Order statistics for histogram data and a box plot visualization tool

Rosanna Verde*, Antonio Balzanella*, Antonio Irpino*

*Second University of Naples, Caserta, Italy
rosanna.verde@unina2.it, antonio.balzanella@unina2.it, antonio.irpino@unina2.it

Abstract. This paper deals with new descriptive statistics for histogram data, in the framework of symbolic data analysis. A main contribution consists in defining the main order statistics (median and quartiles) of a histogram variable using the quantile functions associated with the corresponding empirical distribution functions of the observed histograms. The definition of an order relationship between quantile functions is based on an appropriate probabilistic metric: the ℓ^p Wasserstein distance. Starting from the median and quartile functions definition, we extend the classic box-plot representation for set of quantile functions. Finally, we propose new measures of variability and skewness for a histogram variable associated with this representation. An application on real data allows us to corroborate the proposed measures and the new box-plot visualization tool.

1 Introduction

The advance of technology is making possible to observe and to collect very large datasets. The analysis of such data is often performed after a summarization step whose aims are to obtain a more manageable information, in size and in terms of computational resources, while preserving as much as possible the information of the entire data set. The representation of data through histograms is a common practice in data summarization. In fact, a histogram is parsimonious representation, with respect to storage requirements, and it provides an idea of the underlying distribution of the observed data or of subsets of values observed for a single attribute.

Symbolic Data Analysis (in short SDA) (Boch and Diday, 2000; Billard and Diday, 2006; Diday and Noirhomme-Fraiture, 2008) provides a formalization of a new symbolic descriptor, the *histogram variable* which is a particular case of symbolic multi-valued modal variable. Several techniques (Clustering, Regression, PCA,...) have been proposed in Billard and Diday (2003) to analyze histogram data. Some basic statistics like the *sample mean* and the *standard deviation* for a histogram variable have been introduced in Bertrand and Goupil (2000), Billard and Diday (2003), Billard and Diday (2006) and Irpino et al. (2006). Graphical tools for visualizing symbolic data (including histogram data) have been also presented by Noirhomme-Fraiture and Rouard (1997) and by Noirhomme-Fraiture and Nahimana (2008).

The aim of this paper is to introduce order statistics for histogram variables, especially the Median and the other quartiles. Generally, the definition of order statistic requires to establish an order relationship among data. In data analysis, the ordering definition problem is not a trivial issue. A meaningful example is provided in multivariate data analysis where it is not possible to define a natural ordering in R^d when $d > 1$. A well known approach is based on the concept of *data depth* (Tukey (1975), Liu et al. (1999), Zuo and Serfling (2000)), which is based on center outward ordering criterion. In this sense, a depth measure associates a high degree of *centrality* to an observation, with respect to a dataset, if it is close to the center of data cloud. Here, we assume histogram data as empirical distribution functions. Each histogram is uniquely associated with a cumulative distribution function (*cdf*), which is a piece-wise linear function, and with a quantile function (the inverse function of the *cdf*, which is still a piece-wise linear function). In this paper, we propose a definition of ordering for histogram data by means of their quantile functions. In order to find an ordering within a set of quantile functions associated with observed histogram data, a possibility can be the statistical depth introduced for functional data (Ramsay and Silverman (2005), López-Pintado and Romo (2009)). Following this approach, the median function is defined as the function having the highest depth. However, the functional data approaches do not guarantee that the central function (i.e., the function having the highest depth) is unique.
Our proposal consists in determining the Median histogram according to a natural ordering of the piece-wise quantile functions in sub-intervals of their domain. We select the segments of the quantile function having a central position in each sub-interval that does not intersect the other piece-wise functions. Therefore, the *median level-wise* quantile function does not present any intersection with other observed functions and it is completely inside the other quantile functions, differently from the median obtained using depth functions. The median piece-wise quantile function associated to the median histogram data must respect the proprieties of the median in descriptive statistics and so, it has to minimize the sum of ℓ_1 distances from all the other quantile functions. In order to compare quantile functions we make reference to a family of metrics (Rüshendorff (2001)), here denoted as ℓ^p Wasserstein distances. In particular, according to Arroyo (2008), Arroyo and Maté (2009), Arroyo et al. (2011), the median can be defined as the quantile function which minimizes the ℓ_1 Wasserstein distance.

Finally, after having defined the most common quantile functions-order statistics (1-st and 3-rd Quartiles, Median, Minimum and Maximum quantile functions) we propose a box-plot-like tool for quantile functions which is an extension of the classical visualization tool. Furthermore, in order to improve the description of a set of histogram data, we propose some variability (as the interquartile range IQR) and skewness measures of the quantile functions distribution.

The paper is organized as follows: In the section 2, we introduce the ℓ_p Wasserstein metrics to compare histogram data; in section 3, we give: a detailed description of the procedure for computing the order statistics for quantile functions, the algorithm scheme, and some computational evaluations; in section 4, we illustrate the procedure to construct the box-plot; in section 5, we present some skewness indexes. The section 6 ends the paper with an application on real data.

2 Histogram data

Let us consider a continuous variable **Y** with support $\mathbf{S} = [\underline{y}; \overline{y}]$ where $\underline{y}$ and $\overline{y}$ are the minimum and maximum observed values. The support **S** can be divided in a set of contiguous and non overlapping intervals (or bins). Given N observations of **Y**, a histogram H is a representation of **Y** consisting of a finite number of pairs $\{(I_k, f_k); k = 1, \ldots, K\}$ where $I_k = [\underline{y}_k, \overline{y}_k) \subseteq \mathbf{S}$ (with $\underline{y}_k \leq \overline{y}_k$) are the K bins of the histogram and f_k are the associated relative frequencies (that is, the number of observed values contained in I_k normalized by N).

A *Histogram Variable* **H** is a symbolic multi-valued variable whose realizations are histograms (Bertrand and Goupil (2000)). We indicate with $H_i = \{(I_{ik}, f_{ik}); k = 1, \ldots, K_i\}$ $(i = 1, \ldots, N)$ a set of N realizations of **H**. Since it is assumed that the values are uniformly distributed within each interval $I_{ik} = [\underline{y}_{ik}, \overline{y}_{ik})$, and having indicated with w_{ik} the cumulative frequencies f_{ik}, as:

$$w_{i0} = 0; \quad w_{ik} = \sum_{\ell=1}^{k} f_{i\ell} \qquad k = 1, \ldots, K_i, \tag{1}$$

it is easy to define the cumulative distribution function *cdf* $F_i(y)$ associated to H_i as follows:

$$F_i(y) = \begin{cases} 0 & \text{if } y < \underline{y}_{i1} \\ w_{ik} + \frac{y - \underline{y}_{ik}}{\overline{y}_{ik} - \underline{y}_{ik}} f_{ik}, & \text{if } \underline{y}_{ik} \leq y < \overline{y}_{ik} \quad (k = 1, \ldots, K_i - 1) \\ 1 & \text{if } y \geq \overline{y}_{iK_i}. \end{cases} \tag{2}$$

Reminding that the quantile function *(qf)* of a probability distribution is the inverse of its cumulative distribution function *(cdf)*, the *(qf)* associated to each H_i is:

$$F_i^{-1}(t) = \begin{cases} \underline{y}_{i1} & \text{if } t = 0 \\ \underline{y}_{ik} + \frac{t - w_{ik-1}}{w_{ik} - w_{ik-1}} (\overline{y}_{ik} - \underline{y}_{ik}) & \text{if } w_{ik-1} \leq t < w_{ik} \ (\text{for } k = 1, \ldots, K_i) \\ \overline{y}_{iK_i} & \text{if } t = 1 \end{cases} \tag{3}$$

Graphically, *(cdf)* and *(qf)* are piecewise linear functions as shown in Fig. 1.

