Clustering Analysis

Hervé Guillon

11 May 2020

Contents

Clustering Analysis

Countries in the study area are clustered in relation with socio-hydrologic variables. Two methods of clustering are used: *k*-means clustering (Hartigan and Wong 1979) and hierarchical clustering (Murtagh 1983). The clustering is performed with Euclidian distances and following Ward's criterion. The optimal number of clusters is investigated by evaluating the evolution with the number of clusters of the total within sums of square and of the average silhouette width (Rousseeuw 1987). In addition, the following four validation metrics are used to assess the stability of the clustering under the complete set of clustering variables and a iterative procedure where one variable is removed from the set, an approach akin to leave-one-out cross-validation:

1. the average proportion of (APN) measures the proportion of observations not placed in the same cluster under both cases and evalutes how robust are the clusters under cross-validation (S. Datta and Datta 2003);
2. the average distance between means (ADM) measures the variation of the cluster center and evaluates the stability of the localization of the cluster in the multi-dimensional clustering variable space (S. Datta and Datta 2003);
3. average distance (AD) measures the distance between observations placed in the same cluster and evaluates within-cluster stability (S. Datta and Datta 2003);
4. the figure of merit (FOM) estimates the predictive power of the clustering algorithm by measuring the within-cluster variance of the removed variable (Yeung, Haynor, and Ruzzo 2001).

Both clustering methods yield similar results. The total within sum of squares exhibits a shift in the evolution of the total within sum of square after two clusters are chosen. Similarly, the average silhouette width strongly exhibits a peak for two clusters. Further inspection of clustering in PCA dimensions indicates that the cluster with Mexico and Brazil is significantly distinct from all other countries, explaining the observation of a sharp peak in average silhouette width. However, validation metrics exhibits optimal null values of APN and ADM for two or three clusters. In addition, AD and FOM are lower for three clusters than for two. Based on this results, we chose three clusters to describe the grouping of countries based on their socio-hydrologic variables.

References

Datta, Susmita, and Somnath Datta. 2003. "Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data." *Bioinformatics* 19 (4). Oxford University Press: 459–66.

Hartigan, J. A., and M. A. Wong. 1979. "A K-means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28 (1): 100–108. http://www.jstor.org/stable/2346830?

origin=JSTOR-pdf.

Murtagh, Fionn. 1983. “A Survey of Recent Advances in Hierarchical Clustering Algorithms.” *The Computer Journal* 26 (4). Oxford University Press: 354–59.

Rousseeuw, Peter J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20. North-Holland: 53–65.

Yeung, Ka Yee, David R. Haynor, and Walter L. Ruzzo. 2001. “Validating Clustering for Gene Expression Data.” *Bioinformatics* 17 (4). Oxford University Press: 309–18.