NEW STATISTICS FOR NEW DATA: A PROPOSAL FOR COMPARING MULTIVALUED NUMERICAL DATA

Rosanna Verde, Antonio Irpino

Department of European and Mediterranean Studies, Second University of Naples, Via del Seticio 15, 81100 – Caserta, Italy
rosanna.verde@unina2.it antonio.irpino@unina2.it

Abstract

In data mining, it is usually to describe a set of individuals using some summaries (means, standard deviations, histograms, confidence intervals) that generalize individual descriptions into a typology description. In this case, data can be described by several values. In this paper, we propose an approach for computing basic statics for such data, and, in particular, for data described by numerical multi-valued variables (interval, histograms, discrete multi-valued descriptions).We propose to treat all numerical multi-valued variables as "modal numerical variables". To obtain new basic statistics for measuring the variability and the association between such variables, we extend the classic measure of inertia, calculated with the Euclidean distance, using the L2 Wasserstein distance defined between probability measures. The distance is a generalization of the Wasserstein distance, that is a distance between quantile functions of two distributions. Some properties of such a distance are shown. Among them, we prove the Huygens theorem of decomposition of the inertia. We show the use of the Wasserstein distance and of the basic statistics presenting a k-means like clustering algorithm, for the clustering of a set of data described by modal numerical variables, on a real data set.

Key words: Wasserstein distance, inertia, dependence, multi-valued numerical data

1. INTRODUCTION

In many real experiences, data are collected and/or represented by multi-valued descriptions: intervals, frequency distributions, histograms, density distributions, and so on. Several approaches have been presented in the literature for to treat such data. In particular, when data are described by subsets of $\Re$, the main approaches are the Interval arithmetic approach, for treating interval data [19], the fuzzy approach, and the Symbolic Data Analysis approach [8]. When data descriptors domain is categorical, one of

the main interesting approach is the Compositional data one [1]. Without a loss of generality, when necessary, we refer to the Symbolic Data Analysis approach that can be considered as a good generalization of the other ones, including in its data definitions interval, multi-valued discrete, multi-categorical, histogram and modal descriptors [8, 7, 2]. The last ones can model the description of an individual, or of a concept, by distribution of probabilities, frequencies or, in general, by random variables. In the last years, several authors are proposing and defining new statistics and new techniques for the analysis of a particular case of modal data description: the histogram-valued data. An interesting approach is due to Billard and Diday [7]. They define new elementary statistics, association measures and a linear regression technique for the analysis of this kind of data. Further, [15] proposed an extension of the dynamic clustering algorithm and hierarchical clustering for data described by histograms.

After presenting histogram data and histogram variables in section **??**, in section 3.1, we suggest using a distance based on the Wassertein metric [14] for comparing two distributions, that can be considered as an extension of the Euclidean distance between quantile functions. All the obtained results are generalizable to data described by density functions where the first two moments are finite.

Data can be described by several (histogram) variables. The first problem to solve in the analysis of multivariate data is the standardization of such data in order to balance their contribution to the results of the analysis. Other approaches dealing with the computation of the variability of a set of complex data can be found in [6], [5] and [9]. Billard [6] and Bertrand and Goupil [5] apply the concept of variability to interval-valued data considering an interval-valued realizations on $[a, b]$ that is uniformly distributed $U \sim (a, b)$. On this basis, Bertrand and Goupil [5] developed some basic statistics to interval data and Billard [6] extends them for the computation of dependence and interdependence measures for interval-valued data. ...Final organization of the paper to be inserted here....

2. DATA DEFINITION AND BASIC STATISTICS FOR NUMERICAL MODAL-DATA

In this paper, we follow the Symbolic Data Analysis (SDA) approach for the definition of data introducing some generalizations [8, 7]. SDA aims to extend classical data analysis and statistical methods to more complex data called symbolic data. Considering the classical situation with a set of units $E = \{1, ..., i, ..., N\}$ and a set of p variables $y_1, ..., y_p$ Bock and Diday [8] define symbolic variables as follows:

Definition 1. A variable y is termed *set-valued* with domain Y, if for all $i \in E$,

$$\begin{array}{c} y : E \to D \\ i \mapsto y(i) \end{array} \tag{1}$$

where the description D is defined by $D = \mathcal{P}(Y) = \{U \neq \emptyset | U \subseteq Y\}$. A set-valued variable y is called *multi-valued* if its description set D_c is the set of all finite subsets of the underlying domain Y; such that $|y(i)| < \infty$, for all $i \in E$.

A set-valued variable y is called *categorical multi-valued* if it has a finite set Y of categories and *quantitative multi-valued* if the values $y(i)$ are finite sets of real numbers.

A set-valued variable y is called *interval-valued* if its description set D_I is the set of intervals of $\Re$.

Definition 2. A modal variable y on a set E of objects with domain Y is a mapping

$$y(i) = (S(i), \pi_i), \forall i \in E \tag{2}$$

where π_i is a measure or a (frequency, probability or weight) distribution on the domain Y of possible observation values (completed by a σ-field), and $S(i) \subseteq Y$ is the support of π_i in the domain Y. The description is denoted by D_m.

In the present paper, we do not treat the *multi-categorical* case, but only those descriptions based on numerical support. Indeed, if the support is categorical, we are in presence of a compositional data description [1], expressed by vectors of nonnegative real components having a constant sum.

We propose to treat all numerical (single-valued or set-valued) description as particular cases of the modal description. Considering the definition 2, we treat it in a probabilistic perspective.

Definition 3. Given a set E of objects with domain Y and support $S(i)$ partitioned into n_i subsets, a probability measure defining a density function ψ and the respective distribution function Ψ, such that

$$\Psi_i(y = S_h(i)) = \int_{S_h(i)} \psi_i(y) dy \tag{3}$$

where $h = 1, ..., n_i$, a modal variable y is a mapping

$$y(i) = \{S_h(i), \Psi_i(S_h(i))\}, \forall i \in E. \tag{4}$$

In the following, we consider the main types of symbolic numeric descriptors. Once defined the support $S(i)$, the density function ψ and the distribution function Ψ, we propose how to consider them as particular modal-numeric descriptor.

Classic single valued data $S(i) = y_i$ such that $y_i \in \Re$, and $\pi_i = 1$

An individual $i \in E$ is described by a single value y_i. It is possible to consider it as a modal-numeric variable with associated a density function that follows as Dirac delta function shifted in y_i:

$$\psi_i(y) = \delta(y - y_i) = \begin{cases} +\infty & if \quad y = y_i \\ 0 & otherwise \end{cases}$$

subject to the constraint that $\int_{-\infty}^{+\infty} \delta(y - y_i)dx = 1$.

