Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 25 (2009) 192–207
international journal of forecasting
www.elsevier.com/locate/ijforecast

Forecasting histogram time series with k-nearest neighbours methods

Javier Arroyo[a,*], Carlos Maté[b]

[a] Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense, Profesor García-Santesmases s/n, 28040 Madrid, Spain
[b] Instituto de Investigación Tecnológica, ETSI (ICAI), Universidad Pontificia Comillas, Alberto Aguilera 25, 28015 Madrid, Spain

Abstract

Histogram time series (HTS) describe situations where a distribution of values is available for each instant of time. These situations usually arise when contemporaneous or temporal aggregation is required. In these cases, histograms provide a summary of the data that is more informative than those provided by other aggregates such as the mean. Some fields where HTS are useful include economy, official statistics and environmental science.

This article adapts the k-Nearest Neighbours (k-NN) algorithm to forecast HTS and, more generally, to deal with histogram data. The proposed k-NN relies on the choice of a distance that is used to measure dissimilarities between sequences of histograms and to compute the forecasts. The Mallows distance and the Wasserstein distance are considered. The forecasting ability of the k-NN adaptation is illustrated with meteorological and financial data, and promising results are obtained. Finally, further research issues are discussed.

Keywords: Density forecast; Finance; Nonlinear time series models; Non-parametric forecasting; Symbolic data analysis; Weather forecast

1. Introduction

Time series where observations are single values serve well for representing many practical and theoretical situations. In fact, they are almost the only type of time series that is usually considered. However, they do not faithfully describe phenomena where a set of realizations of the observed variable is available for each time point, or where the observed variable has a certain degree of variability. There are two typical situations where this happens:

(1) If a variable is measured through time for each individual of a group, and the interest does not lie in the individuals but in the group as a whole. In this case, a time series of the sample mean of the observed variable over time would be a weak representation.
(2) When a variable is observed at a given frequency (say minutes), but has to be analyzed at a lower

* Corresponding author.
E-mail address: javier.arroyo@fdi.ucm.es (J. Arroyo).

0169-2070/$ - see front matter © 2008 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2008.07.003

frequency (say days). In this case, if the variable is just sampled at the lower frequency, or if only the highest and lowest values between consecutive lower-frequency instants are represented (as is done when summarizing the intra-daily behaviour of the shares in the stock markets), a lot of information is neglected.

These two situations describe contemporaneous and temporal aggregation, respectively. In each case, a time series of distributions would offer a more informative representation than other forms of aggregated time series. As Schweizer (1984) states, "distributions are the numbers of the future". Thus, instead of simplifying them, it seems better to propose methods which deal with distributions directly.

In order to do this, one has to determine how to represent the observed distributions. Distributions can be estimated either by a parametric method, e.g., a mixture of Gaussian distributions; or by a nonparametric method, e.g., a histogram or a kernel based density estimator. In this article, it is proposed to represent them using histograms, because histograms offer a good tradeoff between simplicity and accuracy.

The analysis of histogram data belongs to the field of symbolic data analysis (Billard & Diday, 2003), which is an emerging paradigm of statistics that works on data sets which are represented by intervals, histograms, lists of values, and so on. Billard and Diday (2006) and Diday and Noirhomme (2008) offer an up-to-date review of this field.

The use of histograms to characterize time series of distributions gives rise to histogram time series (HTS). It is important to note the differences between this approach and density forecasting. In HTS, the aim is to forecast a time series of distributions observed through time and, consequently, the forecast will be a distribution (represented by a histogram). On the other hand, in density forecasting, the observed time series is a time series of point values, but a forecast of the future density is given to provide information about the uncertainty associated with the point forecast.

HTS are suitable for representing aggregated data. It is obvious that if the interest lies in the original data, aggregation should not be considered. However, if this data is not manageable or if the analysis is focused on aggregated behaviour, then HTS are worth considering, as they retain more information than other single aggregates such as the mean or the total. Fig. 1 shows the two typical situations where HTS arise, and they are also detailed below.

In the case of high-frequency data, HTS provides a way of summarizing them, taking into account the intra-period variability. This kind of data typically arises in finance and in data stream mining. As Engle and Russell (in press) point out, high-frequency financial data have some characteristics that pose new challenges for forecasters: irregular temporal spacing, diurnal patterns, price discreteness, and complex dependence. These features make it difficult to forecast these data. However, representing them using HTS provides a way of dealing with these difficulties that takes advantage of the large amount of information available. If HTS are applied, the focus is not on obtaining forecasts for the next intra-period tick of the original time series, but on obtaining a forecast of the overall distribution for the next period. The information provided by a forecast in histogram form complements the information provided by a forecast of the close value, and can be very useful when one is setting trading strategies.

Regarding contemporaneous and spatial data, HTS seems a natural way to manage this kind of information, especially if a huge number of observations are available for each instant of time. González-Rivera, Lee, and Mishra (2008) display a time series of the weekly returns of the firms in the SP500, summarized by histograms. This is an interesting precedent that shows that, in some cases, histograms are preferable to averages or totals. Moreover, it is worth mentioning that sometimes contemporaneous aggregation is the only way to analyze temporal data: for example, consider the observations of a variable measured through time in a population (e.g. the annual incomes of the inhabitants of a region). If the individuals considered in each instant are not the same through time, then it is not possible to deal with the disaggregated data. Moreover, in these cases, histograms provide more information than statistics such as the mean.

In order to forecast HTS, the adaptation of methods already used for classic time series seems a reasonable approach. Arroyo, Maté, Muñoz San Roque and Sarabia (2008) propose exponential smoothing methods based on histogram arithmetic, and obtain good results in the data sets considered. The present article follows the same course and

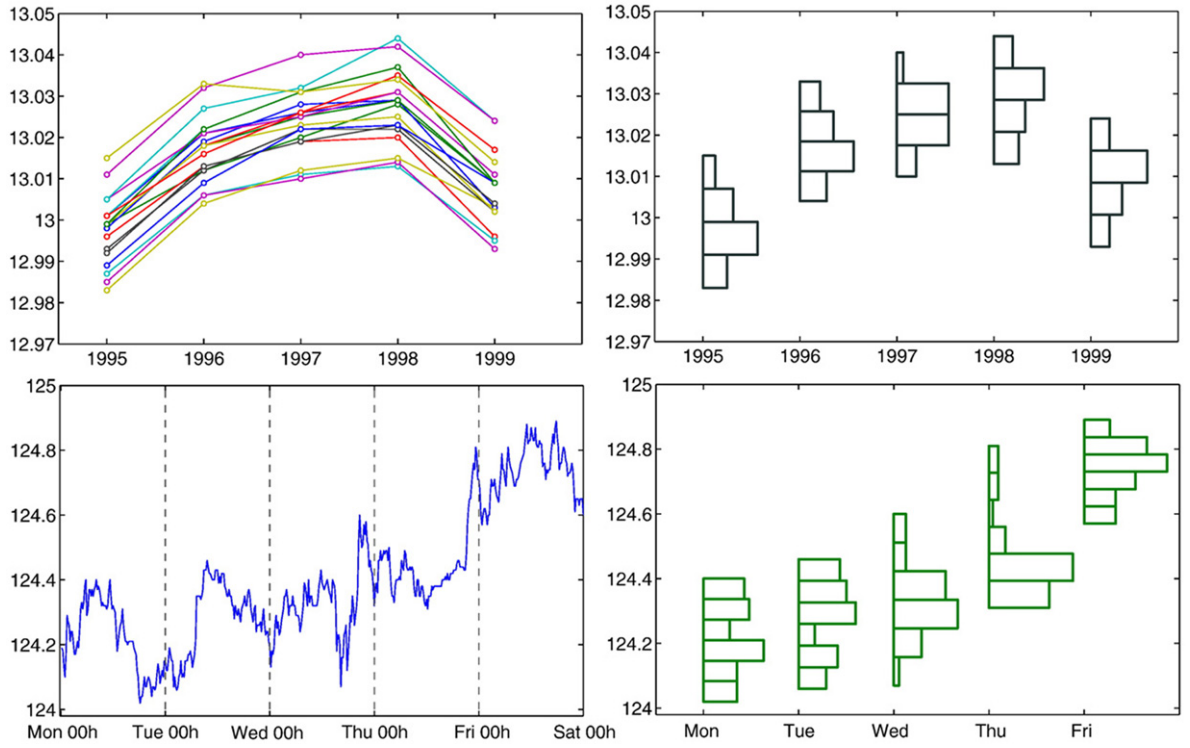

Fig. 1. Top: HTS (right) obtained from a set of distributions observed through time (left). Bottom: HTS (right) obtained from a time series of higher frequency (left).

adapts the k-Nearest Neighbours (k-NN) algorithm to forecast HTS. The forecasting ability and simplicity of k-NN makes its adaptation for HTS suitable. Another strength of the k-NN algorithm is its versatility: it can be applied to density estimation, classification, function approximation, and also time series forecasting (Yakowitz, 1987).

