Time Series Modeling of Histogram-valued Data
The Daily Histogram Time Series of SP500 Intradaily Returns

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

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

*Corresponding author: gloria.gonzalez@ucr.edu, tel (951) 827-1590, fax (951) 827-5685. G. González-Rivera acknowledges the financial support of the UCR University Scholar award and the Academic Senate grants. J. Arroyo acknowledges the financial support of grant TIN2008-06464-C03-01, sponsored by the Spanish Council for Science and Innovation, and the "Programa de Creación y Consolidación de Grupos de Investigación UCM-BSCH" (GR58/08).

ABSTRACT

Histogram time series (HTS) and interval time series (ITS) are examples of symbolic data sets. Though there are methodological developments in a cross-sectional environment, those are scarce in a time series setting. Arroyo, González-Rivera, and Maté (2010) analyze forecasting methods for HTS and ITS adapting smoothing filters and nonparametric algorithms like the k-NN. Though these methods are very flexible, they may not be the true underlying data generating process (DGP). We present the first building block towards the search for a DGP by focusing on the autocorrelation functions (ACF) of HTS and ITS. We analyze the ACF of the daily histogram of 5-minute intradaily returns to the SP500 index in 2007 and 2008. There are clusters of high/low activity that generates a strong, positive, and persistent autocorrelation pointing towards some autoregressive process for HTS. Though smoothing and k-NN may not be the true DGPs, we find that they are very good approximations because they are able to capture almost all the original autocorrelation. However, there seems to be some structure left in the data that will require new modelling techniques. As a byproduct, we also analyze the [90,100%] quantile interval. By using the full information contained in the histogram, we find that there are advantages in the estimation and prediction of a specific interval.

Key Words: Symbolic data, Interval-valued data, Histogram-valued data, Autocorrelation, Intradaily returns

JEL Classification: C22, C53

1 Introduction

Histogram-valued time series (HTS) and interval-valued time series (ITS) are examples of symbolic data sets as opposed to classical data sets. The sample information of classical data sets, cross-sectional or/and time series, consists of a collection of single-valued observations. In symbolic data sets the sample information is a collection of more complex and richer objects. For instance, in a time series context, the datum at time t could be an interval, like the high/low price interval of a given stock in a given day, so that the collection of intervals indexed by t constitutes an interval time series. In the same fashion, we can think of a daily histogram of intradaily prices or returns, so that the collection of histograms indexed by t form a histogram time series. The analysis of symbolic data sets is a very promising tool to deal with massive information sets. Billard and Diday (2006) and Diday and Noirhomme (2008) provide an extensive review of this new field.

Most of the current analysis in symbolic data focuses on cross-sectional data sets. There are developments in descriptive statistics and in the adaptation of classical multivariate methods, like principal components, to symbolic data. Developments for time series data are very scarce. See Han et al. (2008) and Maia et al. (2008) who deal with time series concepts applied to interval data. Arroyo and Maté (2009b) and Arroyo, González-Rivera, and Maté (2010) focus on time series data by providing forecasting methods for interval and histogram time series, which are adapted from classical algorithms such as smoothing filters and non-parametric k-NN methods. These authors show that these forecasting techniques can be very successful with financial data.

The implementation of smoothing and k-NN methods only require the construction of suitable averages. Thus, these methods work under the implicit assumption that, if the world is more or less stable, the future should not be very far from some average (weighted or unweighted) of the past. However, going further, it is of interest to uncover the data generating process behind the dynamics of HTS and ITS. This is a more difficult exercise because, even in the most simple model, at least a notion of regression with intervals and/or histograms will be needed. In this paper we present the first building block towards the search for a DGP. We aim to understand the linear dependence of financial HTS and ITS

by analyzing their empirical autocorrelation functions.

The autocorrelation function of ITS is based on the adoption of the already developed descriptive statistics (Billard and Diday, 2006). An interval is uniquely defined by its center and radius or by its lower and upper bounds. Then, the autocovariance between any two time-indexed intervals will be the result of the interaction between the variability of the centers/radii and the variability between and within intervals. The autocorrelation function of HTS is more complex due to the inherent complexity of histograms. Variance and autocovariance for HTS will be defined as functions of distances with respect to a "barycentric" histogram, which can be considered as a measure of centrality of the time series (Verde and Irpino, 2008).

We apply these new tools to the time series of daily histograms for the 5-minute returns to the SP500 index for 2007 and 2008. There are strong autocorrelations for both years that come mainly from the autocorrelations found in the extreme quantile intervals. The profile of the autocorrelation functions seem to point out to a relatively persistent autoregressive DGP, for which an exponential smoothing filter could be a good approximation. We investigate the performance of smoothing (linear dependence) and k-NN (linear and nonlinear dependence) methods to model the dynamics of the HTS and ITS data. Eventually we produce a one-step-ahead forecast of the series and evaluate the one-step-ahead forecast errors. Overall, the performance is very satisfactory, and until we have more developed methods for estimation and testing in HTS and ITS, the aforementioned classical methods are valuable tools of analysis and prediction.

The article is organized in two main sections. In the first section we review the empirical first and second moments for ITS and HTS, which are necessary to construct the autocorrelation functions. In the second section, we analyze the linear and nonlinear dependence of the daily histogram and interval time series corresponding to intradaily returns to the SP500 index. A conclusion section closes the article.

2 Empirical autocorrelation functions

In this section we deal with the concepts of linear dependence for interval-valued and histogram-valued data. First, we analyze the time dependence of an interval time series, which will be characterized by its autocorrelation function. We calculate the autocorrelations based on the descriptive statistics proposed by Bertrand and Goupil (2000) and Billard and Diday (2006). Secondly, we analyze the time dependence of a histogram time series. A key concept to construct the autocorrelation function is the barycentric histogram, which it is understood as a measure of centrality of a set of histograms, see Arroyo and Maté (2009a). The calculation of the barycenter requires the introduction of a distance measure. Both concepts, barycenter and an appropriate distance measure, are the foundations for the calculation of the empirical moments of a histogram time series.

We start with the introduction of some basic concepts and we follow with the description of the first and second moments for interval and histogram-valued data.

2.1 Interval random variable and stochastic process

Let $(E, \leq)$ be a partially ordered set. An interval $[x]$ over the base set $(E, \leq)$ is an ordered pair $[x] = [x_L, x_U]$ where $x_L, x_U \in E$ are the endpoints or bounds of the interval such that $x_L \leq x_U$. The interval is the set of elements bounded by the endpoints, these ones included, namely $[x] = \{e \in E \mid x_L \leq e \leq x_U\}$. 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 $[x] = \langle x_C, x_R \rangle$ where $x_C = (x_L + x_U)/2$ and $x_R = (x_U - x_L)/2$.

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

Definition 2. 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 1.

