Stability of Grouping the EU Countries in Terms of Sustainable Development Levels

Dorota Rozmus

Abstract The stability of a taxonomy algorithm against minor changes in a data set (e.g. subtraction from a dataset, small changes in variable values) or algorithm parameters (e.g. random selection of parameter values) is a desired property of the method. There is an opinion in the literature that, when properly selected, multiple uses of a given algorithm should give rise to little or no difference in the final clusters (i.e. it should give stable results) and reveal the actual structure present in the data. This criterion is particularly applicable when selecting the number of groups (parameter *k*). The purpose of the paper will be to examine the stability of grouping the EU countries in terms of sustainable development levels.

1 Introduction

The main problem in taxonomy is to determine whether the groups that we detected reflect the actual structure of the groups present in the data. This involves the problem of selecting a “clustering model”, e.g. the number of

Dorota Rozmus
University of Economics, 1 Maja 50, 40-226 Katowice, Poland
✉ dorota.rozmus@ue.katowice.pl

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 4, No. 1, 2018

DOI: 10.5445/KSP/1000085951/19
ISSN 2363-9881

groups *k*, a distance metric, the control parameters of a clustering algorithm. Recently, the stability criterion increasingly gains in popularity in response to these problems (e.g. Ben-Hur and Guyon, 2003; Fang and Wang, 2012; Hennig, 2007; Koepke and Clarke, 2013; Ryazanov, 2016).

Informally, this criterion states that if a cluster algorithm is repeatedly used for independent samples (with unchanged parameters of the algorithm), resulting in similar grouping results, it can be considered as stable and reflecting the actual structure of the groups (Shamir and Tishby, 2008). Volkovich et al (2010) even state that the number of groups that maximizes the stability of clustering can serve as an estimate of the "true" number of groups.

These authors propose a number of different ways for measuring stability. Theoretical considerations have also led to the development of computer tools for the practical implementation of the proposed ways to study stability. The practical tools are available within several **R** packages, for example `clv`, `clValid`, `ClusterStability`, `fpc`, `pvclust`.

This article is a continuation of the author's earlier publications aimed at identifying the methods of measuring cluster stability proposed in the literature (e.g. Rozmus, 2017). In the earlier articles of the author the `clv`, `clvalid` and `fpc` packages were presented. This study will be concentrated on the `pvclust` package, which is the only one dedicated exclusively to hierarchical grouping methods.

Due to the hypothesis that cluster stability can be an answer to the question about the right number of groups in clustering, the purpose of this paper will be to examine the stability of grouping the EU countries in terms of sustainable development levels. The choice of such a research objective is dictated by the fact that the study of the similarity of territorial units with regard to different aspects (e.g. level of economic development, standard of living of the population) is the subject of a number of scientific papers, articles and reports (e.g. Kronthaler, 2005; Repkine, 2012; Shubat et al, 2016; Simpach, 2013). Little attention is devoted in these papers to the study of the stability of the proposed solutions.

2 `Pvclust` Package in R

The `pvclust` package (Suzuki and Shimodaira, 2006a) is designed to measure the stability of hierarchical methods (e.g. average, single, complete, median, centroid, Ward). It is measured by the *p*-value for each group and uses a bootstrap resampling for calculating this probability. Two types of probability are available:

- Bootstrap probability value (BP; Efron, 1979; Felenstein, 1985) and
- Approximately unbiased probability value (AU; Shimodaira, 2002; Shimodaira, 2004; Suzuki and Shimodaira, 2006b).

Based on the articles of Shimodaira (2002, 2004) as well as Suzuki and Shimodaira (2006b) the steps for calculating the bootstrap probability presented in Figure 1 can be formulated as follows:

1. Create bootstrap samples.
2. For each of them use the hierarchical method to get the so-called bootstrap replications of dendrograms.
3. Among all bootstrap dendrograms, calculate the percentage of those that found the specific group.