It is also interesting to note that Pasley and Austin (2004) propose a nearest-neighbours-based method to obtain density forecasts of high-frequency time series. The Pasley and Austin approach uses only a subset of relevant past values of the high frequency time series to construct density forecasts in histogram form. The identification of the relevant values is made by a nearest neighbours algorithm. However, this is a density forecasting method and cannot be applied to forecasting HTS.

To make this article self-contained, Section 2 offers an overview of the fundamentals of HTS. Section 3 adapts the k-NN algorithm for forecasting HTS; an appropriate distance for histograms is suggested, and a method of combining histograms, which will allow us to obtain the forecasts, is proposed. Section 4 analyzes the forecasting performance of the algorithm in two examples from different contexts. Finally, Section 5 concludes.

2. Fundamentals of histogram time series

2.1. Definitions: Histogram time series and histogram variables

A histogram time series, $\{h_{X_t}\}$, is a sequence of distributions observed at times $t = 1, \ldots, n$, where each distribution is represented as a histogram h_{X_t}.

In symbolic data analysis (Billard & Diday, 2006), interval and histogram variables are applied to describe the variability of the quantitative attributes of concepts or groups of individuals. From this perspective, HTS can be considered as time series where the observations are realizations of a histogram variable. In a histogram variable h_X, all elements k

of the set E take values in the domain $\mathcal{B}$ of possible observed values according to the mapping

$$h_{X_k} = \{([I]_{k,1}, \pi_{k,1}), \ldots, ([I]_{k,p_k}, \pi_{k,p_k})\}, \quad \text{for } k \in E, \tag{1}$$

where $\{\pi_{k,i}\}$, with $= 1, \ldots, p_k$, is a frequency or probability distribution on the domain $\mathcal{B}$ that satisfies $\pi_{k,i} \geq 0$ and $\sum_{i=1}^{pk} \pi_{k,i} = 1$; and where $[I]_{k,i} \subseteq \mathcal{B}, \forall k, i$, is an interval (also known as a bin) defined as $[I]_{k,i} = [\underline{I}_{k,i}, \bar{I}_{k,i})$ with $-\infty < \underline{I}_{k,i} \leq \bar{I}_{k,i} < \infty$ and $\bar{I}_{k,i-1} \leq \underline{I}_{k,i}$ $\forall k, i$, with $i \geq 2$.

2.2. Why deal with histograms?

First, it is important to observe that, despite its simplicity, the histogram definition given in Eq. (1) covers all possible binned density estimators. The most popular ones are histograms with a fixed bin width, which are extensively used in practice, but equifrequency histograms are also included. A particular case of the latter kind are boxplots, which offer a streamlined and meaningful display of a batch of data. The aforementioned definition also covers the discretized density representations used in some density forecast applications, i.e., a grid of regular intervals partitioning the range of the variable of interest and a sequence of quantiles of specific interest (Tay & Wallis, 2000).

Secondly, histograms can also provide accurate representations of the underlying true density. In this regard, Wand (1997) proposes a family of rules to determine the optimal bin width, irrespective of the underlying density, for histograms with a fixed bin width. The computational effort required by this method is similar to that required to construct a kernel density estimate.

It can be argued that kernel-based density estimators and Gaussian mixtures offer smoother representations of the underlying density than histograms, and, consequently, that it would be better to consider time series of these estimators. However, dealing with histograms reduces the computational effort enormously. The proposed adaptation of the k-NN algorithm can also be applied to non-binned estimators, but the use of binned estimators simplifies the density representation and the subsequent operations, as is shown in the Appendix A. In addition, one question arises: Is smoothness really needed? If the answer is *no*, histograms can suffice. The reasons for using histograms can be summarized as follows:

- they are close enough to the data at hand, without imposing any distribution law,
- they can describe the essential features of the data with reasonable accuracy, and
- their simple and flexible structure simplifies their use.

2.3. Error measures for histogram time series

In classic time series, accuracy measures are based on the difference between the observed and forecasted values, i.e., $e_t = X_t - \hat{X}_t$, where X_t is the observation and $\hat{X}_t$ is the forecasted value. However, due to the complexity of histograms, HTS require a different approach to quantifying the differences between histograms.

In density forecasting, Diebold, Gunther, and Tay (1998) propose the use of goodness-of-fit tests to evaluate whether the density prediction corresponds to the true density or not. Unfortunately, this approach is not useful for HTS accuracy measurement, as it does not quantify the deviation between the two densities. A different approach is proposed by Hall and Mitchell (2007), who use a dissimilarity measure, the Kullback-Leibler Information Criterion (KLIC), to combine density forecasts and to quantify the differences between the forecasted and true densities. Dissimilarity measures provide objective values representing the deviation between densities; therefore they seem more suitable for HTS accuracy measurement. However, KLIC is not appropriate for HTS, because it requires the support of one of the considered densities to be the same as, or contained in, the support of the other one, which is a condition frequently violated in HTS. The χ^2-divergence suffers from the same problem. Thus, other dissimilarity measures should be considered.

Bock (2000) and Gibbs and Su (2002) review the dissimilarity measures for probability distributions. Most of these measures can, in principle, be considered for defining error measures for HTS. However, this article will only consider the Wasserstein distance and the Mallows distance (Mallows, 1972), which are

related, as will be seen below. Given two density functions, $f(x)$ and $g(x)$ on $\mathbb{R}$, the Wasserstein (or Kantorovich) and Mallows distances between $f(x)$ and $g(x)$ are defined as

$$D_W(f,g) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|\mathrm{d}t \quad \text{and} \tag{2}$$

$$D_M(f,g) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 \mathrm{d}t}, \tag{3}$$

respectively, where $F^{-1}(t)$ and $G^{-1}(t)$ with $t \in [0,1]$ are the inverse cumulative distribution functions of $f(x)$ and $g(x)$, respectively.

The reasons for choosing these dissimilarity measures are the following. First, they are distances, and thus present interesting properties for error measurement: positive definiteness, symmetry, and the triangle inequality condition. Second, the Mallows distance has been successfully applied to cluster histogram data by Irpino and Verde (2006) and Verde and Irpino (2007). Third, according to Levina and Bickel (2001), these distances are conceptually equivalent to the Earth Mover's Distance (EMD) proposed by Rubner, Tomasi, and Guibas (2000). The EMD is a well-known computer vision distance that is used for measuring dissimilarities between colour and texture histograms. This distance can be seen as a solution to the Monge-Kantorovich mass transference problem, i.e., the EMD between two histograms is the least amount of work needed to transform one histogram into the other by means of transportation (further details are given by Rubner et al. (2000)). The effectiveness of the EMD in image retrieval seems a good endorsement for the use of the Mallows and Wasserstein distances in other contexts. Interestingly, as the graphical examples from Rubner et al. (2000) show, the EMD matches the human notion of dissimilarity between histograms. If these examples are extended to the Mallows and Wasserstein distances, the same conclusion is obtained.