An interval-valued time series (ITS) 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 $t = 1, 2, ...T$.

Definition 3. An interval-valued stochastic process is said to be weakly stationary if the bounds of the interval $\{X_{Lt}\}$ and $\{X_{Ut}\}$ are weakly stationary processes. As a consequence, the center and radius processes $\{X_{Ct}\}$ and $\{X_{Rt}\}$ will also be weakly stationary since they are linear combinations of the bound processes.

We also assume that $\{X_{Lt}\}$ and $\{X_{Ut}\}$ (or $\{X_{Ct}\}$ and $\{X_{Rt}\}$) are ergodic so that the sample moments described in the forthcoming sections are consistent estimators of the population moments of the processes.

2.2 Empirical moments of an interval time series

A key assumption behind the calculations of the following sample moments is that the values in a given interval, i.e. $x_{Lt} \leq x_t \leq x_{Ut}$, are uniformly distributed within the interval. To simplify notation, we first proceed to describe the sample moments of an interval random variable X. Afterwards we will generalize them to the set of random variables X_t that form the interval process $\{X_t\}$. The empirical density function of X is defined as

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

where m is a sample of observations ($i = 1, 2, ..., m$) and it is assumed that each observation has the same probability $1/m$ of being realized. The notation $i : x \in [x_i]$ means that the sum runs for those observations for which $x \in [x_i]$.

Based on the density function (1), the sample mean is calculated as

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

so 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

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

to obtain

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

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.

For an interval time series, we will be observing one and only one interval per period of time, i.e. $[x_t] = [x_{Lt}, x_{Ut}] = \langle x_{Ct}, x_{Rt} \rangle$ for each $t = 1, 2, ...T$. Based on the properties of weakly stationarity and ergodicity of the interval stochastic process, then we can define the sample mean and the sample variance of the process as

$$\bar{X} = \frac{1}{2T} \sum_t (x_{Ut} + x_{Lt}) = \frac{1}{T} \sum_t x_{C_t} \tag{3}$$

$$S_X^2 = \frac{1}{3T} \sum_t (x_{Ut}^2 + x_{Ut} x_{Lt} + x_{Lt}^2) - \frac{1}{4T^2} \left[\sum_t (x_{Ut} + x_{Lt}) \right]^2 \tag{4}$$

Following similar reasoning, we can quantify the dependence between two interval random variables Y and X by computing their covariance. For a sample of intervals $([x_i], [y_i])$ for $i = 1, 2, ...m$, and analogously to the univariate case (1), the bivariate empirical density function is defined as

$$f_{X,Y}(x, y) = \frac{1}{m} \sum_{i: x \in [x_i], y \in [y_i]} \frac{I((x, y) \in ([x_i], [y_i]))}{|| \ ([x_i], [y_i]) \ ||} \qquad x, y \in \mathbb{R}. \tag{5}$$

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 (5), Billard and Diday (2006) propose the empirical covariance between two interval random variables defined as

$$Cov(X, Y) = \frac{1}{3m} \sum_i G_{Xi} G_{Yi} (Q_{Xi} Q_{Yi})^{1/2}, \tag{6}$$

where

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

and

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

The measure (6) reduces to (4) when $X = Y$. In addition, if all intervals are degenerate so that the sample becomes a collection of single points, the expression (6) collapses to the classical covariance measure.

When both random variables belong to the same stochastic process, we define the auto-covariance of order k as

$$\gamma_k \equiv Cov(X_t, X_{t-k}) = \frac{1}{3T} \sum_t G_{X_t} G_{X_{t-k}} (Q_{X_t} Q_{X_{t-k}})^{1/2}, \tag{7}$$

where

$$\begin{aligned} Q_{X_t} &= (x_{Lt} - \bar{X})^2 + (x_{Lt} - \bar{X})(x_{Ut} - \bar{X}) + (x_{Ut} - \bar{X})^2 \\ Q_{X_{t-k}} &= (x_{L,t-k} - \bar{X})^2 + (x_{L,t-k} - \bar{X})(x_{U,t-k} - \bar{X}) + (x_{U,t-k} - \bar{X})^2 \end{aligned}$$

and

$$G_{X_t} = \begin{cases} -1, & \text{if } X_{Ct} \leq \bar{X}, \\ 1, & \text{if } X_{Ct} > \bar{X}, \end{cases} \qquad G_{Yi} = \begin{cases} -1, & \text{if } X_{C,t-k} \leq \bar{X}, \\ 1, & \text{if } X_{C,t-k} > \bar{X}, \end{cases}$$

When $k = 0, \gamma_0 = S_X^2$. Consequently, the empirical autocorrelation function of an interval time series is defined by

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{8}$$

This measure has the properties of a well-defined correlation coefficient, and as such it is bounded between $[-1, 1]$.

2.3 Histogram random variable and stochastic process

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

$$h_{X_i} = \{([x]_{i1}, \pi_{i1}), ..., ([x]_{in_i}, \pi_{in_i})\}, \text{for } i \in S, \tag{9}$$

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 $\{h_{X_i}, i = 1....m\}$.

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. Define a partition of Ω into sets $A_X(x)$ such $A_X(x) = \{\omega \in \Omega \imath\ X(\omega) = x\}$, where $x \in \{h_{X_i}, i = 1....m\}$.

Definition 4. A mapping $h_X : \mathcal{F} \to \{h_{X_i}\}$, such that, for all $x \in \{h_{X_i}, i = 1....m\}$ there is a set $A_X(x) \in \mathcal{F}$, is called a histogram random variable.

Then, the definition of stochastic process follows as:

Definition 5. 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 4.

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

Definition 6. A histogram-valued stochastic process is said weakly stationary if every interval in h_{X_t}, i.e. $[x]_{t1}, [x]_{t2}...[x]_{tn_t}$, satisfies Definition 3.

2.4 Empirical moments of a histogram time series

Assuming that the histogram stochastic process is stationary and ergodic, we can calculate consistent sample moments based on time series information. First, we introduce the notion of barycentric histogram, which will play the role of a central measure or first moment, and afterwards we will define the second empirical moments by following Verde and Irpino (2008).

Let $\{h_{X_t}\}$ with $t = 1, ..., T$ be a histogram time series. The barycentric histogram $\hat{h}_c$ is the histogram that minimizes the distances between itself and all T histograms in the time series. It is obtained by solving the following optimization problem

$$\hat{h}_c = \arg\min_{h_c} (\sum_{t=1}^{T} D(h_{X_t}, h_c)^p)^{1/p}, \tag{10}$$

where $D(h_{X_t}, h_c)$ is a dissimilarity or distance measure for histogram-valued data and $p \geq 1$.