The corresponding distribution function is:

$$\Psi_i(y = y_i) = \Psi_i(y < y_i^+) - \Psi_i(y < y_i^-) = \int_{y_i^-}^{y_i^+} \delta(y - y_i)dy = 1$$

In this case the modal-numeric description is:

$$y(i) = (y_i, \Psi_i(y = y_i)) = (y_i, 1).$$

Interval description $S(i) = [a_i, b_i]$ such that $a_i \leq y_i \leq b_i$, and assuming a uniform distribution in $S(i) = [a_i, b_i]$, we can rewrite π_i as

$$\psi_i(y) = \begin{cases} \frac{1}{b_i - a_i} & if \quad a_i \leq y \leq b_i \\ 0 & otherwise \end{cases}$$

The corresponding distribution function is:

$$\Psi_i(y) = \begin{cases} 0 & if \quad y < a_i \\ \int_{a_i}^{y} \frac{1}{b_i - a_i} dy & if \quad a_i \leq y \leq b_i \\ 1 & if \quad y > b_i \end{cases}$$

In this case the modal-numeric description is:

$$y(i) = ([a_i, b_i], \Psi_i(a_i \leq y \leq b_i)) = ([a_i, b_i], 1).$$

If we have information about the distribution of the data in the interval we may consider $\Psi_i(y)$ as a the (cumulative) distribution function corresponding to $\psi_i(y)$.

Histogram valued description We assume that $S(i) = [\underline{z}_i; \overline{z}_i]$ (the support is bounded), where $z_i \in \Re$. The support is partitioned into a set of n_i intervals $S(i) = \{I_{1i}, ..., I_{ui}, ..., I_{n_i i}\}$, where $I_{li} = [\underline{z}_{li}, \overline{z}_{li})$ and $l = 1, ..., n_i$, i.e.

$$\begin{array}{ll} i. & I_{li} \cap I_{mi} = \emptyset;\ l \neq m\ ; \\ ii. & \bigcup\limits_{l=1,...,n_i} I_{li} = S(i) \end{array}$$

Histograms suppose that the values observed in each interval are a uniformly distributed. It is possible to define the modal description of i as follows:

$$y(i) = \{(I_{li}, \pi_{li}) \mid \forall I_{li} \in S(i);\ \pi_{li} = \Psi_i(\underline{y}_{li} \leq y \leq \overline{y}_{li}) = \int\limits_{I_{ui}} \psi_i(z)dz \geq 0\}$$

where $\int\limits_{S(i)} \psi_i(y)dy = 1$.

Each element of the support is associated with a π_{li}, such that

$$\sum_{l=1}^{n_i} \pi_{li} = 1.$$

Given the generical interval $I_{li} = [\underline{z}_{li}, \overline{z}_{li}]$ where $\underline{z}_{li} < \overline{z}_{li}$, and $U(y|I_{li}) = U(y|\underline{z}_{li}, \overline{z}_{li})$ as the Uniform continuous function defined between $\underline{z}_{li}$ and $\overline{z}_{li}$, we may rewrite an histogram as a linear combination of Uniform distribution (a mixture) as follows:

$$\psi_i(y) = \sum_{l=1}^{n_i} \pi_{li} U(y|I_{li})$$

where $\psi_i(y)$ is a density function associated to the description of i and the corresponding distribution function is:

$$\Psi_i(y \leq b) = \sum_{l=1}^{n_i} \left(\pi_{li} \int_{-\infty}^{b} U(y|I_{li})dy \right).$$

Multi-valued discrete description Like in the histogram description, modal multi-valued discrete description can be considered as a mixture of Delta dirac distributions, where $S(i)$ is a set of distinct single values.

The support can be written also as $S(i) = \{y_{1i}, ..., y_{li}, ..., y_{n_i i}\}$. Each element of the support is associated with a π_{li}, such that $\sum\limits_{l=1}^{n_i} \pi_{li} = 1$.We then consider the function:

$$\psi_i(y) = \sum_{l=1}^{n_i} \pi_{li}\delta(y - y_{li})$$

where $\psi_i(y)$ is a density function associated to the description of i and the corresponding distribution function is:

$$\Psi_i(y \leq b) = \sum_{l=1}^{n_i} \left(\pi_{li} \int_{-\infty}^{b} \delta(y - y_{li})dy \right).$$

The same definition can be adapted when we need to describe an individual by mean of a continuous random variable.

Continuous random variable $S(i)$ correspond to the support of the random variable, $\psi_i(y)$ correspond to its density function.

We can consider, then, the density as

$$\psi_i(y) = f_i(y|\mathbf{\Theta}),$$

where $\mathbf{\Theta}$ is a vector of parameters, and the distribution function as

$$\Psi_i(y \leq b) = \int_{-\infty}^{b} f_i(y|\mathbf{\Theta})dy.$$

In order to define new basic statistics (mean, standard deviation), we need to do some assumption about data. Also, it is important to define a way for computing distances or inertia measures among data.

Further, we need a way to define an equivalence relation and, if it is possible an order relation among data. Further, we need to measure the dissimilarity between two multi-valued descriptions.

3. THE MEAN OF A SET OF DISTRIBUTION AS A MINIMIZER OF THE INERTIA

In order to define the mean of a modal numerical variable, we need to introduce an operator that satisfies some invariance properties. It is known, that the arithmetic mean is a statistics that holds the following properties:

Invariance with respect the sum i.e. given a set of n elements described by the variable y the mean M_y respect to the following equation

$$\sum_{i=1}^{n} y_i = nM_y$$

The mean is the value that minimize the inertia The (Moment of) inertia of a set of distribution is defined as the sum of squared distances between all the elements of set and its barycenter. Given a set of n elements described by the variable X the mean (barycenter) M is the argmin of the following minimization problem:

$$\sum_{i=1}^{n} (y_i - M)^2 = \sum_{i=1}^{n} d^2(y_i, M)$$

Two main issues are invoked from such conditions: the definition of the sum of distributions, and the definition of a consistent distance between distributions. The first problem is strictly related to the last one. Indeed, we may define the sum (or linear combination) of distributions once defined a distance function between two distributions. In [8], it is proposed a review of distances that can be used for comparing distributions and for defining the mean (barycenter) element which minimizes the inertia. First of all, when we treat data represented like random variables, we observed that it is preferable to work with their distribution functions. We observed that in two cases it is possible to define a barycenter element that can be represented as a distribution: using the $L2$ norm and the Wasserstein-Kantorovich-Monge-Gini-Mallows $L2$ distance. In the first case, the barycenter random variable of a set of data described as random variables is their mixture. In the last case, the barycenter of a set of data described as random variables can be represented by a random variable where the quantiles of such barycenter variable correspond to the mean of the corresponding quantiles of the data: i.e., the quantile function of the barycenter random variable is the mean of the quantile functions associated with the data distributions. In the next paragraph we present the Wasserstein-Kantorovich-Monge-Gini-Mallows $L2$ Wasserstein-Kantorovich-Monge-Gini-Mallows $L2$ distance (we call it simply Wasserstein distance) and its properties.