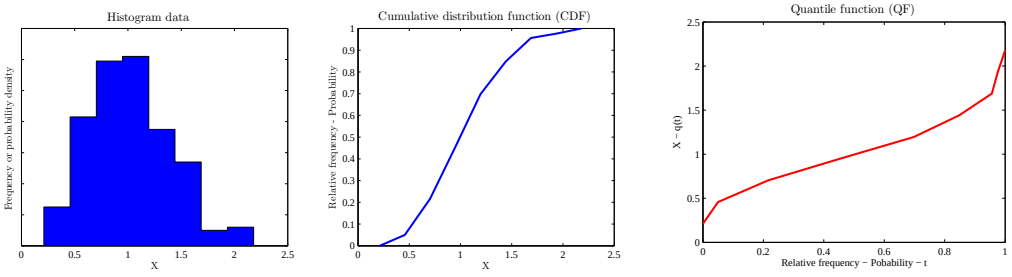

FIG. 1 – *From the left to the right: a histogram datum, its cumulative distribution function (*cdf*) and the corresponding quantile function (*qf*).*

2.1 Metrics for histogram data

An interesting challenge concerns the choice of the metric to compare histogram data. A possibility is to use one of the metrics for probability distributions. Furthermore, the dissimilarity between two histogram data can be computed considering their corresponding cumulated distribution functions.
Among the metrics proposed for matching probability distributions, we can mention: the f-divergence based measures, the discrepancy metric, the Kolmogorov (or Uniform metric), the Prokhorov-Lévi distance and the Wasserstein-Kantorovich-Monge-Gini distance (an overview is available in Gibbs and Su (2002)). In particular, we focus our attention on the latter family of metrics.

According to Rüshendorff (2001), the ℓ^p Wasserstein distance between two distribution functions is expressed by:

$$d_p^p(i,j) = \int_0^1 \left|F_j^{-1}(t) - F_j^{-1}(t)\right|^p dt \tag{4}$$

where $F_i(y)$ and $F_j(y)$ are the cumulative distribution functions (*cdf*s) associated to the $i-th$ and $j-th$ histograms (empirical density functions) and the $F_i^{-1}(t)$ and $F_j^{-1}(t)$ are the corresponding quantile functions (*qf*s). It is worth noting that the closed form distance depends on the possibility of expressing the quantile functions in closed form. According to the formula in equation (4), Irpino et al. (2006) introduced the following closed form of the squared Wasserstein distance (ℓ_2) between two histograms:

$$d_2^2(H_i,H_j) = \int_0^1 \left(F_i^{-1}(t) - F_j^{-1}(t)\right)^2 dt. \tag{5}$$

Being the H_i and H_j two observed histograms (empirical distribution functions), the corresponding distribution functions F_i and F_j, as well as the quantile functions F_i^{-1} and F_i^{-1}, are piece-wise linear functions with angular points in w_l (for $l = 1,\ldots,m$).
The w_l are the cumulated relative frequencies associated to the elementary intervals I_l; by considering the union of the two sequences of w_l related to the cumulate frequencies of the histograms $H_i = \{(I_{ik}, f_{ik}) \mid k = \ldots, K_i\}$ and $H_j = \{(I_{jk}, f_{jk}) \mid k = \ldots, K_j\}$:

$$\{w_{i0},\ldots,w_{iu},\ldots,w_{iK_i}\} \bigcup \{w_{i0},\ldots,w_{ju},\ldots,w_{jK_j}\}$$

Sorting the w_l values and erasing the equal values, we get the set of m distinct levels:

$$\{w_0,\ldots,w_l,\ldots,w_m\}$$

where: $w_0 = 0$, $w_m = 1$ and $\max(K_i,K_j) \leq m \leq (K_i + K_j - 1)$.

Expressing each bin I_k by its center $c_k = \frac{\overline{y}_k + \underline{y}_k}{2}$ and its radius $c_k = \frac{\overline{y}_k - \underline{y}_k}{2}$, and the intervals I_k in normal form: $I_k(t) = c_k + r_k(2t-1)$ for $0 \leq t \leq 1$, the distance between two

histograms expressed by the equation (5) can be written as:

$$d_2^2(H_i, H_j) := \sum_{k=1}^{m} f_k \left[(c_{ik} - c_{jk})^2 + \frac{1}{3} (r_{ik} - r_{jk})^2 \right]. \quad (6)$$

Using this distance proposed in Irpino et al. (2006) and Verde and Irpino (2008), it is proved that the *Average histogram* of a set of histogram data is the histogram $\bar{H}$ having in its description the bins $I_k = [\bar{c}_k - \bar{r}_k; \bar{c}_k + \bar{r}_k]$ and the relative frequencies f_k, with $\bar{c}_k$ and $\bar{r}_k$ the means of the centers c_{ik} and radius r_{ik} of I_k (for $k = 1, \ldots, m$ and $i = 1, \ldots, N$). In fact:

$$\begin{aligned} min_{\overline{H}} f(\overline{H}|H_1, \ldots, H_N) &= min_{\overline{H}} \sum_{i=1}^{N} d_2^2(H_i, \overline{H}) = \quad (7) \\ &= min_{\bar{c}_k, \bar{r}_k} \sum_{i=1}^{N} \sum_{k=1}^{m} f_k \left[(c_{ik} - \bar{c}_k)^2 + \frac{1}{3} (r_{ik} - \bar{r}_k)^2 \right]. \end{aligned}$$

In particular, (7) is solved by $\overline{H}$ with:

$$\bar{c}_k = N^{-1} \sum_{i=1}^{N} f_k c_{ik} \quad ; \quad \bar{r}_k = N^{-1} \sum_{i=1}^{N} f_k r_{ik}.$$

An interesting decomposition of the ℓ_2 Wasserstein distance between two continuous quantile functions associated to random variables Y_i and Y_j was proposed by Irpino and Romano (2007):

$$d_2^2(H_i, H_j) = \left(\overline{y}_i - \overline{y}_j\right)^2 + (s_i - s_j)^2 + 2 s_i s_j (1 - \rho(i, j)) \quad (8)$$

where considering that the mean values of the distributions can be obtained using the quantile function Gilchrist (2000) as follows:

$$\overline{y}_i = \int_{-\infty}^{+\infty} x \, dF(x) = \int_0^1 F_i^{-1}(t) dt, \quad (9)$$

while the standard deviation is similarly obtained as

$$s_i = \sqrt{\int_{-\infty}^{+\infty} x^2 \, dF(x) - [\overline{y}_i]^2} = \sqrt{\int_0^1 [F_i^{-1}(t)]^2 dt - [\overline{y}_i]^2} \quad (10)$$

$\overline{y}_i$ and $\overline{y}_j$,and s_i and s_j are respectively the means and the standard deviations of the two distributions is the correlation coefficient between the two quantile functions, defined as:

$$\rho(i, j) = \frac{\int_0^1 y_i(t) y_j(t) dt - \overline{y}_i \cdot \overline{y}_j}{s_i \cdot s_j}. \quad (11)$$