In this sense, the barycenter is the histogram that reflects the central tendency of a collection of histograms. We choose the Mallows distance D_M, which is defined as

$$D_M(h_1, h_2) = \sqrt{\int_0^1 (H_1^{-1}(r) - H_2^{-1}(r))^2 dr}, \tag{11}$$

where $H^{-1}(r)$ is the inverse of the distribution function, i.e. the quantile function, of the histogram h. In addition, we choose $p = 2$, so that the barycenter (10) is calculated as

$$\begin{aligned} \hat{h}_c &= \arg\min_{h_c} (\sum_{t=1}^{T} D_M(h_{X_t}, h_c)^2)^{1/2} \\ &= \arg\min_{h_c} \sum_{t=1}^{T} \int_0^1 (H_{X_t}^{-1}(r) - H_c^{-1}(r))^2 dr, \end{aligned}$$

The choice of the Mallows distance is not arbitrary. As Irpino and Verde (2006) and Arroyo and Mate (2009) show the minimization with Mallows reduces to a least squares problem for which the barycentric solution involves a collection of averages. 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 T histograms,

$$\begin{aligned} x^*_{Cj} &= \frac{\sum_{t=1}^{T} x_{Ctj}}{T} \\ x^*_{Rj} &= \frac{\sum_{t=1}^{T} x_{Rtj}}{T}. \end{aligned}$$

More interestingly, the r^{th} quantile $\hat{H}_c^{-1}(r)$ of the barycenter $\hat{h}_c$ is the mean of the r^{th} quantiles $H_{X_t}^{-1}(r)$ of all histograms h_{X_t} in the set. Thus, the barycenter behaves as a mean histogram that averages the location, the range and the shape of the set of histograms. For more details, see Arroyo and Maté (2009a).

A dispersion statistics associated with $\hat{h}_c$ measures the average of the squared Mallows-distances from the barycenter $\hat{h}_c$

$$S^2_{h_{X_t}, \hat{h}_c} = \frac{\sum_{t=1}^{T} D_M(h_{X_t}, \hat{h}_c)^2}{T}. \tag{12}$$

We call $S^2_{h_{X_t}, \hat{h}_c}$ the sample variance of the histogram time series with respect to the barycenter $\hat{h}_c$.

Following similar arguments, Verde and Irpino (2008) propose a covariance measure for two histogram variables that we adapt to the concept of autocovariance. Given a histogram time series the sample autocovariance of order k is defined as

$$\gamma_k \equiv Cov(h_{X_t}, h_{X_{t-k}}; \hat{h}_c) = \frac{\sum_{t=1}^{T} \int_0^1 (H_{X_t}^{-1}(r) - \hat{H}_c^{-1}(r))(H_{X_{t-k}}^{-1}(r) - \hat{H}_c^{-1}(r))dr}{T}. \quad (13)$$

Consequently, the empirical autocorrelation function of a histogram time series with respect to the barycenter $\hat{h}_c$ is defined by

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad (14)$$

As in the case of interval time series, this measure has the properties of a well-defined correlation coefficient, and as such it is bounded between $[-1, 1]$. In the Appendix 1, we provide some illustrative examples of correlation between histograms that should be helpful to understand formulas (12) and (13).

3 Autocorrelation of the distribution of SP500 intradaily returns

In this section we analyze the daily histograms of the 5-minute SP500 returns for each trading day in 2007 and 2008. After introducing the empirical autocorrelation function of the histogram time series, we aim to explain its dependence by passing the data through an exponential smoothing filter, which aims to model some linear dependence, and by estimating a nonparametric model adopting the k-NN method, which aims to model linear and nonlinear dependence in the HTS. Both of these methods rely on the barycentric approach for the computation of averages. We will also construct the one-step-ahead histogram forecast with both methods.

3.1 Daily histogram of SP500 intradaily returns

For each day in 2007 and 2008 we construct a daily histogram with the 5-minute SP500 returns. There are 77 returns in each day. Each histogram consists of 10 bins, each containing 10% of the intradaily returns. We analyze 2007 and 2008 separately. In 2007 there are 251

daily histograms. We split the sample into an estimation period with 200 observations (January to mid-October) and a forecasting period with 51 observations (mid-October to December). In 2008 we have 253 observations, from which 200 are used for estimation and 53 are reserved for assessment of the one-step-ahead forecasts. In Figures 1 and 2 we plot the daily histograms for 2007 and 2008 respectively, and in Figure 3, the daily closing prices for both years.

[FIGURES 1, 2, 3]

Overall, the year 2008 is substantially more volatile than 2007. A very prominent feature in the series is the clustering of observations. There are clusters of low volatility, such as the early months of 2007, and clusters of high volatility, such as the months of July/August 2007 when the worldwide liquidity crisis began, and the late summer and fall of 2008 when the economy went into financial meltdown. In Figure 3, we observe that the seven most volatile periods correspond to the seven largest price corrections in the market: August and November 2007, and January, March, July, October, and November 2008.

With the barycentric approach proposed in the previous section, we compute the autocorrelation function of the daily histograms for the estimation samples of 2007 and 2008. These functions are pictured in Figure 4.

[FIGURE 4]

The profile of the autocorrelation functions is similar for both years: positive and strong autocorrelations, starting around 0.5, with relatively slow decay towards zero. In 2008 the autocorrelations are slightly higher and more persistent than those in 2007. This is not surprising since in 2008 the clusters of volatile periods are of longer duration than those in 2007. Given this strong linear dependence, we aim to analyze how much of it can be modelled by passing the data through an exponential smoothing filter and what forecasting performance the filter can deliver. However, there may be additional nonlinear dependence in the series that the exponential method will not address. For that purpose, we also implement a variant of the k-NN method adapted to histogram data, which will be able to detect any

	year 2007		year 2008	
Models	estimation	prediction	estimation	prediction
Naive	3.98E-04	4.73E-04	8.09E-04	13.07E-04
Smoothing	3.35E-04	4.29E-04	6.38E-04	11.84E-04
k-NN	3.14E-04	3.93E-04	6.3E-04	10.68E-04

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

neglected linear and nonlinear dependence. A brief summary on the implementation of both methods for histogram and interval time series is provided in the Appendix 2.

The following Table 1 summarizes the performance of the exponential smoothing and the k-NN methods in the estimation and prediction periods for both years 2007 and 2008. 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. $\hat{h}_{X_{t+1|t}} = h_{X_t}$. For each method, we report the mean distance error (MDE) defined as

$$MDE^q(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \left(\frac{\sum_{t=1}^{T} (D(h_{X_t}, \hat{h}_{X_t}))^q}{T} \right)^{\frac{1}{q}}, \tag{15}$$

where $D(h_{X_t}, \hat{h}_{X_t})$ is the Mallows distance.