Given these reasons, the Mallows and Wasserstein distances seem suitable choices for error measurement in HTS. The adaptation of these distances to histogram data requires the definition of the cumulative distribution function (CDF) of a histogram. In order to do this, it will be assumed that data points are uniformly distributed within each bin of the histogram, as is usually assumed when considering histograms as density estimators, and also in symbolic data analysis (Billard & Diday, 2003). Given this, the CDF $H_A(x)$ of a histogram $h_A = \{([I]_{l_A}, \pi_{l_A})\}$, with $l_A = 1, \ldots, p_A$, is defined as

$$H_A(x) = \int_{-\infty}^{x} h_A(x)\mathrm{d}x = \begin{cases} 0, & \text{if } x < \underline{I}_1; \\ w_{l_A-1} + \dfrac{x - \underline{I}_{l_A}}{\bar{I}_{l_A} - \underline{I}_{l_A}}\pi_{l_A}, & \text{if } x \in [\underline{I}_{l_A}, \bar{I}_{l_A}); \\ 1, & \text{if } x \geq \bar{I}_{p_A}, \end{cases} \tag{4}$$

where $w_l = \sum_{i=1}^{l} \pi_i$ is the cumulative weight associated with the interval l. Given Eq. (4) Appendix A shows how to easily compute the Wasserstein and Mallows distances, given in Eqs. (2) and (3), when dealing with histograms.

Regarding the interpretation of these distances, it can be seen from their definitions (2–3) that both are based on the values $\delta_t = |F^{-1}(t) - G^{-1}(t)|, \forall t \in [0,1]$. These values can be seen as the L_1 distance between the t-quantiles of the considered distribution functions. However, in the Mallows distance an L_2 norm is used. Fig. 2 shows the CDFs of two histograms with the δ_t values. From Eq. (2) and this figure, it is clear that the Wasserstein distance measures the area contained between the two CDFs. However, as the δ_t values are squared in the Mallows distance, this distance resembles a Euclidean metric.

Arroyo et al. (2008) propose the following error measure for HTS. Let $\{h_{X_t}\}$ be the observed HTS, and let $\{\hat{h}_{X_t}\}$ be the forecast of this HTS, where $t = 1, \ldots, n$. The Mean Distance Error (MDE) is defined as

$$MDE(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \frac{1}{n}\sum_{t=1}^{n} D(h_{X_t}, \hat{h}_{X_t}), \tag{5}$$

where $D(h_{X_t}, \hat{h}_{X_t})$ is the Wasserstein or Mallows distance. As distances cannot be negative, the MDE simply averages them across time.

Following the ideas of Hyndman and Koehler (2006), a unit-free version of the MDE can be proposed. If the error in t is scaled by the in-sample error of the naïve method, the Mean Scaled Distance Error (MSDE) is defined as

$$MSDE(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \frac{1}{n}\sum_{t=1}^{n} \frac{D(h_{X_t}, \hat{h}_{X_t})}{MDE_m}$$

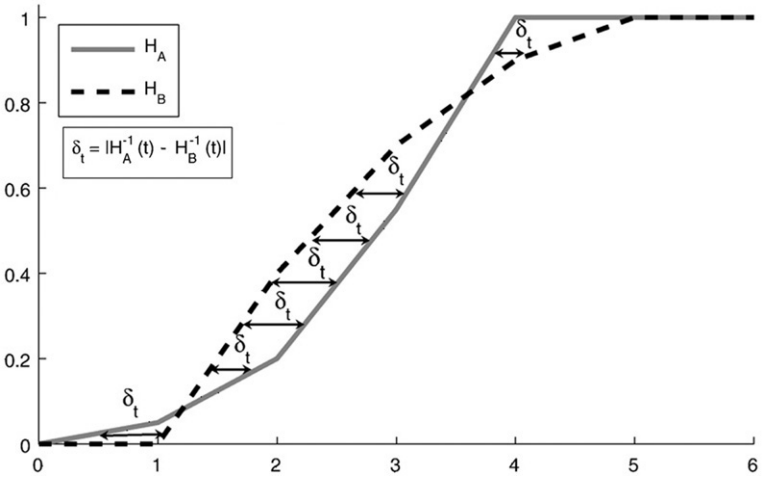

Fig. 2. CDFs of $h_A = \{([0, 1), 0.05); ([1, 2), 0.15); ([2, 3), 0.35); ([3, 4], 0.45)\}$; and $h_B = \{([1, 2), 0.4); ([2, 3), 0.3); ([3, 4), 0.2); ([4, 5], 0.1)\}$.

$$= \frac{\frac{1}{n}\sum_{t=1}^{n} D(h_{X_t}, \hat{h}_{X_t})}{MDE_m}, \qquad (6)$$

where $MDE_m = \frac{1}{m-1}\sum_{i=2}^{m} D(h_{X_i}, h_{X_{i-1}})$ is the MDE of the naïve model in the in-sample period and m is the length of this period. If the MSDE is less than 1 for the period considered, the applied method obtains better forecasts, on average, than the average one-step naïve forecast computed in-sample. However, if it is greater than 1, the method performs worse.

3. k-nearest neighbours for histogram time series

3.1. The k-NN in classic time series

The k-Nearest Neighbours algorithm is a simple, yet powerful and versatile, pattern recognition method. If the target variable is categorical, it can be used for classification tasks (Cover & Hart, 1967). When the target variable is continuous, it can be used as a function approximator that carries out local non-parametric regressions. One of the applications of k-NN as a function approximator is in time series forecasting (Yakowitz, 1987). The idea behind forecasting with k-NN is to identify, in the time series, the past sequences which are the most similar to the current one, and to combine their future values to predict the next value of the current sequence. This idea is closely related to the algorithm for forecasting chaotic data proposed by Farmer and Sidorowich (1987).

The k-NN method is used in finance to model the nonlinear dynamics of the series, see for example Aparicio, Pozo, and Saura (2002), Fernández-Rodríguez, Sosvilla-Rivero, and Andrada-Félix (1999) and Meade (2002). Other application fields are hydrology (Brath, Montari, & Toth, 2002) and the energy sector (Sorjamaa, Reyhani, & Lendasse, 2005).

The algorithm is briefly shown below.

(1) The time series considered, $\{X_t\}$ with $t = 1, \ldots, n$, is transformed into a series of d-dimensional vectors

$$X_t^{d,\tau} = (X_t, X_{t-\tau}, \ldots, X_{t-(d-1)\tau}), \qquad (7)$$

where $d, \tau \in \mathbb{N}$, with d being the number of lags and τ the delay parameter. If $\tau = 1$, as is assumed in many cases (e.g. Fernández-Rodríguez et al. (1999), Meade (2002) and Sorjamaa et al. (2005)), the resulting time series of vectors is denoted by $\{X_t^d\}$, with $t = d, \ldots, n$, where

$$X_t^d = (X_t, X_{t-1}, \ldots, X_{t-(d-1)}) \qquad (8)$$

is a vector of d consecutive observations that can be represented as a point in a d-dimensional space.

(2) The distance between the last vector $X_n^d = (X_n, X_{n-1}, \ldots, X_{n-d+1})$ and each vector in the time series $\{X_t^d\}$ with $t = d, \ldots, n-1$ is

computed, and the k vectors closest to X_n^d are selected. They are denoted by $X_{T_1}^d, X_{T_2}^d \ldots, X_{T_k}^d$. The Euclidean distance is typically applied in this step.

(3) Given the k neighbouring vectors, $X_{T_1}^d, X_{T_2}^d \ldots, X_{T_k}^d$, their subsequent values, $X_{T_1+1}, X_{T_2+1} \ldots, X_{T_k+1}$, are averaged to obtain the forecast, X_{n+1}.