When the *qfs* are piece-wise linear functions associated to the histogram data H_i and H_j then:

$$\rho(i,j) = \frac{\sum_{k=1}^{m} f_k \left[c_{ik} \cdot c_{jk} + \frac{1}{3} r_{ik} \cdot r_{jk}\right] - \overline{y}_i \cdot \overline{y}_j}{s_i \cdot s_j}. \tag{12}$$

3 The Median quantile function and other order statistics

At first we find the *Median qf* and the *Median histogram* for a set of histogram data $\{H_i\}_{i=1,\ldots,N}$. Taking into account the proprieties of the median, in descriptive statistics, and according to Arroyo (2008), Arroyo and Maté (2009), Arroyo et al. (2011), the *Median histogram* can be defined as the histogram H_{ME} which minimizes the ℓ_1 Wasserstein distance in (4):

$$\min_{H_{ME}} \sum_{i=1}^{N} d_1(H_i, H_{ME}) = \min_{F^{-1}(t)} \sum_{i=1}^{N} \int_0^1 \left|F_i^{-1}(t) - F_{ME}^{-1}(t)\right| dt, \tag{13}$$

where F_i^{-1} and F_{ME}^{-1} are the *qfs* associated to H_i and H_{ME} respectively.
It is noteworthy that H_{ME} is the center histogram of the set of histogram data H_i according to the ℓ_1 Wasserstein distance so as the Average histogram is the center according to the ℓ_2 distance as shown by Verde and Irpino (2008), Irpino et al. (2006).

According to the nature of the data and the minimization problem in eq. (13), the *level-wise median function* $ME(t)$ is a quantile function that, for each $t \in [0;1]$, takes the middle value of the sequence of ordered values $F_i(t)$ (for $i = 1, \ldots, N$); in such a way, $ME(t)$ is the quantile function that leaves $N/2$ quantiles at level t before and $N/2$ after its taken values. Like in the classic case, Arroyo et al. (2011) highlight that the definition of Median histogram, if the number of histograms is even, is not unique. However, similarly then the classic way, the median is taken as the value that is half-way between the two central values at position $N/2$ and $N/2+1$.

Thus, we have a level-wise order (for a given t we may order the *qfs*) but not a full order or semi-order relation among *qfs*. Naturally, if for each $t \in [0;1]$ the order of the *qfs* is always the same, we can extend the level-wise order to a full order relation (in all the support interval $[0,1]$ but generally the *qfs* tend to intersect each other like shown in Figure 2.

For our scope, we refer to each histogram $H_i,\ i = 1, \ldots N$ through the set of couples $\{(I_{ik}, f_{ik})\ k = 1, \ldots, K_i\}$ with $w_{ik} = \sum_{l=1}^{k} f_{il},\ k = 1, \ldots, K_i$ the cumulative relative frequencies or *levels*. Our strategy to find a *piece-wise level median quantile function* consists to reduce the number of points t in $[0,1]$ to the level set $\mathbf{w}$ of common w_k levels to all the the *qfs* associated to the histograms H_i, and to look for the median values $ME(w_k)$ only in correspondence of these levels w_k (for $k = 1, \ldots, K$). For that, a first step (here called *homogenization step*) to find the minimum set of w_k cumulative frequency levels (as union of all the sets $\mathbf{w}_i$ for $i =, \ldots, N$), is needed.

Hereafter are described the two main steps to determine the *piece-wise level Median quantile function*:

- **Homogenization step.** It detects the minimum set of level values w_{ik} (for $i = 1, \ldots, N$ and $k = 1, \ldots, K_i$, allowing to define a set of elementary intervals $[w_{k-1}, w_k]$ of levels

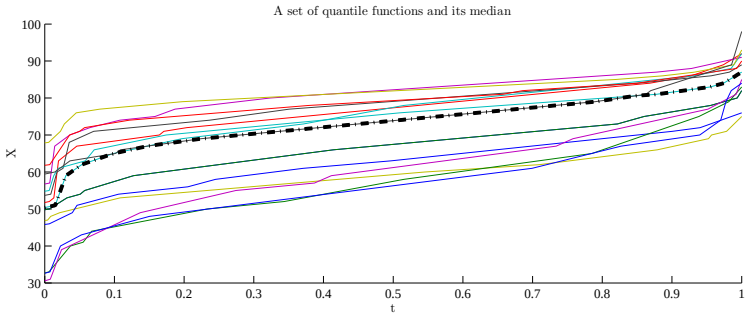

FIG. 2 – *The black dotted curve represents the Median-qf obtained with the proposed method.*

that do not contain angular points. Thus,

$$\{w_{10}, \ldots, w_{1K_1}, \ldots, w_{i1}, \ldots, w_{iK_i}, \ldots, w_{N0}, \ldots, w_{NK_N}\} \tag{14}$$

is the set of the cumulated relative frequencies associated to all histograms $H_i,\ i = 1, \ldots, N$. After sorting the elements of **w** and eliminating the replicated values, the set **w** is now:

$$\{w_0, \ldots, w_k, \ldots, w_K\}, \tag{15}$$

where $w_0 = 0,\ w_K = 1$ and, denoting with $\overline{K} = N^{-1} \sum_{i=1}^{N} K_i$, the value of K varies in

$$\max_{1 \le i \le N} K_i \le K \le (N(\overline{K} - 1) + 1).$$

For each $w_k,\ k = 0, \ldots, K$ the values of $F_i^{-1}(w_k) = y_k,\ i = 1, \ldots, N$ are known or they can be easily computed by a linear interpolation (because the qfs are piece-wise linear functions). So, each $H_i,\ i = 1, \ldots N$ is constituted by a new set of K couples $\{(I_{ik}^*, f_{ik}^*);\ k = 1, \ldots, K\}$, where: $I_{ik}^* = [y_{k-1}, y_k]$ and $f_{ik}^* = w_k - w_{k-1}$.

– **Median level piece-wise selection step.** This step is repeated for each k elementary interval of cumulative frequency levels. Each elementary interval of levels $[w_{k-1}; w_k]$ contains N segments $F_i^{-1}(t)$ (with $w_{k-1} \le t < w_k$) associated to the N qfs $F_i^{-1}(t)$. Let $F_{(i)}^{-1}(t)$ (with $w_{k-1} \le t < w_k$) be the $i-th$ piece-quantile function, with (i) its order with respect to all the others pieces qfs $F_{(j)}^{-1}(t)$ (with $w_{k-1} \le t < w_k$). The order (i) of $F_{(i)}^{-1}(t)$ is kept in all the level interval (w_{k-1}, w_k) only if there are not intersections between pieces quantile functions $F_{(i)}^{-1}(t)$ and $F_{(j)}^{-1}(t)$ (with $w_{k-1} \le t < w_k$) in the interval (as simply shown in figure 3), that changes the order of the sub-pieces of the quantile functions. It requires to check in each level interval (w_{k-1}, w_k) the intersection points and then, to perform a further splitting of the interval in sub-intervals of cumulate frequency levels. The set of w_k's will be increased and the final **w** set is updated by the new levels, so the selection of the median pieces quantile $F_{\left(\frac{N}{2}\right)}^{-1}(t)$ (with $w_{l-1} \le t < w_l$) is performed on an updated number of levels w_k with $k = 1, \ldots, m$ (with $m \ge K$). If

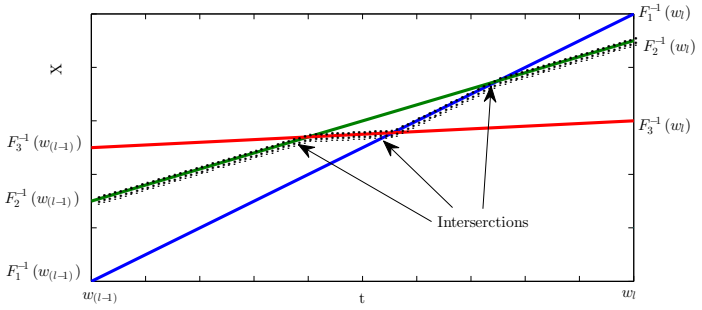

FIG. 3 – *Selection of the pieces quantile level of the Median-qf (the dotted path) in the elementary interval of levels* $[w_{k-1}, w_k]$ *with* $N = 3$.