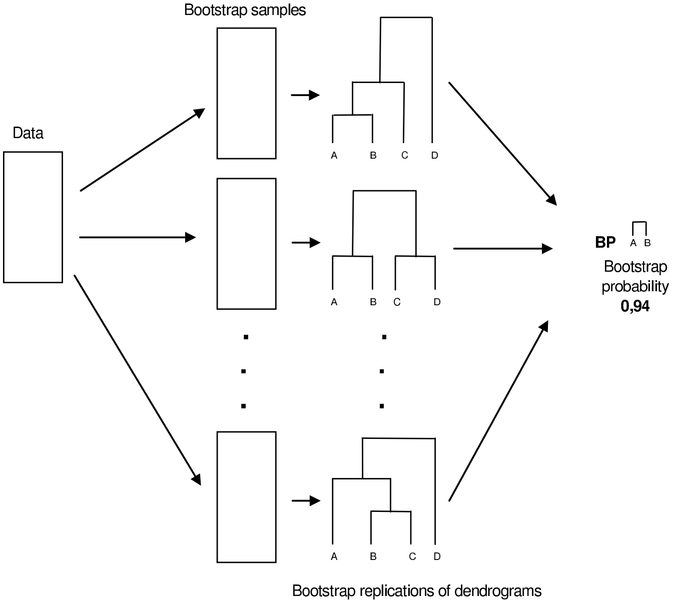

Figure 1: Scheme of bootstrap probability calculation.

The above procedure gives the bootstrap probability that is used to determine the probability of occurrence of a given group. It was shown that the p-value calculated in this way is biased (Hillis and Bull, 1993; Zharkikh and Li, 1992; Sanderson and Wojciechowski, 2000).

Several methods for bias correction exist. In the following we concentrate on the multiscale bootstrap method, which results in approximately unbiased bootstrap probability. In conventional bootstrap analysis the size of the bootstrap sample is identical to the original sample size.The multiscale bootstrap varies the bootstrap sample size in order to infer a correction formula for the biased p-value on the basis of the variation of the results for different sample sizes (Suzuki and Shimodaira, 2006b).

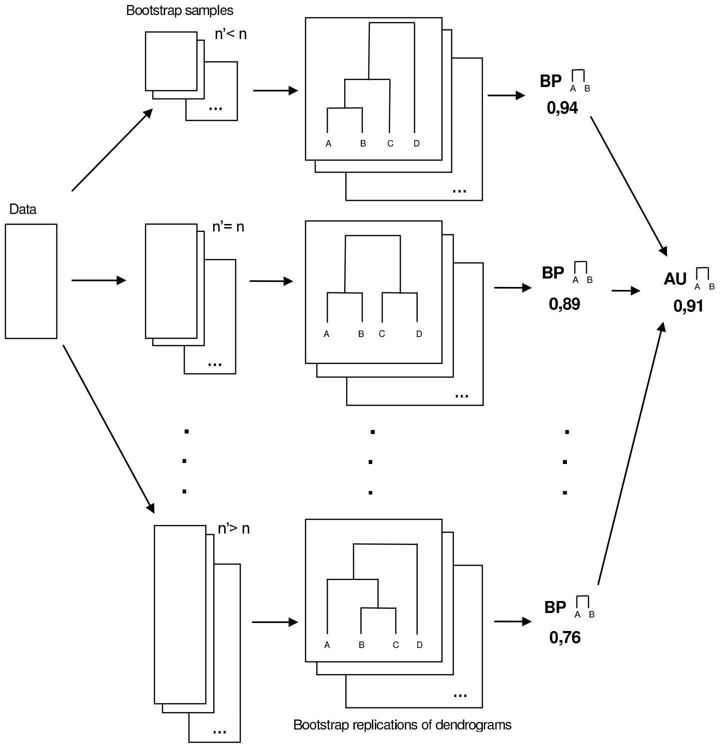

Figure 2: Scheme of multiscale bootstrap and approximately unbiased probability calculation.

Based on the articles of Shimodaira (2002, 2004) as well as Suzuki and Shimodaira (2006b) the steps for calculating the approximately unbiased probability presented in Figure 2 can be formulated follows:

1. Generate bootstrap samples for each sample size. The authors of the `pvclust` package suggest:

   $$n' \in \langle 0.5 \cdot n, 1.4 \cdot n \rangle\,, \tag{1}$$

   where n is the number of observations in the original data set (Suzuki and Shimodaira, 2006a).

2. Apply hierarchical clustering to each bootstrap sample to obtain the sets of bootstrap replications of dendrograms.

3. Compute the bootstrap probability for each sample size.

4. Using the bootstrap probabilities estimate the approximately unbiased p-value by fitting a theoretical equation to them:

   $$AU = 1 - \Phi\left(d - c\right), \tag{2}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. We estimate c and d by fitting the theoretical curve

$$BP(\tau) = 1 - \Phi\left(\frac{d}{\tau} + c \cdot \tau\right) \tag{3}$$

to the $BP(\tau)$ observed values calculated by the multiscale bootstrap method, with $\tau = \sqrt{n/n'}$ and where n' is the number of observations in the bootstrap samples in the multiscale bootstrap method. More details about the calculation of $BP(\tau)$ can be found in Shimodaira (2002).

