DETERMINING THE OPTIMAL NUMBER OF CLUSTERS IN CLUSTER ANALYSIS

Tomáš Löster

Abstract

Cluster analysis is the multivariate method which objective is to classify the objects. In current literature there are many methods and many distances measures, which can be mutually combined. There is no manual and rule which would clearly identify the appropriate combination method and distance measures during clustering. Simultaneously, in cluster analysis it is often necessary to determine the optimal number of clusters in to which the objects are to be classified. The aim of this paper is to illustrate the possibilities of the process of determining the number of clusters and to evaluate selected coefficients for determining the number of clusters in combination with clustering different methods and with different distance measures. For example CHF coefficient is more suitable to be used with combination with Mahalanobis distance, where the success is higher in comparison with Euclidean distance. For example using average linkage method the success is higher by 21.88%. On the other hand, coefficient D-B is more successful while using Euclidean distance measures. In the case of Ward's method the success is higher by 15.63%.

Key words: clustering, evaluating of clustering, methods, number of clusters

JEL Code: C 38, C 40

Introduction

Cluster analysis is multivariate method which objective is to classify the objects into groups called clusters. It is very often used statistical method, see e.g. (Halkidi et al., 2001; Löster at al., 2010; Řezanková et al., 2013; Žambochová, 2012). The need for creation of the groups of objects is an integral part of many disciplines. In practical tasks which are dealing with the classification of objects is crucial for selecting the right multivariate classification methods if they are priory known or unknown the affiliations of the objects to clusters. Objects may be customers, patients, clients, documents, etc. Very often is used to classification of regions. Authors of papers very often used wages to describe regions. The problem of wages and poverty is described e.g. in (Bílková, 2011, 2012; Marek, 2013;

Pavelka, 2012; Miskolczi, 2011, Želinský, 2012). Other demographic variables, which are very often used in cluster analysis, are described in (Megyesiova, et al. 2011, 2012).

In the case when the investigated objects have known inclusion in the group, for classification is used the discriminant analysis, which aims to create a rule by which the new objects of unknown affiliation are classified. This is useful for example in medicine, where based on the properties of the patients the other patients are to be classified into groups known in advance. Second situation, i.e. when the classification of the objects is not known in advance is solved by cluster analysis. Currently there are many methods and approaches in scholar literature, which enable the analyst to classify number of objects set beforehand to clusters. Selection of possible combinations of methods is dependent on many factors. One of them is the type of variables, by which the objects are characterized. In the case when the objects are characterized exclusively by quantitative variables, the analyst has the possibility to choose from many possible combinations of method. In the case when the objects are characterized exclusively by qualitative variables or by mixed variables (combination of quantitative and qualitative) the possibility of choice of clustering methods is limited.

Key role in cluster analysis play the similarity characteristics, resp. distances measures. Also in this case, the variable type, which characterizes each object, is critical. In case of quantitative variables the distance measures are used. There are many distance measures between objects. Linkage clustering methods and distance measures a whole series of combinations emerge, the choice is up to the analyst. Various combinations bring different results. In the current literature there are numbers of comparative studies that seek to evaluate various combinations of clustering methods and measure distances in a variety of conditions. However, there is not a clear rule that would strictly determine what combinations use in what situations. Although they are indicated for instance situations in which different distance measures are unsuitable (for example in case of a strong correlation between the input variables), but the actual effect of breaking of this assumption is usually not analyzed. In the same way the advantages and disadvantages of different clustering algorithms are indicated.

As stated above the part of cluster analysis is also very often the determination of the number of groups to which the objects should be classified. This number is not usually known in advance in cluster analyses. The aim of the paper is to show the possibilities of setting the number of clusters in cluster analysis and to compare the success of selected coefficients on real data files in the case when the objects are characterized by an only quantitative variables.

1 Clustering methods

The aim of cluster analysis is the classification of objects, see (Gan et al., 2007). There are various methods and procedures to do that. These methods and procedures can be categorized according to various criteria see e.g. (Gan et al., 2007; Řezanková et al., 2009). Mostly they are divided on traditional methods and new approaches in the literature. Traditional methods are well developed and they are applied in many software products.

In current scholar literature there are numbers of clustering algorithms, which are implemented to many specialized software products. Among mostly used classification of "traditional" clustering method stated in majority of sources is the division on *hierarchical* and *nom-hierarchical* clustering methods.