N is even, we consider the value that is half-way between two middle quantile functions values, like in the classic case.
The Median qf may correspond to an observed quantile function $F^{-1}_{(\frac{N}{2})}(t)$ or it is obtained by the selected $F^{-1}_{(\frac{N}{2})}(t)$ (with $w_{l-1} \leq t < w_l$) segments of position $(\frac{N}{2})$ in each interval quantile level (w_{l-1}, w_l). In order to distinguish the Median-qf to an observed qf we denote it $ME(t)$. Obviously, the Median histogram is a histogram associated to the Median-qf of the set of N qfs.

Using the same algorithm it is possible to compute the generic $p \cdot N$ ($p \in [0;1]$) order qf statistic, because they may be not observed qfs they are denoted $Q_{(pN)}(t)$. The proposed algorithm for order quantile functions searching, guarantees a unique correspondence between the histograms and the qfs, so the First Quartile-histogram H_{Q_1} is associated with the qf $Q_{(0.25 \cdot N)}(t)$ (or $Q_1(t)$), the Third Quartile-histogram H_{Q_3} with the quantile qf $Q_{(0.75 \cdot N)}(t)$ (or $Q_3(t)$) as well as the Median histogram H_{ME} with the quantile qf $ME(t)$.

Computational cost of building a order statistics for a set of qfs >From a computational point of view we evaluate the time complexity. The selection step is performed K times. The maximum number of intersection between the N segments to be evaluated in $O(N^2)$. Thus, taking into account the number of bins and the number of potential intersections, in the worst case, the computational cost of the whole processing is of order:

$$O\left([N(\overline{K}-1)+1]N^2\right) = O(\overline{K}N^3).$$

4 Box and Whiskers plot for qfs

A new representation tool, similar to the *box-plot*, is here introduced for showing the histogram data set distribution. The *box* is defined as the region bounded by the piecewise quantile functions $Q_1(t)$ and $Q_3(t)$.

For the choice of the whiskers we can consider different criteria for selecting the piecewise qfs corresponding to the Q_{Low} lower and Q_{Upp} upper bounds. We take into consideration three possible ways.

Min-Max qf The lower and the upper bounds are chosen as the first $Q_{(0)}(t)$ and the last $Q_{(N)}(t)$ qfs. However, this solution presents, as disadvantage, the possibility to include extreme or *outlying qfs*.

90% **most central quantiles** A second way consists in choosing $Q_{(0.05\cdot N)}(t)$ and the $Q_{(0.95\cdot N)}(t)$ bounded qfs. This solution is less sensible to *outlying* qfs.

1.5 **times the Inter Quartile Range** this third way needs to define, firstly, an extension of the Inter Quartile Range (IQR) measure. In this analysis context, we define the *Inter Quartiles Range* (or *IQR*) as the area between the qfs (similarly the cdfs) associated with the first $Q_1(t)$ and the third $Q_3(t)$ computed through the ℓ_1 Wasserstein distance:

$$IQR = d_1(H_{Q_1}, H_{Q_3}) = \int_0^1 |Q_3(t) - Q_1(t)|dt = \overline{Q_3} - \overline{Q_1}, \tag{16}$$

where $\overline{Q_3}$ and $\overline{Q_1}$ are the means of the densities described by the histograms H_{Q_1} and H_{Q_1}, computed according to Eq. (9).
The choice of ℓ_1 distance seems to be consistent with the metric used for defining the order statistics like the Median and the Quartile functions. Considering that $Q_3(t) \geq Q_1(t)\ \forall t \in [0,1]$, IQR can be interpreted as the difference between the mean values of the H_{Q_1} and the H_{Q_3} distributions [1]. Therefore, a third way for defining the upper and the lower bounds of the whiskers consists in translating $Q_3(t)$ and $Q_1(t)$ of 1.5 times the Inter Quartile Range IQR. Thus, H_{Low} is the histogram associated with $Q_1(t) - 1.5IQR$, while H_{Upp} is the histogram associated with $Q_3(t) + 1.5IQR$. In this case, the qfs-bounds have the same shape of Q_1 and Q_3. However that can represent a limit in the interpretation of the final results, since it does not take into consideration eventual different shapes of the extreme qfs.

4.1 Variability and shape measures

The *qfs box-plot* shown in Fig. 4 generalizes the classic box and whisker plot to quantile functions. It is composed by the qfs related to the Median $ME(t)$ (in the center), the $Q_1(t)$, the Third Quartile $Q_3(t)$, for the box, and by the lower $Q_{Low}(t)$ and upper $Q_{Upp}(t)$ that delimit the whiskers.

The interpretative features of a classic box-plot can be generalized to the box-plot of qf*s*. However, the complex nature of the histogram-valued data, even observed through their qf*s*, does not allow a direct generalization. We here present some variability and shape-related

1. Because $Q_3(t) \geq Q_1(t)$ for each $t \in [0,1]$, and thus, $Q_3(t) - Q_1(t) \geq 0$, we obtain that

$$IQR = \int_0^1 |Q_3(t) - Q_1(t)|dt = \int_0^1 (Q_3(t) - Q_1(t))\, dt = \int_0^1 Q_3(t)dt - \int_0^1 Q_1(t)dt = \overline{Q_3} - \overline{Q_1}.$$

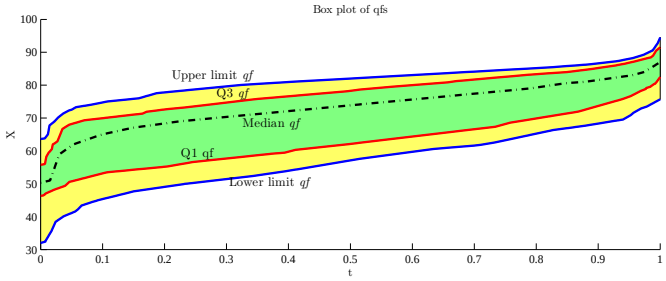

FIG. 4 – *The quantile function box-plot consisting in five qfs: the median, the First and the Third quartile qfs delimiting the box, an Upper and a Lower bound qf are the extremes of the whiskers.*

measures, like the skewness, that must be considered a first tentative to extend some classic measures for point-valued valued to histogram-valued ones.
As measures of variability we propose the IQR introduced in (16). Alternatively, we can define a further interquartile range IQR_2 based on the ℓ_2 Wasserstein distance as follows:

$$\begin{aligned} IQR_2 &= d_2^2(H_{Q_1}, H_{Q_3}) = \int_0^1 \left(Q3(t) - Q1(t)\right)^2 dt = \\ &= \underbrace{\left(\overline{Q_3} - \overline{Q_1}\right)^2}_{IQR^2} + \underbrace{\left(s_{Q_3} - s_{Q_1}\right)^2 + 2 s_{Q_3} s_{Q1} \left(1 - \rho(Q_3, Q_1)\right)}_{\Delta IQR_2}. \end{aligned} \tag{17}$$

As shown in the eq. (8), the ℓ_2 Wasserstein distance can be decomposed in the two components related to the location and the variability respectively. Thus, the (17) takes into account the location and the variability of the set of the *qfs* respectively.
Moreover, the distance of the Median *qf* with respect to the First and the Third Quartile *qfs* can indicate a global degree of skewness of the set of the *qfs*, even if it takes into account only a partial information (around the 50%) of the distribution of the *qfs* around the Median *qf*.