3 Results

Sustainable development supports the environment-friendly economy based on a rational, economical and more competitive use of resources. The EU Sustainable Development Strategy (launched by the European Council in 2001 and renewed in June 2006) aims for the continuous improvement in the quality of life for current and future generations.

The survey has been conducted on the data set obtained from the Sustainable Development Indicators application developed by the Central Statistical Office in Poland. The variables in this application are grouped in four domains: social, economic, environmental, institutional-political. This study used 51 indicators with completed data. All observations are from the year 2015. The presentation of the results will start with presenting a line of code from **R** and discussing the most important parameters.

```
stability <- pvclust( data, method.hclust="ward",
  method.dist="correlation", nboot=1000, r=seq(.5,1.4,by=.1) )
```

The study opted for the Ward method as a clustering method, where the distance was calculated by the correlation coefficient, `nboot=1000` means that 1000 bootstrap samples were used, and `r=seq(.5,1.4,by=.1)` shows that the sample size was changed by 10 percentage points from 50 up to 150 percent.

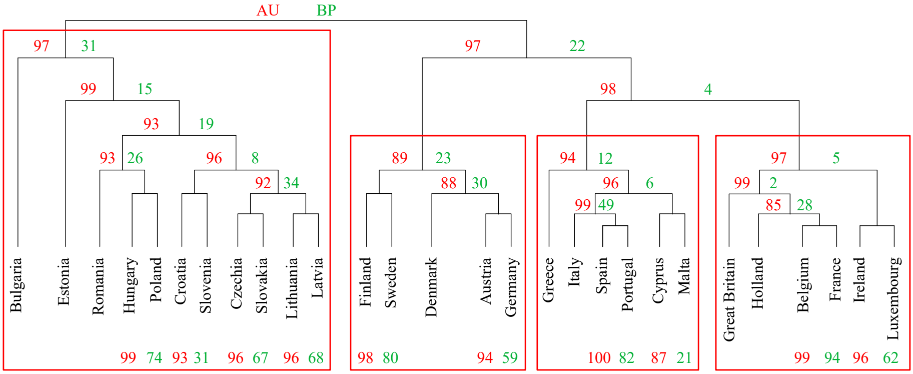

Figure 3: Result of grouping and values of stability measures (green: BP values, red: AU values).

Figure 3 presents a dendrogram with groups and values of stability measures. The green values (“BP” on the right side of each branch) show the probability calculated by the bootstrap probability method, while the red values (“AU” on the left side of each branch) are calculated with the approximately unbiased method. Only values of “AU” are taken into consideration. One can separate four groups of states. The first contains: Bulgaria, Estonia, Hungary, Poland, Croatia, Slovenia, Czechia, Slovakia, Lithuania and Latvia. These are former communist states, which belong to the EU countries with a lower development level, where the implementation of the postulates of sustainable development policy often encounters some difficulties. The probability of occurrence for this group is 0.97. Removing Bulgaria from this cluster would increase the probability of existence of this group to 0.99. The second group consists mainly of Scandinavian countries, i.e. Finland, Sweden, Denmark to which Austria and Germany were attached. It is a group of leading economies in the EU, where the postulates of sustainable development policy are largely and quite easily implemented. The surprise is that this group has the lowest probability of occurrence, only 0.89, which may indicate a large heterogeneity and internal diversity of this group. The third group includes countries from the south of Europe: Greece, Italy, Spain, Portugal, Cyprus, Malta. The probability of occurrence for this group is 0.94. The last group occurs with high probability of 0.97 and it contains Great Britain, Holland, Belgium, France, Ireland and Luxembourg.

4 Summary

In this contribution we used the Shimodaira (2002) approximately unbiased test of cluster stability for checking the validity of a cluster solution of European countries with regard to their sustainable development level (Shimodaira, 2002). In this article we used the approximately unbiased *p*-value, which is part of the `pvclust` package in **R**. Using this method for measuring the stability of the cluster solution of EU countries we obtained four highly reliable groups.

References

Ben-Hur A, Guyon I (2003) Detecting Stable Clusters Using Principal Component Analysis. Functional Genomics. Methods in Molecular Biology 224:159–182, Humana Press. DOI: 10.1385/1-59259-364-X:159.

Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 7(1):1–26, Institute of Mathematical Statistics.

Fang Y, Wang J (2012) Selection of the Number of Clusters via the Bootstrap Method. Computational Statistics and Data Analysis 56(3):468–477, Elsevier Science Publishers B. V., Amsterdam (The Netherlands). DOI: 10.1016/j.csda.2011.09.003.

