Forecasting with interval and histogram data Some financial applications*

Javier Arroyo
Universidad Complutense de Madrid
Department of Computer Science and Artificial Intelligence
28040 Madrid, Spain

Gloria González-Rivera
University of California, Riverside
Department of Economics
Riverside, CA 92521

Carlos Maté
Universidad Pontificia de Comillas
Institute of Technological Research ETSI (ICAI)
28015 Madrid, Spain

*Corresponding author: Gloria González-Rivera, Department of Economics, University of California, Riverside, CA 92521-0427, email gloria.gonzalez@ucr.edu, tel (951) 827-1590, fax (951) 827-5685.

ABSTRACT

Data sets across many disciplines are becoming consistently large and they bring with them the need of new methods for processing information. We introduce the analysis of interval-valued and histogram-valued data sets as an alternative to classic single-valued data sets and we show the promise of this approach on dealing with economic and financial data. Being our current focus the prediction problem we explore two different venues to produce a forecast with interval time series (ITS) and histogram time series (HTS). For ITS, we adapt classical regression methods and time series strategies for model building and prediction. For ITS and HTS, we implement filtering techniques, such as smoothing, and non-parametric methods such as the k-NN. We need interval arithmetic in ITS and the concept of a barycentric histogram in HTS to compute the appropiate averages required by smoothing and k-NN. The assessment of the forecast error also requires the introduction of dissimilarity measures like a kernel based distance for ITS and the Wasserstein and Mallows distances for HTS. We apply the proposed methods to predict the daily interval-valued dispersion for the level of SP500 index and the weekly cross-sectional histogram of the returns to the constituents of the SP500 index. Overall, k-NN methods perform very well.

Key Words: Interval-valued data, histogram-valued data, interval arithmetic, dissimilarity measures, exponential smoothing, k-NN, Wasserstein distance, Mallows distance.

JEL Classification: C22, C53

Article Outline

1 Introduction

In Economics we customarily deal with classical data sets. When we collect information on a set of variables of interest, either in a cross-sectional or/and time series framework, our sample information is a collection of data points $\{y_i\}$, $i = 1 \dots n$ or $\{y_t\}$, $t = 1 \dots T$ where y_i or $y_t \in \mathbb{R}$ takes a single value in $\mathbb{R}$. In many instances, the single value is the result of an aggregation procedure, spatial or temporal, over information collected at a very disaggregated level. Some pertinent examples follow.

In financial markets the price of an asset (stocks, bonds, exchange rates, etc.) is observed at a very high frequency, i.e. tick-by-tick, however there is a huge number of studies where the analysis is performed at the daily frequency using the closing price, or even at lower frequencies such as weekly or monthly. It may be claimed that tick-by-tick pricing will generate a huge amount of data from which it will be difficult to discriminate information from noise, but on the other extreme, by analyzing just closing prices we will be discarding valuable intraday information. We can think of alternative ways of collecting information, for instance, we can gather the maximum and minimum prices in a day so that the information to be analyzed will come in an interval format; or the daily interquartile prices such that the interval will run from the price at the 25% quartile to the price at the 75% quartile; or we can construct daily histograms with all the intraday prices. In these cases the data point is not longer a single value but a collection of values represented by the daily low/high interval, or the interquartile interval, or the daily histogram. The intervals or the histograms, when indexed by time, will constitute an interval time series or a histogram time series.

Another instance refers to the information collected by national statistical institutes in relation to income and population dynamics. The census surveys provide socio-economic information on all individuals in a nation that it is customarily disseminated in an aggregated format, for instance a time series of average income per capita. The objective of these national surveys is not to follow the dynamics of single individuals, which most likely will be different from one period to the next, but the dynamics of a collective. However summarizing national information by averages, though informative, is a poor approach that throws away the internal variation provided by the disaggregated information about the single units. Once

more, disseminating the data in a richer format such as intervals or histograms will provide a more complete picture of income and population dynamics. There are many other areas such as marketing, environmental sciences, quality control, medical sciences, etc. in which the information is rich enough to make the object of analysis not the single-valued variable but the interval-valued or the histogram-valued variable.

Interval and histogram-valued data can be classified as symbolic data sets as opposed to classical data sets. Symbolic data is a proposal to deal with the massive information contained in nowadays super large data sets found across many disciplines. While the analysis of these data sets requires some summary procedure to bring them to a manageable size, the objective is to retain as much of their original knowledge as possible. An extensive review of this new field, which started in the late 1980s and early 1990s, is provided by Billard and Diday (2003, 2006), who define the complexity of symbolic data, review the current methods of analysis, and state the challenges that lie ahead.

Economics and Business are disciplines in which data sets are becoming consistently larger due to sophisticated information systems that collect and storage huge amount of data. However the development of new methodologies to deal with the characteristics of large data sets is moving at a slower pace. A case on time is the aforementioned high-frequency financial data and the challenges brought by it such as irregularly spaced observations with strong intraday patterns and a complex dependence structure. There are other examples in the economics literature that emphasize the richness of the data, though eventually the analysis is performed within the boundaries of classical inferential methods. For instance, the article by Zellner and Tobias (2000) provides the time series of the median and interquartile range of the industrial production growth rates of 18 countries but eventually the authors focus on the single-valued time series of the median growth rates. The article by González-Rivera et al. (2008) presents a stylized time series of cross-sectional returns of the constituents of the SP500 index grouped in histograms. However, the authors focus on the dependence structure of the single-valued time series of the time-varying cross-sectional ranks (VCR). Both of these instances could be viewed from the perspective of symbolic data: in Zellner and Tobias (2000) the data is an interval-valued time series and in González-Rivera et al. (2008) is a histogram-valued time series.

[FIGURES 1 & 2]

This chapter focuses on the forecasting of interval and histogram-valued data. The surveys and review articles by Diday and his co-authors focus on descriptive and multivariate methods of analysis adapted from the classical statistical methodology. To our knowledge, the development of forecasting methods for interval and histogram-valued data is in its infancy so that this article is a contribution to that end. We start with a preliminary section defining the structure of the data and basic descriptive statistics such that the mean, variance, covariance, and correlation. There are two main sections, one for interval data and another for histogram data. In the first, we review how classical regression methods can be adapted to analyze intervals. The main insight is that the interval can be uniquely defined by its center and radius or by its minimum and maximum, so that we construct two time series to which classical methods can be applied. In this vein, we build a system, either VAR or VEC models, from which an interval forecast will be obtained. In a different approach based on the arithmetic of intervals and on notions of distances between intervals, we adapt classical filtering techniques like the exponential smoothing and non-parametric techniques like the k-NN algorithm to produce the interval forecast. In the second main section, we deal with histogram-valued data. In this case the object of analysis is considerably more difficult to analyze and we focus exclusively on the adaptation of smoothing techniques and the k-NN. To construct a histogram forecast, we will not base our operations on the arithmetic of histograms but on the key idea of the "barycentric" histogram as the "average" measure. We should stress that no attempt has been made, either with a time series of intervals or histogram, to uncover the data generating mechanism but rather to forecast the future under the premise that it should not be very far from some average (weighted or unweighted) of the past.

2 Interval data

In this section we will define interval data and the interval random variable. As a foundation for the forthcoming analysis we succinctly introduce the algebra of intervals. We will focus on the empirical first and second moments of the interval random variable. The main objective

of this section is to discuss (i) regression analysis with interval data, and (ii) the forecasting problem. A financial application will showcase the contribution of (i) and (ii) to the modeling of economic and financial data. While we will not discuss the nature of interval data, we acknowledge that there are many reasons why interval data may arise. Among others, interval data is generated when the data collection process genuinely produces intervals, or when there are not exact numerical values to quantify a variable, or when there is uncertainty of any kind in the values of the variable, or when variability of a variable is the focus of analysis, or when the measurement tools produce measurement errors. Regardless of the origin, the researcher will be facing data that comes with an interval format and this is the primary object of analysis.

2.1 Preliminaries

We start with the basic notion of an interval following Kulpa (2006). Let $(E, \leq)$ be a partially ordered set. An interval is generally defined as follows:

Definition 1 . An interval $[a]$ over the base set $(E, \leq)$ is an ordered pair $[a] = [a_L, a_U]$ where $a_L, a_U \in E$ are the endpoints or bounds of the interval such that $a_L \leq a_U$.

The interval is called degenerate when $a_L = a_U$, in which case the interval reduces to a point. An interval is the set of elements bounded by the endpoints, these ones included, namely $[a] = \{e \in E \mid a_L \leq e \leq a_U\}$. When the base set E is the set of real numbers $\mathbb{R}$, the intervals are subsets of the real line $\mathbb{R}$.

An equivalent representation of an interval is given by the center (midpoint) and radius (half range) of the interval, namely $[a] = \langle a_C, a_R \rangle$ where $a_C = (a_L + a_U)/2$ and $a_R = (a_U - a_L)/2$.

Basic interval arithmetic. In order to proceed with our analysis we need an algebra to operate with intervals. Basic interval arithmetic (Moore, 1966; Moore et al., 2009) is based on the following principle: let $[a]$ and $[b]$ be two intervals and $\square$ be an arithmetic operator, then $[a]\square[b]$ is the smallest interval which contains $a\square b$, $\forall a \in [a]$ and $\forall b \in [b]$. Interval addition, subtraction, multiplication and division are particular cases of this principle and

are defined by

$$[a] + [b] = [a_L + b_L, a_U + b_U] \tag{1}$$

$$[a] - [b] = [a_L - b_U, a_U - b_L] \tag{2}$$

$$\begin{aligned} [a] \cdot [b] &= [\min\{a_L \cdot b_L, a_L \cdot b_U, a_U \cdot b_L, a_U \cdot b_U\}, \\ &\quad \max\{a_L \cdot b_L, a_L \cdot b_U, a_U \cdot b_L, a_U \cdot b_U\}] \end{aligned}$$

$$[a]/[b] = [a] \cdot (1/[b]), \text{with } 1/[b] = [1/b_U, 1/b_L] \tag{3}$$

It is worth noting that interval arithmetic subsumes the classical one, in the sense that, if the operands are degenerate intervals, the result of interval operations will be equal to the result obtained by the single number arithmetic. In interval arithmetic, addition and multiplication satisfy the associative and commutative properties. The distributive property does now always hold, but the subdistributive property is satisfied, which is defined as

$$[a]([b] + [c]) \subseteq [a][b] + [a][c]. \tag{4}$$

If $[a]$ is a degenerate interval, then this property becomes the distributive property. The interval arithmetic is key for the development of regression techniques and for the adaptation of forecasting methods to interval data.

Interval random variable. We proceed with the definition of an interval random variable. Let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is the set of elementary events, $\mathcal{F}$ is the σ-field of events and $P : \mathcal{F} \to [0, 1]$ the σ-additive probability measure; and define a partition of Ω into sets $A(x)$ such $A_X(x) = \{\omega \in \Omega | X(\omega) = x\}$, where $x \in [x_L, x_U]$, then

Definition 2. A mapping $X : \mathcal{F} \to [x_L, x_U] \subset \mathbb{R}$ such that for each $x \in [x_L, x_U]$ there exists a set $A_X(x)$ is called an interval random variable.

Descriptive statistics. The descriptive statistics of an interval random variable are proposed by Bertrand and Goupil (2000). For an interval random variable X, suppose that we have a sample of m individuals ($i = 1, 2, ..., m$) and for each i, an interval data point $[x_i] \equiv [x_{Li}, x_{Ui}]$. A key assumption for the forthcoming descriptive statistics is that the values in a given interval, i.e. $x_{Li} \leq x \leq x_{Ui}$, are uniformly distributed within the interval. Furthermore, we assume that each individual has the same probability $1/m$ of

being observed. Then, the *empirical distribution function* $F_X(x)$ is a mixture of m uniform distributions

$$F_X(x) = \frac{1}{m}\left\{ \sum_{i:x\in[x_i]} \frac{x - x_{Li}}{x_{Ui} - x_{Li}} + N_X(x \geqslant x_{Ui}) \right\} \qquad x \in \mathbb{R}$$

where the notation $i : x \in [x_i]$ in the summation sign means that the sum runs for those individuals for which $x \in [x_i]$, if $x \notin [x_i]$ we only count the number of intervals $N_X(x \geqslant x_{Ui})$ for which the condition $x \geqslant x_{Ui}$ is met. As usual, by taking the derivative with respect to x we will obtain the empirical density function

$$f_X(x) = \frac{1}{m} \sum_{i:x\in[x_i]} \frac{1}{x_{Ui} - x_{Li}} = \frac{1}{m} \sum_{i:x\in[x_i]} \frac{I(x \in [x_i])}{||\ [x_i]\ ||} \qquad x \in \mathbb{R}. \tag{5}$$

where $I(x \in [x_i])$ is an indicator function that takes the value 1 when $x \in [x_i]$ and zero otherwise; and $||\ [x_i]\ ||$ is the length of the interval $[x_i]$.