We propose the following skewness indices:

$$A_1 = \frac{d_1(H_{Q_3}, H_{ME})}{d_1(H_{Q1}, H_{ME})} = \frac{\overline{Q_3} - \overline{ME}}{\overline{ME} - \overline{Q1}}; \tag{18}$$

$A_1 > 0$ being the averages of the Third Quartile, Median and First Quartile *qfs* in following relations: $\overline{Q_3} \geq \overline{ME}$ and $\overline{ME} \geq \overline{Q1}$.

$A_1 = 0$ when $\overline{Q_3} = \overline{ME}$ and $\overline{ME} \neq \overline{Q1}$. If the averages of $Q_3(t)$ and $ME(t)$ *qfs* are the same, then the 25% of the *qfs* upper the half are coincident with the ME(t); that means the IQR is equal to the difference between the $ME(t)$ and the $Q_1(t)$ *qfs*. Then, the qfs distribution presents a negative asymmetry. In the case, that also $\overline{ME} = \overline{Q1}$, the A_1 is indeterminate but it happens only if there is not an interquartile variability of the *qfs*.

$A_1 \to +\infty$ if $(\overline{ME} - \overline{Q1}) \gg (\overline{Q_3} - \overline{ME})$; that can be interpreted as a strong positive asymmetry of the qfs distribution.

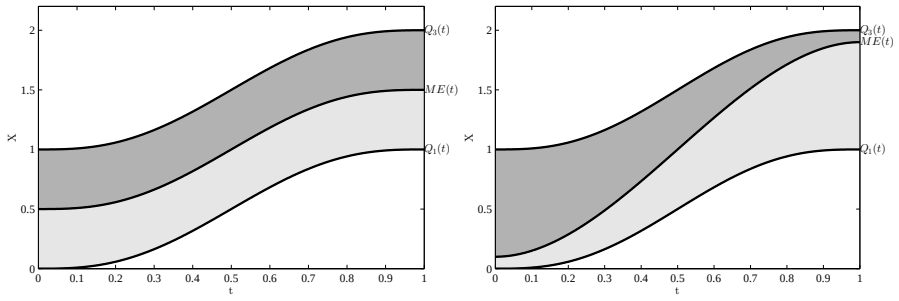

FIG. 5 – *The averages of the $Q_1(t)$, $ME(t)$ and $Q_3(t)$ quantile functions are respectively: $\overline{Q_1} = 0.5$, $\overline{ME} = 1$, $\overline{Q_3} = 1.5$. The value of $A_2 = 0$ in both the configurations, that corresponds to a symmetry of the distributions of the qfs whereas the shape of the areas between $Q_3(t)$ and $ME(t)$, and between $ME(t)$ and $Q_3(t)$ are not the same.*

$A_1 = 1$ if $(\overline{Q_3} - \overline{ME}) = (\overline{ME} - \overline{Q_1})$. That can correspond to a symmetry of the *qfs* distribution of the quartiles $Q_3(t)$ and $Q_1(t)$ *qfs* with respect to $ME(t)$ *qf* based on the differences between the averages of $Q_3(t)$, $ME(t)$ and $Q_1(t)$.

A limit of this index is that it is expressed as a ratio between two distances and to assume an infinite value when the measure at denominator is equal to 0. Therefore, an alternative formulation of A_1 can be provided by a log transformation, as follows:

$$LA_1 = log(d_1(H_{Q_3}, H_{ME})) - log(d_1(H_{Q1}, H_{ME})) = log(\overline{Q_3} - \overline{ME}) - log(\overline{ME} - \overline{Q1}); \tag{19}$$

LA_1 takes values in all real number domain $\Re$; in particular, it assumes value 0 when there is a symmetry of the quartiles $Q_3(t)$ and $Q_1(t)$ *qf* distributions with respect to $ME(t)$ *qf*.

Let us define a second skewness index, expressed by:

$$A_2 = d_1(H_{Q_3}, H_{ME}) - d_1(H_{Q_1}, H_{ME}) = \overline{Q_1} + \overline{Q_3} - 2\overline{ME}. \tag{20}$$

We can interpret the asymmetry of the *qfs* distribution, according to the values that the A_2 index takes:

$A_2 < 0$ if $(\overline{Q_3} - \overline{ME}) < (\overline{ME} - \overline{Q_1})$, that corresponds to a negative asymmetry

$A_2 > 0$ if $(\overline{Q_3} - \overline{ME}) > (\overline{ME} - \overline{Q_1})$, that corresponds to a positive asymmetry

$A_2 = 0$ if $(\overline{Q_3} - \overline{ME}) = (\overline{ME} - \overline{Q_1})$, that corresponds to symmetric distribution of the *qfs*, but always based on the difference between the averages of the Quartiles functions ($(\overline{Q_3} - \overline{ME})$ and $(\overline{ME} - \overline{Q_1})$). We notice that $A_2 = 0$ does not take into account eventual different shapes of the $Q_3(t)$, $ME(t)$ and $Q_1(t)$ *qfs* (see Fig. 5).

This can be instead considered by using the $A_2(\ell)$ index (hereafter detailed) which allows to evaluate the differences between the average values of the Quartile functions ($(\overline{Q_3(\ell)} - \overline{ME(\ell)})$ and $(\overline{ME(\ell)} - \overline{Q_1(\ell)})$) for every frequency level of the *qfs*.

Therefore, an extension of the Bowley skewness index (Kenney and Keeping, 1962), also known as quartile skewness coefficient) is:

$$A_3 = \frac{d_1(H_{Q_3}, H_{ME}) - d_1(H_{Q1}, H_{ME})}{d_1(H_{Q_3}, H_{Q1})} = \frac{A_2}{IQR}, \tag{21}$$

it is the normalized A_2 index. A_3 takes values in $[-1, 1]$. Because $IQR \geq (d_1(H_{Q_3}, H_{ME}) - d_1(H_{Q1}, H_{ME})$ then:

$A_3 = -1$ if $(\overline{Q_3} = \overline{ME})$, that corresponds to a negative asymmetry

$A_3 = 1$ if $(\overline{Q_1} = \overline{ME})$; that corresponds to a positive asymmetry

$A_3 = 0$ if $d_1(H_{Q_3}, H_{ME}) \cong d_1(H_{ME}, H_{Q_1})$, the distribution of *qfs* can be considered symmetrical in mean. In fact, as shown in Fig. 5, we obtain $d_1(H_{Q_3}, H_{ME}) \cong d_1(H_{ME}, H_{Q_1})$ also when the Median *qfs* is not always at the center between the two Quartile *qfs* .