3.1 WASSERSTEIN DISTANCE BETWEEN DISTRIBUTIONS

If $\Phi_i(y)$ and $\Phi_j(y)$ are the distribution functions of two random variables $\phi_i(y)$ and $\phi_j(y)$ respectively, with first moments μ_i and μ_j, and s_i and s_i their standard deviations, the Wasserstein L2 metric is defined as [14]

$$d_W(\phi_i(y), \phi_j(y)) := \left[\int_0^1 \left(\Phi_i^{-1}(t) - \Phi_j^{-1}(t) \right)^2 dt \right]^{1/2} \tag{5}$$

where $\Phi_i^{-1}(t)$ and $\Phi_j^{-1}(t)$ are the quantile functions of the two distributions. It is possible to prove (see A) that the distance can be decomposed as:

$$d_W^2(\phi_i(y), \phi_j(y)) = \underbrace{(\mu_i - \mu_j)^2}_{Location} + \underbrace{(s_i - s_j)^2}_{Size} + \underbrace{2s_i s_j(1 - \rho_{QQ}(\Phi_i^{-1}, \Phi_j^{-1}))}_{Shape} \quad (6)$$

where

$$\rho_{QQ}(\Phi_i^{-1}, \Phi_j^{-1}) = \frac{\int_0^1 \left(\Phi_i^{-1}(t) - \mu_i\right)\left(\Phi_j^{-1}(t) - \mu_j\right) dt}{s_i s_j} = \frac{\int_0^1 \Phi_i^{-1}(t)\Phi_j^{-1}(t)dt - \mu_i\mu_j}{s_i s_j} \quad (7)$$

is the correlation coefficient of the quantiles of the two distributions as represented in a classical QQ plot. It is worth noting that $0 < \rho_{QQ} \leq 1$ differently from the classical range of variation of the Bravais-Pearson's correlation coefficient ρ. This decomposition allows us to take into consideration three aspects in the comparison of distribution function. The first aspect is related to the location: two distributions can differ in position and this aspect is explained by the distance between the mean values of the two distributions. The second aspect is related to the different variability of the compared distribution. This aspect is related to the different standard deviations of the distributions and to the different shapes of the density functions. While the former sub-aspect is taken into account by the distance between the standard deviations, the latter sub-aspect is taken into consideration by the value of ρ_{QQ}. Indeed, ρ_{QQ} is equal to one only if the two (standardized) distributions are of the same shape. Using this distance, we introduce an extended concept of inertia for a set of distributions.

This distance allow us to introduce the sum of a set of quantile functions. A quantile function is a non-decreasing function $f : [0, 1] \to \Re$ such that $\Phi_i^{-1}(t) = y(i)$. It is known that the sum of non-decreasing functions is itself a non-decreasing function[1], i.e., having n quantile functions, the $S^{-1}(t)$ function defined as follows is itself a quantile function:

$$S^{-1}(t) = \sum_{i=1}^{n} \Phi_i^{-1}(t) \quad \forall t \in [0, 1] \quad (8)$$

defining the product between a scalar $k \in \Re^+$ and a quantile function $F^{-1}(t)$ as:

[1] *In general, the difference between two non-decreasing functions is not a non-decreasing function. Then, we are not able to define a difference operator between quantile functions.*

$$k\Phi_i^{-1}(t) \quad \forall t \in [0,1] \tag{9}$$

we can define the mean quantile function (or barycenter) $\bar{M}^{-1}(t)$ as:

$$\bar{\Phi}_i^{-1}(t) = \frac{1}{n} S^{-1}(t) \quad \forall t \in [0,1]. \tag{10}$$

To this function can be associated the distribution function of the barycenter that we denote as $\bar{\Phi}(y)$ and its density function as $\bar{\phi}(y)$. We can compute also the mean of such modal description as

$$\mu_{\bar{y}} = \int_{-\infty}^{+\infty} y \cdot \bar{\phi}(y) dy \tag{11}$$

or its standard deviation as

$$s_{\bar{y}} = \sqrt{\int_{-\infty}^{+\infty} (y - \mu_{\bar{y}})^2 \cdot \bar{\phi}(y) dy} \tag{12}$$

The last result is very interesting. Indeed, it states that the barycenter of a set data has the same typology of the data, i.e.: the barycenter of a set of data, described as random variables, is a random variable itself. If we have single valued data (points), the barycenter is a point (i.e. it generalizes the arithmetic mean of a set of standard data), if we have interval-valued data, the barycenter is an interval valued description, if we have histogram-valued data, the barycenter is a histogram, and so on.

4. THE INERTIA OF A SET OF DATA DESCRIBED BY MODAL NUMERIC VARIABLES

A representative (prototype, barycenter) $\bar{y}_E$ associated with a set E of n elements described by a random variables y defined on $D \subset \Re$ is an element of the space of description of E. Extending the inertia concept of a set of points to a set of distributions, we may define such inertia as:

$$\begin{aligned} &Inertia_E = \textstyle\sum_{i=1}^n d_W^2(y(i), \bar{y}) = \sum_{i=1}^n \int_0^1 \left(\Phi_i^{-1}(t) - \bar{\Phi}^{-1}(t)\right)^2 dt = \\ &= \textstyle\sum_{i=1}^n \left[\left(\mu_{y(i)} - \mu_{\bar{y}}\right)^2 + \left(s_{y(i)} - s_{\bar{y}}\right)^2 + 2 s_{y(i)} s_{\bar{y}} (1 - \rho_{QQ}(\Phi_i^{-1}, \bar{\Phi}^{-1})) \right]. \end{aligned} \tag{13}$$

The $\bar{y}_E$ barycenter is obtained by minimizing the inertia criterion in (13), in the same way as the mean is the best least squares fit of a constant function to the given data points.

$\bar{F}_E$ is a distribution where its $t-th$ quantile is the mean of the $t-th$ quantiles of the n distributions belonging to E. In this paper we introduce new measures of variability consistent with the classical concept of variability of a set of elements, without discarding any characteristics of the complex data (bounds, internal variability, shape, etc.).

It is interesting to note that the Wasserstein distance allows the Huygens theorem of decomposition of inertia for clustered data. Indeed, we showed [16, 15] that it can be considered as an extension of the Euclidean distance between quantile functions.