Based on the density function (5), the sample mean is obtained by solving the following integral

$$\bar{X} = \int_{-\infty}^{\infty} x f(x) dx = \frac{1}{m} \sum_{i:x\in[x_i]} \frac{1}{x_{Ui} - x_{Li}} \int_{x_{Li}}^{x_{Ui}} x dx = \frac{1}{2m} \sum_i (x_{Ui} + x_{Li}) = \frac{1}{m} \sum_i x_{C_i} \tag{6}$$

concluding that the sample mean of an interval random variable is the average of the centers of the intervals in the sample. Analogously, the sample variance is calculated by solving the integral

$$S_X^2 = \int_{-\infty}^{\infty} (x - \bar{X})^2 f(x) dx = \left(\int_{-\infty}^{\infty} x^2 f(x) dx \right) - \bar{X}^2 \tag{7}$$

with

$$\int_{-\infty}^{\infty} x^2 f(x) dx = \frac{1}{m} \sum_{i:x\in[x_i]} \frac{1}{x_{Ui} - x_{Li}} \int_{x_{Li}}^{x_{Ui}} x^2 dx = \frac{1}{3m} \sum_i \frac{(x_{Ui}^3 - x_{Li}^3)}{x_{Ui} - x_{Li}} \tag{8}$$

and substituting (8) in (7), the sample variance becomes

$$\begin{aligned} S_X^2 &= \frac{1}{3m} \sum_i \frac{(x_{Ui}^3 - x_{Li}^3)}{x_{Ui} - x_{Li}} - \frac{1}{4m^2} \left[\sum_i (x_{Ui} + x_{Li}) \right]^2 = \\ &= \frac{1}{3m} \sum_i (x_{Ui}^2 + x_{Ui}x_{Li} + x_{Li}^2) - \frac{1}{4m^2} \left[\sum_i (x_{Ui} + x_{Li}) \right]^2 \end{aligned} \tag{9}$$

The sample variance combines the variability of the centers as well as the variability within each interval. When the interval is degenerate, both sample moments, the mean and the variance, collapse to the sample mean and variance of the classical data.

Now suppose that we have two interval random variables Y and X for which we collect a sample of intervals $([x_i],[y_i])$ for $i = 1, 2, ...m$. The interval data point i is a rectangle centered in the centers of $[x_i]$ and $[y_i]$ and whose sides are equal to the length of the respective intervals. A graphical representation of this data is provided in Figure 3.

[FIGURE 3]

The question of interest is to quantify the dependence of the two interval random variables. Following the classical analysis, we will focus on the covariance and correlation coefficient of Y and X. Analogously to the univariate case (5), we define the bivariate empirical density function as

$$f_{X,Y}(x,y) = \frac{1}{m} \sum_{i:x\in[x_i],y\in[y_i]} \frac{I((x,y) \in ([x_i],[y_i]))}{||\ ([x_i],[y_i])\ ||} \qquad x, y \in \mathbb{R}. \tag{10}$$

where $I((x,y) \in ([x_i],[y_i]))$ is an indicator function that takes the value one when the point (x,y) is inside the rectangle $([x_i],[y_i])$ and zero otherwise; and $||\ ([x_i],[y_i])\ ||$ is the area of the rectangle $([x_i],[y_i])$. Based on (10), Billard and Diday (2006) define the empirical covariance between two interval random variables as

$$Cov(X,Y) = \frac{1}{3m} \sum_i G_{Xi} G_{Yi} (Q_{Xi} Q_{Yi})^{1/2}, \tag{11}$$

where

$$\begin{aligned} Q_{Xi} &= (x_{Li} - \bar{X})^2 + (x_{Li} - \bar{X})(x_{Ui} - \bar{X}) + (x_{Ui} - \bar{X})^2 \\ Q_{Yi} &= (y_{Li} - \bar{Y})^2 + (y_{Li} - \bar{Y})(y_{Ui} - \bar{Y}) + (y_{Ui} - \bar{Y})^2 \end{aligned} \tag{12}$$

and

$$G_{Xi} = \begin{cases} -1, & \text{if } X_{Ci} \le \bar{X}, \\ 1, & \text{if } X_{Ci} > \bar{X}, \end{cases} \qquad G_{Yi} = \begin{cases} -1, & \text{if } Y_{Ci} \le \bar{Y}, \\ 1, & \text{if } Y_{Ci} > \bar{Y}, \end{cases} \tag{13}$$

The measure (11) reduces to (9) when $X = Y$. In addition, if all intervals are degenerate so that the sample becomes a collection of single points, the expression (11) collapses to the classical covariance measure. The empirical correlation coefficient follows accordingly as

$$\rho_{XY} = \frac{Cov(X,Y)}{S_X S_Y} \tag{14}$$

The reader is referred to Billard and Diday (2006) for a variety of examples to illustrate the numerical calculations of the sample mean, variance, covariance, and correlation coefficient.

2.2 The regression problem

In this section we will review the analysis of a regression model with interval data. The classical regression model can be adapted to interval data by focusing on the centers of the interval, or on the maximum and minimum of the interval, or on the center and radius of the interval. The advantage of this approach is that statistical inference is readily available.

The simplest approach to estimate a regression model with interval data is provided by Billard and Diday (2000). It consists of fitting a regression line to the centers of the intervals, $y_{Ci} = \beta' x_{Ci} + \varepsilon_{Ci}$, so that the objective function to minimize is

$$\min_{\hat{\beta}} \sum_i \hat{\varepsilon}_{Ci}^2 = \sum_i (y_{Ci} - \hat{\beta}' x_{Ci})^2, \tag{15}$$

the solution to this problem is the classical least squares estimator $\hat{\beta} = (X_C' X_C)^{-1} X_C' Y_C$ and standard statistical inference will apply under the standard assumptions about the error term of the regression. Though this model will provide information about the average centrality of the intervals, it disregards the range of the intervals that is an important feature of interval data.

There are several proposals aimed to incorporate the length of the interval into the analysis. Brito (2007) proposes to minimize the following objective function

$$\min_{\hat{\beta}} \sum_i (\hat{\varepsilon}_{Li}^2 + \hat{\varepsilon}_{Ui}^2) = \sum_i (y_{Li} - \hat{\beta}' x_{Li})^2 + \sum_i (y_{Ui} - \hat{\beta}' x_{Ui})^2, \tag{16}$$

which is equivalent to run two constrained (same regression coefficients) regressions on the lower bounds $y_{Li} = \beta' x_{Li} + \varepsilon_{Li}$ and the upper bounds $y_{Ui} = \beta' x_{Ui} + \varepsilon_{Ui}$ of the intervals. For the case of one regressor model, the OLS estimators have the following expression

$$\begin{aligned} \hat{\beta}_1 &= \frac{\tilde{S}_{XY}}{\tilde{S}_X^2} = \frac{\frac{1}{2m}\sum_i \left[(x_{Li} - \bar{X})(y_{Li} - \bar{Y}) + (x_{Ui} - \bar{X})(y_{Ui} - \bar{Y})\right]}{\frac{1}{2m}\sum_i \left[(x_{Li} - \bar{X})^2 + (x_{Ui} - \bar{X})^2\right]} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned} \tag{17}$$

where $\bar{X}$ and $\bar{Y}$ are given in (6). Brito (2007) calls the numerator $\tilde{S}_{XY}$ the co-dispersion measure and the denominator $\tilde{S}_X^2$ the dispersion measure, which is different from (9). This regression line passes through the average center $(\bar{X}, \bar{Y})$ but the slope is guided by the range of the intervals, whose effect is summarized by the sum of the covariance between the lower bounds of $[x_i]$ and $[y_i]$ and the covariance between the upper bounds of $[x_i]$ and $[y_i]$. In other words, the researcher collects a sample of points as (x_{Li}, y_{Li}) and (x_{Ui}, y_{Ui}) and fits a unique regression line to the full sample. Equivalently, we can understand Brito's proposal as a system of equations

$$\underset{2m\times 1}{\begin{bmatrix} Y_L \\ Y_U \end{bmatrix}} = \underset{2m\times k}{\begin{bmatrix} X_L \\ X_U \end{bmatrix}} \underset{k\times 1}{\beta} + \underset{2m\times 1}{\begin{bmatrix} \varepsilon_L \\ \varepsilon_U \end{bmatrix}} \tag{18}$$

and if the only interest of the researcher is to find a consistent estimator of β, then the OLS estimator will suffice

$$\hat{\beta}_{OLS} = \left[X_L' X_L + X_U' X_U\right]^{-1} \left[X_L' Y_L + X_U' Y_U\right] \tag{19}$$

but if there is a need for statistical inference, then we should take into account the properties of the error term. The vector ε is likely to be heteroscedastic, i.e.

$$\Omega = \begin{pmatrix} \sigma_L^2 I & \sigma_{LU} I \\ \sigma_{LU} I & \sigma_U^2 I \end{pmatrix} \tag{20}$$

where $I_{m\times m}$ is the identity matrix. In this case, we recommend applying the GLS (or FGLS) estimator

$$\hat{\beta}_{GLS} = \left[X'\Omega^{-1}X\right]^{-1} \left[X'\Omega^{-1}Y\right] \tag{21}$$

An alternative proposal by Billard and Diday (2000, 2002) is to estimate two different regression lines, one for the minima and another for the maxima of the intervals with no

restrictions across lines as in

$$\begin{aligned} y_{Li} &= \beta_L' x_{Li} + \varepsilon_{Li} \\ y_{Ui} &= \beta_U' x_{Ui} + \varepsilon_{Ui} \end{aligned} \tag{22}$$

The estimation of the model proceeds by minimizing the following objective function

$$\min_{\hat{\beta}_L, \hat{\beta}_U} \sum_i (\hat{\varepsilon}_{Li}^2 + \hat{\varepsilon}_{Ui}^2) \tag{23}$$

which is equivalent to perform two separate minimizations, $\min_{\hat{\beta}_L} \sum_i \hat{\varepsilon}_{Li}^2$ and $\min_{\hat{\beta}_U} \sum_i \hat{\varepsilon}_{Ui}^2$ because of the absence of cross-equation restrictions. In practice, we can have the case that for some observations the estimated dependent variable is such that $\hat{y}_{Li} > \hat{y}_{Ui}$, which obviously contradicts the logic of interval data. The corresponding estimator to this minimization is the OLS estimator but it will not be the most efficient estimator because very likely ε_{Li} and ε_{Ui} will be correlated given that $y_{Li} \leq y_{Ui}$ and $x_{Li} \leq x_{Ui}$. If statistical inference is of interest, we propose to express this approach as a system of seemingly unrelated regression equations (SURE)

$$\underset{2m\times 1}{\begin{bmatrix} Y_L \\ Y_U \end{bmatrix}} = \underset{2m\times 2k}{\begin{bmatrix} X_L & 0 \\ 0 & X_U \end{bmatrix}} \underset{2k\times 1}{\begin{bmatrix} \beta_L \\ \beta_U \end{bmatrix}} + \underset{2m\times 1}{\begin{bmatrix} \varepsilon_L \\ \varepsilon_U \end{bmatrix}} \tag{24}$$

to take into account the properties of the error term. In this case the GLS estimator $\hat{\beta}_{GLS} = [X'\Omega^{-1}X]^{-1}[X'\Omega^{-1}Y]$ is the most appropriate with Ω given in (20).

The last approach based on classical regression techniques is proposed by Lima Neto and de Carvalho (2008). It consists on running two independent regression models for the center and the radius (or range) of the intervals. Recall that $x_{Ci} = (x_{Li} + x_{Ui})/2$ and $x_{Ri} = (x_{Ui} - x_{Li})/2$. The model is

$$\begin{aligned} y_{Ci} &= \beta_C' x_{Ci} + \varepsilon_{Ci} \\ y_{Ri} &= \beta_R' x_{Ri} + \varepsilon_{Ri} \end{aligned} \tag{25}$$

and the objective function to minimize is

$$\min_{\hat{\beta}_C, \hat{\beta}_R} \sum_i (\hat{\varepsilon}_{Ci}^2 + \hat{\varepsilon}_{Ri}^2) \tag{26}$$

which, in the absence of cross-equation restrictions and with spherical disturbances, is equivalent to perform two separate minimizations, $\min_{\hat{\beta}_C} \sum_i \hat{\varepsilon}_{C_i}^2$ and $\min_{\hat{\beta}_R} \sum_i \hat{\varepsilon}_{Ri}^2$. The corresponding estimator is the classical OLS but the properties of the error term may dictate the choice of a GLS estimator, within a SURE system, as more appropriate than the OLS estimator. Other estimators as MLE or QMLE can also be implemented. However the radius, being strictly positive, will not be normally distributed and a MLE estimator based on multivariate normality of the vector $(\varepsilon_{Ci}, \varepsilon_{Ri})'$ will be at least highly inefficient.

Figures 4, 5, and 6 describe the graphical differences among the three regression lines proposed by Billard and Diday (2000, 2002) and Brito (2007). The proposal by Lima Neto and de Carvalho (2008) cannot be graphed in the same set of coordinates (X, Y).

[FIGURES 4, 5, & 6]

2.3 The prediction problem

In this section we define an interval-valued time series (ITS), we propose an approach to measure dissimilarities between intervals in ITS, and we implement forecasting methods for ITS based on smoothing filters and non-parametric estimators like the k-NN. Neither of these two approaches aims to specify a model for an ITS that may approximate a hidden data generating mechanism but rather they should be viewed as automatic procedures to extract information from a noisy signal from which eventually we can extrapolate a future value.

Definition 3. An interval-valued stochastic process is a collection of interval random variables that are indexed by time, i.e. $\{X_t\}$ for $t \in T \subset \mathbb{R}$, with each X_t following Definition 2.

An interval-valued time series is a realization of an interval-valued stochastic process and it will be equivalently denoted as $\{[x_t]\} = \{[x_{Lt}, x_{Ut}]\} = \{\langle x_{Ct}, x_{Rt}\rangle\}$ for $t = 1, 2, ...T$.