The A_1 and A_2 indices can be extended to a wider domain of the distributions considering the quantile functions at 5% and 95% of the distributions (excluding extreme quantile functions which can be outliers) rather than the Q_1 and Q_3. The index A_3 extended to a $5th$ and $95th$ quantile requires a different normalization than the previous one. In such way we propose:

$$A_3' = \frac{d_1(H_{0.95N}, H_{ME}) - d_1(H_{0.05N}, H_{ME})}{N^{-1} \sum_{i=1}^{N} d_1(H_i, H_{ME})} \tag{22}$$

where the normalizing term $N^{-1} \sum_{i=1}^{N} d_1(H_i, H_{ME})$ is a sort of *simple median deviation* measure computed according to the Wasserstein ℓ_1 distance.

In order to take into account the skewness of the set of the qfs in correspondence of the different level intervals $[w_{l-1}, w_l]$, we propose to express A_2 as a level-wise function assuming constant values in $[w_{l-1}, w_l]$ as follows:

$$\begin{aligned} A_2(l) &= \int_{w_{l-1}}^{w_l} |Q_3(t) - ME(t)|dt - \int_{w_{l-1}}^{w_l} |Q1(t) - ME^{-1}(t)|dt = \\ &= \int_{w_{l-1}}^{w_l} [Q1(t) + Q_3(t) - 2 \cdot ME(t)]\, dt \end{aligned} \tag{23}$$

for $w_{l-1} \leq t \leq w_l$ with $l = 1, \ldots L'$ where the number L' is computed during the homogenization step for the Q_1, ME, Q_3 and $max\{K_{Q1}, K_{ME}, K_{Q_3}\} \leq L' \leq (K_{Q1} + K_{ME} + K_{Q_3} - 2)$ with K_{Q1}, K_{ME}, K_{Q_3} the number of quantile levels of the First quartile, Median and Third quartile functions. Obviously, the index A_2 can be retrieved by the sum of the $A_2(l), l = 1, \ldots, L'$, that is $A_2 = \sum_{l=1}^{L} A_2(l)$ [2]. It provides an information about the skewness of the distribution of qfs around the Median qf for each frequency level interval. In

2. $\sum_{l=1}^{L} A_2(l) = \sum_{l=1}^{L} \left\{ \int_{w_{l-1}}^{w_l} [Q1(t) + Q_3(t) - 2 \cdot ME(t)]\, dt \right\} = \sum_{l=1}^{L} \int_{w_{l-1}}^{w_l} Q_1(t)dt + \sum_{l=1}^{L} \int_{w_{l-1}}^{w_l} Q_3(t)dt - 2\sum_{l=1}^{L} \int_{w_{l-1}}^{w_l} ME(t) = \int_0^1 Q_1(t)dt + \int_0^1 Q3(t)dt - 2\int_0^1 ME(t)dt = \overline{Q_1} + \overline{Q_3} - 2 \cdot \overline{ME} = A_2$

fact, $A_2(\ell) = 0$ means a symmetry of the qfs (50%) distribution in the interval w_{l-1}, w_l; while $A_2(\ell) < 0$ ($A_2(\ell) > 0$) means an higher concentration of qfs pieces distribution between the $Q_1(l)$ and the $ME(l)$ than between the $ME(l)$ and $Q_3(l)$, for $w_{l-1} \leq t \leq w_l$. Some examples of $A_2(l)$ functions associated to four groups of four histogram variables are shown in Fig. 9.

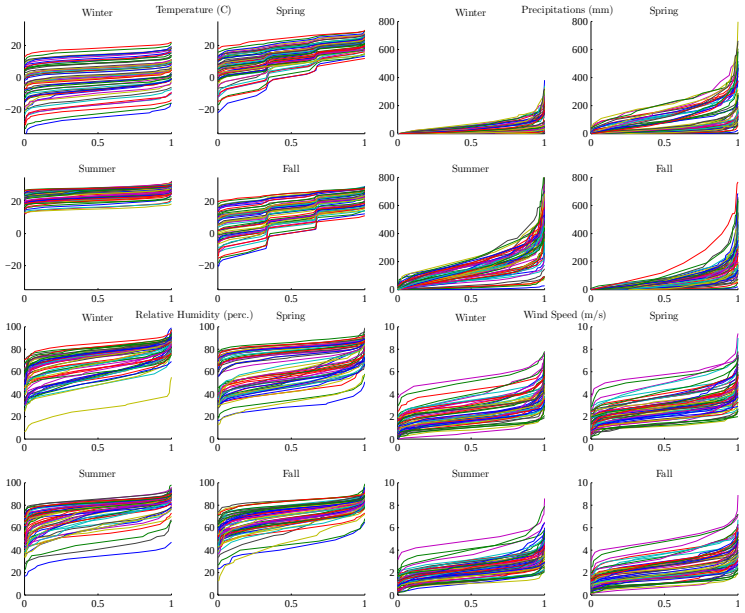

FIG. 6 – *The quantile functions of the histogram data in the China dataset.*

Similarly the indices IQR in (16) and (17) may be expressed as a piece-wise function $IQR(l)$, $l \in \{1, \ldots, L'\}$ for a fixed L', as follows:

$$IQR(l) = \int_{w_{l-1}}^{w_l} |Q_3(t) - Q_1(t)| dt, \quad IQR_2(l) = \int_{w_{l-1}}^{w_l} (Q_3(t) - Q_1(t))^2 \, dt. \tag{24}$$

The previous functions are useful for obtaining information about variability of the set of qfs in each level interval $[w_{l-1}, w_l]$.
Conversely, the A_1 and A_3 are assumed only as global indices: the A_1 being expressed as a ratio cannot be summarized by the sum of $A_1(l)$ values at each level l; the $A_3(l)$ $l \in \{0, \ldots, L\}$ is a normalized $A_2(l)$ index, the denominator IQR scales the $A_2(l)$ in smaller values which are difficult to interpret with respect to the values -1 and 1 that the global A_3 index can assume in case of negative and positive asymmetry.

5 An example

To corroborate the proposed order basic statistics, the variability and the skewness measures as well as the box-plot tools, we analyze a dataset of histograms obtained summarizing

climatic information recorded by 60 meteorological Chinese stations. The histogram dataset is