Reasoning by analogy with the classic single-valued numerical data, the inertia of a set of n points described by single valued real variable $y_i \in \Re$ $(i = 1, \ldots, n)$ is given by the sum of the squared Euclidean distance of each pair of observations:

$$Inertia(y) = \sum_{i=1}^{n}\sum_{j=1}^{n} d_E^2(y_i, y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} (y_i - y_j)^2. \tag{14}$$

It can be proved that:

$$Inertia(y) = 2n \sum_{i=1}^{n} (y_i - \bar{y})^2 = 2n \cdot ss_y = 2n^2 \cdot s_y^2$$

Where ss_y is the sum of squares difference from the mean and s_y^2 is the variance. In our case, we generalize such statistics to the case of modal numerical variables as follows:

$$Inertia(y) = \sum_{i=1}^{n}\sum_{j=1}^{n} d_W^2(y(i), y(j)) = \sum_{i=1}^{n}\sum_{j=1}^{n} \int_0^1 \left(\Phi_i^{-1}(t) - \Phi_j^{-1}(t)\right)^2 dt. \tag{15}$$

Also in this case, it is possible to prove that:

$$Inertia(y) = 2n \sum_{i=1}^{n} \int_0^1 \left(\Phi_i^{-1}(t) - \bar{\Phi}^{-1}(t)\right)^2 dt.$$

We define ss_y^F as the sum of squares difference from the barycenter $\bar{\Phi}^{-1}$ and $s_y^{2_F}$ as the variance of a modal numerical variable and we can write:

$$Inertia(y) = 2n \cdot ss_y^F = 2n \cdot \left(s_y^F\right)^2. \tag{16}$$

The definition of the variance and of the standard deviation allow us to define a measure of variability of a set of modal multi-valued numerical data. We define the standard deviation as:

$$s_y^F = \left[\frac{\sum_{i=1}^{n} \int_0^1 \left(\Phi_i^{-1}(t) - \bar{\Phi}^{-1}(t) \right)^2 dt}{n} \right]^{1/2} = \left[\frac{\sum_{i=1}^{n} d_W^2 \left(y(i), \bar{y} \right)}{n} \right]^{1/2} \tag{17}$$

The main properties of this measures are the same of a classical variability measures:

Non-negativity $s_y^F \geq 0$.

Constant data description If all data have the same modal multi-valued numerical description $s_y^F = 0$.

Shrinking Given two real numbers $h \neq 0$ and k:

$$s^F_{(h \cdot y + k)} = |h| s_y^F$$

.

Comparing this measure with those introduced by [5] and by [7], the main differences are related to the value of standard deviation when the data have the same description. In that case, the standard deviation proposed by [7] is generally grater than zero also when data have the same modal numerical description.

5. MEASURES OF INTERPENDENCE BETWEEN MODAL NUMERICAL VARIABLES

In this section, we introduce new statistics for measuring the interdependence between two modal multi-valued numerical variables. We start introducing a new measure for the covariance between two variables denoted by y and z. We propose to extend the covariance measure for modal multi-valued numerical data as:

$$s_{yz}^F = \frac{s s_{yz}^F}{n} \tag{18}$$

For each individual we know only the marginal distributions (the modal multi-valued numerical description for each variable) of the multivariate distribution that has generated it, and it is not possible to known the dependency structure between two modal multi-valued numerical descriptions observed for two variables. We assume that each individual is described by

independent descriptions for each variables. This is commonly used in the analysis of symbolic data [6]. On this assumption, given two modal numerical variables y and z, a set E of n modal numerical data with distribution $F_i(y)$ and $F_i(z)$ $(i = 1, \ldots, n)$, and considering the barycenter distributions $\bar{F}(y)$ and $\bar{F}(z)$ of E for the two variables, we propose to extend the classical codeviance (ss_{xy}) measure to modal numerical variables as:

$$ss_{yz}^F = \sum_{i=1}^{n} \int_0^1 \left(F_{iy}^{-1}(t) - \bar{F}_y^{-1}(t)\right)\left(F_{iz}^{-1}(t) - \bar{F}_z^{-1}(t)\right) dt. \tag{19}$$

Recalling equation (7), we may express it as:

$$\begin{aligned} ss_{yz}^F = &\sum_{i=1}^{n} \left[\alpha_i \cdot s_{iy}s_{iz} - \beta_i \cdot s_{\bar{y}}s_{iz} - \gamma_i \cdot s_{iy}s_{\bar{z}}\right] + n \cdot \delta \cdot s_{\bar{y}}s_{\bar{z}} + \\ &+ \left(\sum_{i=1}^{n} \mu_{iy}\mu_{iz} - n\mu_{\bar{y}}\mu_{\bar{z}}\right) \end{aligned} \tag{20}$$

where

- $\alpha_i = \rho_{QQ}\left(F_{iy}^{-1}, F_{iz}^{-1}\right)$ is the QQ-correlation between the quantile function for the y variable and the quantile function for the y variable observed for the $i-th$ individual,
- $\beta_i = \rho_{QQ}\left(F_{iz}^{-1}, \bar{F}_y^{-1}\right)$ is the QQ-correlation between the quantile function of the barycenter of the y variable and the quantile function for the z variable observed for the $i-th$ individual,
- $\gamma_i = \rho_{QQ}\left(F_{iy}^{-1}, \bar{F}_z^{-1}\right)$ is the QQ-correlation between the quantile function of the barycenter of the z variable and the quantile function for the y variable observed for the $i-th$ individual,,
- $\delta = \rho_{QQ}\left(\bar{F}_y^{-1}, \bar{F}_z^{-1}\right)$ is is the QQ-correlation between the quantile function of the barycenter of the y variable and the quantile function of the barycenter of the y variable.

As a particular case, if all the distributions have the same shape (for example, they follow Gaussian distributions) then ρ_{QQ}'s are equal to 1 and ss_{yz}^F can be simplified as

$$ss_{yz}^F = \left(\sum_{i=1}^{n} \mu_{iy}\mu_{iz} - n\mu_{\bar{y}}\mu_{\bar{z}}\right) + \left(\sum_{i=1}^{n} s_{iy}s_{iz} - ns_{\bar{y}}s_{\bar{z}}\right)$$

It is interesting to note that this approach is fully consistent with the classical decomposition of the codeviance. Indeed, we may consider a distribution as an information related to a group of individuals. It can be proven that having a set of individuals grouped into k classes, the total codeviance can be decomposed in two additive components, the codeviance within and the codeviance between groups. With minimal algebra it is possible to prove that $|ss_{yz}^F|$ cannot be greater than $\sqrt{ss_y^F \cdot ss_z^F}$. Then, we introduce the correlation measure for two modal multi-valued numerical variables as:

$$r_{yz}^F = \frac{ss_{yz}^F}{\sqrt{ss_y^F \cdot ss_z^F}} = \frac{s_{yz}^F}{s_y^F \cdot s_z^F} \tag{21}$$

It is worth noting that if all modal multi-valued numerical descriptions have are identically distributed except for the first moments, r_{yz}^F depends only on the correlation of their first two moments, and that if $r_{yz}^F = 1$ (resp. -1) then all the histograms have their first moment aligned along a positive (resp. negative) sloped line and are identically distributed (except for the first two moments).