2.3.1 Accuracy of the forecast

As it is customary in classical time series, the assessment of a forecast is a function of the difference between the realized value and the forecast value. In ITS, one may be tempted to calculate the difference $[x]_{t+1} - [\hat{x}]_{t+1}$ but, because the interval difference bounds all the

possible results when considering single real numbers in the two operands, see property (2), the resulting interval will have an extreme width and thus, it will not be deemed appropriate to measure the accuracy of a forecast (Arroyo et al., 2008). The following example will clarify this point.

Suppose that $[x]_{t+1} = [\hat{x}]_{t+1} = [a_L, a_U], a_L < a_U$. Since the realized value is identical to the forecast, the forecast error must be zero $[x]_{t+1} - [\hat{x}]_{t+1} = [0, 0]$. If this difference is the interval difference (2), then it must be the case that $[A] = [a, a]$ with $a \in \mathbb{R}$, which is a contradiction with our assumption $a_L < a_U$. If $[a_L, a_U]$ is a non-degenerate interval, the result of the difference is an interval with the center in zero and with a length twice the length of the interval $[a_L, a_U]$, i.e. if $[a_L, a_U] = [1, 2]$, $[x]_t - [\hat{x}]_t = [-1, 1]$.

Given these shortcomings, Arroyo and Maté (2006) propose the use of distances to quantify the dissimilarity (the forecast error) between the realized and the forecast intervals. The properties of distances, i.e., non-negativity, symmetry, and triangle inequality, make them a suitable tool for this purpose.

Arroyo and Maté (2006) consider different distances, among them, the Hausdorff distance that is very popular for interval data and it is defined as

$$D_H([x], [y]) = \max(|x_L - y_L|, |x_U - y_U|) = |x_C - y_C| + |x_R - y_R|. \tag{27}$$

However, according to González et al. (2004) , the Hausdorff distance is not able to discriminate between some intervals. For example, if $A = [0, 1]$ and $B = [2 - h, 3]$ with $h \in [0, 4]$, then $D_H(A, B) = 2, \ \forall h \in [0, 4]$. Hence, these authors propose a kernel based distance defined as

$$\begin{aligned} D_K([x], [y]) &= \frac{1}{\sqrt{2}} \sqrt{(x_L - y_L)^2 + (x_U - y_U)^2} \\ &= \sqrt{(x_C - y_C)^2 + (x_R - y_R)^2} \end{aligned} \tag{28}$$

which can be understood as an Euclidean-like distance considering the description of the intervals by their minimum and their maximum or, alternatively, by their center and by their radii. There is a large number of distances proposed in the literature, each with its advantages and disadvantages so that their use will depend on the needs of the researcher.

In the forthcoming sections we will implement the Euclidean-type distance because of its intuitive and mathematical appeal.

Now, the assessment of a forecast will proceed by the choice of a distance measure and a loss function. Given a realized and a forecast ITS, $\{[x]_t\}$ and $\{[\hat{x}]_t\}$ with $t = 1, ..., T$, Arroyo et al. (2008) propose the Mean Distance Error to quantify the accuracy of the forecast

$$MDE^q(\{[x]_t\}, \{[\hat{x}]_t\}) = \left(\frac{\sum_{t=1}^{T} (D_X([x]_t, [\hat{x}]_t))^q}{T} \right)^{\frac{1}{q}}, \tag{29}$$

where D_X is a distance measure, for instance D_K , and q is the order of the distance, such that for $q = 1$ the mean distance error is similar in spirit to the mean absolute error (MAE) loss function, and for $q = 2$ to the root mean squared error (RMSE) loss function. Other loss functions, statistical or economic/business based, can also be chosen to evaluate a forecast. The important point is that the quantification of the error should be based on a distance measure.

2.3.2 Smoothing methods

Smoothing is a filtering technique that consists on averaging values of a time series, and by doing that, removing noise. These methods are easy to implement and they constitute a benchmark to evaluate the forecasting ability of more sophisticated methods (Gardner, 2006). The simplest smoothing is either a moving average or a weighted moving average. With the help of the arithmetic of intervals, it is relatively easy to adapt these smoothing procedures to ITS.

Moving average. Given an ITS from an interval-valued stochastic process $\{X_t\}$ with $t = 1, ..., T$, the forecast for the $t + 1$ period of a moving average of order q is a weighted average of the last q intervals of the ITS

$$[\hat{x}]_{t+1} = \omega_1 [x]_t + \omega_2 [x]_{t-1} + \ldots + \omega_q [x]_{t-(q-1)}, \tag{30}$$

where $\sum_{i=1}^{q} \omega_i = 1$ and $\omega_i \geq 0, \forall i$.

In a simple moving average, equal weights are assigned to each interval, i.e. $\omega_i = \frac{1}{q}$. In a weighted moving average, the weights may decrease arithmetically so that $\omega_i = \frac{q-i+1}{\sum_{i=1}^{q} i}$; or

the weights may decrease exponentially so that $\omega_i = \alpha(1-\alpha)^{i-1}$ with $\alpha = \frac{2}{q+1}$, for q large enough.

Exponential smoothing. The moving average with exponentially decreasing weights is equivalent to the simple exponential smoothing, which in recursive form is written as

$$[\hat{x}]_{t+1} = \alpha[x]_t + (1-\alpha)[\hat{x}]_t. \tag{31}$$

where $\alpha \in [0,1]$. This representation weights the most recent observation and its forecast. In classic time series, the simple exponential smoothing can be equivalently represented in error correction form. However, with ITS both representations are not equivalent due to the properties of the interval arithmetic. To understand this difference, let us write the error correction representation

$$[\hat{x}]_{t+1} = [\hat{x}]_t + \alpha[e]_t, \tag{32}$$

where $[e]_t$ would be the interval error in t, $[e]_t = [x]_t - [\hat{x}]_t$. Due to the subdistributive property (4) of interval arithmetic, the relation between both expressions is the following

$$\alpha[x]_t + (1-\alpha)[\hat{x}]_t \subseteq \alpha[x]_t - \alpha[\hat{x}]_t + [\hat{x}]_t = [\hat{x}]_t + \alpha([x]_t - [\hat{x}]_t) \tag{33}$$

which means that the recursive form yields tighter intervals than the error correction form. Due to this fact, the error correction form should not be considered in ITS forecasting. In addition, the error correction representation is not equivalent to the moving average with exponentially decreasing weights, while the recursive form is. By backward substitution in (31), the exponentially weighted moving average becomes

$$[\hat{x}]_{t+1} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1}[x]_{t-(j-1)}. \tag{34}$$

Since the interval arithmetic subsumes the classical arithmetic, the smoothing methods for ITS subsume those for classic time series, so that if the intervals in the ITS are degenerated then the smoothing results will be identical to those obtained with the classical smoothing methods. When using (31), all the components of the interval –center, radius, minimum, and maximum– are equally smoothed, i.e.

$$\hat{x}_{\Gamma,t+1} = \alpha x_{\Gamma,t} + (1-\alpha)\hat{x}_{\Gamma,t} \quad \text{where } \Gamma \in \{L, U, C, R\} \tag{35}$$

Additional smoothing procedures, like exponential smoothing with trend, or damped trend, or seasonality, can be adapted to ITS following the same principles presented in this section.

In Figure 7, we present an example of an ITS that has been exponentially smoothed as in (31) with $\alpha = 0.4$. As expected, in the smoothed ITS both the position and the width of the intervals show less variability than in the original ITS.

[FIGURE 7]

2.3.3 k-NN method

The k-Nearest Neighbors (k-NN) method is a classic pattern recognition procedure that can be used for time series forecasting (Yakowitz, 1987). The k-NN forecasting method in classic time series consists of two steps: identification of the k sequences in the time series that are more similar to the current one, and computation of the forecast as the weighted or unweighted average of the k-closest sequences determined in the previous step.

The adaptation of the k-NN method to forecast ITS consists of the following steps:

1. The ITS, $\{[x]_t\}$ with $t = 1, ..., T$, is organized as a series of d-dimensional interval-valued vectors

$$[x]_t^d = ([x]_t, [x]_{t-1}, ..., [x]_{t-(d-1)})', \tag{36}$$

   where $d \in \mathbb{N}$ is the number of lags.

2. We compute the dissimilarity between the most recent interval-valued vector $[x]_T^d = ([x]_T, [x]_{T-1}, ..., [x]_{T-d+1})'$ and the rest of the vectors in $\{[x]_t^d\}$. We use a distance measure to assess the dissimilarity between vectors, i.e.

$$D_t([x]_T^d, [x]_t^d) = \left(\frac{\sum_{i=1}^{d} \left(D^q([x]_{T-i+1}, [x]_{t-i+1}) \right)}{d} \right)^{\frac{1}{q}}, \tag{37}$$

   where $D([x]_{T-i+1}, [x]_{t-i+1})$ is a distance such as the kernel based distance shown in (28), q is the order of the measure that has the same effect that in the error measure shown in (29).

3. Once the dissimilarity measures are computed for each $[x]_t^d$, $t = T-1, T-2, ..., T-d$, we select the k-closest vectors to $[x]_T^d$. These are denoted by $[x]_{T_1}^d, [x]_{T_2}^d, ..., [x]_{T_k}^d$.

4. Given the k-closest vectors, their subsequent values, $[x]_{T_1+1}, [x]_{T_2+1}..., [x]_{T_k+1}$, are averaged to obtain the final forecast

$$[\hat{x}]_{T+1} = \sum_{p=1}^{k} \omega_p \cdot [x]_{T_p+1}, \tag{38}$$

where $[x]_{T_p+1}$ is the consecutive interval of the sequence $[x]_{T_p}^d$, and ω_p is the weight assigned to the neighbor p, with $\omega_p \geq 0$ and $\sum_{p=1}^{k} \omega_p = 1$. The average (38) is computed according to the rules of interval arithmetic. The weights are assumed to be equal for all the neighbors $\omega_p = 1/k \ \ \forall p$, or inversely proportional to the distance between the last sequence $[x]_T^d$ and the considered sequence $[x]_{T_p}^d$

$$\omega_p = \frac{\psi_p}{\sum_{l=1}^{k} \psi_l}, \tag{39}$$

with $\psi_p = (D_{T_p}([x]_T^d, [x]_{T_p}^d) + \xi)^{-1}$ for $p = 1, ..., k$. The constant $\xi = 10^{-8}$ prevents the weight to explode when the distance between two sequences is zero.

2.4 Interval-valued dispersion

In this section, we apply the aforementioned interval regression and prediction methods to the daily interval time series of low/high prices of the SP500 index. We will denote the interval as $[p_L, p_U]_t$. There is strand in the financial literature, Parkinson (1980), Garman and Klass (1980), and Yang and Zhang (2000) among others, that deals with functions of the range of the interval, $p_U - p_L$, in order to provide an estimator of the volatility σ of asset returns. In this paper we do not pursue this route. The object of analysis is the interval $[p_L, p_U]_t$ itself and our goal is the construction of the one-step-ahead forecast $[\hat{p}_L, \hat{p}_U]_{t+1}$. Obviously such a forecast can be an input to produce a forecast $\hat{\sigma}_{t+1}$ of volatility. One of the advantage of forecasting the low/high interval versus forecasting volatility is that the prediction error of the interval is based on observables as opposed to the prediction error for the volatility forecast for which "observed" volatility may a problem.

The sample period goes from January 3, 2000 to September 30, 2008. We consider two sets of predictions:

i. Low volatility prediction set (year 2006): estimation period that goes from January 3, 2000 to December 30, 2005 (1508 trading days) and prediction period that goes from January 3, 2006 to December 29, 2006 (251 trading days).

ii. High volatility prediction set (year 2008): estimation period that goes from January 2, 2002 to December 31, 2007 (1510 trading days) and prediction period that goes from January 3, 2008 to September 30, 2008 (189 trading days).

A plot of the ITS $[p_L, p_U]_t$ is presented in Figure 8.

[FIGURE 8]

Following the classical regression approach to ITS we are interested in the properties and time series regression models of the components of the interval, i.e. p_L, p_U, p_C, and p_R. We present the most significant and unrestricted time series models for $[p_L, p_U]_t$ and $\langle p_C, p_R \rangle_t$ in the spirit of the regression proposals of Billard and Diday (2000, 2002) and Lima Neto and de Carvalho (2008) reviewed in the previous sections. To save space we omit the univariate modelling of the components of the interval but these results are available upon request. However, we need to report that for p_L and p_U, we cannot reject a unit root, which is expected because these are price levels of the SP500, and that p_C has also a unit root because is the sum of two unit root processes. In addition, p_L and p_U are cointegrated of order one with cointegrating vector $(1, -1)$, which implies that p_R is a stationary process given that $p_R = (p_U - p_L)/2$. Following standard model selection criteria and time series specification tools, the best model for $\langle \Delta p_C, p_R \rangle_t$ is a VAR(3) and for $[p_L, p_U]_t$ a VEC(3). The estimation results are presented in Tables A-1 and A-2 in the Appendix A.