Hierarchical clustering represents that way of clustering, the process aimed at creating a treelike structure of clusters. Their output is besides others a so-called *dendrogram* which represents graphical presentation of the clustering process according to the selected metrics. An important feature of hierarchical clustering method is that the results of the previous step are always assigned to the results obtained in the next step and thus the tree structure is created. The advantage of hierarchical method is that it is not necessary to know in advance the number of clusters, which is considered a major advantage over their non-hierarchical clustering method. They are relatively fast, but are not suitable for large data files.

Non-hierarchical clustering does not focus on the creation of dendrograms, but concentrate on classifying of the objects into the previously known number of clusters. First it is needed to establish the initial decomposition of objects into clusters and then using iterative procedures and methods to improve the initial decomposition. In this method the gradual improving of the decomposition of objects may cause a shift of an object from one cluster to another. The quality of this method depends on the ability of the user to select the initial decomposition.

Application of various methods of clustering on same objects described by identical properties can produce different results. As stated by Gan et al. (2007) and Halkida (2002) "It cannot be a priori said which method is the best for a given problem. Usually, the method of the nearest neighbour is the least suitable and method of average distance or Ward's method suits in many cases the best". But it is important also those practical experience researchers with the type of job are used. Among the methods hierarchical clustering can be included, for example, the nearest neighbour method, method of the farthest neighbour, method of the average distance, centroid method.

Method of the nearest neighbour was firstly described in year 1957 by P. H. A. Sneath. It is the oldest and the simplest method. There are searched two objects, between which the distance is the shortest and they are joined to the cluster. Another cluster is created by linking the third closest object. Distance between two clusters is defined as the shortest distance of any point in cluster in relation to any point in another cluster, see Gan et al. (2007). As one of crucial disadvantage of this methods is stated that occurs so-called *chaining*, when two objects, which are the closest in relation to each other, but not in relation to majority of other objects, are sorted to one cluster.

Method of the farthest neighbour is based on the opposite principle than the method of the nearest neighbour. Its author is Sörensen. Its essence is in linking those clusters, which distance between the furthest objects in minimal. The advantage of this method is that it creates small, compact and clearly separated clusters. Contrary to the nearest neighbour method there is no problem with clusters' chaining.

Using **method of average distance** the criterion for emerge of the clusters represents the average distance of all objects in one cluster to all objects in second cluster. Results of this method are not influenced by extreme values as in the case of method of the nearest and furthers neighbour. Emerge of the cluster is dependent on all objects. Two clusters are merged to new cluster, if there is minimal distance between them.

Centroid method was firstly used by Sokal and Michener under name "weighted group method". For expression of the dissimilarity of clusters is used Euclidean distance of their centres of gravity (centroids). This method does not use between-cluster distances of the objects. To new cluster those two clusters are merged, between what is minimal distance of their centroids, while the centroid is understood as an average of the variables in particular clusters. The advantage of this method is that it is not that significantly influenced by remote objects. In this method so-called *void clusters* can appear that means that the distance between centroid of one pair is smaller than the distance between centroid of different pair created in previous step.

Median method was firstly introduced by Gower under name "unweighted group method". The aim of the method is the effort to eliminate the disadvantages of centroid methods, see above. Gower proclaimed that "... different number of objects of clusters cause different weight of first two parts of the recursive prescription of centroid method and thus it happens that the characteristics of small clusters disappears in final linkage”. Median method is an analogy of centroid method and the difference is that instead of the distance between centroid clusters is used the distance between medians of those clusters. To one cluster are merged two clusters between which medians is the closest distance. The advantage of this method is in removing of different weights which are in centroid method assigned to differently sized clusters.

Ward’s method solves the clustering procedure differently than above stated methods that are optimizing the distances between particular clusters. Method minimizes the heterogeneity of clusters, i.e. clusters are formed using maximization of intragroup homogeneity. As the measure of homogeneity of clusters is understood intragroup sum of squares of the deviations of values from the average of the clusters and it is called *Ward’s criterion*. Criterion for linking the clusters is based on the idea that in each step of clustering there is minimal increment of Ward’s criterion. Ward’s method has tendency to remove small clusters and create clusters of approximately same size.

Among non-hierarchical methods of clustering it is possible to include for example the method of *k*-means. **Method of *k*-means** is suitable in case when variables that characterize the objects are only quantitative and is based on moving particular objects between clusters. It is a method which belongs to the group of so-called optimization methods.