The optimal values of α (exponential smoothing) and k and d (k-NN method) are obtained by minimizing the MDE function. In the exponential smoothing, this function happens to be rather flat around the minimum so that the optimal value is $\hat{\alpha} \in (0.2, 0.5)$ for 2007 and $\hat{\alpha} \in (0.2, 0.4)$ for 2008. In the k-NN, the optimal values are $k = 11$ and $d = 2$ for 2007, and $k = 8$ and $d = 2$ for 2008. Comparing the three methods, the naive approach is outperformed by both smoothing and k-NN methods. The latter delivers the smallest MDE in the estimation and prediction samples for both years 2007 and 2008. In the year 2008 the MDE's are substantially larger than those in 2007. As we have seen in Figure 2, the year 2008 is much more volatile than 2007 and consequently any estimation method based on averages will necessarily deliver larger errors. We should also note that the prediction sample in 2008 (mid-October to December) is very different –highly volatile– from the estimation sample, rendering forecasting a very difficult exercise. It is fair to say that we are evaluating the performance of these methods in a "worst-case" scenario. Given this fact, the overall

results are quite satisfactory.

We measure the in-sample performance of these methods by evaluating the classical autocorrelation function of the distance error $D(h_{X_t}, \hat{h}_{X_t})$ in (15), which is the dissimilarity between the realized and the estimated observations. This measure plays the role of a "residual" and as such, we should not expect any autocorrelation if the aforementioned methods are successful on modeling the autocorrelation of the original data. In Figure 5, we present the autocorrelation functions of the "residual" for the three methods and for both years 2007 and 2008.

[FIGURE 5]

Comparing these autocorrelations with those in Figure 4 we observe that, with the exception of the naive model, they are much smaller than those in the original data. In addition, and as we have already mentioned, the year 2008 is notoriously more difficult to model than 2007. The naive model is the worst performer because the autocorrelation function has an almost identical profile to that of the original series. On the other hand, the k-NN method is the most successful, confirming the results of Table 1. With k-NN the autocorrelation of the residual has been brought down to values around 0.2 when the autocorrelation in the original data was around 0.5. Though this is a very good improvement and the residual autocorrelations are mostly small and statistically insignificant, there seems to be some structure left in the data that will require new modelling techniques.

We assess the out-of-sample performance by analyzing the autocorrelation of the one-step-ahead forecast distance error, which is measured as the distance between the realized histogram and the one-step-ahead predicted histogram, i.e. $D(h_{X_{t+1}}, \hat{h}_{X_{t+1|t}})$. We call $\varepsilon_{t+1|t} \equiv D(h_{X_{t+1}}, \hat{h}_{X_{t+1|t}})$ the one-step-ahead forecast distance error. This error will not have mean zero because conceptually it is a distance, but nevertheless it should behave as an innovation with respect to the set containing information up to time t . As such, we should not expect any autocorrelation in the forecast error if the forecast $\hat{h}_{X_{t+1|t}}$ exploits the information set efficiently. Consequently, we have calculated the classical autocorrelation function of $\varepsilon_{t+1|t}$ for both years 2007 and 2008 and for the three models considered in Table 1. In Figure 6 we present the autocorrelation functions.

[FIGURE 6]

There are similar features to those encountered in Figure 5. The naive model is outperformed by the exponential smoothing and the k-NN. In the naive model, there seems to be a systematic error given the profile of the autocorrelation function. This is not the case in the autocorrelation functions of the one-step-ahead forecast distance error corresponding to the smoothing and the k-NN methods. As before, 2008 is more unpredictable than 2007.

The overall assessment of these methods is very satisfactory. They are able to capture a substantial amount of the autocorrelation in the histogram time series and they forecast very well mainly when we consider that we are dealing with financial returns, which are notoriously difficult to model and forecast.

3.2 Quantile intervals

An advantage of estimating and predicting a histogram is that we can analyze each bin separately. The researcher may be just interested in the dynamics of the extreme quantiles or in the center quantiles or in any combination of quantiles. As an illustration, we focus on the upper tail of the histogram of the SP500 intradaily returns and we analyze the [90,100%] quantile interval that contains the largest 10% of the intradaily returns. The interval time series (ITS) is plotted in Figures 7 (2007) and 8 (2008). As we have seen in the HTS, the most extreme events correspond to September to December 2008. We proceed in similar fashion as in the analysis of HTS. For each year, we split the sample into an estimation sample (first 200 observations) and a prediction sample (next 51 observations in 2007, and 53 in 2008).

[FIGURES 7 AND 8]

We calculate the empirical autocorrelation functions according to (7). These are plotted in Figure 9 together with those corresponding to the [50,60%] quantile interval, which contains the 10% of the observations around the middle of the distribution.

[FIGURE 9]

	year 2007		year 2008	
Models	estimation	prediction	estimation	prediction
Smoothing (ITS)	0.0011	0.0009	0.0027	0.0028
Smoothing (HTS)	0.0011	0.0009	0.0027	0.0028
k-NN (ITS)	0.0010	0.0012	0.0028	0.0035
k-NN (HTS)	0.0010	0.0008	0.0027	0.0027

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

A very interesting finding is that the [90,100] quantile interval has very strong autocorrelations that mimic those found in the HTS, while the [50,60] quantile interval has no correlation whatsoever. We calculate the interval-autocorrelations of all the 10 bins in the histogram and we find that the central bins have much lower autocorrelation than the extreme bins. This means that the linear dependence in HTS comes from the very extreme quantiles and those close to them.

In Table 2, we present the performance of the smoothing and k-NN methods measured by the MDE for interval data, i.e.

$$MDE^q(\{[x]_t\}, \{[\hat{x}]_t\}) = \left(\frac{\sum_{t=1}^{T} (D([x]_t, [\hat{x}]_t))^q}{T} \right)^{\frac{1}{q}}, \tag{16}$$

where $D([x], [\hat{x}]) = \sqrt{(x_C - \hat{x}_C)^2 + (x_R - \hat{x}_R)^2}$ is an Euclidean-type distance, see González et al. (2004). In the table we distinguish between smoothing and k-NN for ITS and HTS. When we write ITS, we mean that we analyze the interval time series according to the aforementioned methodology for intervals; and when we write HTS, we mean that the object of analysis is the histogram time series and we extract the corresponding interval to the [90,100] quantile from the estimated or predicted histogram. According to the MDE criterion, there is similar performance across methods in the estimation period for both years 2007 and 2008. However, out-of-sample, there is an advantage on using the k-NN with HTS as it produces the smallest mean prediction error.

Similarly, observing the in-sample autocorrelation functions of the distance error $D([x]_t, [\hat{x}]_t)$ in Figure 10, the k-NN (HTS) is the best performer in 2007 as it is able to capture all the autocorrelation of the original [90,100] quantile interval. In 2008, there is some small positive autocorrelation left but nevertheless the residual autocorrelation is substantially smaller

than that of the original autocorrelation.