6. USING BASIC STATISTICS FOR MODAL NUMERIC DATA: MAHALANOBIS-WASSERTEIN DISTANCE

The proposed statistics can be useful for extending several algorithms of data analysis from classical single-valued numerical data to modal multi-valued numerical data.

Using standard deviation and covariance measure we can, for example, introduce the Mahalanobis version of the Wasserstein distance as follows.

Given a set E of n individuals described by p modal numerical variables, each individual can be described as a vector $\mathbf{y}(i) = [y_1(i), \ldots, y_p(i)]$. Let the variance-covariance matrix be denoted as $\Sigma^F = [ss_{hk}^F]_{p\times p}$, its corresponding inverse $\Sigma^{-1_F} = [a_{hk}]$ we can introduce the Mahalanobis-Wasserstein distance as follows:

$$d_{MW}(\mathbf{y}(i), \mathbf{y}(i')) = \tag{22}$$
$$= \sqrt{\sum_{h=1}^{p}\sum_{k=1}^{p}\int_0^1 a_{hk}\left(F_{ih}^{-1}(t) - F_{i'k}^{-1}(t)\right)\left(F_{ih}^{-1}(t) - F_{i'k}^{-1}(t)\right)dt}$$

the squared distance can be written:

$$
\begin{aligned}
&d_{MW}(\mathbf{y}(i),\mathbf{y}(i')) = \sum_{k=1}^{p} a_{kk} d_W^2\left(F_{ik}, F_{i'k}\right) + \\
&+2\sum_{h=1}^{p-1}\sum_{k=h}^{p} a_{hk}\left[\int_0^1 \left(F_{ih}^{-1}(t) - F_{i'h}^{-1}(t)\right)\left(F_{ik}^{-1}(t) - F_{i'k}^{-1}(t)\right) dt\right] = \\
&= \sum_{k=1}^{p} a_{kk} d_W^2\left(F_{ik}, F_{i'k}\right) + \\
&+2\sum_{h=1}^{p-1}\sum_{k=h}^{p} a_{hk}\left[\left(\alpha_{hk} - \beta_{hk} - \gamma_{hk} + \delta_{hk}\right) + \left(\mu_{ih} - \mu_{i'h}\right)\left(\mu_{ik} - \mu_{i'k}\right)\right]
\end{aligned}
\tag{23}
$$

where:
$\alpha_{hk} = \rho_{QQ}\left(F_{ih}^{-1}, F_{ik}^{-1}\right) \cdot s_{ih} s_{ik}$; $\beta_{hk} = \rho_{QQ}\left(F_{ih}^{-1}, F_{i'k}^{-1}\right) \cdot s_{ih} s_{i'k}$;
$\gamma_{hk} = \rho_{QQ}\left(F_{ik}^{-1}, F_{i'h}^{-1}\right) \cdot s_{ik} s_{i'h}$; $\delta_{hk} = \rho_{QQ}\left(F_{i'h}^{-1}, F_{i'k}^{-1}\right) \cdot s_{i'k} s_{i'k}$.

If all distributions have the same shape (i.e., the distributions differ only for their first two moments) the distance can be simplified as:

$$
\begin{aligned}
&d_{MW}(\mathbf{y}(i),\mathbf{y}(i')) = \sum_{k=1}^{p} d_W^2\left(y_k(i), y_k(i')\right) a_{kk} + \\
&+2\sum_{h=1}^{p-1}\sum_{k=h}^{p}\left[\left(s_{ih} - s_{i'h}\right)\left(s_{ik} - s_{i'k}\right) + \left(\mu_{ih} - \mu_{i'h}\right)\left(\mu_{ik} - \mu_{i'k}\right)\right] a_{hk}.
\end{aligned}
\tag{24}
$$

7. AN APPLICATION ON A CLIMATIC DATASET

In this section, we show some results of clustering of data describing the mean monthly temperature, pressure, relative humidity, wind speed and total monthly precipitations of 60 meteorological stations of the People's Republic of China[2], recorded from 1840 to 1988. For the aims of this paper, we have considered the distributions of the variables for January (the coldest month) and July (the hottest month), so our initial data is a 60×10 matrix where the generic (i, j) cell contains the distribution of the values for the $j-th$ variable of the $i-th$ meteorological station. Figure 1 shows the geographic position of the 60 stations, while in Table 1 we have the basic statistics as proposed in section 4, and in Table 2 we show the interdependency measures as proposed in section 5. In particular, the upper triangle of the matrix contains the $COVAR_F$'s, while the bottom triangle contains the $CORR_F$'s for each couple of the histogram variables.

[2] *Dataset URL: http://dss.ucar.edu/datasets/ds578.5/*

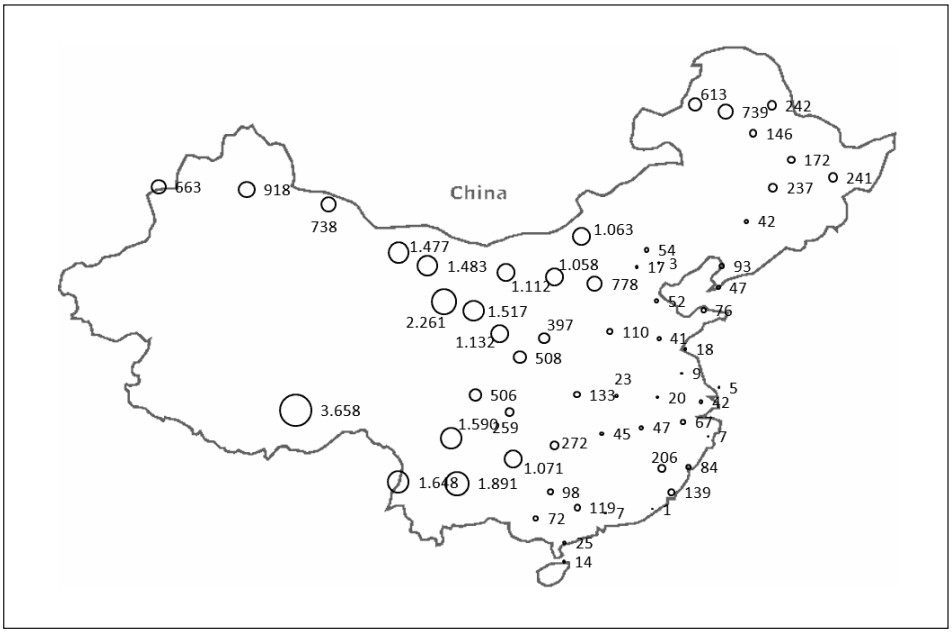

Fig. 1: The 60 meteorological stations of the China dataset; beside each point there is the elevation in meters.

Tab. 1: Basic statistics of the histogram variables: μ_j and σ_j are the mean and the standard deviation of the barycenter distribution of the *j-th* variable, while $VAR_F(X_j)$ and $STD_F(X_j)$ are the variability measures as presented in this paper.