Besides above stated ways of clustering there is also so-called fuzzy clustering. This clustering is based on assumption that there are n objects and k clusters. For each $i^{\text{-th}}$ object and $h^{\text{-th}}$ cluster is set the *measure of affiliation* u_{ih} which represents the probability that set object i is classified to $h^{\text{-th}}$ cluster. Fuzzy clustering is process that in contrary to above stated procedures enables inclusion of one object to more clusters what is considered to be the advantage of this method. The outcome of this method is matrix of affiliation of particular objects to clusters.

For clustering objects also **two step cluster analyses** can be used. The method can be utilized for clustering objects that are characterized by exclusively nominal variables, or by variables of various types. As a measure of the distance, or dissimilarity can be used either Euclidean measure (only in the case of quantitative variables) or likelihood measure (for variables of various types). This method consists of two phases. In the first phase the objects are clustered to sub-clusters (small clusters) that number is much lower than the number of original objects. During this step is used so-called incremental clustering when the objects are either put to some of the created clusters or new cluster is created. In both phases of clustering the same dissimilarity measure is used, i.e. likelihood measure as it is in software SPSS.

Besides the clustering methods themselves and important (key) role is played also by the measures of dissimilarity. Similarity is used as the criterion for the creation of clusters. For measuring of similarity it is necessary to distinguish by what types of variables are characterized the features of particular objects. Majority of methods and procedures is usable for the situation when all objects are characterized by quantitative variables. Measurement of the similarity of objects when they are characterized by quantitative variables is based on the distances of the objects. Transformation of the distance measures to similarity (dissimilarity) measures is done according to simple rules. Very important are the measures of similarities, resp. the distance levels. There are a number of distance levels and in the practice they are combined with various clustering methods, see e.g. (Gan et al., 2007; Řezanková et al. 2009). For measurement of the distance are frequently used:

Euclidean distance (also geometric metric) represent the length of hypotenuse of a rectangular triangle. Calculation of this measure is based on Pythagoras theorem. Using so-called Ward's clustering method the squared Euclidean distances are usually applied.

Hamming distance is not suitable when the variables characterizing the objects are mutually correlated. If the variables were correlated, the resulting clusters would be wrong.

Also **Minkowski distance** can be used. Similarly to Hamming distance it is not suitable when the variables characterizing the objects are mutually correlated.

Eventually it is possible to use **chordal distance, Mahalanobis distance**. It diminishes the problem that occurs while using non-standardized data that can cause differences among clusters due to different measurement units. This measure is usable in the case when all the variables characterizing the objects are mutually correlated.

Detailed descriptions and formulas of particular distance measures can be found e.g. in Řezanková (2009).

2 Coefficients for clustering evaluation and setting the number of clusters

In this section will be presented the comprehensive summary of the coefficients that serves to the evaluation of the clustering in the case when we assume clustering of the objects with fixed affiliation to the clusters. There were elaborated many coefficients for the evaluation of the separation of the set of *n* objects to *k* disjunctive clusters regardless the ways how the separation of the objects was done. Hence it does not matter whether the clusters are the results of the decomposition method or whether they are the results of hierarchical clustering. Their calculation will not be described, only the review in which software product it is possible to found those coefficients and by which way they are evaluated (see Tab. 1). Detailed description of the criteria can be found e.g. in Löster (2011) or Řezanková (2009).

Tab. 1: Selected criteria for evaluation of the results of disjunctive clustering

Coefficient	Searched extreme	Software
Silhouette coefficient	maximum	S-PLUS, SPSS
CHF index (pseudo F)	maximum	SAS, SYSTAT
PTS index (T-square)	minimum	SAS, SYSTAT
RS (R-square, RSQ)	maximum	SAS
SPRS (SPRSQ)	minimum	SAS
BIC, AIC	minimum	SPSS
RMSSTD	minimum	SYSTAT
Davies-Bouldin (DB)	minimum	SYSTAT
Dunn’s separation index	maximum	SYSTAT