[FIGURE 10]

Out-of-sample, the autocorrelation functions of the one-step-ahead forecast distance error in Figure 11, i.e. $\varepsilon_{t+1|t} \equiv D([x]_{t+1}, [\hat{x}]_{t+1|t})$ reveals that these errors are basically white noise, and this is the case across methods and in both years 2007 and 2008.

[FIGURE 11]

Given this evidence, it is fair to say that, even if the researcher is interested in a specific interval, modelling the histogram time series is a strategy that should be considered because the global information that HTS provides is helpful on improving the estimation and forecasting of ITS.

4 Conclusions

Histogram and interval time series are symbolic data sets that require new methods of analysis. Though there are methodological advances within cross-sectional symbolic data sets, from a time series perspective we are in the early stages of development. Arroyo, González-Rivera, and Maté (2010) and Arroyo and Maté (2009) have shown that classical algorithms, such as smoothing filters and k-NN methods, can be adapted for HTS and ITS forecasting. The methods that they entertained are appealing because of the relatively low number of parameters to estimate, and their simplicity in implementation as they are based on the construction of suitable averages. Therefore, forecasting with these methods is based on the implicit assumption that, in a relatively stable world, the future should not be very far from some average of the past. In other words, they do not attempt to uncover the data generating process of HTS and ITS, though the methods are quite flexible on adapting to changes in the underlying true process.

In this paper we have taken the view that the first building block towards the search for a DGP is to understand the dynamics of HTS and ITS. We have focused on the empirical autocorrelation functions of HTS but, since a histogram is a collection of bins (or intervals)

with specific frequency, it is also relevant to understand the autocorrelation functions of interval time series. In HTS, the main ingredient that percolates through the analysis is the notion of distance with respect to a barycentric histogram, which should be understood as a central histogram. The empirical second moments, variance and covariances, are functions of quantile Mallows-distances between histograms and their barycenter. We have applied these concepts to the daily histogram of the 5-minute intradaily returns to the SP500 index for 2007 and 2008. The main finding is that there are clusters of high/low activity that produces a strong, positive, and persistent autocorrelation in the HTS. This profile points towards some autoregressive process for HTS. We have investigated the performance of exponential smoothing and k-NN methods on modelling the HTS autocorrelation. Understanding that these models may not be the true DGPs, we have found that they are very good approximations because they are able to capture almost all the original autocorrelation. However, there seems to be some structure left in the data that will require new modelling techniques. We have also analyzed the [90,100%] quantile interval following similar techniques for interval data, and we have compared it with our findings based on histogram data. By using the full information contained in the histogram, we find that there are some advantages in the estimation and prediction of a specific interval.

References

Arroyo, J., G. González-Rivera, and C. Maté (2010). *Handbook of Empirical Economics and Finance (forthcoming)*, Chapter :Forecasting with interval and histogram data. Some financial applications. Chapman & Hall/CRC.

Arroyo, J. and C. Maté (2009a). Descriptive distance-based statistics for histogram data. *Working paper*.

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

Becketti, S. and W. Gould (1987). Rangefinder box plots: A note. *The American Statistician 41*(2), 149.

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

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

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.

Han, A., Y. Hong, K. Lai, and S. Wang (2008). Interval time series analysis with an application to the Sterling-Dollar exchange rate. *Journal of Systems Science and Complexity 21*(4), 558–573.

Irpino, A. and R. Verde (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In *Data Science and Classification, Proceedings of the IFCS 2006*, Berlín, pp. 185–192. Springer.

Maia, A. L. S., F. d. A. de Carvalho, and T. B. Ludermir (2008). Forecasting models for interval-valued time series. *Neurocomputing 71*(16–18), 3344–3352.

Verde, R. and A. Irpino (2008). Comparing histogram data using a Mahalanobis-Wasserstein distance. In *COMPSTAT 2008. Proceedings in Computational Statistics*, pp. 77–89. Springer.

Appendix 1

In Figure 12 we illustrate the meaning of correlation between two histogram random variables, say h_X and h_Y, contained in the formulas (12) and (13).

[FIGURE 12]

To simplify the graphical exposition let us assume that each histogram is represented by a box plot, which consists of five elements: the minimum, the 25, 50, 75 quantiles, and the maximum. Call these $(x_{\min}, x_{25}, x_{50}, x_{75}, x_{\max})$ for h_X, and $(y_{\min}, y_{25}, y_{50}, y_{75}, y_{\max})$ for h_Y. In the plane (X, Y), the observation i is a pair of histograms (h_{X_i}, h_{Y_i}) that gives rise to two rectangles, the inner and the outer. The inner rectangle is the area limited by the vertexes $(x_{25}, y_{25})_i, (x_{25}, y_{75})_i, (x_{75}, y_{25})_i$, and $(x_{75}, y_{75})_i$. The median is the interior point $(x_{50}, y_{50})_i$. The outer rectangle is limited by $(x_{\min}, y_{\min})_i, (x_{\min}, y_{\max})_i, (x_{\max}, y_{\min})_i$, and $(x_{\max}, y_{\max})_i$. All points $(x, y)_i$ within the i observation (h_{X_i}, h_{Y_i}) will be enclosed within the perimeter of the outer rectangle. This type of graph is similar to the rangefinder box plot, see Becketti and Gould (1987). Our interest is the empirical correlation between h_X and h_Y when we have observations like (h_{X_i}, h_{Y_i}). In other words, the scatter plot is a cloud of of observations (h_{X_i}, h_{Y_i}) where each is displayed as a two-rectangle object, and our aim is to find the correlation of these objects. Note that the distribution of the points $(x, y)_i$ within each observation (h_{X_i}, h_{Y_i}) is not taken into account.

In Case 1, we have three observations. Observe that we can trace a line connecting all median points $(x_{50}, y_{50})_i$, the points $(x_{\min}, y_{\min})_i, (x_{25}, y_{25})_i, (x_{75}, y_{75})_i$, and $(x_{\max}, y_{\max})_i$ for $i = 1, 2, 3$. In this case, the correlation coefficient is equal to one. In Case 2, we also have three observations but now the vertex $(x_{\max}, y_{\max})$ in the middle observation is not longer aligned. The resulting correlation drops to 0.9872. In Case 3, we consider the same observations as in Case 1 and we add a fourth observation for which the inner and outer rectangles are clearly misaligned in relation to the remaining three points. In this case, the correlation coefficient is 0.4559. In Case 4, there are four observations spread all over the plane with no obvious alignment around the median points or around any vertexes, as a result, the correlation is zero. These are illustrative examples on how the correlation varies

depending upon the alignments of the median points and the vertexes that defined the inner and the outer rectangles of each observation.