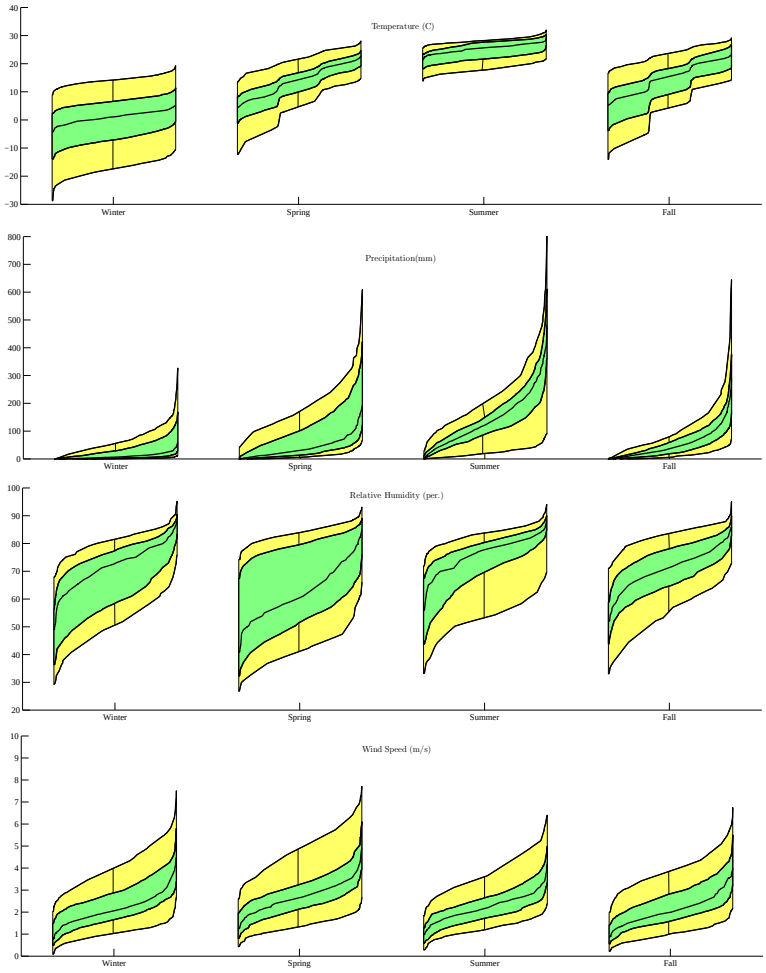

FIG. 7 – *The Box and Whisker plot for quantile functions of the China dataset.*

constructed starting from a public available repository of climatic data [3] containing values of: Mean monthly Temperatures, Precipitations, Relative Humidity and Wind Speed for 60 stations in China from 1930 to 1988.
In order to deal with histogram data, we have considered the Mean monthly measurements for each station collected for each meteorological season (Winter, Spring, Summer and Fall), carrying out a dataset of 60 stations described by 4×4 (variables $\times$ seasons) histogram variables. The quantile functions associated with each histogram are shown in Figure 6.
We computed the following piece-wise qfs: $Q_{0.05}(t)$, $Q_1(t)$, $ME(t)$, $Q_3(t)$ and $Q_{0.95}(t)$ for each histogram variable. The relative box-plots drawn by using these qfs are shown in Figure

3. Dataset URL: `http://dss.ucar.edu/datasets/ds578.5/`

7. With each qfs, it is possible to associate its density function. However, the box-plot of qfs offers a more simple interpretation of the characteristics of the observed variable, while, the interpretation using the density functions is complicated by an absence of a natural ordering for histograms. In Fig. 8, we present a direct comparison between the box-plots for the variable temperature and a representation of the densities that are associated with the qfs of the main elements of the box-plot. For adding a bit of readability, we have not represented on the same plane, but in a sliced way.

It is worth of noting (Fig. 6) that the qfs of the *Temperatures in the Summer* show almost

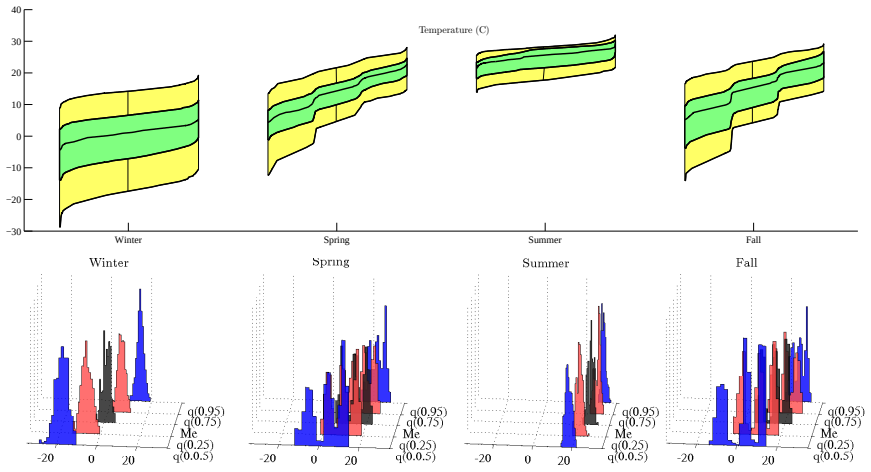

FIG. 8 – *The Box and Whisker plot for quantile functions of Temperature and the corresponding histograms. Please note that, even if the histograms have the same support, they were represented in a sliced way in order to improve the readability of the graph.*

constant values in the coldest and in the warmest days (the corresponding qf curves are quite parallel to horizontal axis) and a few variability in the range of values around 15-25 °C. Similarly, the shape of the qfs of the *Temperatures in the Winter* observed in the 60 meteo-stations, in particular the warmest regions, do not point out relevant variations of the temperatures between the coldest and the warmest days (from 15°C to 20 °C) while we note that the coldest regions have a stronger change of the temperatures from the coldest and the less cold days (from -30°C to 15 °C). However, comparing the distribution of the 60 qfs of the *Temperatures in the Winter* with the qfs of the *Temperatures in the Summer* it presents a higher variability as shown by the wideness of the bands between the five piecewise qfs ($Q_{0.05}(t)$, $Q_1(t)$, $ME(t)$, $Q_3(t)$ and $Q_{0.95}(t)$).

The shape of the qfs of the *Temperatures in the Spring* and in the *Fall* is different. They are "step functions" with quite constant trends in correspondence of each of the three months. For reason of brevity, among the other nine (3×3) variables, we comment only some of them: the *Precipitations*, for the specific characteristics, presents a strong variability from the lowest to the highest values of precipitations in the same regions (that is typical of precipitation series). The qfs distribution of the *Relative Humidity* and *Wind speed*, in the different seasons, have intermediate configurations with a more considerable variability of the ones related to the *Rel-*

ative Humidity values especially in the *Spring* and in the *Summer*, while a qf "outlier curve" is learnt in the Winter (the lowest in the graphic representation).

TAB. 1 – *Variability and skewness indices.*

Variable	Season	IQR	IQR2	$\frac{\Delta IQR2}{IQR_2}$	A_1	A_2	A_3
Temperature	Winter	13.6	186.9	0.004	0.707	-2.3	-0.172
(C)	Spring	7.0	50.6	0.045	0.634	-1.6	-0.224
	Summer	5.6	30.9	0.003	0.465	-2.0	-0.366
	Fall	9.8	100.4	0.042	0.524	-3.1	-0.312
Precipitations	Winter	35.1	1939.8	0.364	4.626	22.6	0.644
(mm)	Spring	100.3	13630.6	0.263	3.866	59.1	0.589
	Summer	77.7	8095.5	0.255	0.761	-10.6	-0.136
	Fall	46.0	3253.4	0.350	0.944	-1.3	-0.029
Relative	Winter	18.2	346.8	0.044	0.370	-8.4	-0.459
Humidity	Spring	28.0	804.9	0.025	1.621	6.6	0.237
(perc.)	Summer	12.0	165.6	0.137	0.444	-4.6	-0.385
	Fall	12.6	161.4	0.023	0.983	-0.1	-0.009
Wind speed	Winter	1.30	1.7	0.024	1.887	0.400	0.307
(m/s)	Spring	1.21	1.5	0.019	1.175	0.097	0.081
	Summer	1.00	1.0	0.026	0.991	-0.004	-0.004
	Fall	1.23	1.6	0.041	1.694	0.318	0.258

Looking at the box-plots in Fig. 7, it is interesting to comment the "skewness" of the qfs distribution of the *Precipitations in the Spring*: the half of the qfs, under the Median qf, show a small variability of lower values of the precipitations, proper of this season in the less rainy regions, while the most rainy regions present a higher variability of the milliliters of precipitation values during the season.