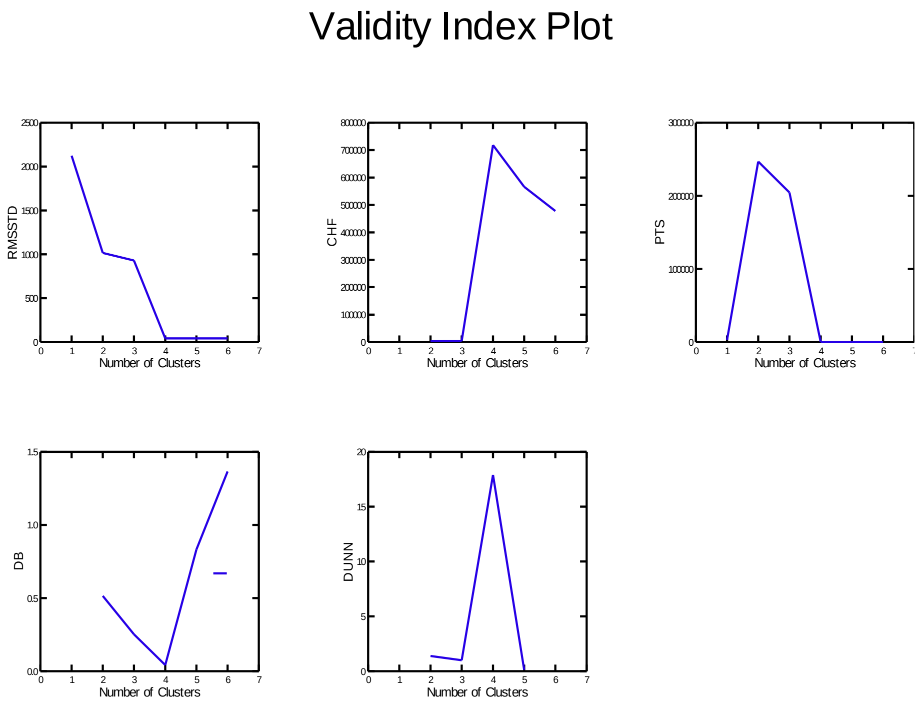
Source: Own elaboration

In practical use it is suitable to set the values of more coefficients at once because there is no criterion, which would surely and uniquely evaluated the clustering results (method or the number of clusters). As it was stated above, in literature there is not uniquely determined during what requirements is any of the coefficients the most suitable for the evaluation of particular clustering. In the case that the results of clustering match based on the values more coefficients at once, it is possible to consider those conclusions as “true”. As stated by the authors of the coefficients themselves, see Löster (2011), in some cases it is even necessary to evaluate the results of clustering by more coefficients at once.

2.1 Example of the number of clusters evaluation

In the following section are presented the examples how to evaluate the number of clusters base on various coefficients in different conditions. From the graphs at Fig. 1 are obvious the values of selected coefficients (RMSSTD, CHF, PTS, DB, Dunn’s coefficient). Those coefficients were obtained from SYSTAT. On the first file where the clusters are well separated (the distances between centroids are sufficiently big) was used the method of average linkage and Euclidean distance. Based on the values of the majority of coefficients (from the curves of stated coefficients) it is obvious that the majority of them agreed on the value 4, and therefore the optimal value was set at 4 clusters.

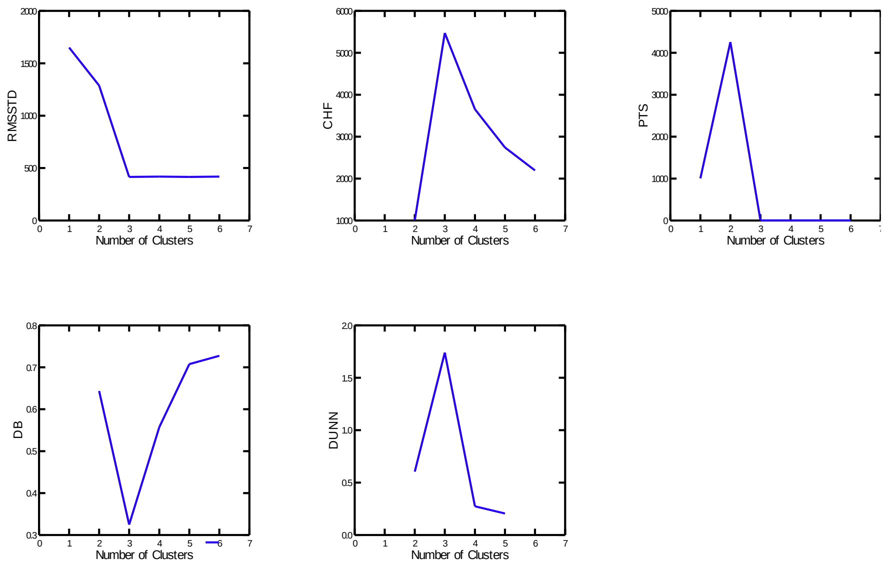
Fig. 1: Evaluation coefficients for the setting of the number of clusters – 4 clusters