Appendix 2

We summarize the implementation of exponential smoothing and k-NN methods for estimation and prediction of histogram and interval-valued time series. For a more detailed explanation, please see Arroyo et al. (2010).

Exponential smoothing

The exponential smoothing is a weighted average of the most recent observation and its forecast. For a histogram time series, the exponentially smoothed forecast is given by the following equation

$$\hat{h}_{X_{t+1}} = \alpha h_{X_t} + (1-\alpha)\hat{h}_{X_t}, \tag{17}$$

where $\alpha \in [0,1]$, which, by backward substitution, it can also be expressed as a weighted average of all past observations

$$\hat{h}_{X_{t+1}} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1} h_{X_{t-(j-1)}}. \tag{18}$$

Since this is an average of histograms, the histogram forecast $\hat{h}_{X_{t+1}}$ can be understood as a barycenter, which will be obtained by solving the following optimization problem

$$\hat{h}_{X_{t+1}} \equiv \arg\min_{\hat{h}_{X_{t+1}}} \left(\sum_{j=1}^{t} \omega_i D^q(\hat{h}_{X_{t+1}}, h_{X_{t-(j-1)}}) \right)^{1/q}, \tag{19}$$

with $D(\cdot,\cdot)$ as the Mallows distance. The estimation of α is performed by minimizing a mean distance error as

$$MDE^q(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \left(\frac{\sum_{t=1}^{T} D^q(h_{X_t}, \hat{h}_{X_t})}{T} \right)^{\frac{1}{q}}, \tag{20}$$

where, for a chosen α, $\hat{h}_{X_t}$ is obtained from (18), $D(\cdot,\cdot)$ is the Mallows distance, and $q = 1$.

Analogously, for an interval-valued time series, the exponential smoothed forecast is written as

$$[\hat{x}]_{t+1} = \alpha[x]_t + (1-\alpha)[\hat{x}]_t. \tag{21}$$

where $\alpha \in [0,1]$, which in its recursive form reads as

$$[\hat{x}]_{t+1} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1}[x]_{t-(j-1)}. \tag{22}$$

As before, the estimation of α proceeds by minimizing a mean distance error

$$MDE^q(\{[x]_t\}, \{[\hat{x}]_t\}) = \left(\frac{\sum_{t=1}^{T}(D^q([x]_t, [\hat{x}]_t))}{T} \right)^{\frac{1}{q}}, \tag{23}$$

where the distance is an Euclidean-type measure as

$$D([x], [\hat{x}]) = \sqrt{(x_C - \hat{x}_C)^2 + (x_R - \hat{x}_R)^2} \tag{24}$$

k-NN method

The k-NN method adapted to histogram and interval time series consists of the following steps:

1. The histogram time series, $\{h_{X_t}\}$ or the interval time series $\{[x]_t\}$ with $t = 1, ..., T$, is organized as a series of d-dimensional histogram-valued vectors $\{h^d_{X_t}\}$ or interval-valued vectors $\{[x]^d_t\}$ where

$$h^d_{X_t} = (h_{X_t}, h_{X_{t-1}}, ..., h_{X_{t-(d-1)}})', \tag{25}$$

$$[x]^d_t = ([x]_t, [x]_{t-1}, ..., [x]_{t-(d-1)})', \tag{26}$$

   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

$$D_t(h^d_{X_T}, h^d_{X_t}) = \left(\frac{\sum_{i=1}^{d}\left(D^q(h_{X_{T-i+1}}, h_{X_{t-i+1}})\right)}{d} \right)^{\frac{1}{q}}, \tag{27}$$

   where $D^q(h_{X_{T-i+1}}, h_{X_{t-i+1}})$ is the Mallows distance of order q. With interval data we assess the dissimilarity between the most recent interval-valued vector $[x]^d_T = ([x]_T, [x]_{T-1}, ..., [x]_{T-d+1})'$ and the rest of the vectors in $\{[x]^d_t\}$ using the Euclidean-type measure (24) in $D_t([x]^d_T, [x]^d_t)$.

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

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

$$\hat{h}_{X_{T+1}} \equiv \arg\min_{\hat{h}_{X_{T+1}}} \left(\sum_{p=1}^{k} \omega_p D^q(\hat{h}_{X_{T+1}}, h_{X_{T_p+1}}) \right)^{1/q}, \tag{28}$$

where $D(\hat{h}_{X_{T+1}}, h_{X_{T_p+1}})$ is the Mallows 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 $\sum_{p=1}^{k} \omega_p = 1$. For interval time series, the subsequent values, $[x]_{T_1+1}, [x]_{T_2+1}..., [x]_{T_k+1}$, to the k-closest vectors are averaged to obtain the final forecast

$$[\hat{x}]_{T+1} = \sum_{p=1}^{k} \omega_p \cdot [x]_{T_p+1}, \tag{29}$$

where $[x]_{T_p+1}$ is the consecutive interval of the sequence $[x]^d_{T_p}$. The average (29) is computed according to the rules of interval arithmetic.