Felenstein J (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39(4):783–791. DOI: 10.1111/j.1558-5646.1985.tb00420.x

Hennig C (2007) Cluster-Wise Assessment of Cluster Stability. Computational Statistics and Data Analysis 52:258–271.

Hillis D, Bull J (1993) An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology 42(2):182–192, Oxford University Press, Society of Systematic Biologists. DOI: 10.2307/2992540.

Koepke H, Clarke B (2013) A Bayesian Criterion for Cluster Stability. Statistical Analysis and Data Mining: The ASA Data Science Journal (Special Issue on Statistical Learning) 6(4):346–374, John Wiley & Sons (on behalf of the American Statistical Association). DOI: 10.1002/sam.11176.

Kronthaler F (2005) Economic Capability of East German Regions: Results of a Cluster Analysis. Regional Studies 39(6):739–750. DOI: 10.1080/00343400500213630.

Repkine A (2012) How Similar are the East Asian Economies? A Cluster Analysis Perspective on Economic Cooperation in the Region. Journal of International and Area Studies 19(1):27–44, Institute of International Affairs, Graduate School of International Studies, Seoul National University.

Rozmus D (2017) Using R Packages for Comparison of Cluster Stability. Acta Universitatis Lodziensis. Folia Oeconomica 4(330):77–86. DOI: 10.18778/0208-6018.330.05.

Ryazanov V (2016) About Estimation of Quality of Clustering Results via its Stability. Intelligent Data Analysis – Intelligent Computing for Pattern Recognition, Image Processing and Computer Vision Papers from CIARP 2014 20(1):5–15, Puerto Vallarta, Jalisco (Mexico). DOI: 10.3233/IDA-160842.

Sanderson MJ, Wojciechowski M (2000) Improved Bootstrap Confidence Limits in Large-Scale Phylogenies, with an Example from Neo-Astragalus (Leguminosae). Systematic Biology 49(4):671–685, Oxford University Press, Society of Systematic Biologists. DOI: 10.1080/106351500750049761.

Shamir O, Tishby N (2008) Cluster Stability for Finite Samples. Advances in Neural Information Processing Systems 20:1297–1304, Platt JC, Koller D, Singer Y, Roweis ST (eds), Curran Associates, Inc. URL: `http://papers.nips.cc/paper/3227-cluster-stability-for-finite-samples.pdf`.

Shimodaira H (2002) An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology 51(3):492–508, Oxford University Press, Society of Systematic Biologists. DOI: 10.1080/10635150290069913.

Shimodaira H (2004) Approximately Unbiased Tests of Regions Using Multistep-Multiscale Bootstrap Resampling. Annals of Statistics 32(6):2616–2641, Institute of Mathematical Statistics. DOI: 10.1214/009053604000000823.

Shubat O, Bagirova A, Makhabat A, Ivlev A (2016) The Use of Cluster Analysis for Demographic Policy Development: Evidence from Russia. 30th European Conference on Modelling and Simulation (ECMS), pp. 159–165, European Council for Modeling and Simulation, Regensburg (Germany). DOI: 10.7148/2016-0159.

Simpach O (2013) Application of Cluster Analysis on the Demographic Development of Municipalities in the Districts of Liberecky Region. Conference Proceedings of the 7th International Days of Statistics and Economics, pp. 1390–1399, Prague (Czech Republic). ISBN: 978-8-086175-87-4, URL: https://msed.vse.cz/msed_2013/en/toc.

Suzuki R, Shimodaira H (2006a) Hierarchical Clustering with p-values via Multiscale Bootstrap Resampling. URL: https://www.researchgate.net/publication/230710851_Hierarchical_clustering_with_P-values_via_multiscale_bootstrap_resampling.

Suzuki R, Shimodaira H (2006b) Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering. Bioinformatics 22(12):1540–1542, Oxford University Press, International Society for Computational Biology. DOI: 10.1093/bioinformatics/btl117.

Volkovich Z, Barzily Z, Toledano-Kitai D, Avros R (2010) The Hotteling's Metric as a Cluster Stability Measure. Computer Modelling and New Technologies 14(4):65–72. ISSN: 1407-5814, URL: http://www.cmnt.lv/en/on-line-journal/2010/2010-volume-14-4.

Zharkikh A, Li W (1992) Statistical Properties of Bootstrap Estimation of Phylogenetic Variability from Nucleotide Sequences. I. Four Taxa with a Molecular Clock. Molecular Biology and Evolution 9(6):1119–1147, Oxford University Press, Society for Molecular Biology and Evolution. DOI: 10.1093/oxfordjournals.molbev.a040782.