This is the most simple k-NN, but more sophisticated versions can be proposed. For example, Meade (2002) uses a geometrically weighted Euclidean distance to put greater emphasis on the similarity between more recent observations, and obtains the forecast as a weighted average that assigns more weight to the closest neighbours. In addition, more attention can be given to choosing the values of k, d and τ; see Jayawardena, Li, and Xu (2002) for an interesting revision of the subject. It is also recommended that only a subset of the d past values be selected as input variables, because, as in other classification methods, the performance of k-NN is poor if irrelevant inputs are considered. Sorjamaa et al. (2005) propose three methods to guide the selection of the inputs.

3.2. The adaptation of k-NN to HTS

The adaptation of the k-NN algorithm to deal with HTS relies heavily on the choice of a distance for the histograms. This distance will be used to measure dissimilarities between histogram vectors, to obtain forecasts, and to measure errors. As these three issues are closely related, it is recommended to use the same distance for all three.

In Section 2.3, the Wasserstein and Mallows distances were considered suitable for the measurement of errors in HTS. The same reasoning can be used to argue that they are appropriate for dissimilarity measurement.

Regarding forecast obtention, this article proposes, as will be shown below, to replace the averaging with the estimation of a distance-based barycenter. Verde and Irpino (2007) analyze a set of dissimilarity measures in order to analyze their appropriateness for the construction of prototypes (barycenters) of the partitions produced by a dynamic clustering algorithm for histogram data. Verde and Irpino (2007) discard measures such as the Total Variation distance, the Hellinger distance, the Kolmogorov metric and the Prokhorov–Lévy distance, and only find appropriate the prototypes that are obtained using the Mallows distance. The arguments given by Verde and Irpino (2007) remain valid for k-NN forecast obtention. However, for this purpose, we will also consider the Wasserstein distance, because the barycenters obtained with this distance present properties different from those obtained using the Mallows distance, as will be shown in the Appendix A.

The adaptation of the k-NN method for HTS requires the specification of procedures for distance measurement between histogram sequences and for obtaining the forecasts. They are detailed below.

3.2.1. Distance measurement

The HTS $\{h_{X_t}\}$, with $t = 1, \ldots, n$, is transformed into a series of d-dimensional histogram vectors

$$h_{X_t^d} = (h_{X_t}, h_{X_{t-1}}, \ldots, h_{X_{t-d+1}}), \tag{9}$$

with $t = d, \ldots, n$. The distance between the last vector $h_{X_n^d}$ and all of the past vectors $h_{X_t^d}$, with $t = d, \ldots, n-1$, is computed as

$$D(h_{X_n^d}, h_{X_t^d}) = \frac{1}{d} \sum_{i=1}^{d} D(h_{X_{n-i+1}}, h_{X_{t-i+1}}), \tag{10}$$

where $D(h_{X_{n-i+1}}, h_{X_{t-i+1}})$ can be either the Mallows (3) or the Wasserstein (2) distance. Once the distances have been computed, the k closest sequences to $h_{X_n^d}$ are identified. They are denoted by $h_{X_{T_p}^d}$, with $p = 1, \ldots, k$.

3.2.2. Forecast obtention

In the classic k-NN method, forecasts are usually computed as an average (weighted or not) of the subsequent values of the k neighbouring sequences. In order to adapt k-NN to HTS, the averaging procedure can be substituted by the estimation of the barycenter as the two concepts are closely related.

In physics, the barycenter is the center of mass of a system of particles; in other words, it is a specific point at which, for many purposes, the system's mass behaves as if it were concentrated at that point. Analytically, the coordinates of the barycenter can be computed as the weighted (according to the mass of the particle) average of the coordinates of the particles. If all of the particles have the same mass, the barycenter is the point that minimizes the addition of the squared Euclidean distances between itself and

each system's particle; i.e., the barycenter is the mean of the coordinates of the particles.

Given this, in the k-NN for HTS, it seems reasonable to replace the averaging process with the obtention of the *barycentric* histogram that minimizes the addition of the distances between itself and each of the subsequent histograms of the k neighbouring sequences, denoted by $h_{X_{T_p+1}}$, with $p = 1, \ldots, k$. The forecast $\hat{h}_{X_{n+1}}$ will be optimal in the sense that it will be the solution of

$$\underset{\hat{h}_{X_{n+1}}}{arg\ min} \sum_{p=1}^{k} \omega_p D(\hat{h}_{X_{n+1}}, h_{X_{T_p+1}}), \tag{11}$$

where $D(\hat{h}_{X_{n+1}}, h_{X_{T_p+1}})$ is the Mallows (3) distance or the Wasserstein (2) distance, $h_{X_{T_p+1}}$ is the subsequent histogram of the neighbouring sequence $h_{X^d_{T_p}}$, and ω_p is the weight assigned to each neighbour and satisfies $\omega_p \geq 0$ and $\sum_{p=1}^{k} \omega_p = 1$.

If all the neighbours have the same weight, then $\omega_p = 1/k, \forall p$. However, more sophisticated weighting schemes can be considered. For example, the weight assigned to the subsequent value of each neighbouring sequence can depend inversely on the distance between the neighbouring and the considered sequences, as follows

$$\omega_p = \frac{\psi_p}{\sum_{l=1}^{k} \psi_l}, \tag{12}$$

where $\psi_p = (D(h_{X^d_n}, h_{X^d_{T_p}}) + \xi)^{-1}$, with $p = 1, \ldots, k$, and where $D(h_{X^d_n}, h_{X^d_{T_p}})$ is given in Eq. (10). The constant $\xi = 10^{-8}$ avoids the infinite values caused by zero distances between sequences.

The Appendix A shows how to estimate the barycentric histogram based on the Wasserstein and Mallows distances without applying optimization techniques. The use of binned estimators drastically reduces the computational effort required. The Appendix A also shows that the Mallows barycenter can be considered as the *mean histogram*, while the Wasserstein barycenter can be considered as the *median histogram* of the given set of histograms.

4. Illustrative examples

This section illustrates the use of HTS to represent time series of distributions in two different contexts: meteorology and finance. In the first example, the HTS will be created by spatial aggregation, and in the second, by temporal aggregation. The features of the resulting HTS are quite different, but it will be shown that the k-NN algorithm performs well in both cases.

In each example, the k-NN algorithms based on the Mallows and the Wasserstein distances have both been applied. However, the results do not differ significantly. Consequently, for the sake of brevity, the first example only reports the results of the Mallows-based k-NN, while the second only shows those obtained using the Wasserstein-based k-NN. For the same reason, the tables only report the results of the k-NN with equal weights, despite the fact that the weighting scheme proposed in Eq. (12) has also been applied.

In each example, the number of neighbours, k, and the length of the sequences, d, have been determined by a grid-search method with the aim of minimizing the forecasting error in the training set.

4.1. Example from a meteorological context

In meteorology, historical data offer a great opportunity to apply HTS. In this case, the data set considered has been obtained from the Long-Term Instrumental Climatic Database of the People's Republic of China.[1] It contains 14 monthly-recorded meteorological variables observed at 60 stations, where each station is representative of a particular climatic region of China. The 60 stations form a network with a relatively uniform spatial distribution.

These features make this data set suitable for spatial aggregation using histograms. The resulting HTS describe the monthly distribution of the considered weather variables in the whole of China over time. The time series of the distribution of a weather variable in a country will offer a different view of meteorological phenomena, such as the possible global warming effect.

[1] Available for registered users from http://dss.ucar.edu/datasets/ds578.5/data/ (registration is free).