In Table A-1, the estimation results for $\langle \Delta p_C, p_R \rangle_t$ in both periods are very similar. The radius $p_{R,t}$ exhibits high autoregressive dependence and it is negatively correlated with the previous change in the center of the interval $\Delta p_{C,t-1}$ so that positive surprises in the center tend to narrow down the interval. On the other hand $\Delta p_{C,t}$ has little linear dependence and it is not affected by the dynamics of the radius. There is Granger-causality from the center to the radius but not viceversa. The radius equation enjoys a relative high adjusted R-squared

of about 40% while the center equation is basically not linearly predictable. In general terms, there is a strong similarity between the modeling of $\langle \Delta p_C, p_R \rangle_t$ and the most classical modeling of volatility with ARCH models for financial returns. The processes $p_{R,t}$ and the conditional variance of an ARCH model, i.e. $\sigma^2_{t|t-1} = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \alpha_2 \varepsilon_{t-1} + \beta \sigma^2_{t-1|t-2}$, share the autoregressive nature and the well-documented negative correlation of past innovations, which are $\Delta p_{C,t-1}$ in the center process and ε_{t-1} in the return process, with the radius process $p_{R,t}$ and the volatility process $\sigma^2_{t|t-1}$respectively. The unresponsiveness of the center to the information in the dynamics of the radius is also similar to the findings of ARCH-in-mean process where it is difficult to find significant effects of the volatility in the return process.

In Table A-2, we report the estimation results for $[p_L, p_U]_t$ for both periods 2000-2005 and 2002-2007. In general, there is much less linear dependence in the short-run dynamics of $[p_L, p_U]_t$, which is expected as we are modeling financial prices. There is Granger-causality running both ways, from Δp_L to Δp_U and viceversa. Overall, the 2002-2007 period seems to be noisier (R-squared of 14%) than the 2000-2005 (R-squared of 20 to 16%).

Based on the estimation results of the VAR(3) and VEC(3) models, we proceed to construct the one-step-ahead forecast of the interval $[\hat{p}_L, \hat{p}_U]_{t+1|t}$.We also implement the exponential smoothing methods and the k-NN method for ITS proposed in the above sections and compare their respective forecasts. For the smoothing procedure, the estimated value of α is $\hat{\alpha} = 0.04$ in the estimation period 2000-2005 and $\hat{\alpha} = 0.03$ in 2002-2007. We have implemented the k-NN with equal weights and with inversely proportional as in (39). In the period 2000-2005, the numbers of neighbors is $\hat{k} = 24$ (equal weights) and $\hat{k} = 25$ (proportional weights); in 2002-2007 $\hat{k} = 22$ in both cases. The length of the vector is $\hat{d} = 2$ in both periods. The estimation of α, k, and d has been performed by minimizing the mean distance MDE (29) with $q = 2$. In both methods, smoothing and k-NN, the centers of the intervals have been first-differenced to proceed with the estimation and forecasting. However, in the following comparisons, the estimated differenced centers are transformed back to present the estimates and forecasts in levels. In Table 1 we show the performance of the five models measured by the MDE ($q = 2$) in the estimation and prediction periods. We have also added a "naive" model that does not entail any estimation and whose forecast is the observation in the previous period, i.e. $[\hat{p}_L, \hat{p}_U]_{t+1|t} = [p_L, p_U]_t$.

	period 2000-2006		period 2002-2008	
Models	estimation 2000-2005	prediction 2006	estimation 2002-2007	prediction 2008
VAR(3)	9.359	6.611	7.614	15.744
VEC(3)	9.313	6.631	7.594	15.766
k-NN (eq.w.)	9.497	6.382	7.621	15.892
k-NN (prop.w.)	9.511	6.367	7.623	15.877
Smoothing	9.833	6.698	7.926	16.274
Naive	10.171	7.056	8.233	16.549

Table 1: Performance of the forecasting methods: MDE ($q = 2$)

	t-test for $H_0 : E(D^2_{(naive)} - D^2_{(other)}) = 0$	
Models	2006	2008
VAR(3)	2.86	2.67
VEC(3)	2.26	2.46
k-NN(eq.w.)	3.67	2.26
k-NN(prop.w.)	3.64	2.26
Smoothing	5.05	1.15

Table 2: Results of the Diebold and Mariano test

For both low and high volatility periods the performance ranking of the six models is very similar. The worst performer is the naive model followed by the smoothing model. In 2006, the k-NN procedures are superior to the VAR(3) and VEC(3) models but in 2008 the VAR and VEC systems perform slightly better that the k-NN's. The high-volatility year 2008 is clearly more difficult to forecast, the MDE in 2008 is twice as much as the MDE in the estimation period 2002-07. On the contrary, in the low volatility year 2006, the MDE in the prediction period is about 30% lower than the MDE in the estimation period 2000-2005. A statistical comparison of the MDE's of the five models in relation to the naive model is provided by the Diebold and Mariano test of unconditional predictability (Diebold and Mariano, 1995). The null hypothesis to test is the equality of the MDE's, i.e. $H_0 : E(D^2_{(naive)} - D^2_{(other)}) = 0$ versus $H_1 : E(D^2_{(naive)} - D^2_{(other)}) > 0$. If the null hypothesis is rejected the other model is superior to the naive model. The results of this test are presented in Table 2.

	Asymptotic (corrected) t-test $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$ $p_t = c + \beta \hat{p}_t + \sum_i \gamma_i \Delta \hat{p}_{t+i} + e_t$			
	2006		2008	
	min: $p_{L,t}$	max:$p_{U,t}$	min: $p_{L,t}$	max:$p_{U,t}$
VAR(3)	3.744	-1.472*	3.024	-2.712
VEC(3)	1.300*	0.742*	2.906	-2.106**
k-NN (eq.w.)	0.961*	-4.385	1.914*	-3.402
k-NN (prop.w.)	0.8654*	-3.369	2.032**	-3.478
Smoothing	-3.542	-2.544	2.739	-3.449

*Fail to reject the null at 5% level and ** fail to reject the null at 1% level

Table 3: Results of the t-test for cointegrating vector (1,-1)

In 2006, all the five models are statistically superior to the benchmark naive model. In 2008 the smoothing procedure is statistically equivalent to the naive model while the remaining four models outperform the naive.

We also perform a complementary assessment of the forecasting ability of the five models by running some regressions of the Mincer-Zarnowitz type. For the minimum p_L and maximum p_U, in the prediction periods, we run separate regressions of the realized observations on the predicted observations as in $p_{L,t} = c + \beta \hat{p}_{L,t} + \varepsilon_t$ and $p_{U,t} = c + \beta \hat{p}_{U,t} + \upsilon_t$. Under a quadratic loss function, we should expect an unbiased forecast, i.e. $\beta = 1$ and $c = 0$. However, the processes $p_{L,t}$ and $\hat{p}_{L,t}$ are I(1) and, as expected, cointegrated, so that these regressions should be performed with care. The point of interest is then to test for a cointegration vector of (1, -1). To test this hypothesis using an OLS estimator with the standard asymptotic distribution, we need to consider that in the I(1) process $\hat{p}_{L,t}$, i.e. $\hat{p}_{L,t} = \hat{p}_{L,t-1} + \nu_t$ the innovations ε_t and ν_t are not independent; in fact because $\hat{p}_{L,t}$ is a forecast of $p_{L,t}$ the correlation $\rho(\nu_{t+i}, \varepsilon_t) \neq 0$ for $i > 0$. To remove this correlation, the cointegrating regression will be augmented with some terms to finally estimate a regression as $p_{L,t} = c + \beta \hat{p}_{L,t} + \sum_i \gamma_i \Delta \hat{p}_{L,t+i} + e_t$ (the same argument applies to $p_{U,t}$). The hypothesis of interest is $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$. A t-statistic for this hypothesis will be asymptotically standard normal distributed. We may need to correct the t-test if there is some serial correlation in e_t . In Table 3 we present the testing results.

We reject the null for the smoothing method for both prediction periods and for both $p_{L,t}$ and $p_{U,t}$ processes. Overall the prediction is better for 2006 than for 2008, which is in agreement with the results based on the MDE, and we find better forecasting performance on predicting $p_{L,t}$ than on predicting $p_{U,t}$. The VEC(3) and the k-NN methods deliver better forecasts across the four instances considered. For those models in which we fail to reject $H_0 : \beta = 1$, we also calculate the unconditional average difference between the realized and the predicted values, i.e $\bar{p} = \sum_t (p_t - \hat{p}_t)/T$. For 2006, the forecasts of $p_{L,t}$ with the k-NN methods are slightly understated with $\bar{p} \simeq 1$ ($t = 2.6$ for $H_0 : \mu_p = 0$), and overstated for 2008 with $\bar{p} \simeq -4$ ($t = -3.7$ for $H_0 : \mu_p = 0$). With VEC(3), we find that the predictions of $p_{L,t}$ are slightly overstated with $\bar{p} \simeq -2$ ($t = -4.8$ for $H_0 : \mu_p = 0$). However, these differences are almost insignificant given that the level of the index is in the thousands. In Figure 9 we show the k-NN based forecast of the interval low/high of the SP500 index for the two last months of 2006.

[FIGURE 9]

3 Histogram data

In this section, our premise is that the data is presented to the researcher as a frequency distribution, which may be the result of an aggregation procedure, or the description of a population or any other grouped collective. We start by describing histogram data and some univariate descriptive statistics. Our main objective is to present the prediction problem by defining a histogram time series (HTS) and implementing smoothing techniques and non-parametric methods like the k-NN algorithm. As we have seen in the section on interval data, these two methods require the calculation of suitable averages. To this end, instead of relying on the arithmetic of histograms, we introduce the barycentric histogram that is an average of a set of histograms. The choice of appropriate distance measures is key to the calculation of the barycenter, and eventually of the forecast of a HTS.

3.1 Preliminaries

Given a variable of interest X we collect information on a group of individuals or units that belong to a set S. For every element $i \in S$, we observe a datum such as

$$h_{X_i} = \{([x]_{i1}, \pi_{i1}), ..., ([x]_{in_i}, \pi_{in_i})\}, \text{for } i \in S, \tag{40}$$

where π_{ij} , $j = 1, ..., n_i$ is a frequency that satisfies $\pi_{ij} \geq 0$ and $\sum_{j=1}^{n_i} \pi_{ij} = 1$; and $[x]_{ij} \subseteq \mathbb{R}$, $\forall i, j$, is an interval (also known as bin) defined as $[x]_{ij} \equiv [x_{Lij}, x_{Uij})$ with $-\infty < x_{Lij} \leq x_{Uij} < \infty$ and $x_{Ui\,j-1} \leq x_{Lij}$ $\forall i, j$, for $j \geq 2$. The datum h_{X_i} is a histogram and the data set will be understood as a collection of histograms $\{h_{X_i}, i = 1....m\}$. As in the case of interval data, we could summarize the histogram data set by its empirical density function from which the sample mean and the sample variance can be calculated (Billard and Diday, 2006). The sample mean is

$$\bar{X} = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{Uij} + x_{Lij}) \pi_{ij}, \tag{41}$$

which is the average of the weighted centers for each interval; and the sample variance is

$$S_X^2 = \frac{1}{3m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{Uij}^2 + x_{Uij} x_{Lij} + x_{Lij}^2) \pi_{ij} - \frac{1}{4m^2} \left[\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{Uij} + x_{Lij}) \pi_{ij} \right]^2,$$

which combines the variability of the centers as well as the intra-interval variability. Note that the main difference between these sample statistics and those in (6) and (9) for interval data is the weight provided by the frequency $\pi_{i,j}$ associated with each interval $[x]_{i,j}$.

Next, we proceed with the definition of a histogram random variable. Let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is the set of elementary events, $\mathcal{F}$is the σ-field of events and $P : \mathcal{F} \to [0, 1]$ the σ-additive probability measure; and define a partition of Ω into sets $A_X(x)$ such $A_X(x) = \{\omega \in \Omega \mid X(\omega) = x\}$, where $x \in \{h_{X_i}, i = 1....m\}$.

Definition 5. A mapping $h_X : \mathcal{F} \to \{h_{X_i}\}$ such that for each $x \in \{h_{X_i}, i = 1....m\}$ there exists a set $A_X(x)$ is called a histogram random variable.

Then, the definition of stochastic process follows as:

Definition 6. A histogram-valued stochastic process is a collection of histogram random variables that are indexed by time, i.e. $\{h_{X_t}\}$ for $t \in T \subset \mathbb{R}$, with each h_{X_t} following Definition 5.

A histogram-valued time series is a realization of a histogram-valued stochastic process and it will be equivalently denoted as $\{h_{X_t}\} \equiv \{h_{X_t},\ t = 1, 2, ...T\}$.

3.2 The prediction problem

In this section, we propose a dissimilarity measure for HTS based on a distance. For this purpose, two distance measures will be presented. These distances will play a key role in the estimation and prediction stages with histogram-valued data. They will help us to define a barycentric histogram, which will be used as the average of a set of histograms. Finally, we will present the implementation of the prediction methods.

3.2.1 Accuracy of the forecast

Suppose that we construct a forecast for $\{h_{X_t}\}$, which we denote as $\{\hat{h}_{X_t}\}$. It is sensible to define the forecast error as the difference $h_{X_t} - \hat{h}_{X_t}$. However, the difference operator based on histogram arithmetic does not provide information on how dissimilar the functions h_{X_t} and $\hat{h}_{X_t}$ are. In order to avoid this problem, Arroyo and Maté (2009) propose the Mean Distance Error (MDE), which in its most general form is defined as

$$MDE^q(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \left(\frac{\sum_{t=1}^{T} (D_X(h_{X_t}, \hat{h}_{X_t})^q}{T} \right)^{\frac{1}{q}}, \tag{42}$$

where $D(h_{X_t}, \hat{h}_{X_t})$ is a distance measure such as the Wasserstein or the Mallows distance to be defined shortly and q is the order of the measure, such that for $q = 1$ the resulting accuracy measure is similar to the MAE and for $q = 2$ to the RMSE.