#	Variable	μ_j	σ_j	$s^{2F}(y_j)$	$s^F(y_j)$
y_1	Mean Relative Humidity (percent) Jan	67.9	7.0	127.9	11.3
y_2	Mean Relative Humidity (percent) July	73.9	4.5	114.2	10.7
y_3	Mean Station Pressure(mb) Jan	968.3	3.6	5864.7	76.5
y_4	Mean Station Pressure(mb) July	951.1	3.0	5084.4	71.3
y_5	Mean Temperature (Cel.) Jan	-1.2	1.7	114.8	10.7
y_6	Mean Temperature (Cel.) July	25.2	1.0	11.3	3.4
y_7	Mean Wind Speed (m/s) Jan	2.3	0.6	1.1	1.0
y_8	Mean Wind Speed (m/s) July	2.3	0.5	0.6	0.8
y_9	Total Precipitation (mm) Jan	18.2	14.3	519.6	22.7
y_{10}	Total Precipitation (mm) July	144.6	80.8	499.9	70.7

Tab. 2: Covariances and correlations (in bold) of the ten histogram variables.

Vars	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
y_1	128.0	49.1	510.2	486.1	34.0	20.0	0.7	1.6	109.9	97.9
y_2	**0.41**	114.2	392.6	376.4	53.5	11.2	4.2	1.2	72.3	475.6
y_3	**0.59**	**0.48**	**5,864.7**	5,455.2	162.9	198.3	32.0	24.3	672.4	1,570.2
y_4	**0.60**	**0.49**	**1.00**	5,084.4	158.9	183.1	29.9	22.5	634.8	1,504.6
y_5	**0.28**	**0.47**	**0.20**	**0.21**	114.8	22.5	0.0	-1.5	119.7	305.6
y_6	**0.52**	**0.31**	**0.77**	**0.76**	**0.62**	11.3	0.4	0.3	41.4	56.9
y_7	**0.06**	**0.38**	**0.40**	**0.40**	**0.00**	**0.13**	1.1	0.7	2.9	17.0
y_8	**0.17**	**0.14**	**0.39**	**0.39**	**-0.18**	**0.11**	**0.82**	0.6	1.6	-0.7
y^9	**0.43**	**0.30**	**0.39**	**0.39**	**0.49**	**0.54**	**0.12**	**0.09**	519.6	426.0
y_{10}	**0.12**	**0.63**	**0.29**	**0.30**	**0.40**	**0.24**	**0.23**	**-0.01**	**0.26**	4,999.3

7.1 DYNAMIC CLUSTERING

The Dynamic Clustering Algorithm (DCA) [13] represents a general reference for partitioning algorithms. Let E be a set of n data described by p histogram variables y_j $(j = 1, \ldots, p)$. The general DCA looks for the partition $P \in P_k$ of E in k classes, among all the possible partitions P_k, and the vector $L \in L_k$ of k prototypes representing the classes in P, such that, the following Δ fitting criterion between L and P is minimized:

$$\Delta(P^*, L^*) = Min\{\Delta(P, L) \mid P \in P_k, L \in L_k\}. \tag{25}$$

Such a criterion is defined as the sum of dissimilarity or distance measures $\delta(x_i, G_h)$ of fitting between each object x_i belonging to a class $C_h \in P$ and the class representation $G_h \in L$:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{x_i \in C_h} \delta(x_i, G_h).$$

A prototype G_h associated to a class C_h is an element of the space of the description of E, and it can be represented as a vector of histograms. The algorithm is initialized by generating k random clusters or, alternatively, k random prototypes. We here present the results of two dynamic clustering using $k = 5$. The former considers δ as the squared Wasserstein distance among standardized data, while the latter uses the proposed squared Mahalanobis-Wasserstein distance. We have performed 100 initializations and we have considered the two partitions allowing the best quality index as defined in Chavent et al. (2003):

$$Q(P_k) = 1 - \frac{\sum_{h=1}^{k} \sum_{x_i \in C_h} \delta(x_i, G_h)}{\sum_{i \in E} \delta(x_i, G_E)}$$

where G_E is the prototype of the set E. $Q(P_k)$ can be considered as the generalization of the ratio between the inter-cluster inertia and the total inertia of the dataset. Comparing the two clustering results, we may observe that the two clustering agree only on the 65% of the observations (see Table 3): while DCA using Wassertein distance on standardized data allows a 61.53% of intra cluster inertia, DCA using Mahalanobis-Wassertein distance allows a 91.64%, but, considering that Mahalanobis distance removes redundancy between the variables, allows the definition of five clusters that collect stations at different elevations: the cluster 3 contains those stations between 0 and 140 meters, cluster 5 between 140 and 400 meters, cluster 1 between 500 and 900 meters, cluster 2 between 1000 and 1800, and cluster 4 between 2,000 and 3,500 meters. Observing a physical map of China, the obtained clusters seems more representative of the different typologies of meteorological stations for their location and elevation. It is interesting to note that, also in this case, the use of a Mahalanobis metric for clustering data gives the same advantages of a clustering after a factor analysis (for example, a Principal Components Analysis), because it removes redundant information (in terms of linear relationships) among the descriptors.

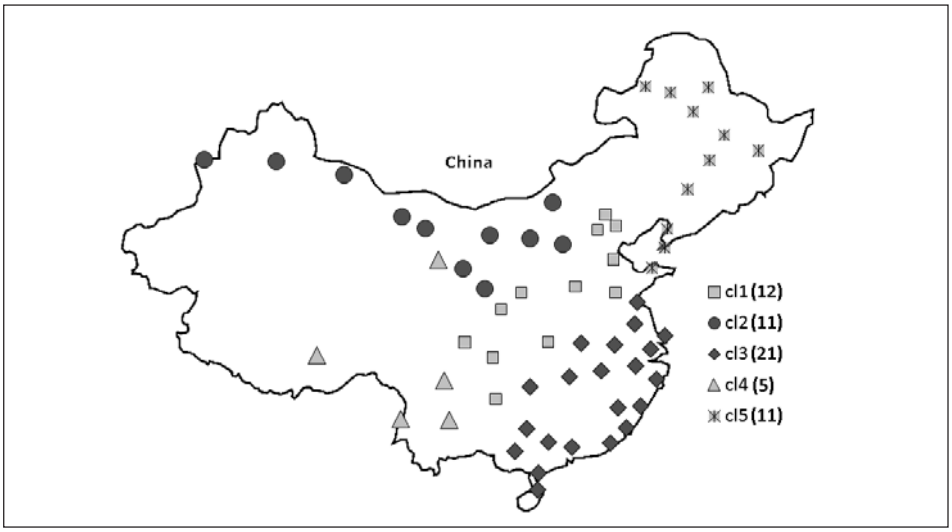

Fig. 2: Dynamic Clustering of the China dataset into 5 clusters (in brackets there is the cardinality of the cluster) using the Wasserstein distance on standardized data $Q(P_5) = 0.6253$.