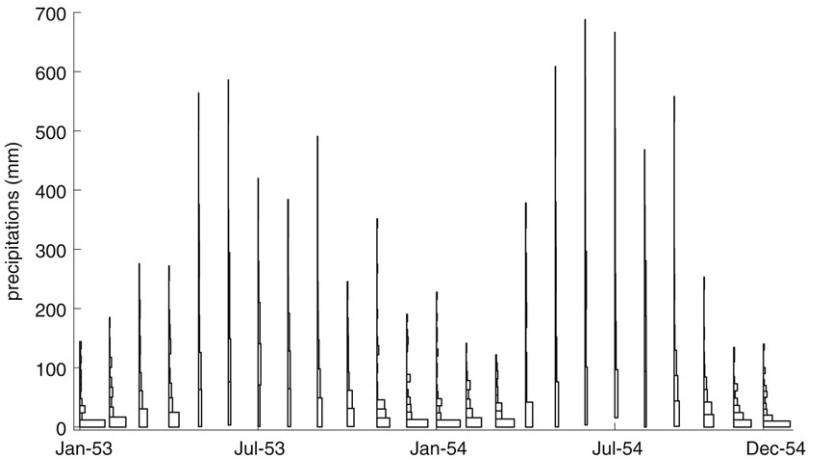

Fig. 3. HTS of the distribution of the monthly total precipitation in China.

This example focuses on the distribution of the monthly total precipitation throughout China in mm. The HTS considered will cover the period 1951–1988, i.e., 456 monthly histograms. Fig. 3 shows that the HTS has a seasonal component that affects the shape and range of the histograms. Seasonality is due to the monsoon, as most of China's rain falls during the period from May to October. A forecast of the distribution of the precipitation could help the Chinese Government to plan for issues related to water reserves, e.g., water supply, crops, energy production, etc.

The rule proposed by Wand (1997) has been used to estimate the optimal bin width for obtaining histograms that accurately represent the true underlying density. For this reason, the number of bins in the histograms varies along the HTS. This fact makes it necessary to set a fixed number of bins in the forecasted histograms. In this case, it was decided to set this value to 10, because 10 is the median and 10.65 the mean of the number of bins in the original HTS.

The k-NN based on the Mallows distance has been applied to forecast the HTS. The naïve method with seasonality, $\hat{h}_{X_{t+1}} = h_{X_{t+1-s}}$, where s is the length of the seasonal cycle, was used as a benchmark. The first 144 periods have been used as the initialization set required by the k-NN to look for the neighbours, the next 156 periods as the training set, and the last 156 periods as the test set.

Table 1 shows the forecasting errors of the methods along with the parameters that were determined by minimizing the MDE_M in the training set. The table displays the MDE_M (5) and the $MSDE_M$ (6), where the errors have been scaled using the MDE_M from the benchmark method in the training set. The k-NN clearly produces better results than the benchmark method. According to the $MSDE_M$, the MDE of the k-NN in the test set is more than a 30% smaller than the in-sample MDE of the benchmark method. Fig. 4 shows that the forecasted HTS accurately describes the actual behaviour of the HTS.

Table 1
Errors of the different forecasting methods in the precipitation HTS.

Method	Training set		Test set	
	MDE_M	$MSDE_M$	MDE_M	$MSDE_M$
Naive method with seasonality	33.5	1	31.77	0.948
k-NN equal weights ($k = 8, d = 16$)	25.33	0.756	23.01	0.687

4.2. Example from a financial context

In finance, the continuous monitoring of stocks, indexes and exchange rates yields very long tick-by-tick time series. These high-frequency data are often reduced to the daily closing values, but they can also be temporally aggregated using histograms, giving rise to HTS, where histograms describe the overall behaviour during each period considered.

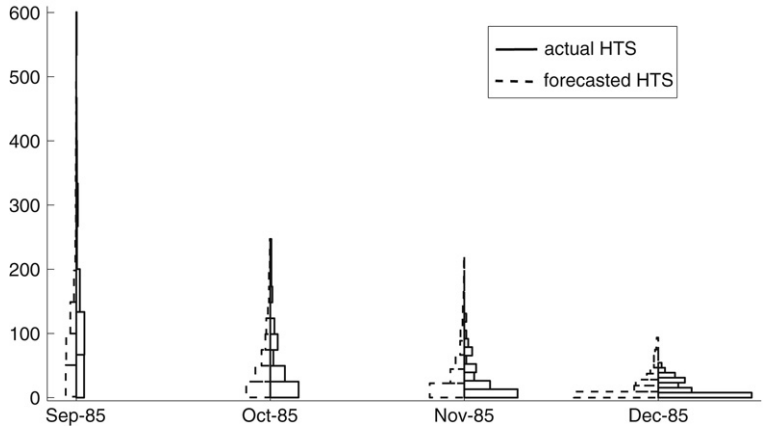

Fig. 4. Excerpt of the actual and forecasted HTS in the test set (using the k-NN with equal weights).

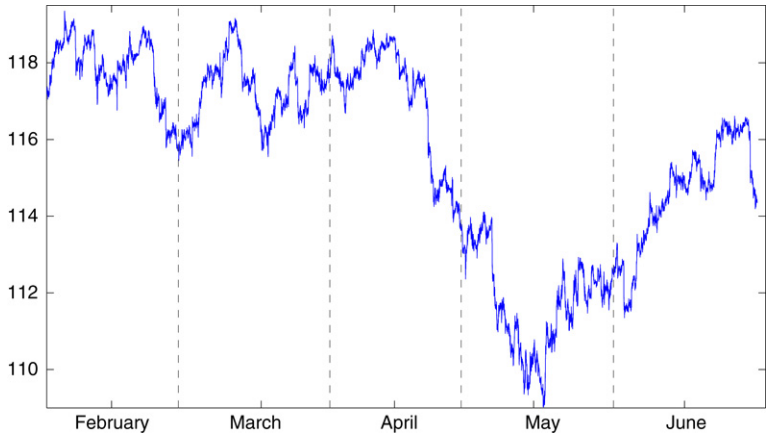

Fig. 5. Intra-daily values of the $-¥ currency rate from 2-1-2006 to 6-30-2006.

It could be argued that the summary provided by the HTS neglects information contained in the original high-frequency data, which is true. However, it is also true that, as Engle and Russell (in press) state, high-frequency financial time series present some features that make them hard to forecast with standard methods. These features are, primarily, irregular temporal spacing, diurnal patterns, price discreteness, and complex dependence. Engle and Russell (in press) propose several methods for modelling intra-daily data.

Even with extremely robust methods, it seems a titanic task to accurately forecast the whole tick-by-tick time series one day ahead. On the other hand, obtaining one-day-ahead forecasts of an aggregated time series is more feasible. In this sense, closing values are useful aggregates, but they do not report information about the intra-daily behaviour. This information is provided by histograms, as they describe the intra-daily volatility in each session, which can be very helpful when setting a trading strategy, for example.

In this example, the time series considered consists of the intra-daily values of the dollar-yen currency rate from February 1, 2006, to June 30, 2006. There are 107 trading days, and 288 values per trading day. In Fig. 5, it can be seen that the time series presents shifts in the level from April to June. As these shifts have no precedent in the past of the time series being considered, the k-NN method is not expected to perform well for this period. In order to remove the shifts, the time series will be transformed into

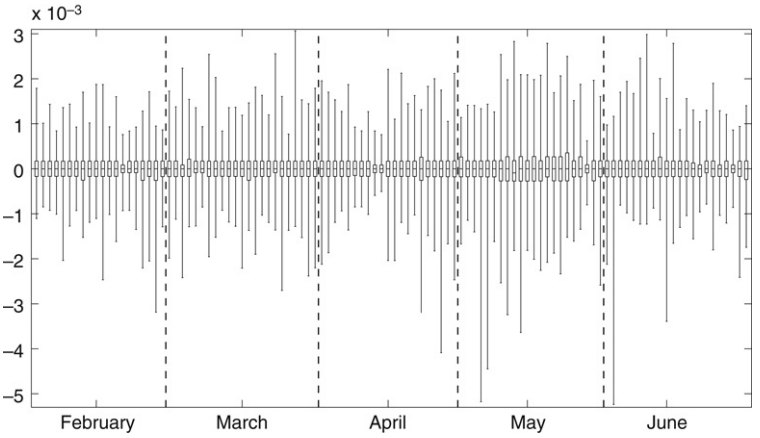

Fig. 6. Boxplots of the daily $−¥ arithmetic returns from 2-1-2006 to 6-30-2006.

arithmetic returns

$$Q_t = \frac{X_t - X_{t-1}}{X_{t-1}}, \tag{13}$$

where X_t is the current value of the original series and X_{t-1} is the immediately prior value. The arithmetic returns of each day will be summarized by a boxplot, giving rise to the daily HTS shown in Fig. 6. Note that boxplots are histograms with four bins, where each bin has a relative frequency of 0.25. Boxplots divide the batch of exchange rates into four regions, offering an intuitive representation of the volatility in each trading day.