Consider two density functions, $f(x)$ and $g(x)$, with their corresponding cumulative distribution functions (CDF), $F(x)$ and $G(x)$, the Wasserstein distance between $f(x)$ and $g(x)$ is defined as

$$D_W(f, g) = \int_0^1 |F^{-1}(t) - G^{-1}(t)| dt, \tag{43}$$

and the Mallows as

$$D_M(f,g) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}, \text{ and} \tag{44}$$

where $F^{-1}(t)$ and $G^{-1}(t)$ with $t \in [0,1]$ are the inverse CDFs of $f(x)$ and $g(x)$, respectively. It is easy to see the relation between both distances. The dissimilarity between two functions is essentially measured by how far apart their t-quantiles are, i.e. $F^{-1}(t) - G^{-1}(t)$. In the case of Wasserstein, the distance is defined in the L_1 norm and in the Mallows in the L_2 norm. When considering (42), $D(h_{X_t}, \hat{h}_{X_t})$ will be calculated by implementing the Wasserstein or Mallows distance. To this end we need to define the CDF of a histogram. Following Billard and Diday (2006), the CDF $H_X(x)$ of a histogram $h_X = \{((x_{Lj}, x_{Uj}], \pi_j)\}$ for $j = 1, ..., n$ is defined as

$$H_X(x) = \int_{-\infty}^{x} h_X(x)dx = \begin{cases} 0 & \text{if } x \leq x_{L1} \\ w_{j-1} + \frac{x - x_{Lj}}{x_{Uj} - x_{Lj}} \pi_j & \text{if } x \in (x_{Lj}, x_{Uj}] \\ 1 & \text{if } x > x_{Un} \end{cases} \tag{45}$$

where $w_l = \sum_{j=1}^{l} \pi_j$, is the cumulative weight associated with the interval l. With (45), the Wasserstein and Mallows distances between two histograms h_X and h_Y can be written as

$$D_W(h_X, h_Y) = \sum_{j=1}^{n} \pi_j \left| x_{Cj} - y_{Cj} \right| \tag{46}$$

$$D_M^2(h_X, h_Y) = \sum_{j=1}^{n} \pi_j \left[(x_{Cj} - y_{Cj})^2 + \frac{1}{3}(x_{Rj} - y_{Rj})^2 \right] \tag{47}$$

3.2.2 The barycentric histogram

Given a set of K histograms h_{X_k} with $k = 1, ..., K$, the barycentric histogram h_{X_B} is the histogram that minimizes the distances between itself and all the K histograms in the set. The optimization problem is

$$\min_{h_{X_B}} \sum_{k=1}^{K} D(h_{X_k}, h_{X_B}), \tag{48}$$

where $D(h_{X_k}, h_{X_B})$ is a distance measure. The choice of the distance will determine the properties of the barycenter.

When the chosen distance is Mallows, the optimal barycentric histogram $h^*_{X_B}$ has the following center/radius characteristics. For each suitable bin $j = 1, ..., n$, the barycentric center x^*_{Cj} is the mean of the centers of the corresponding bin in each histogram and the barycentric radius x^*_{Rj} is the mean of the radii of the corresponding bin in each of the K histograms,

$$x^*_{Cj} = \frac{\sum_{k=1}^{K} x_{Ckj}}{K} \tag{49}$$

$$x^*_{Rj} = \frac{\sum_{k=1}^{K} x_{Rkj}}{K}. \tag{50}$$

When the distance is Wasserstein, for each suitable bin $j = 1, ..., n$, the barycentric center x^*_{Cj} is the median of the centers of the corresponding bin in each of the K histograms,

$$x^*_{Cj} = \text{ median}(x_{Ckj}) \quad \text{ for } k = 1, ...K \tag{51}$$

and the radius x^*_{Rj} is the corresponding radius of the bin where the median x^*_{Cj} falls among the K histograms. For more details on the optimization problem, please see Arroyo and Maté (2009).

Other distances like Hellinger, Total Variation, Kolmogorov, Prokhorov-Lévy, etc. are not as suitable as the Mallows or Wassertein to find the barycentric histogram because they produce either non-unique barycenters or mixture-like barycenters, which are not desirable features in the prediction stage. See Verde and Irpino (2007) for further details.

3.2.3 Exponential smoothing

The exponential smoothing method can be adapted to histogram time series by replacing averages with the barycentric histogram, as it was shown in Arroyo and Maté (2008).

Let $\{h_{X_t}\}$ $t = 1, ..., T$ be a histogram time series, the exponentially smoothed forecast is given by the following equation

$$\hat{h}_{X_{t+1}} = \alpha h_{X_t} + (1 - \alpha)\hat{h}_{X_t}, \tag{52}$$

where $\alpha \in [0, 1]$. Since the right hand side is a weighted average of histograms, we can use the barycenter approach so that the forecast is the solution to the following optimization

exercise

$$\hat{h}_{X_{t+1}} \equiv \arg\min_{\hat{h}_{X_{t+1}}} \left(\alpha D(\hat{h}_{X_{t+1}}, h_{X_t}) + (1-\alpha) D(\hat{h}_{X_{t+1}}, \hat{h}_{X_t}) \right), \tag{53}$$

where $D(\cdot,\cdot)$ is the Mallows distance. The use of the Wasserstein distance is not suitable in this case because of the properties of the median, which will ignore the weighting scheme (with the exception of $\alpha = 0.5$) so intrinsically essential to the smoothing technique.

The recursive equation (52) can be easily rewritten as a moving average

$$\hat{h}_{X_{t+1}} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1} h_{X_{t-(j-1)}}, \tag{54}$$

which in turn can also be expressed as the following optimization problem

$$\hat{h}_{X_{t+1}} \equiv \arg\min_{\hat{h}_{X_{t+1}}} \sum_{j=1}^{t} \omega_i D(\hat{h}_{X_{t+1}}, h_{X_{t-(j-1)}}), \tag{55}$$

with $D(\cdot,\cdot)$ as the Mallows distance. The equations (53) and (55) are equivalent.

Figure 10 shows an example of the exponential smoothing using (53) for the histograms $h_{X_t} = \{([19,20), 0.1), ([20,21), 0.2), ([21,22], 0.7)\}$ and $\hat{h}_{X_t} = \{([0,3), 0.35), ([3,6), 0.3), ([6,9], 0.35)\}$ with $\alpha = 0.9$ and $\alpha = 0.1$. In both cases, the resulting histogram averages the location, the support, and the shape of both histograms h_{X_t} and $\hat{h}_{X_t}$ in a suitable way.

[FIGURE 10]

3.2.4 k-NN method

The adaptation of the k-NN method to forecast HTS was proposed by Arroyo and Maté (2009). The method consists of similar steps to those described in the interval section:

1. The HTS, $\{h_{X_t}\}$ with $t = 1, ..., T$, is organized as a series of d-dimensional histogram-valued vectors $\{h^d_{X_t}\}$ where

$$h^d_{X_t} = (h_{X_t}, h_{X_{t-1}}, ..., h_{X_{t-(d-1)}})', \tag{56}$$

   where $d \in \mathbb{N}$ is the number of lags.

2. We compute the dissimilarity between the most recent histogram-valued vector $h^d_{X_T} = (h_{X_T}, h_{X_{T-1}}, ..., h_{X_{T-(d-1)}})'$ and the rest of the vectors in $\{h^d_{X_t}\}$ by implementing the following distance measure

$$D_t(h^d_{X_T}, h^d_{X_t}) = \left(\frac{\sum_{i=1}^{d} \left(D^q(h_{X_{T-i+1}}, h_{X_{t-i+1}}) \right)}{d} \right)^{\frac{1}{q}}, \tag{57}$$

where $D^q(h_{X_{T-i+1}}, h_{X_{t-i+1}})$ is the Mallows or the Wasserstein distance of order q.

3. Once the dissimilarity measures are computed for each $h^d_{X_t}$, $t = T-1, T-2, ...T-d$, we select the k-closest vectors to $h^d_{X_T}$. These are denoted by $h^d_{X_{T_1}}, h^d_{X_{T_2}}..., h^d_{X_{T_k}}$.

4. Given the k-closest vectors, their subsequent values, $h_{X_{T_1+1}}, h_{X_{T_2+1}}..., h_{X_{T_k+1}}$, are averaged by means of the barycenter approach to obtain the final forecast $\hat{h}_{X_{T+1}}$ as in

$$\hat{h}_{X_{T+1}} \equiv \arg\min_{\hat{h}_{X_{T+1}}} \sum_{p=1}^{k} \omega_p D(\hat{h}_{X_{T+1}}, h_{X_{T_p+1}}), \tag{58}$$

where $D(\hat{h}_{X_{T+1}}, h_{X_{T_p+1}})$ is the Mallows or the Wasserstein distance, $h_{X_{T_p+1}}$ is the consecutive histogram in the sequence $h^d_{X_{T_p}}$, and ω_p is the weight assigned to the neighbor p, with $\omega_p \geq 0$ and $\sum_{p=1}^{k} \omega_p = 1$. As in the case of the interval-valued data, the weights may be assumed to be equal for all the neighbors $\omega_p = 1/k \ \ \forall p$, or inversely proportional to the distance between the last sequence $h^d_{X_T}$ and the considered sequence $h^d_{X_{T_p}}$.

3.3 Histogram forecast for SP500 returns

In this section we implement the exponential smoothing and the k-NN methods to forecast the one-step-ahead histogram of the returns to the constituents of the SP500 index. We collect the weekly returns of the 500 firms in the index from 2002 to 2005. We divide the sample into an estimation period of 156 weeks running from January 2002 to December 2004, and a prediction period of 52 weeks that goes from January 2005 to December 2005. The histogram data set consists of 208 weekly equiprobable histograms. Each histogram has four bins, each one containing 25% of the firms' returns.

Models	estimation 2002-2004	prediction 2005
Mall. k-NN (eq.w.)	4.988	4.481
Mall. k-NN prop.w.)	4.981	4.475
Wass. k-NN (eq.w.)	4.888	4.33
Wass. k-NN (prop.w.)	4.882	4.269
Exp. Smoothing	4.976	4.344
Naive	6.567	5.609

Table 4: Performance of the forecasting methods: MDE ($q = 1$)

For the smoothing procedure, the estimated value of α is $\hat{\alpha} = 0.13$. We have implemented the k-NN with equal weights and with inversely proportional as in (39) using the Mallows and Wasserstein distances. With the Mallows distance, the estimated numbers of neighbors is $\hat{k} = 11$ and the length of the vector is $\hat{d} = 9$. With the Wasserstein distance, $\hat{k} = 12$, $\hat{d} = 9$ (equal weights), and $\hat{k} = 17$, $\hat{d} = 8$ (proportional weights). The estimation of α, k, and d has been performed by minimizing the mean Mallows distance MDE (57) with $q = 1$, except for the Wasserstein-based k-NN. In Table 4, we show the performance of the five models measured by the Mallows-based MDE ($q = 1$) in the estimation and prediction periods. We have also added a "naive" model that does not entail any estimation and for which the one-step-ahead forecast is the observation in the previous period, i.e. $\hat{h}_{X_{t+1|t}} = h_{X_t}$.

In the estimation and prediction period, the naive model is clearly outperformed by the rest of the five models. In the estimation period, the five models exhibit similar performance with a MDE of 4.9 approximately. In the prediction period, the exponential smoothing and the Wasserstein-based k-NN seem to be superior to the Mallows-based k-NN. We should note that the MDEs in the prediction period are about 11% lower than the MDEs in the estimation period.

For the prediction year 2005, we provide a statistical comparison of the MDEs of the five models in relation to the naive model by implementing the Diebold and Mariano test of unconditional predictability (Diebold and Mariano, 1995). The null hypothesis to test is the equality of the MDEs, i.e. $H_0 : E(D_{(naive)} - D_{(other)}) = 0$ versus $H_1 : E(D_{(naive)} - D_{(other)}) > 0$. If the null hypothesis is rejected the "other" model is superior to the naive model. The results of this test are presented in Table 5.

	t-test for $H_0 : E(D_{(naive)} - D_{(other)}) = 0$
Models	2005 prediction year
Mall. k-NN(eq.w.)	2.32
Mall. k-NN(prop.w.)	2.69
Wass. k-NN(eq.w.)	2.29
Wass. k-NN(prop.w.)	2.29
Exp. Smoothing	3.08

Table 5: Results of the Diebold and Mariano test

In 2005, all the five models are statistically superior to the benchmark naive model, though the rejection of the null is stronger for the exponential smoothing and the Mallows-based k-NN models with proportional weights.

In Figure 11, we present the 2005 one-step-ahead histogram forecast obtained with the exponential smoothing procedure and we compare it to the realized value. Overall the forecast follows very closely the realized value except for those observations that have extreme returns. The fit can be further appreciated when we zoom in the central 50% mass of the histograms (Figure 12).

[FIGURE 11 & 12]

4 Summary and conclusions

Large databases prompt the need for new methods of processing information. In this article we have introduced the analysis of interval-valued and histogram-valued data sets as an alternative to classical single-valued data sets and we have shown the promise of this approach to deal with economic and financial data.