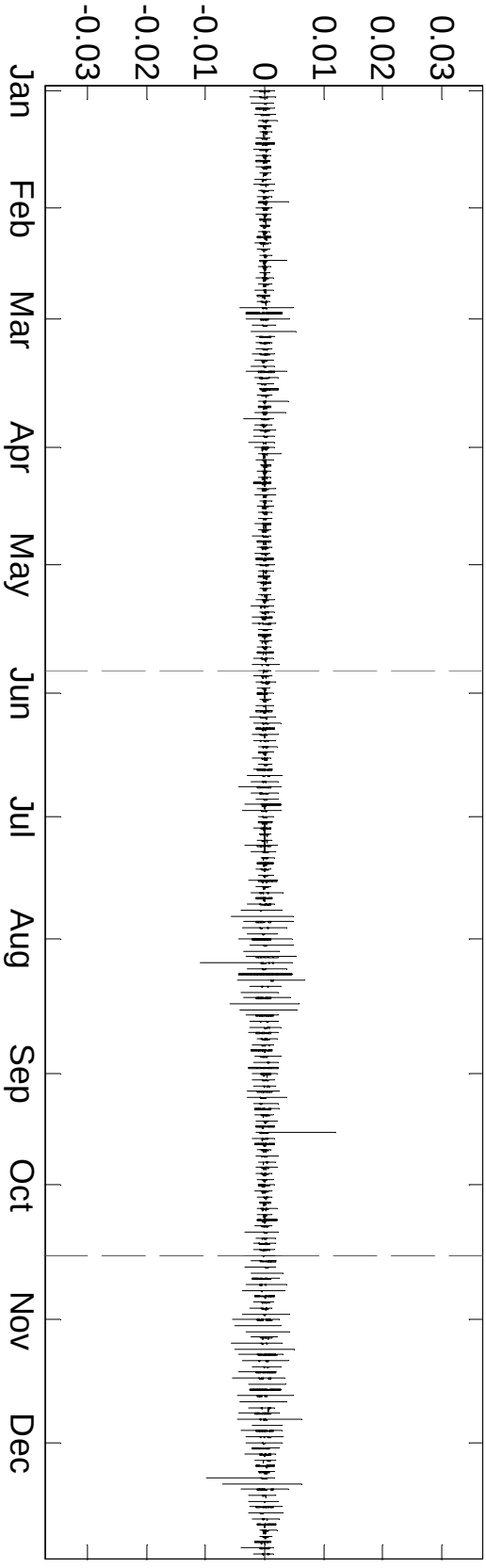


Figure 1. Daily HTS of 5-minute intradaily SP500 returns in 2007

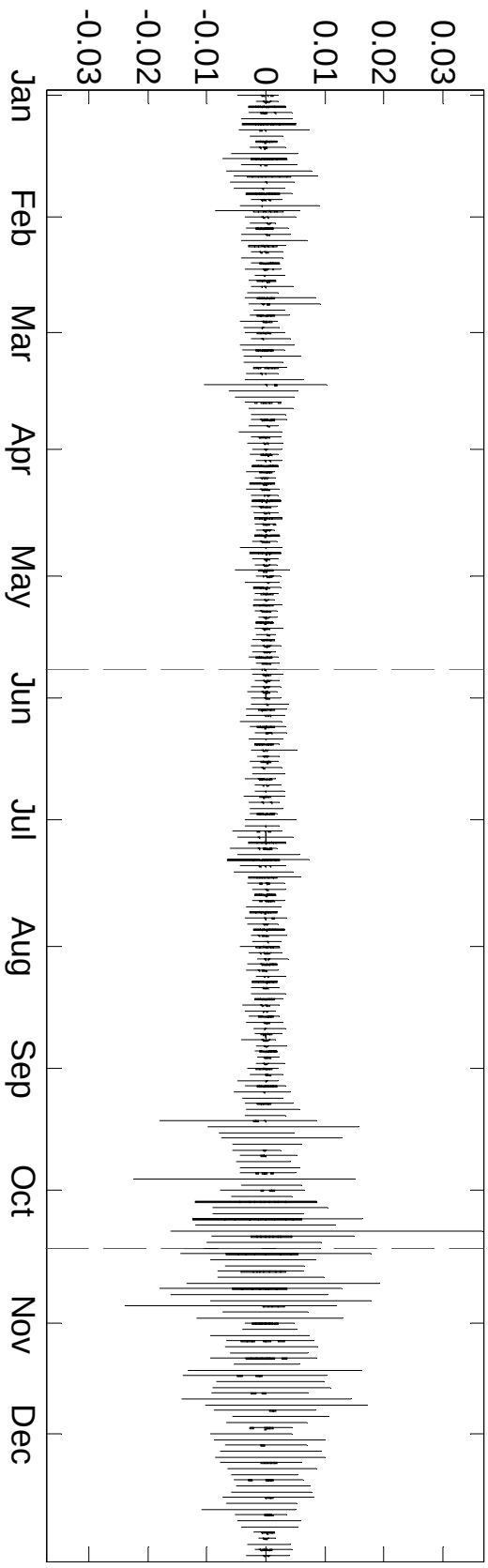


Figure 2. Daily HTS of 5-minute intradaily SP500 returns in 2008

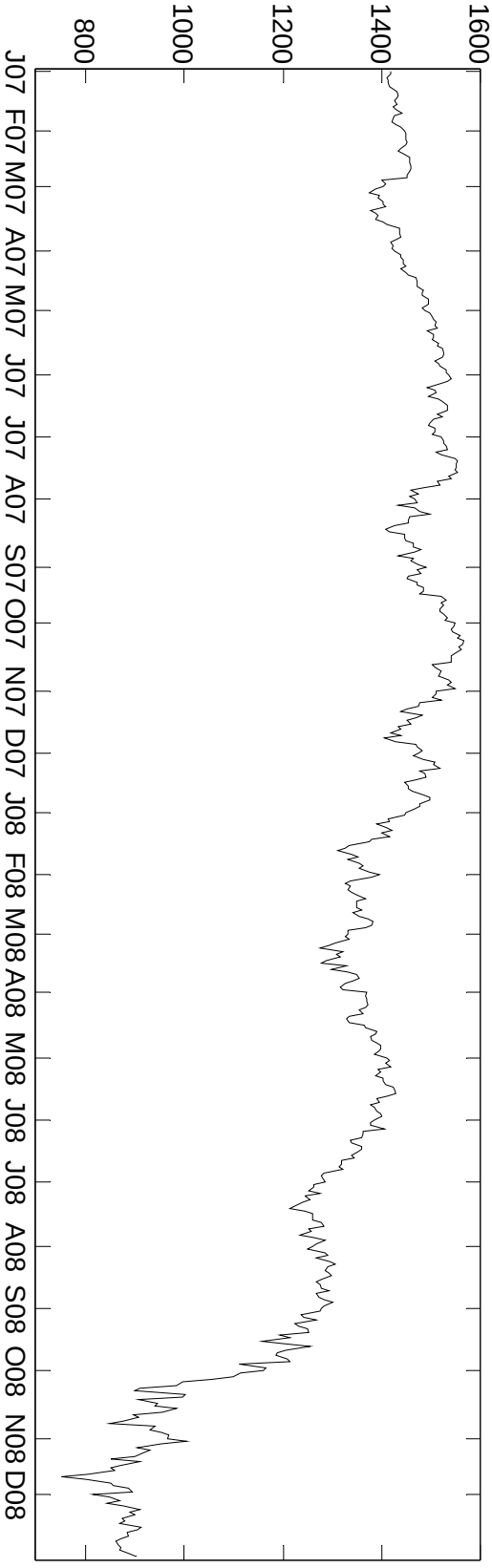
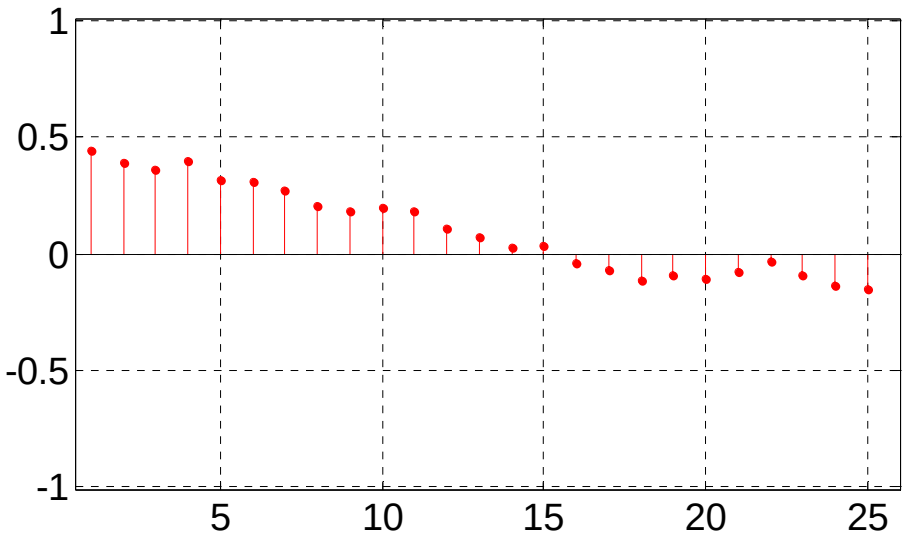