In this case, the k-NN applied is based on the Wasserstein distance. As is shown in the Appendix A, a forecast obtained using this method can be seen as the median of the set of subsequent histograms of the k neighbouring sequences. Thus, the prediction does not take into account histograms distant from the rest of the set, which seems desirable in finance.

The first 35 periods make up the initialization set, the next 35 periods the training set, and the last 37 periods the test set. The optimal parameters have been determined by a grid search minimizing the MDE_W in the training set. The optimal length of the sequences is $d = 10$, i.e., two trading weeks, and $k = 9$ neighbours are required to yield the forecasts.

According to the results in Table 2, the HTS turns out to be more unpredictable in the test set. This is due to the fact that the test set corresponds to an unprecedented rising high-volatility period, as Figs. 5 and 6 show. Even in these unfavourable circumstances, k-NN obtains good results and clearly outperforms the naïve method. This is promising, because, in exchange rate forecasting using classic time series, it is not common to obtain much better results than those yielded by the naïve approach, as Fernández-Rodríguez et al. (1999) and Meade (2002) report.

Table 2
Errors in the different forecasting methods in the $-¥ arithmetic returns HTS.

Method	Training set		Test set	
	MDE_W	$MSDE_W$	MDE_W	$MSDE_W$
Naive method	$1.81 \cdot 10^{-4}$	1	$1.9 \cdot 10^{-4}$	1.05
k-NN equal weights ($k = 9, d = 10$)	$1.31 \cdot 10^{-4}$	0.72	$1.66 \cdot 10^{-4}$	0.913

5. Conclusions

HTS describe the evolution through time of phenomena that have to be represented by a distribution in order not to lose significant amounts of information. Forecasting time series of distributions is not trivial, but the structural simplicity of histograms allows us to propose statistical methods without excessive computational effort.

This article has proposed an adaptation of the k-NN method to forecasting HTS, which yields promising results. The extension of the proposed method to allow the inclusion of lagged values of explanatory time series is direct, and can be very useful in some applications. In addition, a methodology to guide the

choice of the distance, the parameters, the weighting scheme and the relevant inputs in different contexts, could be proposed. It would also be interesting to propose a k-NN variant to deal with HTS with trend.

As HTS is in its early stages, there are many possibilities for future research. For example, new forecasting methods and tools for analyzing HTS need to be proposed. The application of HTS in fields where classic time series involve excessive simplification, such as official statistics, seems promising. In finance, the characterization of the daily distributions of a share or of an index by means of HTS opens challenging perspectives for risk management. The application of HTS would also be interesting in spatial econometrics and in panel data.

Furthermore, despite the fact that HTS and density forecasting are approaches with notable differences, both pursue the same aim: predicting a future density. Throughout this article some connections between these areas have been mentioned, but it would be interesting to analyze the possible synergies between the two fields in issues such as error measurement and forecast combination.

Acknowledgments

This research was supported by the Spanish Ministry of Education and Science (TIN2005-08501-C03-01), and by Universidad Pontificia Comillas (PRESIM project). We are also grateful to the referees and Gloria González–Rivera for their helpful comments.

Appendix A. Estimation of the barycentric histogram

Irpino and Verde (2006) show how to easily estimate the barycenter of a set of histograms using the Mallows distance. This appendix reproduces this procedure, and extends it to estimating the barycenter using the Wasserstein distance.

Verde and Irpino (2007) analyze the barycenters of histogram data obtained with different distances, and consider that only the Mallows-based barycenter is appropriate, although they do not consider the Wasserstein distance. They also show that the barycenters obtained using distances such as the Hellinger and Total variation distances are mixture-like, in the sense that, given a set of unimodal histograms, the barycenter obtained is multi-modal. This feature, which is not desirable for a histogram forecast, is not present in the barycenters obtained using either the Mallows or the Wasserstein distance.

A.1. Rewriting the distances for histogram data

In order to estimate the barycenters for histogram data, the Mallows (3) and Wasserstein (2) distances have to be adapted to deal with histograms. More precisely, these distances have to be expressed in terms of the real-line intervals (bins) for which both histograms are uniformly dense.

Given $h_X = \{([I]_{l_X}, \pi_{l_X})\}$ with $l_X = 1, \ldots, p_X$, and $h_Y = \{([I]_{l_Y}, \pi_{l_Y})\}$ with $l_Y = 1, \ldots, p_Y$, the set of cumulative weights of their density functions is:

$$w = \{w_{0X}, w_{1X}, \ldots, w_{p_X X}, w_{0Y}, w_{1Y}, \ldots, w_{p_Y Y}\}, \tag{A.1}$$

where $w_{0X} = w_{0Y} = 0$, $w_{l_X X} = \sum_{i=1}^{l_X} \pi_{iX}$, and $w_{l_Y Y} = \sum_{i=1}^{l_Y} \pi_{iY}$.

Given h_X and h_Y, the set of values (cumulative weights) where their CDFs, H_X and H_Y, intersect is v. In other words, the elements of v are the points in the interval (0, 1) where $H_X(x) = H_Y(x)$ with $x \in \Re$.

Given these two sets, the vector $\mathbf{z} = [z_0, \ldots, z_l, \ldots, z_m]$, with $z_0 = 0$ and $z_m = 1$, is the result of

- sorting w without repetitions, in the case of the Mallows distance, or
- sorting $w \cup v$ without repetitions, in the case of the Wasserstein distance.

The elements in $\mathbf{z}$ represent cumulative weights. Each pair of consecutive elements (z_{l-1}, z_l) in $\mathbf{z}$ sets the bounds of intervals in the real line for which both histograms are uniformly dense. In order to map these cumulative weights with the real line, the inverse of the CDF of a histogram is required.

Given a histogram $h = \{([I]_l, \pi_l)\}$, with $l = 1, \ldots, p$, and the definition of CDF for histograms shown in Eq. (4), and assuming that the data points are uniformly distributed within histogram bins $[I]_l$, the inverse of the CDF of h is

$$H^{-1}(t) = \underline{I}_l + \frac{t - w_{l-1}}{w_l - w_{l-1}}(\bar{I}_l - \underline{I}_l) \quad \text{if } t \in [w_{l-1}, w_l), \tag{A.2}$$

where $w_0 = 0$, $w_l = \sum_{i=1}^{l} \pi_i$ and $t \in [0, 1]$.

Given Eq. (A.2), each pair of consecutive elements (z_{l-1}, z_l) in $\mathbf{z}$ can be mapped to two uniformly dense intervals (bins), one for h_X and one for h_Y, in the real line. These intervals are defined as follows:

$$I_{lX} = [H_X^{-1}(z_{l-1}), H_X^{-1}(z_l)] \quad \text{and}$$
$$I_{lY} = [H_Y^{-1}(z_{l-1}), H_Y^{-1}(z_l)]. \tag{A.3}$$

As these intervals are uniformly dense, they can be expressed as a function of t, where $t \in [0, 1]$, in terms of its center and radius. The interval $I_l = [\underline{I}_l \ \bar{I}_l]$, can be rewritten as:

$$I_l(t) = c_l + r_l(2t-1) \quad \text{for } 0 \le t \le 1, \text{ with}$$
$$c_l = \frac{\underline{I}_l + \bar{I}_l}{2} \quad \text{and} \quad r_l = \frac{\bar{I}_l - \underline{I}_l}{2}. \tag{A.4}$$

The weight associated with interval I_l is $\pi_l = z_l - z_{l-1}$, with $l = 1, \ldots, m$.