Source: our calculation

From the graphs at Fig. 2 can be seen the differences of selected coefficients (RMSSTD, CHF, PTS, DB, Dunn's coefficient). In the case of the second file the clusters were well separated. A centroid method together with Mahalanobis distance for clustering was used. Based on the values of stated coefficients (from their curves) it is obvious that the majority of them agree at value 3, and therefore the optimal value was set at 3 clusters

Fig. 2: Evaluation coefficients for the setting of the number of clusters – 3 clusters



Source: our calculations

4 Real data sets

For evaluation of the total success of selected coefficients was selected 32 real data sets from known database *The UCI Machine Learning Repository* (see https://archive.ics.uci.edu/ml/datasets.html). This database contains various data files that have before known numbers of the clusters, hence using these data the evaluation of the coefficients is possible. The data sets are following: *Wine, Iris, Abalone, Cardiotocography, German Credit Data, Banknote Authentication, Blood Transfusion Service Center, Climate Model Simulation Crashes, Connectionist Bench (Sonar, Mines vs. Rocks), Ecoli, Echocardiogram, Energy Efficiency, Fertility, Haberman's Survival, Indian Liver Patient, Connectionist Bench (Vowel Recognition - Deterding Data), Ionosphere, Musk (Version 1), Parkinson Speach, Pima Indians Diabetes, QSAR Biodegradation, QSAR Biodegradation NV 1, QSAR Biodegradation NV 2, Seeds, Statlog (Vehicle Silhouettes) a+b, Statlog (Vehicle Silhouettes) a+g, Vertebral Column, Wall-Following Robot Navigation Data, Wholesale Customers, Susy NV 1, Susy NV 2 and Susy NV 3*.

5 Results

Based on a combination of various distances measures and different clustering method were obtained various results of the optimal number of clusters which were provided by particular coefficient. Tab. 1 shows the number of cases in which individual coefficient correctly determined the number of clusters using various clustering method in combination with a Euclidean distance measure. It reveals, for example, that the best results were achieved by using nearest neighbour methods when using Dunn's coefficient. Success in determining the optimal number of clusters was 59.38%. As unusable appears RMSSTD coefficient which success in combination with any method did not exceed 20%.

Tab. 1: Number of correctly set clusters (in %) – Euclidean distance measure

Methods/coefficients	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	9,38	53,13	50,00	59,38	59,38
Farthest neighbour	18,75	31,25	31,25	50,00	31,25
Centroid method	18,75	43,75	25,00	56,25	50,00
Average distance	18,75	31,25	28,13	53,13	56,25
Ward's method	18,75	34,38	53,13	25,00	31,25

Source: our calculations

In summary, the most successful coefficients using Euclidean distance measures might be considered Davies-Bouldin's and Dunn's indexes.

In Tab. 2 there are stated the number of cases in which the particular coefficients correctly determined the number of clusters while using various method of clustering with combination of Mahalanobis distance measure. It reveals, for example, that the best results were achieved by using centroid method with Davies-Bouldin's coefficient. Success in determination the optimal number of clusters was 68.75%. When using this distance measure also RMSSTD coefficient appears as usable.

Tab. 2: Number of correctly set clusters (in %) – Mahalanobis distance measure

Methods/coefficients	RMSSTD	CHF	PTS	D-B	Dunn
Nearest neighbour	6,25	50,00	46,88	56,25	46,88
Farthest neighbour	21,88	37,50	40,63	37,50	37,50
Centroid method	9,38	59,38	50,00	68,75	37,50
Average distance	9,38	53,13	46,88	65,63	53,13
Ward's method	28,13	50,00	37,50	9,38	59,38

Source: our calculations

In summary, the most successful coefficient while using Mahalanobis distance measure might be consider Davies-Bouldin's index.

Conclusion

Cluster analysis is a multivariate statistical method, which is used to classify objects into clusters. There are many clustering methods and there are many measures of the distances between objects. The combination of various method and different distance measures give different results. The current literature does not address the different combinations and there is no indication which combination is successful.

Part of the cluster analysis is usually the setting of the optimal number clusters to which the particular objects should be classified. Even in this case there are many coefficients that can be used for this task. Choice of the coefficient is also affected by the clustering method as well as by the chosen measures of the distances between objects. The aim of this paper was to give the orientation in the difficult issue of determining the number of clusters. There were named selected coefficient, which are applied to various software products and on two examples there was the process of clusters selection outlined. Based on the analysis of 32 real data files there were found suitable combinations, which provided the best results. A total of 5 clustering method and 5 coefficients to determine the optimal number of clusters were compared. Based on various combinations of percentage was investigated the successfulness with link to two clustering methods - Euclidean distance measure and Mahalanobis distance measure. These two measures were chosen because that the first stated is used very often and is often described as very successful, and the second distance measure eliminates a potential problem with correlations of variables that characterize the individual objects.