With interval data, most of the current efforts have been directed to the adaptation of classical regression models as the interval is decomposed into two single-valued variables, either the center/radius or the min/max. The advantage of this decomposition is that classical inferential methods are available. Methodologies that analyze the interval *per se* fall into the realm of random sets theory and though there is some important research on regression analysis with random sets, inferential procedures are almost non-existent. Being

our current focus the prediction problem, we have explored two different venues to produce a forecast with interval time series (ITS). First, we have implemented the classical regression approach to the analysis of ITS, and secondly we have proposed the adaptation of filtering techniques, such as smoothing, and non-parametric methods, such as the k-NN, to ITS. The latter venue requires the use of interval arithmetic to construct the appropriate averages and the introduction of distance measures to assess the dissimilarity between intervals and to quantify the prediction error. We have implemented these ideas with the SP500 index. We modelled the center/radius time series and the low/high time series of what we called interval-valued dispersion of the SP500 index and compared their one-step-ahead forecasts to those of a smoothing procedure and k-NN methods. A VEC model for the low/high series and the k-NN methods have the best forecasting performance.

With histogram data, the analysis becomes more complex. Regression analysis with histograms is in its infancy and the venues for further developments are large. We have focused exclusively in the prediction problem with smoothing methods and non-parametric methods. A key concept for the implementation of these two procedures is the introduction of the barycentric histogram that is a device that works as an average (weighted or unweighted) of a set of histograms. As with ITS, the introduction of the appropriate distances to judge dissimilarities among histograms and to assess forecast errors are fundamental ingredients in the analysis. The collection over time of cross-sectional returns of the firms in the SP500 index provides a nice histogram time series (HTS), on which we have implemented the aforementioned methods to eventually produce the one-step-ahead histogram forecast. Simple smoothing techniques seem to work remarkably well.

There are many research areas to explore with ITS and HTS. A very important question is the search for a model. This will require the understanding of the notion of dependence in ITS and HTS. From an econometric point of view, concurrent stages such as identification, estimation, testing, and model selection should be integral components in the ITS and HTS research agenda for model building. Economic and financial questions will benefit greatly from this new approach to the analysis of large data sets.

References

Arroyo, J., R. Espínola, and C. Maté (2008). Different approaches to forecast interval time series: a comparison in finance. *Computation Statistics and Data Analysis (submitted)*.

Arroyo, J. and C. Maté (2008). Forecasting time series of observed distributions with smoothing methods based on the barycentric histogram. In *Computational Intelligence in Decision and Control. Proceedings of the 8th International FLINS Conference*, pp. 61–66. World Scientific.

Arroyo, J. and C. Maté (2006). Introducing interval time series: Accuracy measures. In *COMPSTAT 2006, Proceedings in Computational Statistics*, Heidelberg, pp. 1139–1146. Physica-Verlag.

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

Bertrand, P. and F. Goupil (2000). *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Chapter Descriptive statistics for symbolic data, pp. 103–124. Springer.

Billard, L. and E. Diday (2000). Regression analysis for interval-valued data. In *Data Analysis, Classification and Related Methods : Proceedings of the 7th Conference of the IFCS, IFCS 2002*, Berlín, pp. 369–374. Springer.

Billard, L. and E. Diday (2002). Symbolic regression analysis. In *Classification, Clustering and Data Analysis: Proceedings of the 8th Conference of the IFCS, IFCS 2002*, Berlín, pp. 281–288. Springer.

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*(462), 470–487.

Billard, L. and E. Diday (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining* (1st ed.). Chichester: Wiley & Sons.

Brito, P. (2007). Modelling and analysing interval data. In *Proceedings of the 30th Annual Conference of GfKl*, pp. 197–208. Springer.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*(3), 253–263.

Gardner, E. S. (2006). Exponential smoothing: The state of the art. Part 2. *International Journal of Forecasting 22*(4), 637–666.

Garman, M. B. and M. J. Klass (1980). On the estimation of security price volatilities from historical data. *The Journal of Business 53*(1), 67–78.

González, L., F. Velasco, C. Angulo, J. A. Ortega, and F. Ruiz (2004). Sobrenúcleos, distancias y similitudes entre intervalos. *Inteligencia Artificial, Revista Iberoamericana de IA 8*(23), 111–117.

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

Kulpa, Z. (2006). A diagrammatic approach to investigate interval relations. *Journal of Visual Languages and Computing 17*(5), 466–502.

Lima Neto, E. A. and F. d. A. T. de Carvalho (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis 52*, 1500–1515.

Moore, R. E. (1966). *Interval Analysis*. Englewood Cliffs, N.J.: Prentice Hall.

Moore, R. E., R. B. Kearfott, and M. J. Cloud (Eds.) (2009). *Introduction to Interval Analysis*. Philadelphia, PA: SIAM Press.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business 53*(1), 61.

Verde, R. and A. Irpino (2007). *Selected Contributions in Data Analysis and Classification*, Chapter Dynamic clustering of histogram data: using the right metric, pp. 123–134. Springer.

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

Yang, D. and Q. Zhang (2000). Drift independent volatility estimation based on high, low, open, and close prices. *The Journal of Business 73*(3), 477–492.

Zellner, A. and J. Tobias (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting 19*, 457–469.

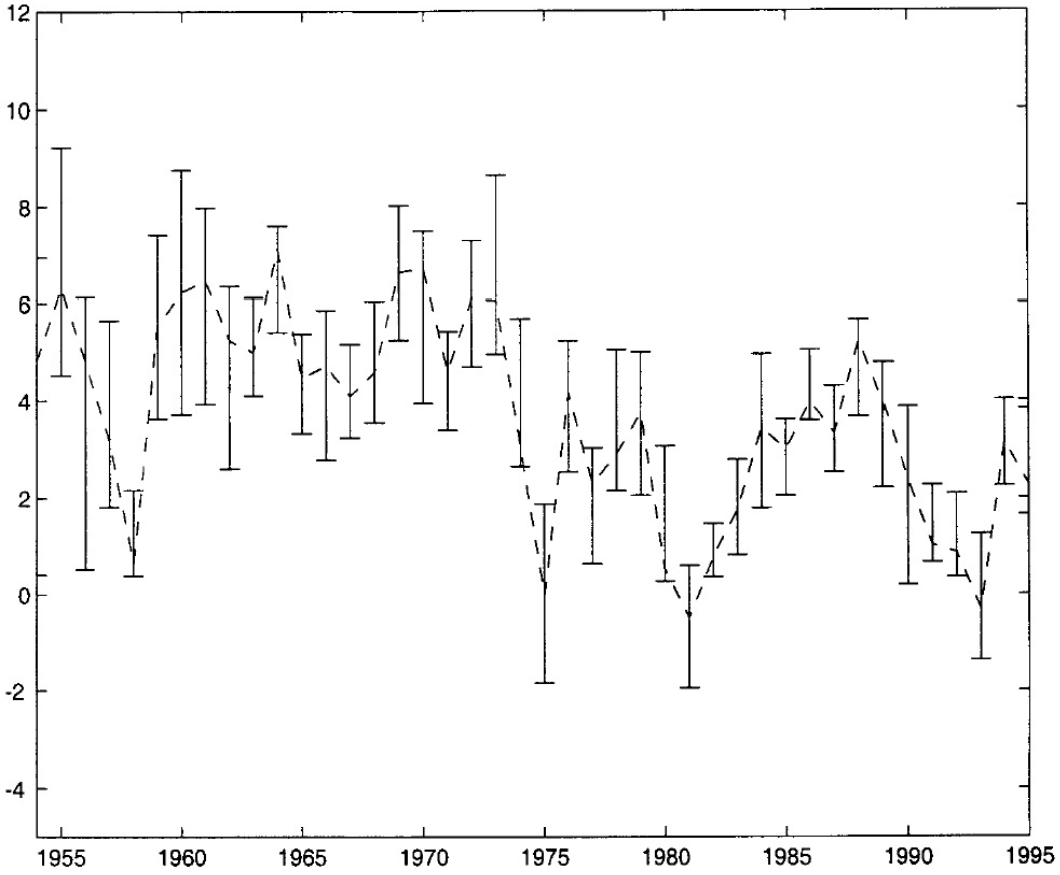

Figure 1. Time series of the interquantile range (solid interval) and time series of the median (dashed line) of the industrial production growth rates of 18 countries (Zellner and Tobias, 2000)

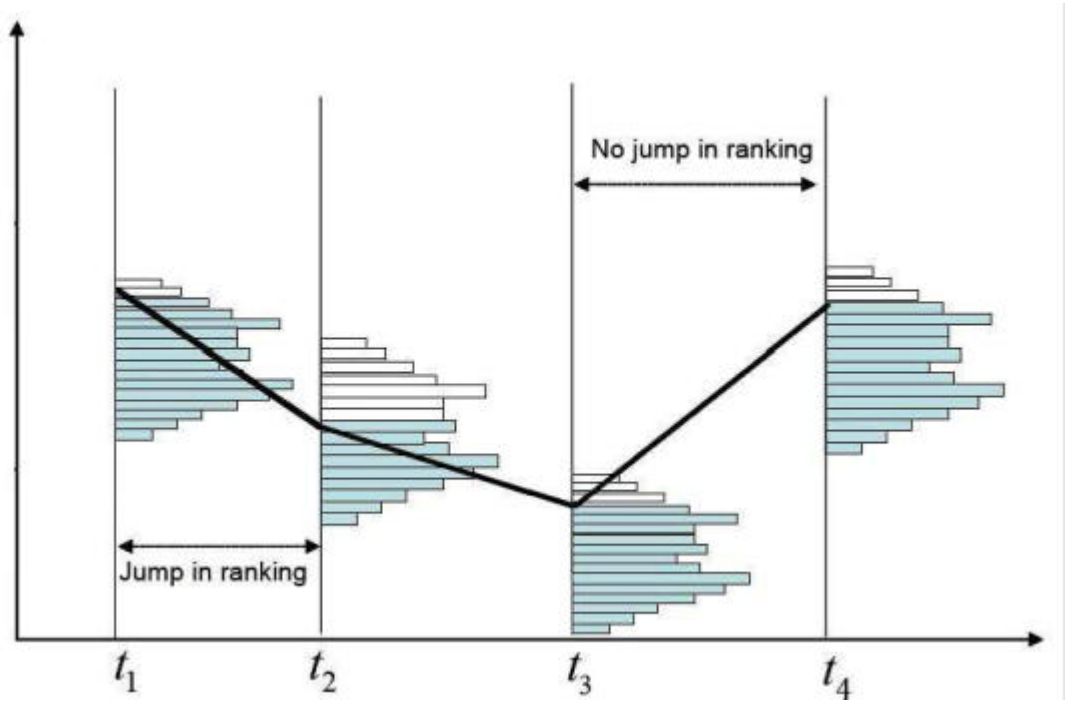

Figure 2. Stylized time series of the histograms of the cross-sectional returns of the constituents of the SP500 index (González-Rivera, Lee, and Mishra, 2008)