Given this, the Mallows distance between a pair of histograms h_X and h_Y can be rewritten in terms of uniformly dense intervals as

$$\begin{aligned} D_M^2(h_X, h_Y) &= \sum_{l=1}^{m} \int_{z_{l-1}}^{Z_l} (H_X^{-1}(t) - H_Y^{-1}(t))^2 \mathrm{d}t \\ &= \sum_{l=1}^{m} \pi_l \int_0^1 [(c_{lX} + r_{lX}(2t-1)) \\ &\quad - (c_{lY} + r_{lY}(2t-1))]^2 \mathrm{d}t \\ &= \sum_{l=1}^{m} \pi_l \left[(c_{lX} - c_{lY})^2 + \frac{1}{3}(r_{lX} - r_{lY})^2 \right]. \end{aligned} \tag{A.5}$$

Similarly, the Wasserstein distance can be rewritten as

$$\begin{aligned} D_W(h_x, h_y) &= \sum_{l=1}^{m} \int_{z_{l-1}}^{z_l} |H_x^{-1}(t) - H_Y^{-1}(t)| \mathrm{d}t \\ &= \sum_{l=1}^{m} \pi_l \int_0^1 |(c_{lX} + r_{lX}(2t-1)) \\ &\quad - (c_{lY} + r_{lY}(2t-1))| \mathrm{d}t \\ &= \sum_{l=1}^{m} \pi_l |c_{lX} - c_{lY}|. \end{aligned} \tag{A.6}$$

A.2. Formulating the minimization problem

The barycentric histogram h_{X_B} of a set of k histograms $h_{X_1}, h_{X_2}, \ldots, h_{X_k}$ is the solution of the minimization problem

$$\underset{h_{X_B}}{arg\ min} \sum_{p=1}^{k} \omega_p D(h_{X_B}, h_{X_p}), \tag{A.7}$$

where $D(h_{X_B}, h_{X_p})$ is either the Mallows or the Wasserstein distance, and ω_p is the weight associated with histogram h_{X_p}, with $\omega_p \ge 0$ and $\sum_{p=1}^{k} \omega_p = 1$.

If the weights ω_p are equal for all values of p, then ω_p is a constant that does not have any effect on the minimization, and thus can be ignored. If the weights are not equal for all values of p, the minimization problem can be reformulated in order to also ignore the weights. The reformulation consists of repeating each histogram h_{X_p} some number of times proportional to its weight ω_p. The barycenter h_{X_B} of the original set can then be computed as the barycenter of the new set of k' histograms $h'_{X_1}, h'_{X_2}, \ldots, h'_{X_{k'}}$, not taking the weights into account. For example, consider a set of $p = 3$ histograms $\{h_{X_1}, h_{X_2}, h_{X_3}\}$ with weights $\omega_1 = 0.2$, $\omega_2 = 0.3$ and $\omega_3 = 0.5$, respectively. The barycenter can be computed as the barycenter of a new set of 10 histograms, where $h'_{X_1} = h'_{X_2} = h_{X_1}$, $h'_{X_3} = h'_{X_4} = h'_{X_5} = h_{X_2}$, and $h'_{X_6} = \ldots = h'_{X_{10}} = h_{X_3}$, and where the weights are ignored. It can be shown that the solutions obtained using this method and those obtained by minimizing Eq. (A.7) with non-constant weights are the same for the two distances considered.

Therefore, the minimization problem in (A.7) can always be reformulated so that it does not take the weights into account, being reduced to

$$\underset{h_{X_B}}{arg\ min} \sum_{p=1}^{k} D(h_{X_B}, h_{X_p}). \tag{A.8}$$

If the squared Mallows distance (A.5) is considered, then the solution to the unweighted minimization is the barycentric histogram, $h_{X_B} = \{([I]_{lB}, \pi_{lB})\}$, where $[I]_{lB} = [c_{lB} - r_{lB}, c_{lB} + r_{lB}]$, with $l = 1, \ldots, m_k$. The barycentric histogram is given by

$$\begin{aligned} &\underset{h_{X_B}}{arg\ min} \sum_{p=1}^{k} D_M^2(h_{X_B}, h_{X_p}) \\ &= \underset{c_{lB}, r_{lB}}{arg\ min} \sum_{p=1}^{k} \sum_{l=1}^{m_k} \pi_l \\ &\quad \times \left[(c_{lp} - c_{lB})^2 + \frac{1}{3}(r_{lp} - r_{lB})^2 \right], \end{aligned} \tag{A.9}$$

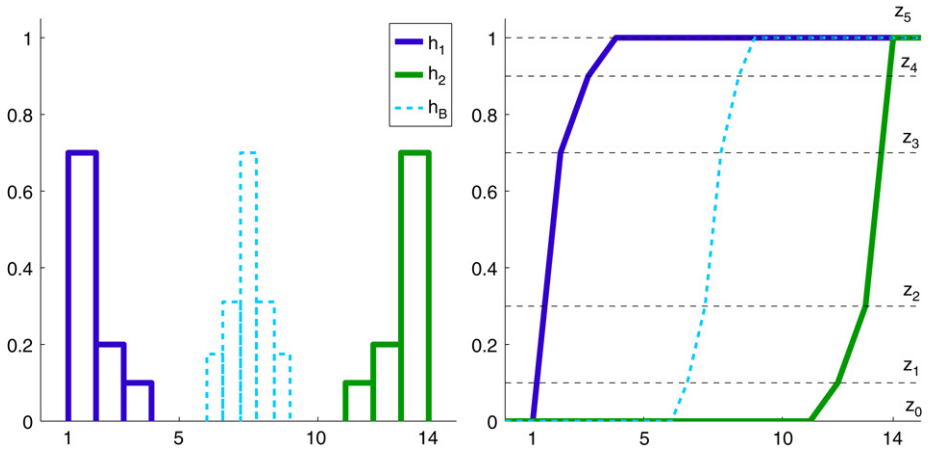

Fig. A.1. Mallows barycenter of $h_1 = \{([1, 2), 0.7); ([2, 3), 0.2); ([3, 4], 0.1)\}$ and $h_2 = \{([11, 12), 0.1); ([12, 13), 0.2); ([13, 14], 0.7)\}$. Density functions (left) and CDFs (right).

where m_k is the length of the vector $\mathbf{z}_k$. This vector is the result of sorting w without repetitions, w being the set of cumulative weights of the density functions of the k histograms considered, as shown in Section A.1.

In the case of the Wasserstein distance (A.6), h_{X_B} is computed as

$$\underset{h_{X_B}}{arg\ min} \sum_{p=1}^{k} D_W(h_{X_B}, h_{X_p}) = \underset{c_{lB}}{arg\ min} \sum_{p=1}^{k} \sum_{l=1}^{m_k} \pi_l |c_{lp} - c_{lB}|, \tag{A.10}$$

where m_k is the length of the vector $\mathbf{z}_k$. In this case, $\mathbf{z}_k$ is the result of sorting $w \cup v$ without repetitions, w being the set of cumulative weights of the density functions of the k histograms considered, and v being the set of the intersections of the CDFs of the histograms, as shown in Section A.1.