The Tab. 1 shows the interquartile index (IQR, IQR2) values in the first two columns; the third column contains the relative importance of the IQR2 due to the variability of the distributions (histogram data) (according to the ℓ_2 Wasserstein distance decomposition as showed above) of the several variables in the different seasons recorded in the 60 meteo-stations. It is easy to note that the stronger effect of the variability of the distributions (histogram data) is observed for the *Precipitations*, especially in the coldest seasons (*Fall* and *Winter*). This result is consistent with the considerations expressed about the box-plots of the *Precipitations in the Fall* and in the *Winter*. Moreover, we observe the highest positive values of the A_2 skewness index and of the normalized A_3 one (in the last columns of the table) for the qf distributions of the *Precipitations in the Winter* and *Spring* compared with the all other ones. That corresponds to the highest variability of the distributions of milliliters of *Precipitations* in the most rainy regions with respect to the less rainy areas like in the *Winter* and in the *Spring*.

Figure 9 shows, for each variable the different $A_2(\ell)$ skewness normalized functions associated with each season. The last discussed results about the skewness of the precipitations distributions in Winter and in Spring, expressed by the high values of A_1, A_2 and A_3 indexes (positive for A_2 and A_3) can be better read on the graphics of the $A_2(\ell)$ curves that highlights the growing of the $A_2(\ell)$ step-wise functions of the *Precipitations in the Winter* and in the

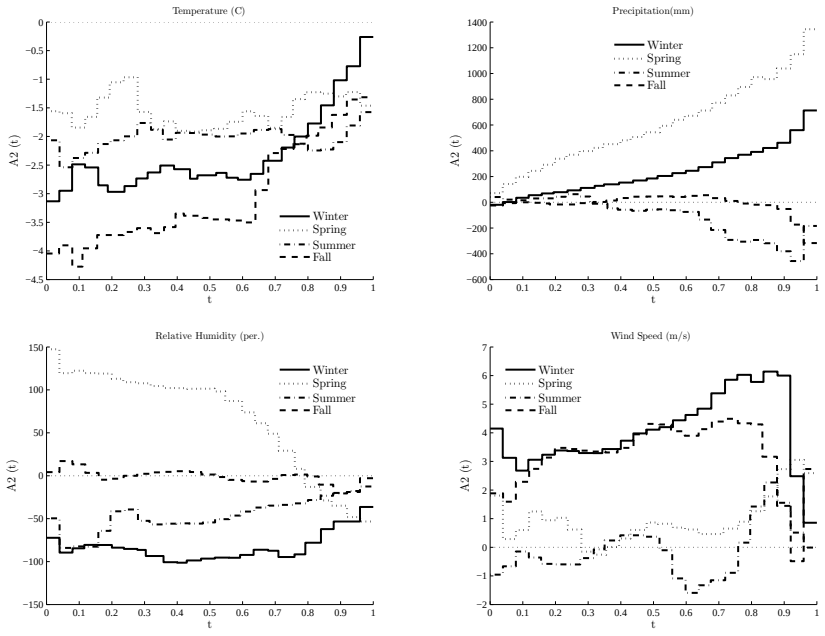

FIG. 9 – *The skew functions $A_2(\ell)$ for the China dataset.*

Spring. In opposite the $A_2(\ell)$ step-wise function of the *Relative Humidity in Winter* presents always negative values for each quantile level, that is synthesized by the negative values of A_2 index, as well as by the normalized A_3 index.

6 Conclusions

In this paper we have proposed order statistics for histogram variables based on the representation of the histograms (realization of the histogram variable) through their corresponding quantile functions. In particular, with the aim of defining an ordering between quantile function values we have used the "ℓ_1" norm Wasserstein distance. Starting from the main order statistics: the median, the first and third quartiles and the upper and lower quartile function, we have also presented a new graphical representation, like box plot, which permits to visualize the characteristics of the distribution of a quantile functions set. Due to the particular nature of the histogram data, realizations of a histogram variable, that are described by a sequence of pairs (interval and frequency), it is very hard to define an order relation among them. Our main contribute consists in proposing a way to define the median (and then, the other quartiles) as the sequence of those values of a quantile function which minimize the ℓ_1 Wasserstein. This corresponds to find an order relation between the quantile function in correspondence of all the values of their common support $[0, 1]$. However, we propose an algorithm which reduces the computational cost because it looks for the median function values on a reduced number of values of the quantile functions support: the elements of the set w. Other approaches which find

an order relation among histogram data are difficult to get. In any way, working on the quantile functions associated to histograms, the nearest research area is the functional data analysis. In such context, *depth function* based on a concept of *internality* of the observed curves, rather than on the *centrality* one, has been proposed. It presents, as main advantage, to be an observed function, the most internal to the other ones but it does not guarantee to be the Median function for all the values of the domain, due to the intersection that it can present with the other functions. Instead, our proposed approach allows to have a piece-wise median quantile function which is always at half-way position with respect to all the others. In prospective, further comparisons between the two approach could be interesting to put in evidence some advantages of each other, especially when the number of functions is very high.

References

Arroyo, J. (2008). *Métodos de predicción para series temporales de intervalos e histogramas.* Phd thesis, Universidad Pontificia Comillas.

Arroyo, J., G. Gonzalez-Rivera, C. Maté, and A. Muñoz San Roque (2011). Smoothing methods for histogram-valued time series: an application to value-at-risk. *Statistical Analysis and Data Mining 4*(2), 216–228.

Arroyo, J. and C. Maté (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting 25*(1), 192–207.

Bertrand, P. and F. Goupil (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*, pp. 103–124. Springer Berlin Heidelberg.

Billard, L. and E. Diday (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association 98*, 470–487.

Billard, L. and E. Diday (2006). *Symbolic Data Analysis: conceptual statistics and data Mining*. Wiley.

Boch, H. H. and E. Diday (2000). *Analysis of Symbolic Data*. Springer.

Diday, E. and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.

Gibbs, L. A. and F. E. Su (2002). On choosing and bounding probability metrics. *International Statistical Review 70*, 419–435.

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. CRC.

Irpino, A. and E. Romano (2007). Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. *Revue des Nouvelles Technologies de l'Information RNTI-E-9*, 99–110.

Irpino, A., R. Verde, and Y. Lechevallier (2006). Dynamic clustering of histograms using wasserstein metric. In *COMPSTAT 2006*, pp. 869–876. Physica-Verlag.

Kenney, J. F. and E. S. Keeping (1962). *Mathematics of Statistics*. Princeton, NJ: Van Nostrand.

Liu, R. Y., J. M. Parelius, and K. Singh (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics 27(3)*, 783–858.

López-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. *Journal of the American Statistical Association 104*(486), 718–734.

Noirhomme-Fraiture, M. and A. Nahimana (2008). Visualization. In E. Diday and M. Noirhomme-Fraiture (Eds.), *Symbolic Data Analysis and the SODAS Software*, pp. 109–120. Wiley.

Noirhomme-Fraiture, M. and M. Rouard (1997). Zoom star: a solution to complex statistical object representation. In *INTERACT*, pp. 100–101.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis (Second Edition)*. Springer.

Rüshendorff, L. (2001). Wasserstein metric. In *Encyclopedia of Mathematics*. Springer.

Tukey, J. W. (1975). *Exploratory Data Analysis*. Addison-Wesley.

Verde, R. and A. Irpino (2008). Comparing histogram data using a *Mahalanobis - Wasserstein* distance. In P. Brito (Ed.), *COMPSTAT 2008*, Chapter 7, pp. 77–89. Heidelberg: Physica-Verlag HD.

Zuo, Y. and R. Serfling (2000). General notions of statistical depth function. *Annals of Statistics 28*, 461–482.