# Distributional Data Clustering and Visualization for Official Statistics

**Rosanna Verde** [1]

*[1] Dipartimento di Matematica e Fisica - Università della Campania "Luigi Vanvitelli", Caserta, Italy, rosanna.verde2@gmail.com*

Distributional Data Analysis (DDA) is a new field of research related to Symbolic Data Analysis (Bock and Diday, 2000). The statistical units, or objects, are described by empirical distribution of values observed for numerical variables. In some cases, the distributions are synthesis or aggregated data, in order to preserve the confidentiality of the information. Many Official Statistics are in form of "histogram data", like the "Financial Characteristics for Housing Units With a Mortgage" data of the ACS American Community Survey 2015, of the UScensus Bureau. In such a case, an interval on proportions is furnished as a sort of "confidence interval".

Starting from previous techniques, proposed for analyzing DDA (e.g., clustering, principal component analysis), we introduce a further information in the data, given by the intervals of proportions. That leads to consider that each statistical unit is described by a set of distributions for each variable, expressed by the convex combination of the interval values on proportions.

A comparison between objects is then performed using a suitable measure between distributions: the L2 Wasserstein distance (Rüschendorf, 2001). This metric is particularly appropriate in DDA, and some of its properties have been demonstrated, especially, the decomposition of the L2 Wasserstein distance in three components related to the position, size and shape characteristics of the distributions.

A hierarchical clustering analysis allows to discover classes of objects according to the characteristics of the distribution of the values, taking also into account their variability, expressed by a set of distributions, carried out as convex combination of the bounded distributions.

A new visualization tool of the characteristics of the distributions, in a reduced subspace, has been furnished by a Principal Component Analysis (PCA) technique for distributional data (Verde et al., 2016).

An extension of PCA to this kind of Official Statistics allows a visualization of the variability of the data related to the several components of DDA.

Finally, DDA presents strong potentiality in the analysis of Big Data, and all that concerns the Volume of values and the Variability.

**Bibliography:**

Bock H.-H., Diday E. (2000) *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis and Knowledge Organization. Springer-Verlag, Berlin.

Dias, S., Brito, P. (2015) Linear Regression Model with Histogram-Valued Variables. *Statistical Analysis and Data Mining*, 8(2), 75-113.

Irpino A., Verde R. (2006) A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. IN: BATANJELI, V., BOCK, H.H., FERLIGOJ, A. & ZIBERNA, A. (Eds.) *Data science and classification, IFCS 2006*. Springer, Berlin, 185–192.

Irpino A., Verde R. (2008a) Dynamic clustering of interval data using a Wasserstein-based distance. *Pattern Recogn Lett*, 29, 1648–1658.

Irpino A., Verde R. (2008) Comparing histogram data using a Mahalanobis-Wasserstein distance. IN: BRITO, P. (Ed.) COMPSTAT 2008. Physica-Verlag, Heidelberg, 77–89.

Kim J., Billard L. (2013) Dissimilarity measures for histogram-valued observations. *Communications in Statistics: Theory and Methods*, 42, 283–303.

Rüschendorf L. (2001) Wasserstein metric. IN: HAZEWINKEL, M. (Ed.) *Encyclopedia of mathematics.* Springer, New York.

Verde R., Irpino A., Balzanella A. (2016). Dimension Reduction Techniques for Distributional Symbolic Data. *IEEE Transactions on Cybernetics*, 46, 344-355.