It is interesting to note that the minimization in Eq. (A.9) is a least squares problem, while the minimization in Eq. (A.10) is a least absolute deviations problem. Consequently, the minimum in the case of the Mallows distance (A.9) is the mean of the parameters

$$c_{lB} = \frac{\sum_{p=1}^{k} c_{lp}}{k} \quad \text{and} \quad r_{lB} = \frac{\sum_{p=1}^{k} r_{lp}}{k}, \tag{A.11}$$

for each bin $l = 1, \ldots, m_k$; while in the case of the Wasserstein distance, it is the median of the parameters

$$c_{lB} = \text{median}_p(c_{lp}), \quad \text{with } p = 1, \ldots, k \tag{A.12}$$

for each $l = 1, \ldots, m_k$. If c_{lq} with $q \in 1, \ldots, k$ is the median of the centers for the bin l of the barycenter, its corresponding radius will be $r_{lB} = r_{lq}$. As the optimal solution is a median, the barycenter is not unique if k is even.

Finally, the barycentric histogram is $h_{X_B} = \{([I]_{lB}, \pi_{lB})\}$, where

$$[I]_{lB} = [c_{lB} - r_{lB}, c_{lB} + r_{lB}], \quad l = 1, \ldots, m_k, \tag{A.13}$$

and where the weights associated with each bin are $\pi_{lB} = \pi_l$ with $l = 1, \ldots, m_k$.

It is important to note that the barycenter obtained using the Mallows distance behaves as a mean, i.e., the barycentric histogram averages the features of the k histograms considered. On the other hand, the one obtained using the Wasserstein distance resembles a median, since it is not unique and takes into account only the central 50% of the features of the k histograms considered.

Fig. A.1 shows the Mallows barycenter of two histograms. The figure also shows the CDFs of the histograms, because they help us to understand how the barycenter is computed. The horizontal lines in the CDFs chart represent the z values of the set

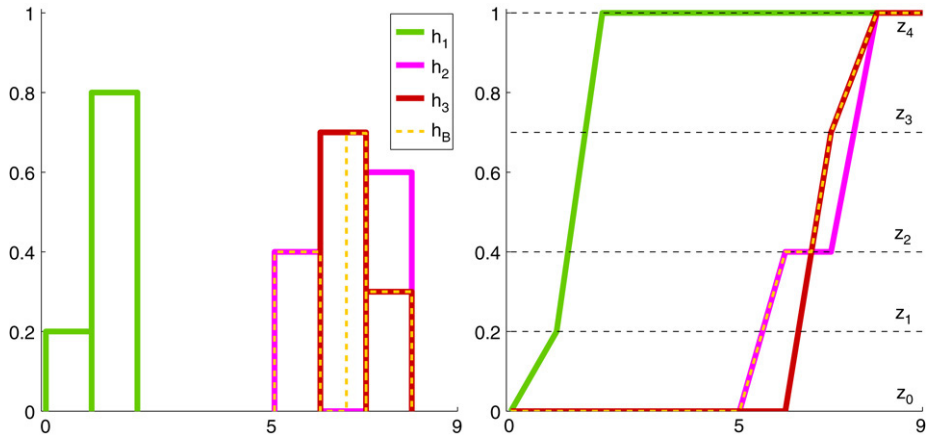

Fig. A.2. Wasserstein barycenter of $h_1 = \{([0, 1), 0.2); ([1, 2], 0.8)\}$, $h_2 = \{([5, 6), 0.4); ([7, 8], 0.6)\}$ and $h_3 = \{([6, 7), 0.7); ([7, 8], 0.3)\}$. Density functions (left) and CDFs (right).

z, which partitions the interval [0, 1] into regions. A bin of the barycenter is estimated in each of these regions, according to the formulae shown. It can be seen that the location and shape of the Mallows barycenter are the result of *averaging* the features of the CDFs considered. In this example, as k is even, the Wasserstein barycenter would not be unique. Consequently, any CDF contained within the CDFs of the histograms considered would be a valid Wasserstein barycenter.

Fig. A.2 shows the Wasserstein barycenter of three histograms. In this case, as k is odd, the Wasserstein barycenter is unique. In the CDF chart, the median-like behaviour of this barycenter can be seen: histogram h_1, whose CDF is located at the other end of the observed range in all the **z**-induced regions, has no effect on the barycenter.

References

Aparicio, T., Pozo, E., & Saura, D. (2002). The nearest neighbour method as a test for detecting complex dynamics in financial series: An empirical application. *Applied Financial Economics*, *12*(7), 517–525.

Arroyo, J., Maté, C., Muñoz San Roque, A., & Sarabia, A. (2008). Exponential smoothing methods for histogram time series based on histogram arithmetic, *Technical report*, Universidad Complutense de Madrid.

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.

Billard, L., & Diday, E. (2006). *Symbolic data analysis: conceptual statistics and data mining*. Chichester: Wiley & Sons.

Bock, H.-H. (2000). Dissimilarity Measures for Probability Distributions. In *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data* (pp. 153–164). Springer.

Brath, A., Montari, A., & Toth, E. (2002). Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. *Hydrology and Earth Systems Sciences*, *6*(4), 627–640.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Diday, E., & Noirhomme, M. (2008). *Symbolic data and the SODAS software*. Chichester: Wiley & Sons.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4), 863–883.

Engle, R. F., & Russell, J. Analysis of high frequency and transaction data. In *Handbook of financial econometrics*. North-Holland. http://home.uchicago.edu/~lhansen/handbook.htm (in press).

Farmer, J. D., & Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical Review Letters*, *8*(59), 845–848.

Fernández-Rodríguez, F., Sosvilla-Rivero, S., & Andrada-Félix, J. (1999). Exchange-rate forecasts with simultaneaous nearest-neighbour methods: evidence from the EMS. *International Journal of Forecasting*, *15*(4), 383–392.

Gibbs, A. L., & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, *70*(3), 419–435.

González-Rivera, G., Lee, T. H., & Mishra, S. (2008). Jumps in cross-sectional rank and expected returns: A mixture model. *Journal of Applied Econometrics*, *23*(5), 585–606.

Hall, S. G., & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*(1), 1–13.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Irpino, A., & Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In *Data science and classification, proceedings of the IFCS 2006* (pp. 185–192). Berlin: Springer.

Jayawardena, A. W., Li, W. K., & Xu, P. (2002). Neighbourhood selection for local modelling and prediction of hydrological time series. *Journal of Hydrology*, *258*, 40–57.

Levina, E., & Bickel, P. J. (2001). The Earth Mover's distance is the Mallows distance: Some insights from statistics. In *Proceedings of the 8th IEEE international conference on computer vision*: *Vol. 2* (pp. 251–256).

Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, *43*(2), 508–515.

Meade, N. (2002). A comparison of the accuracy of short term foreign exchange forecasting methods. *International Journal of Forecasting*, *18*(1), 67–83.

Pasley, A., & Austin, J. (2004). Distribution forecasting of high frequency time series. *Decision Support Systems*, *37*(4), 501–513.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Schweizer, B. (1984). Distributions are the numbers of the future. In *Proceedings of the mathematics of fuzzy systems meeting* (pp. 137–149). Naples: University of Naples.

Sorjamaa, A., Reyhani, N., & Lendasse, A. (2005). Input and structure selection for k-NN approximator. In *Computational intelligence and bioinspired systems*, *8th international work-conference on artificial neural networks, IWANN 2005, Lecture notes in computer science* (pp. 985–992). Springer.

Tay, A. S., & Wallis, K. F. (2000). Density forecasting: a survey. *Journal of Forecasting*, *19*(4), 235–254.

Verde, R., & Irpino, A. (2007). Dynamic clustering of histogram data: using the right metric. In *Selected contributions in data analysis and classification* (pp. 123–134). Springer.

Wand, M. P. (1997). Data-based choice of histogram bin width. *The American Statistician*, *51*(1), 59–64.

Yakowitz, S. (1987). Nearest-neighbour methods for time series analysis. *Journal of Time Series Analysis*, *8*(2), 235–247.