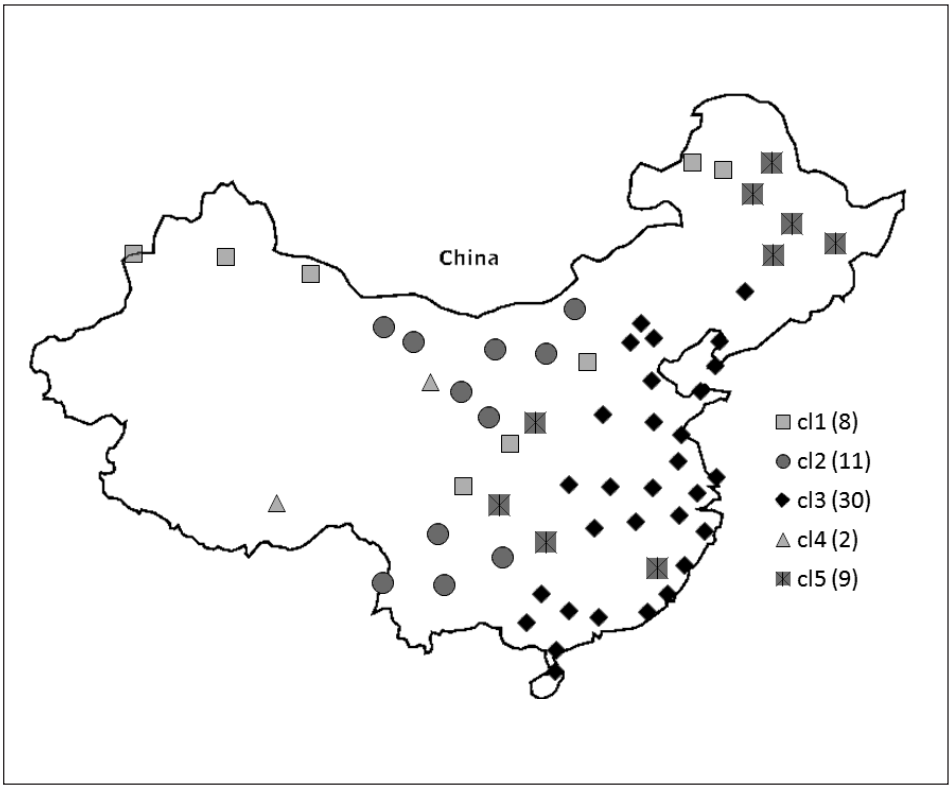

Fig. 3: Dynamic Clustering of the China dataset into 5 clusters (in brackets there is the cardinality of the cluster) using the Mahalanobis-Wasserstein distance, $Q(P5) = 0.9164$.

Tab. 3: Cross-classification table of the clusters obtained from the two dynamic clusterings.

		Clusters using Mahal.-Wass. distance					
		Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	**Total**
	Cl 1	***2***	1	7		2	**12**
Clustering using	Cl 2	4	***7***				**11**
Wasserstein distance	Cl 3			***19***		2	**21**
between standardized	Cl 4		3		***2***		**5**
data	Cl 5	2		4		***5***	**11**
	Total	**8**	**11**	**30**	**2**	**9**	**60**

8. CONCLUSIONS AND FUTURE RESEARCH

In this paper we have presented a new distance for comparing histogram data. The proposed method can be used in the interval data analysis whereas the intervals are considered as uniform densities according to Bertrand and Goupil (2000) and Billard (2007). Using the Wasserstein distance, we showed a way to standardize data, extending the classical concept of inertia for a set of histogram data. The Mahalanobis–Wassestein distance and the proposed interdependency measures between histogram variables can be considered as new useful tools for developing further analysis techniques for histogram data. The next step, considered very hard from a computational point of view (see Cuesta-Albertos and Matrán (1997)), is to find a way of considering the dependencies inside the histogram observations for multivariate histogram data in the computation of the Wasserstein distance.

A. Proof of the decomposition of the Wasserstein distance.

$$d_W^2\left(y(i), y(j)\right) := \int_0^1 \left(F_i^{-1}(t) - F_j^{-1}(t)\right)^2 dt = \\ = \underbrace{\left(\mu_i - \mu_j\right)^2}_{Location} + \underbrace{\left(s_i - s_j\right)^2}_{Size} + \underbrace{2 s_i s_j (1 - Corr_{QQ}(Y(i), Y(j)))}_{Shape} \tag{26}$$

Let us observe two density functions $f_i(y)$ and $f_j(y)$ having finite the first two moments. With each density function can be associated the distribution functions $F_i(y)$ and $F_j(y)$, the means μ_i and μ_j, the standard deviations s_i and s_j where:

$$\mu_i = \int_{-\infty}^{+\infty} y \cdot f_i(y) dy = \int_0^1 F_i^{-1}(t) dt$$

Indeed

$$\int_{-\infty}^{+\infty} y f(y) dy = \int_{-\infty}^{+\infty} y dF(y)$$

if $t = F(y)$ and considering that $yx = F^{-1}(F(y)) = F^{-1}(t)$ by substitution we obtain

$$\mu = \int_0^1 F^{-1}(t) dt$$

And where:

$$s^2(y) = \int_{-\infty}^{+\infty} y^2 f(y)dy - \mu^2 = \int_0^1 \left(F^{-1}(t)\right)^2 dt - \mu^2$$

for the same substitutions adopted above.

Now let assume to center the two distributions using their means such that:

$z(i) = y(i) - \mu_i$ and $F_i^{-1c}(t) = z(i)$ and $F_i^{-1c}(t) = F_i^{-1}(t) - \mu_i$

In [4] is proven that

$$d_W^2\left(y(i), y(j)\right) := \left(\mu_i - \mu_j\right)^2 + d_W^2\left(z(i), z(j)\right) \tag{27}$$

where

$$d_W^2\left(z(i), z(j)\right) := \int_0^1 \left(F_i^{-1c}(t) - F_j^{-1c}(t)\right)^2 dt \tag{28}$$

Developing the square we obtain

$$\begin{aligned} d_W^2\left(z(i), z(j)\right) &:= \int_0^1 \left(F_i^{-1c}(t)\right)^2 dt + \int_0^1 \left(F_j^{-1c}(t)\right)^2 dt - 2\int_0^1 F_i^{-1c}(t)F_j^{-1c}(t)dt = \\ &= \int_0^1 \left(F_i^{-1}(t) - \mu_i\right)^2 dt + \int_0^1 \left(F_j^{-1}(t) - \mu_j\right)^2 dt - 2\int_0^1 \left(F_i^{-1}(t) - \mu_i\right)\left(F_j^{-1}(t) - \mu_j\right) dt = \\ &= s_i^2 + s_j^2 - 2\int_0^1 \left(F_i^{-1}(t) - \mu_i\right)\left(F_j^{-1}(t) - \mu_j\right) dt \end{aligned} \tag{29}$$