Based on stated results, it was found that it is not be clearly stated which of the distance measures is the most successful. It is always necessary to evaluate the combination of clustering methods, distance measures and the coefficient. Generally, when comparing both methods always the better results are obtained using Mahalanobis distance measure with a given combination of clustering method. When using Mahalanobis distance measure the best results when determining the optimal number of clusters are obtained using the Davies-Bouldin's index whose success when using average and centroid method are 68.75 and 65. 63%, resp. Contrary to that, when using the Euclidean distance measure the best results are achieved, both at 59.38% of cases when using nearest neighbour and in the case of Davies-Bouldin's and Dunn's index.

Acknowledgment

This paper was supported by long term institutional support of research activities IP400040 by Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

References

Bílková, D. (2011). Modelling of income and wage distribution using the method of l-moments of parameter estimation. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 40-50). ISBN 978-80-86175-77-5.

Bílková, D. (2012). Development of wage distribution of the Czech Republic in recent years by highest education attainment and forecasts for 2011 and 2012. In Löster T., Pavelka T. (Eds.), 6th International Days of Statistics and Economics (pp. 162-182). ISBN 978-80-86175-86-7.

Gan, G., Ma, Ch., Wu, J. (2007). *Data Clustering Theory, Algorithms, and Applications*, ASA, Philadelphia.

Halkidi, M., Vazirgiannis, M. (2001). *Clustering validity assessment: Finding the optimal partitioning of a data set*, Proceedings of the IEEE international conference on data mining, pp. 187-194.

Löster, T., & Langhamrova, J. (2011). Analysis of long-term unemployment in the Czech Republic. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 307-316). ISBN 978-80-86175-77-5.

Marek, L. (2013). Some Aspects of Average Wage Evolution in the Czech Republic. In: International Days of Statistics and Economics. [online], Slaný: Melandrium, pp. 947–958. ISBN 978-80-86175-87-4. URL: http://msed.vse.cz/files/2013/208-Marek-Lubos-paper.pdf.

Megyesiova, S., & Lieskovska, V. (2011). Recent population change in Europe. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 381-389). ISBN 978-80-86175-77-5.

Megyesiova, S., & Lieskovska, V. (2012). Are europeans living longer and healthier lives?. In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 766-775). ISBN 978-80-86175-86-7.

Meloun, M., Militký, J., Hill, M. (2005): Počítačová analýza vícerozměrných dat v příkladech, Academia, Praha.

Miskolczi, M., Langhamrova, J., & Fiala, T. (2011). Unemployment and GDP. In Löster Tomas, Pavelka Tomas (Eds.), International Days of Statistics and Economics (pp. 407-415). ISBN 978-80-86175-77-5.

Pavelka, T. (2012). The minimum wage in the Czech Republic – the instrument for motivation to work? In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 903-911). ISBN 978-80-86175-86-7.

Řezanková, H., Húsek, D., Snášel, V. (2009). *Cluster analysis dat*, 2. vydání, Professional Publishing, Praha.

Řezanková, H., & Löster, T. (2013). Shlukova analyza domacnosti charakterizovanych kategorialnimi ukazateli. *E+M. Ekonomie a Management*, *16*(3), 139-147. ISSN: 1212-3609.

Šimpach, O. (2012). Statistical view of the current situation of beekeeping in the Czech Republic. In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 1054-1062). ISBN 978-80-86175-86-7.

Stankovičová, I., Vojtková, M. (2007): Viacrozmerné štatistické metódy s aplikáciami, Ekonómia, Bratislava.

Žambochová, M. (2012): *Classification in terms of students' preferences for information sources, Efficiency and responsibility in education.* 9th International Conference on Efficiency and Responsibility in Education, Praha, p. 612-620, ISBN 978-80-213-2289-9.

Želinský, T., & Stankovičová, I. (2012). Spatial aspects of poverty in Slovakia. In Löster Tomas, Pavelka Tomas (Eds.), 6th International Days of Statistics and Economics (pp. 1228-1235). ISBN 978-80-86175-86-7.

Contact

Ing. Tomáš Löster, Ph.D.
University of Economics, Prague,
Dept. of Statistics and Probability
W. Churchill sq. 4,
130 67 Prague 3, Czech Republic
losterto@vse.cz