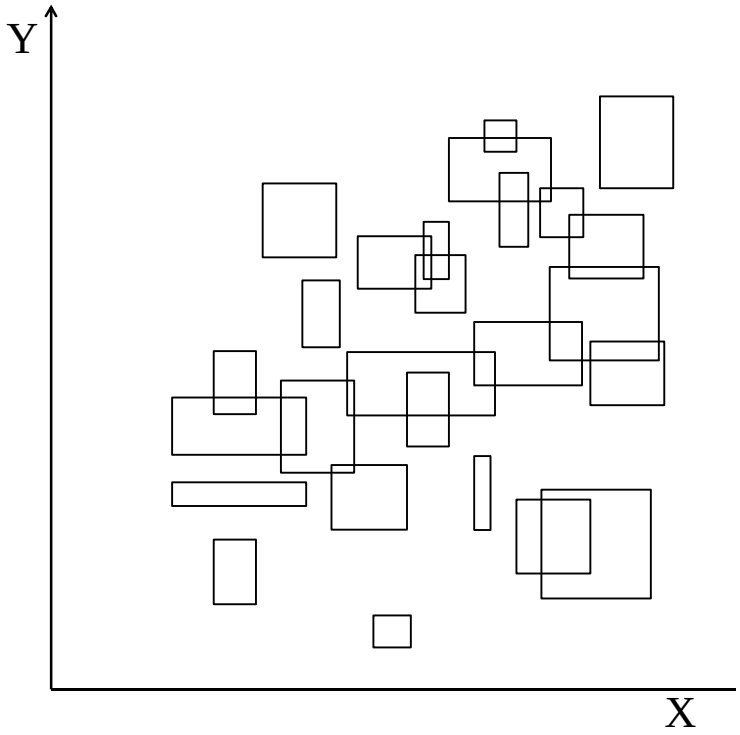

Figure 3. Scatter plot of two interval variables X and Y

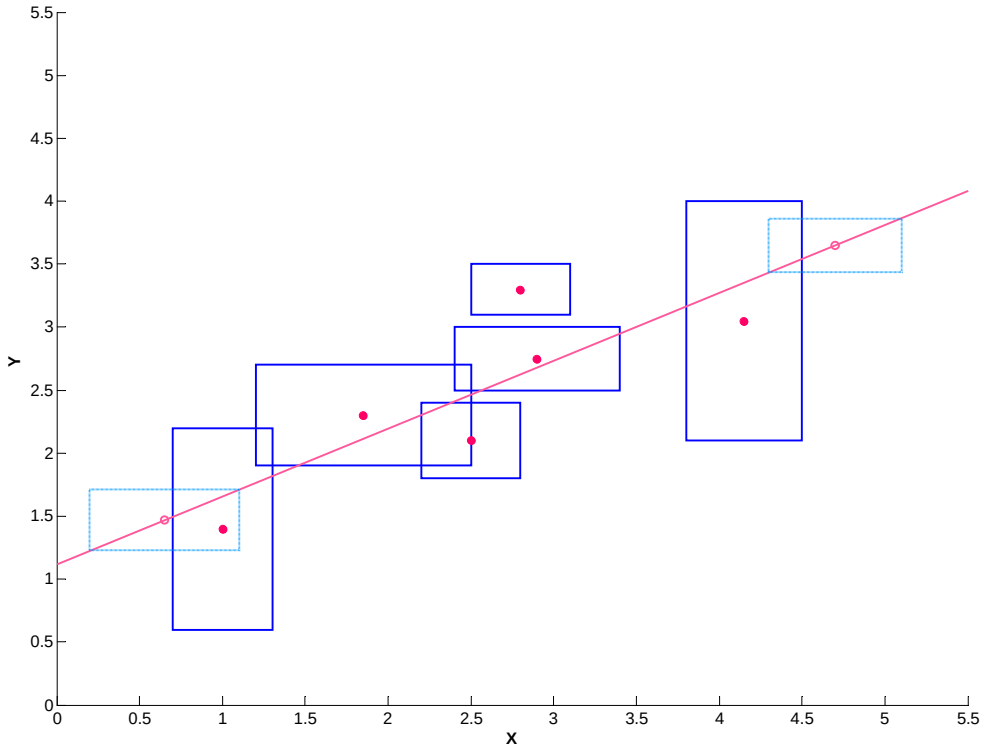

Figure 4. Fitting of a regression line (red) to the centers of the intervals (Billard and Diday, 2000). The estimated rectangles according to the regression line are represented by a blue dashed line.

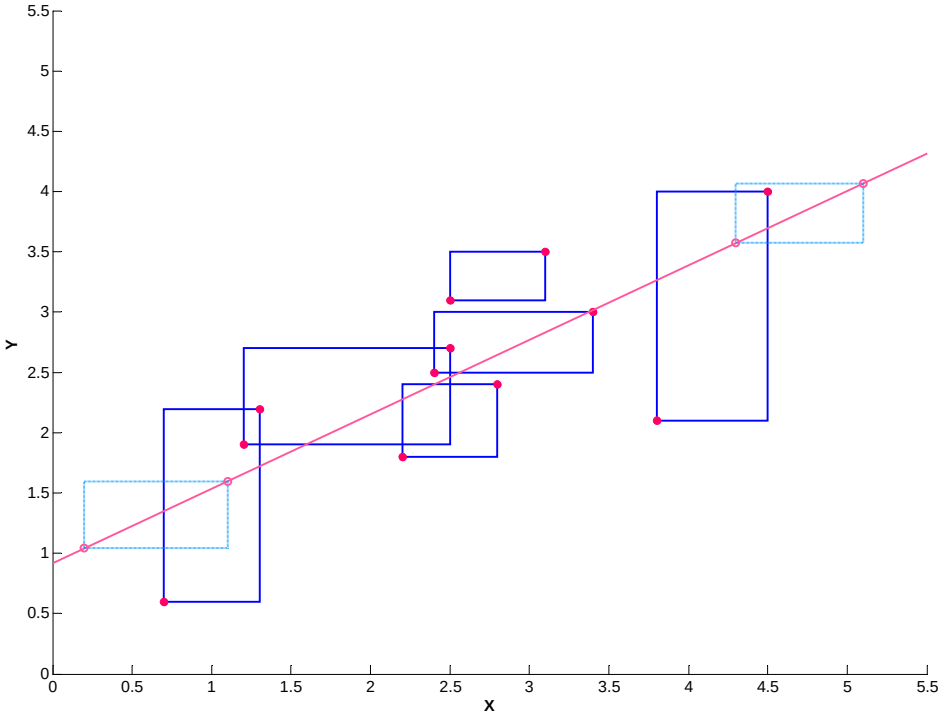

Figure 5. Regression line (red) according to Brito (2007). The estimated rectangles from the regression line are represented by a blue dashed line.

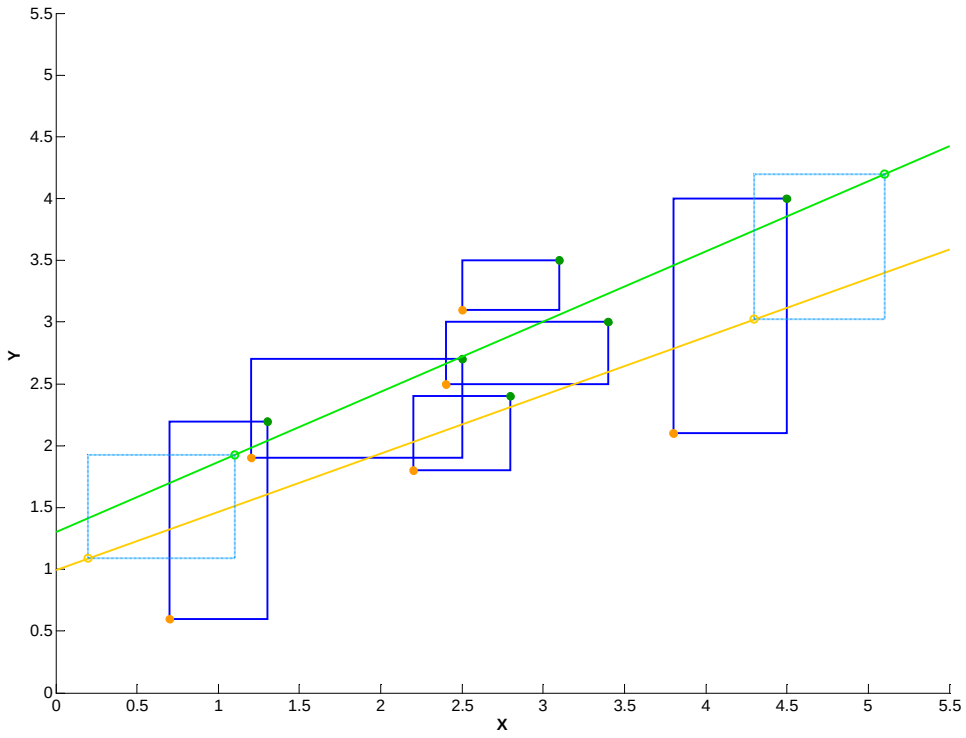

Figure 6. Regression lines fitted to the minima (yellow) and maxima (green) of the intervals (Billard and Diday, 2002). The estimated rectangles according to the regression lines are represented by a blue dashed line.

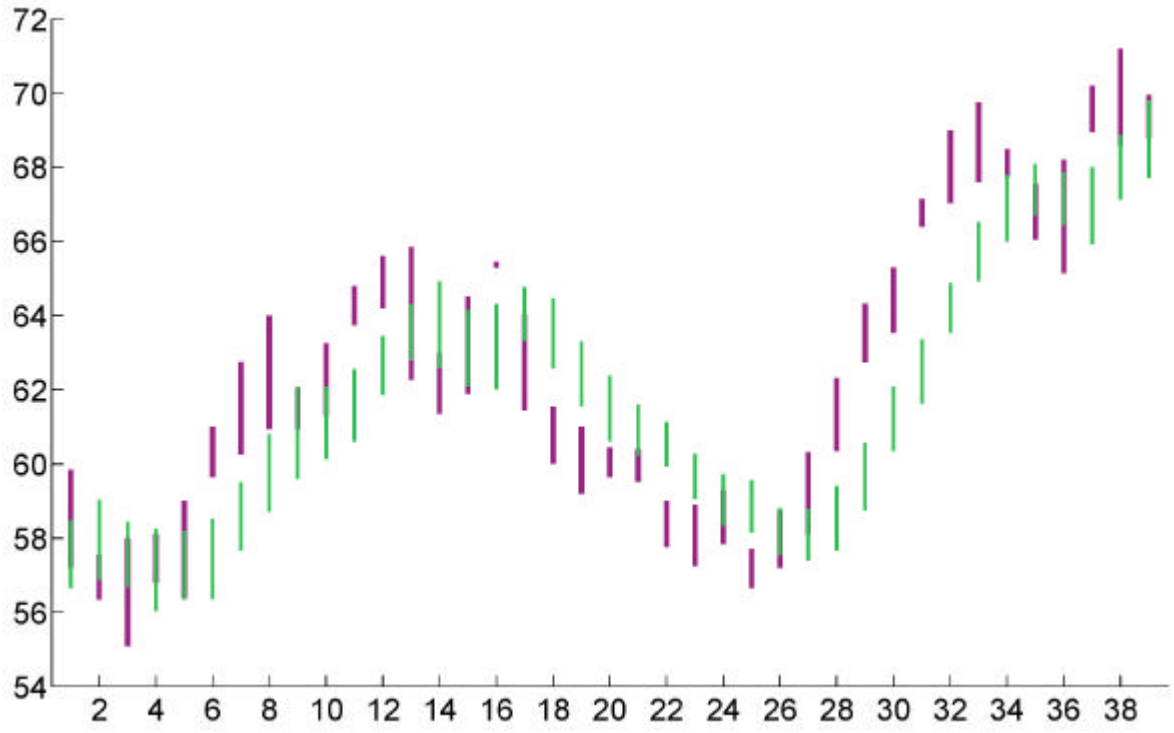

Figure 7. Observed ITS (purple) and exponentially smoothed (green) ITS with α=0.4.

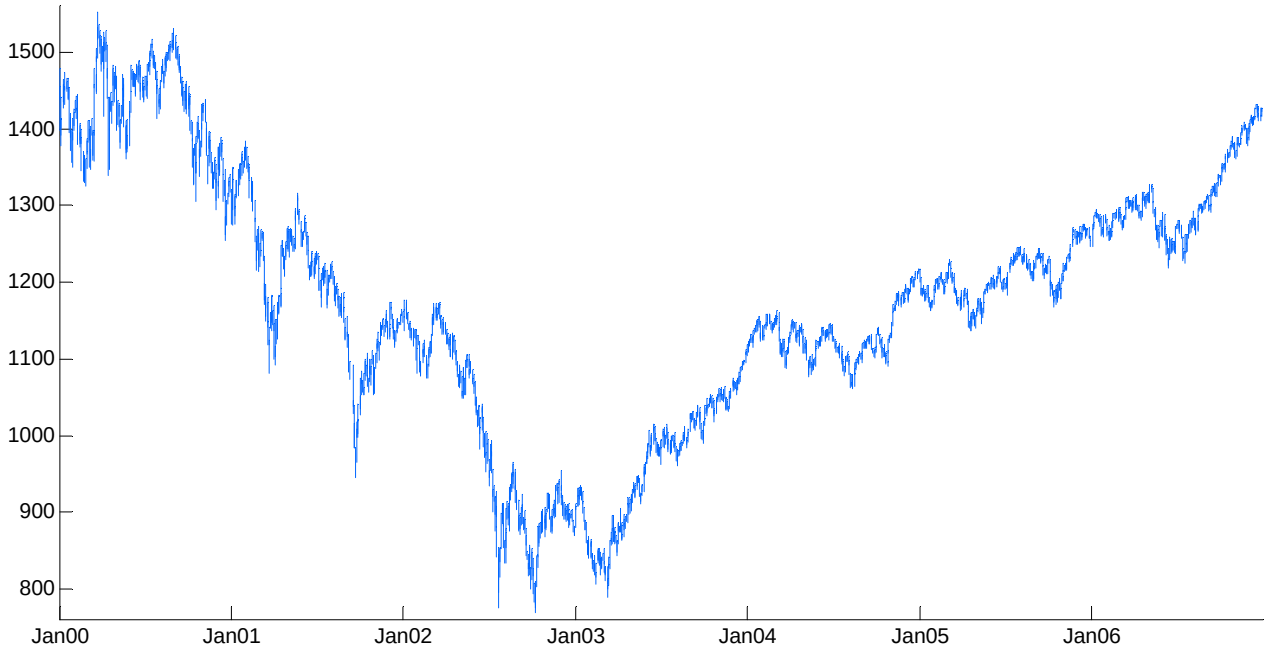

Figure 8. ITS of the weekly low/high prices of the SP500 index.