Let us consider the following quantity

$$\begin{aligned} \rho_{QQ} &= \frac{\int_0^1 F_i^{-1c}(t)F_j^{-1c}(t)dt}{\sqrt{\int_0^1 \left(F_i^{-1c}(t)\right)^2 dt \int_0^1 \left(F_j^{-1c}(t)\right)^2 dt}} = \frac{\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt}{\sqrt{\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)^2 dt \int_0^1 \left(F_j^{-1}(t)-\mu_j\right)^2 dt}} = \\ &= \frac{\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt}{s_i s_j} \end{aligned} \tag{30}$$

It can be considered as the correlation of two series of data where each couple of observations is represented respectively by the $t-th$ quantile of the first distribution and the $t-th$ quantile of the second. In this sense we may consider it as the correlation between quantile functions represented by the curve of the infinite quantile points in a QQ plot. It is worth noting that $0 < \rho_{QQ} \leq 1$ differently from the classical range of variation of the Bravais-Pearson's correlation index $(-1, +1)$. Equation (29) can be rewritten as

$$d_W^2\left(z(i),z(j)\right) := s_i^2+s_j^2-2\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt = s_i^2+s_j^2-2\rho_{QQ}s_is_j \tag{31}$$

Adding and subtracting $2s_is_j$ we obtain

$$d_W^2\left(z(i),z(j)\right) := s_i^2+s_j^2-2s_is_j+2s_is_j-2\rho_{QQ}s_is_j = \left(s_i-s_j\right)^2+2s_is_j\left(1-\rho_{QQ}\right) \tag{32}$$

We may replace this result in (27) obtaining:

$$d_W^2\left(y(i),y(j)\right) := \left(\mu_i-\mu_j\right)^2 + d_W^2\left(z(i),z(j)\right) =$$

$$= \left(\mu_i-\mu_j\right)^2 + \left(s_i-s_j\right)^2 + 2s_is_j\left(1-\rho_{QQ}\right)$$

QED

REFERENCES

[1] AITCHISON, J. (1986), The Statistical Analysis of Compositional Data, New York: Chapman Hall.

[2] BILLARD, L., DIDAY, E. (2003), From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis *Journal of the American Statistical Association*, 98, 462, 470-487

[3] MALLOWS, C. L. (1972), A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43(2), 508-515.

[4] BARRIO, E., MATRAN, C., RODRIGUEZ-RODRIGUEZ, J. and CUESTA-ALBERTOS,J.A. (1999), Tests of goodness of fit based on the L2-Wasserstein distance. *Annals of Statistics* (1999), 27, 1230-1239.

[5] BERTRAND, P. and GOUPIL, F. (2000), Descriptive statistics for symbolic data. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer, Berlin, 103–124.

[6] BILLARD, L. (2007), Dependencies and Variation Components of Symbolic Interval–Valued Data. In: P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.):'*Selected Contributions in Data Analysis and Classication*, Springer, Berlin, 3–12.

[7] BILLARD,L. and DIDAY, E. (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chirchester.

[8] BOCK, H.H. and DIDAY, E. (2000), *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.

[9] BRITO, P. (2007), On the Analysis of Symbolic Data. In: P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.): *Selected Contributions in Data Analysis and Classication*. Springer, Berlin, 13–22.

[10] CHAVENT, M., DE CARVALHO, F.A.T., LECHEVALLIER, Y., and VERDE, R. (2003), Trois nouvelles méthodes de classification automatique des données symbolique de type intervalle. *Revue de Statistique Appliquée*, LI, 4, 5–29.

[11] CUESTA-ALBERTOS, J.A., MATR´AN, C., TUERO-DIAZ, A. (1997), Optimal transportation plans and convergence in distribution. *Journ. of Multiv. An.*, 60, 72–83.

[12] DIDAY, E., and SIMON, J.C. (1976): Clustering analysis, In: Fu, K.S. (Eds.), *Digital Pattern Recognition*, Springer Verlag, Heidelberg, 47–94.

[13] DIDAY, E. (1971), Le m´ethode des nu´ees dynamique, *Revue de Statistique Appliquée, 19, 2*, 19–34.

[14] GIBBS, A.L. and SU, F.E. (2002), On choosing and bounding probability metrics. *Intl. Stat. Rev. 7 (3),* 419{435.

[15] IRPINO, A., LECHEVALLIER, Y. and VERDE, R. (2006), Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A., Vichi, M. (eds.) *COMPSTAT 2006*. Physica-Verlag, Berlin, 869–876.

[16] IRPINO, A. and 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 Classication, IFCS 2006*, Springer, Berlin, 185–192.

[17] VERDE, R. and IRPINO, A. (2007), Dynamic Clustering of Histogram Data: Using the Right Metric.In: P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.): *Selected Contributions in Data Analysis and Classication*, Springer, Berlin, 123–134.

[18] IRPINO, A. and VERDE, R. (2008), Dynamic clustering of interval data using a Wasserstein-based distance.*Pattern Recognition Letters 29*, 1648–1658.

[19] MOORE, R.E. (1966), Interval Analysis. Prentice Hall, Englewood Cliffs, NJ.

NUOVE STATISTICHE PER NUOVI DATI: UNA PROPOSTA PER COMPARARE DATI NUMERICI A VALORI MULTIPLI

Nel Data Mining, spesso un insieme di individui è descritto attraverso opportune sintesi (media, scarto quadratico medio, istogrammi, stime di intervalli di confidenza, …) che generalizzano le descrizioni individuali ad una descrizione, più ampia, di una tipologia. Nel caso preso in esame, i dati possono essere descritti da più valori osservati di una stessa variabile. In questo lavoro, proponiamo il calcolo di statistiche per dati descritti da variabili numeriche a valori multipli (intervalli, istogrammi, più valori discreti). Tutte le variabili numeriche a valori multipli sono trattate come variabili numeriche modali. Al fine di ottenere nuove statistiche di base per misurare la variabilità e l'associazione tra tali tipi di variabili, consideriamo un'estensione della classica misura di inerzia usando, in luogo della distanza euclidea, la distanza L2 di Wasserstein definita tra misure di probabilità. Tale metrica è una generalizzazione della distanza di Wasserstein tra funzioni quantile di due distribuzioni.

Dimostriamo poi alcune proprietà di tale distanza e in particolare, la decomposizione dell'inerzia (teorema di Huygens). Proponiamo infine l'applicazione della distanza di Wasserstein e delle statistiche di base in un algoritmo di cluster di tipo k medie per partizionare un insieme di dati descritti da variabili numeriche modali e riferiti a un caso reale.