest values of all indicators characterizing e-communication level. Less than 60% of households are equipped with computer and covered by the Internet in average, almost 40% of people have never used the Internet and only 43.8% use the Internet frequently. Computer and Internet skills are also at very low level – the average percent of citizens with high computer skills is 22% only and the average percent of those with high Internet skills is less than 10%.

Comparing results

Two various approaches to data preparation were

	Data preparation	
	Dimension reduction on the basis of correlation coefficient	**Principal component analysis**
"Cluster 1 TOP"	Denmark Finland France Luxembourg Netherlands Sweden United Kingdom	Denmark Finland France Luxembourg Netherlands Sweden United Kingdom
"Cluster 2 MIDDLE"	Austria Czech Republic Estonia Hungary Latvia Lithuania Slovenia	Austria Belgium Czech Republic Estonia Germany Hungary Latvia Lithuania Malta Slovenia
"Cluster 3 THE LOWEST"	Belgium Bulgaria Cyprus Germany Greece Ireland Italy Malta Poland Portugal Romania Slovakia Spain	Bulgaria Cyprus Greece Ireland Italy Poland Portugal Romania Slovakia Spain

Source: own working

Table 4: Comparison of the resulting clusters.

used. On the basis of results of hierarchical cluster analysis the states were divided into three clusters.

In both analysis the three clusters found grouped together the states with the highest, middle and the lowest level of e-communication. The table 4 compares the clusters found when using factor scores as input variables to the solution based on the reduced data set.

The seven states that are at the top from the view of e-communication level were grouped together when using correlation as well as principal component analysis for data preparation. There are differences between the two clusters of states with middle and low e communication level.

Germany, Malta and Belgium were included in different clusters. Considering the results of the first cluster analysis, the three states are of higher level in nine of fourteen indicators mentioned at the beginning. So in the group of the lowest thirteen states they are at the top.

Another possibility how to consider the two results is to look at the variability explained by the clusters found. When considering the first result on the basis of eight poorly correlated variables, the proportion of variance accounted for by the clusters is just under 47%.

When the states are grouped into three clusters on the basis of component scores for the first five components, the proportion of variance accounted for by the clusters is almost sixty percent (57.8%).

Therefore it can be stated, that there is higher variability among the clusters found on the basis of principal components and the input variables (component scores) are very poorly correlated as well.

Conclusion

The article introduces two possible approaches to data preparation to avoid high correlation among variables. The aim of the article was to identify states of similar e-communication level. This was realized by cluster analysis which is sensitive to collinearity. Firstly, the original data set was reduced on the basis of correlation coefficient while in case of strong correlation the variable of lower variability was eliminated. The second application of cluster analysis was based on principal components. By the use of five principal components, about 92% of variability can be explained.

In case of both applications, three clusters were assumed on the basis of two criterions: Pseudo t-Squared and Pseudo F Statistic. The group of states of the highest e-communication level has been found the same but there are differences for the rest of states. An important criterion to assess the results is to look at the proportion of variance accounted for by the clusters which is much higher for the results based on principal components.

Therefore it can be stated, that in case of using principal component scores as an input variables for cluster analysis with higher proportion of variance explained, there was lower lack of information compared to data reduction on the basis of correlation analysis.

The results of cluster analysis have confirmed the conclusions published by the authors previously, which is the top position of Nordic European states and Luxembourg together with France or United Kingdom and lower prevalence and use of e-communication tools in southern European states e.g.

Acknowledgement

The authors gratefully acknowledge the support from the Faculty of Economic and Management, Czech University of Life Sciences, via IGA grant, no. 20121049, “Analysis of the Internet and computer literacy using data mining techniques“.