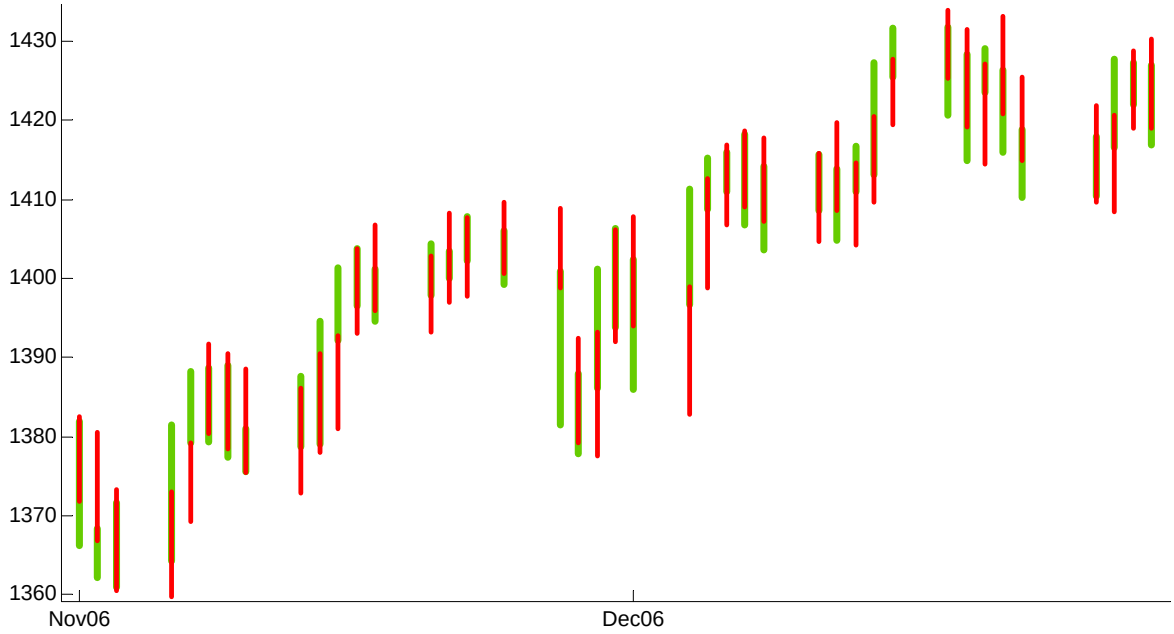

Figure 9. k-NN based forecast (red) of the low/high prices of the SP500; realized ITS (green)

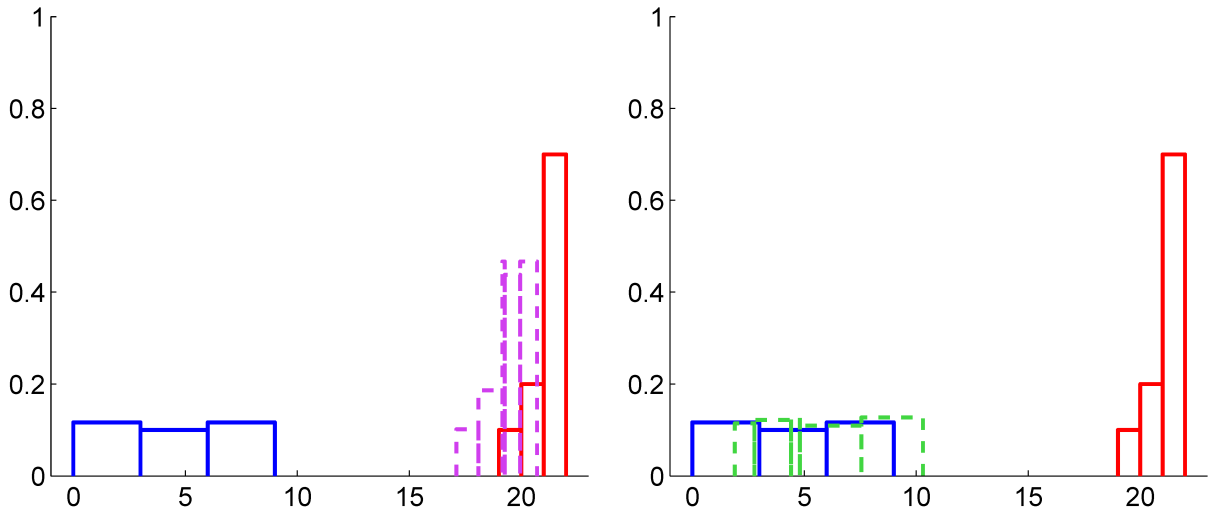

Figure 10. Exponential smoothing of histograms using the recursive formulation with α=0.9 (left) and α=0.1 (right). In each figure, the barycenter is the dash-lined histogram.

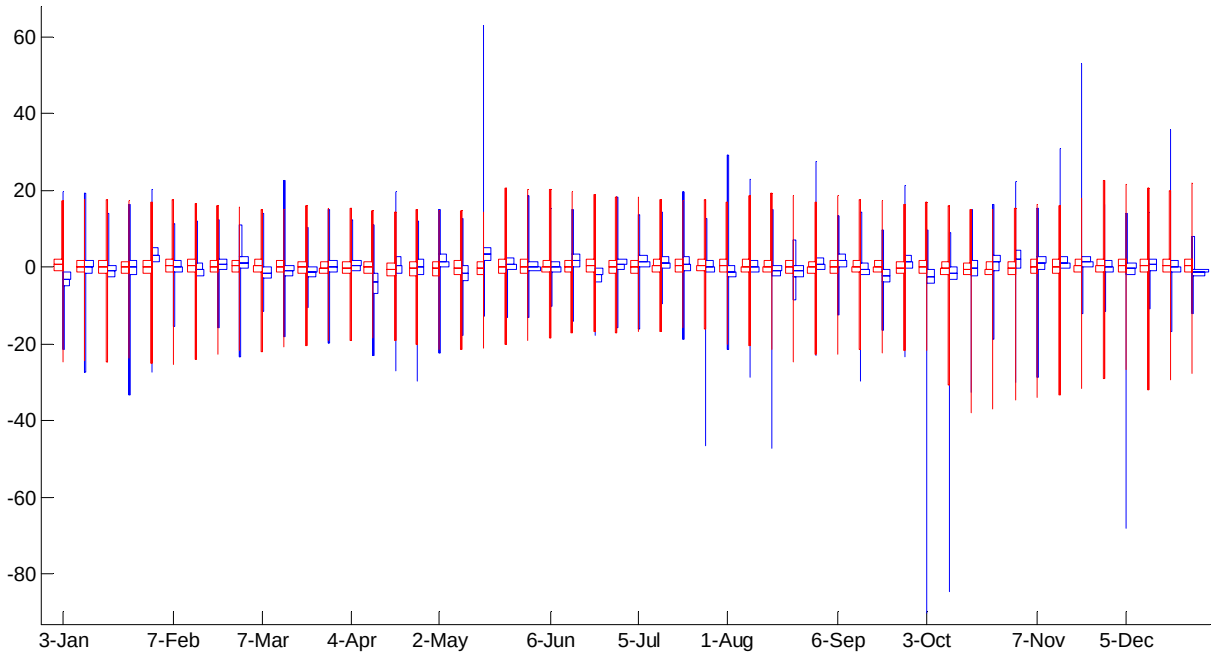

Figure 11. 2005 realized histogram (blue) and exponential smoothed one-step-ahead histogram forecast (red) for the HTS of SP500 returns. Weekly data.

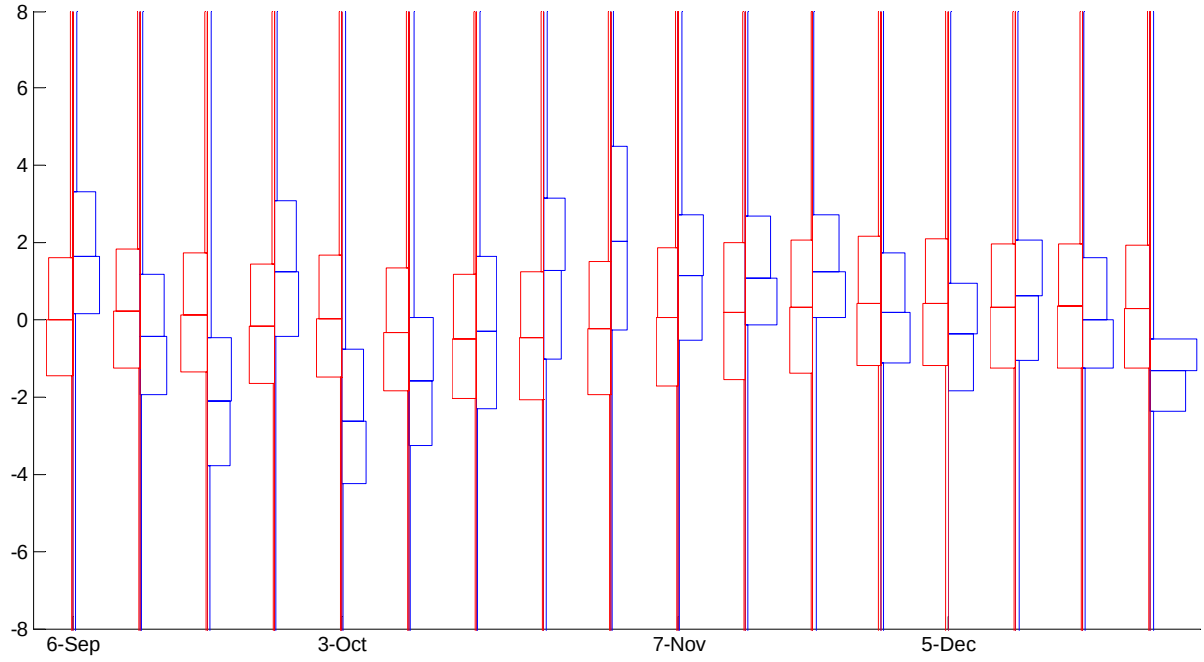

Figure 12. Zoom of Figure 11 from September to December 2005.

APPENDIX A. Estimation Results for ITS SP500 Index

Table A-1. Estimation of the VAR(3) model for the differenced center and radius time series

Estimation sample 2000-2005

VAR	D(Cen)	Rad
D(Cen(-1))	0.33218	-0.09764
	0.0262	0.00997
	[12.6803]	[-9.79410]
D(Cen(-2))	-0.181348	-0.001809
	0.02742	0.01043
	[-6.61378]	[-0.17332]
D(Cen(-3))	0.050564	0.00429
	0.02616	0.00996
	[1.93281]	[0.43091]
Rad(-1)	0.066659	0.150616
	0.06593	0.02509
	[1.01103]	[6.00287]
Rad(-2)	-0.049629	0.313259
	0.06319	0.02405
	[-0.78541]	[13.0270]
Rad(-3)	0.129442	0.285272
	0.0648	0.02466
	[1.99747]	[11.5678]
C	-1.319847	2.088036
	0.60607	0.23064
	[-2.17772]	[9.05315]

Estimation sample 2002-2007

VAR	D(Cen)	Rad
D(Cen(-1))	0.279225	-0.074092
	0.02619	0.00978
	[10.6611]	[-7.57934]
D(Cen(-2))	-0.092471	-0.010534
	0.02713	0.01012
	[-3.40879]	[-1.04037]
D(Cen(-3))	0.006178	-0.013364
	0.02629	0.00981
	[0.23500]	[-1.36214]
Rad(-1)	-0.00284	0.152907
	0.06731	0.02512
	[-0.04219]	[6.08652]
Rad(-2)	0.046537	0.27345
	0.0649	0.02422
	[0.71705]	[11.2886]
Rad(-3)	-0.01386	0.276629
	0.06635	0.02477
	[-0.20888]	[11.1699]
C	-0.045805	2.074405
	0.5355	0.19987
	[-0.08554]	[10.3788]

Table A-2. Estimation of the VEC(3) model for Low/High time series

Estimation sample 2000-2005

Error Correction:	D(Low)	D(High)
CointEq1	-0.438646	0.007023
	0.05364	0.04758
	[-8.17770]	[0.14761]
D(Low(-1))	0.112549	0.515586
	0.05429	0.04816
	[2.07293]	[10.7050]
D(Low(-2))	-0.093605	0.193326
	0.0505	0.0448
	[-1.85344]	[4.31532]
D(Low(-3))	0.026446	0.112943
	0.0396	0.03512
	[0.66790]	[3.21547]
D(High(-1))	0.313542	-0.287591
	0.05905	0.05238
	[5.30959]	[-5.49018]
D(High(-2))	-0.073453	-0.382411
	0.05604	0.04971
	[-1.31078]	[-7.69307]
D(High(-3))	0.04646	-0.065429
	0.04356	0.03864
	[1.06663]	[-1.69337]
C	-0.064365	-0.118124
	0.28906	0.25642
	[-0.22267]	[-0.46068]

Cointegrating Eq:	CointEq1
Low(-1)	1
High(-1)	-1.001255
	0.00268
	[-373.870]
@TREND(1)	-0.012818
	0.00105
	[-12.1737]
C	27.97538

Estimation sample 2002-2007

Error Correction:	D(Low)	D(High)
CointEq1	-0.124897	0.121926
	0.04103	0.03692
	[-3.04419]	[3.30283]
D(Low(-1))	-0.165406	0.425054
	0.0489	0.044
	[-3.38238]	[9.66024]
D(Low(-2))	-0.314249	0.130253
	0.04863	0.04375
	[-6.46233]	[2.97698]
D(Low(-3))	-0.15041	0.061275
	0.0399	0.0359
	[-3.76992]	[1.70691]
D(High(-1))	0.524179	-0.221533
	0.05188	0.04668
	[10.1046]	[-4.74625]
D(High(-2))	0.248088	-0.239401
	0.05323	0.04789
	[4.66085]	[-4.99871]
D(High(-3))	0.182654	-0.073329
	0.04262	0.03835
	[4.28593]	[-1.91234]

Cointegrating Eq:	CointEq1
Low(-1)	1
High(-1)	-1.002284
	0.00318
	[-315.618]
C	16.82467
	3.81466
	[4.41053]