Corresponding author:
Ing. Zuzana Pacáková, Ph.D.
Department of Statistics, Faculty of Economics and Management,
Czech University of Life Sciences in Prague, Kamýcká 129, 165 21 Prague 6, Czech Republic
E-mail: pacakova@pef.czu.cz

References

[1] Berry, M. J. A., Linoff, G. Data Mining Techniques For Marketing, Sales, and Customer Relationship Management. Second Edition. Indianapolis: Wiley Publishing, 2004, 643 p. ISBN: 0-471-47064-3.

[2] Bharti, K. K., Singh, P. K. A two-stage unsupervised dimension reduction method for text clustering, Advances in Intelligent Systems and Computing, 202 AISC/2, 2013, pp. 529-542, ISSN 1615-3871.

[3] Bock, H. H. On Some Significance Tests in Cluster Analysis, Journal of Classification, 1985, 2, pp. 77–108, ISSN 0176-4268.

[4] Calinski, T., Harabasz, J. A Dendrite Method for Cluster Analysis, Communications in Statistics, 1974, 3, pp. 1–27.

[5] Collica, R. S. CRM Segmentation and Clustering Using SAS Enterprise Miner. Cary, NC: SAS Institute. 2007, ISBN 978-1-59047-508-9.

[6] Cooper, M. C., Milligan, G. W. The Effect of Error on Determining the Number of Clusters in Data, Expert Knowledge and Decisions, pp. 319–328, 1988, ed. W. Gaul and M. Schrader, London: Springer-Verlag.

[7] Duda, R. O., Hart, P. E. Pattern Classification and Scene Analysis, New York: John Wiley & Sons. 1973.

[8] Field, A. Discovering Statistics Using SPSS. Second Edition. London: SAGE Publication, 2005, 779 p. ISBN: 978-0761944522.

[9] Garcia-Cuesta, E., Galvan, I. M., De Castro, A. J. Supervised clustering via principal component analysis in a retrieval application, Proceedings of the 3rd International Workshop on Knowledge Discovery from Sensor Data, SensorKDD'09 in Conjunction with the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD-09, 28 June, 2009, Paris, France, pp. 97-104.

[10] Meloun, M., Militký, J. Kompendium statistického zpracování dat. Praha: Academia, 2006, 984 p., ISBN: 80-200-1396-2.

[11] Milligan, G. W., Cooper, M. C. An Examination of Procedures for Determining the Number of Clusters in a Data Set, Psychometrika, 1985, 50, pp. 159–179, ISSN: 0033-3123 (Print) 1860-0980 (Online).

[12] Lavine, K. B. Clustering and Classification of Analytical Data. Encyclopedia of Analytical Chemistry, John Wiley & Sons Ltd, Chichester, 2000, p. 9689-9710. ISBN: 9780470027318.

[13] Rencher, A. Methods of Multivariate Analysis. Second Edition. New York: Wiley Publishing, 2002, 738 p., ISBN 978-0471418894.

[14] Řezanková, H., Húsek, D., Snášel, V. Shluková analýza dat. Příbram: Professional Publishing, 2007, 220 p., ISBN 978-80-86946-26-9.

[15] SAS/STAT® 9.2 User's Guide, The CLUSTER Procedure, 2008. Cary, NC: SAS Institute.

[16] Sembiring, R. W., Mohamad Zain, J., Embong, A. Alternative model for extracting multidimensional data based-on comparative dimension reduction, 2nd International Conference on Software Engineering and Computer Systems, ICSECS 2011, 2011, 27 – 29 June, Kuantan, Malaysia, pp. 28-42.

[17] Shamsinejadbabki, P., Saraee, M., A new unsupervised feature selection method for text clustering based on genetic algorithms, Journal of Intelligent Information Systems, 2012, 38/3, pp. 669-684, ISSN: 0925-9902 (Print) 1573-7675 (Online).

[18] Xu, Z. J., Zheng, J. J., Zhang, J., Ma, Q. Application of cluster analysis and factor analysis to evaluation of loess collapsibility, Yantu Lixue/Rock and Soil Mechanics, 2010, 31, pp. 407-411. ISSN: 10007598.