Figure 3. Time series of the daily SP500 closing prices in 2007 and 2008

2007



2008

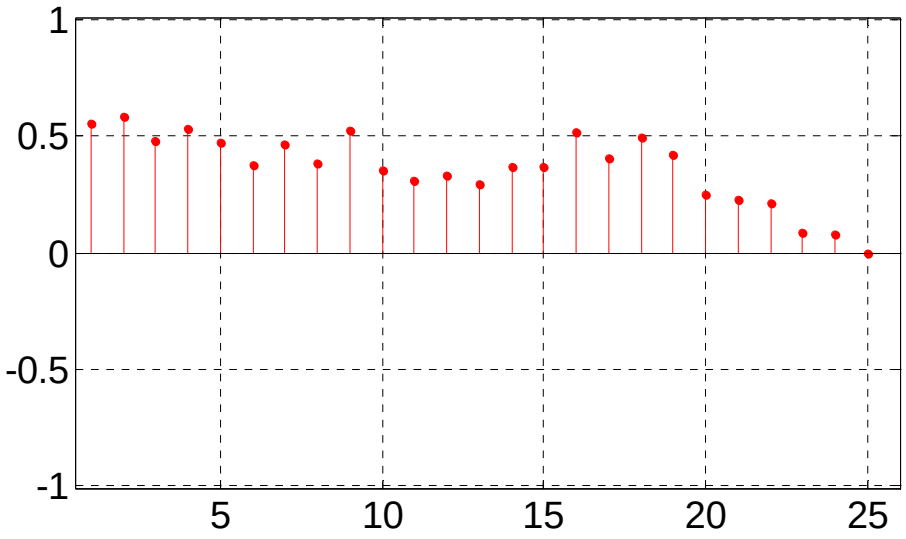


Figure 4. Empirical ACFs of the daily HTS of SP500 intradaily returns in the estimation periods (January to mid-October) of 2007 and 2008

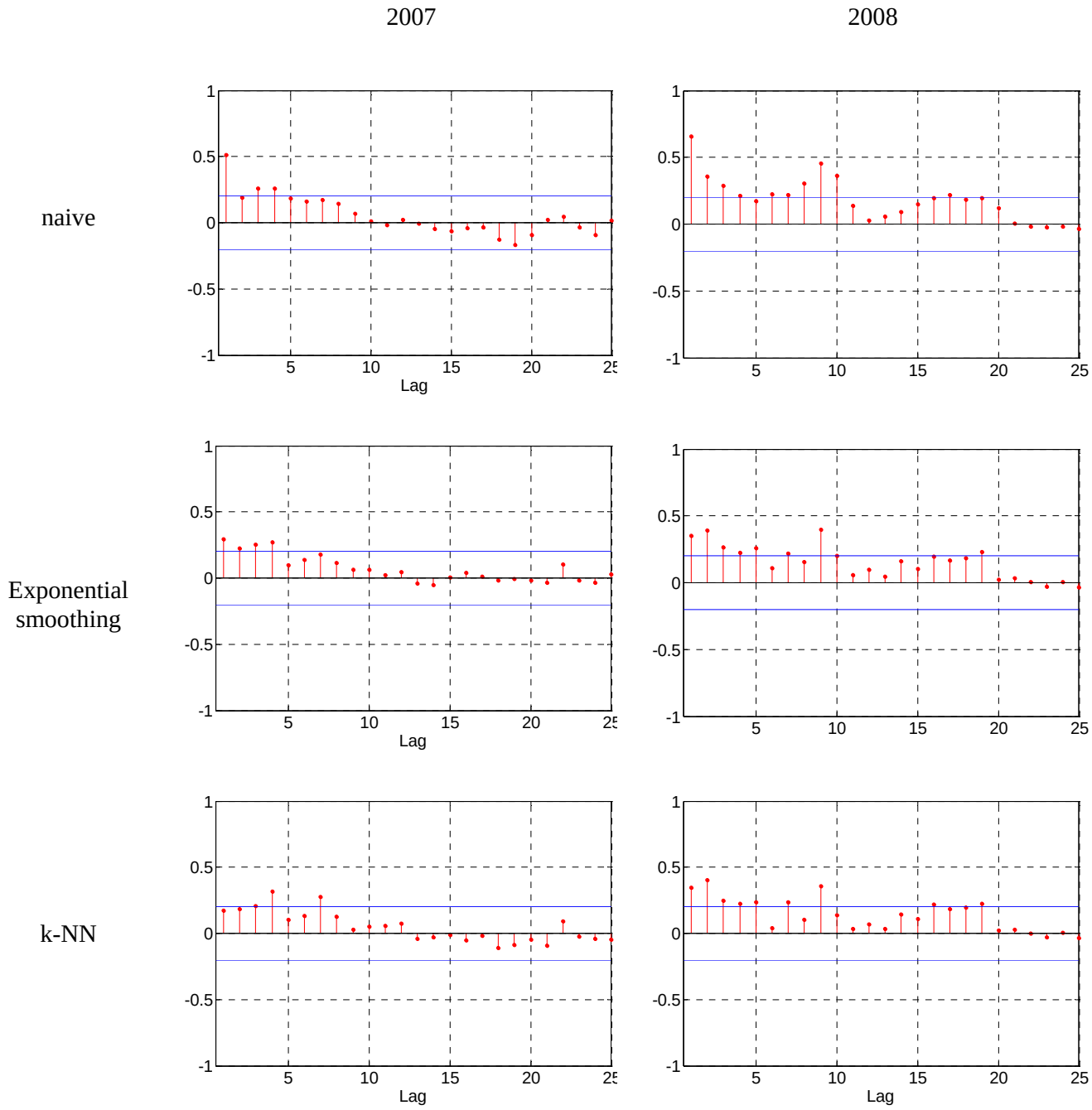


Figure 5. Empirical ACFs of the “residuals” from the estimation of naïve, exponential smoothing, and k-NN models in 2007 and 2008

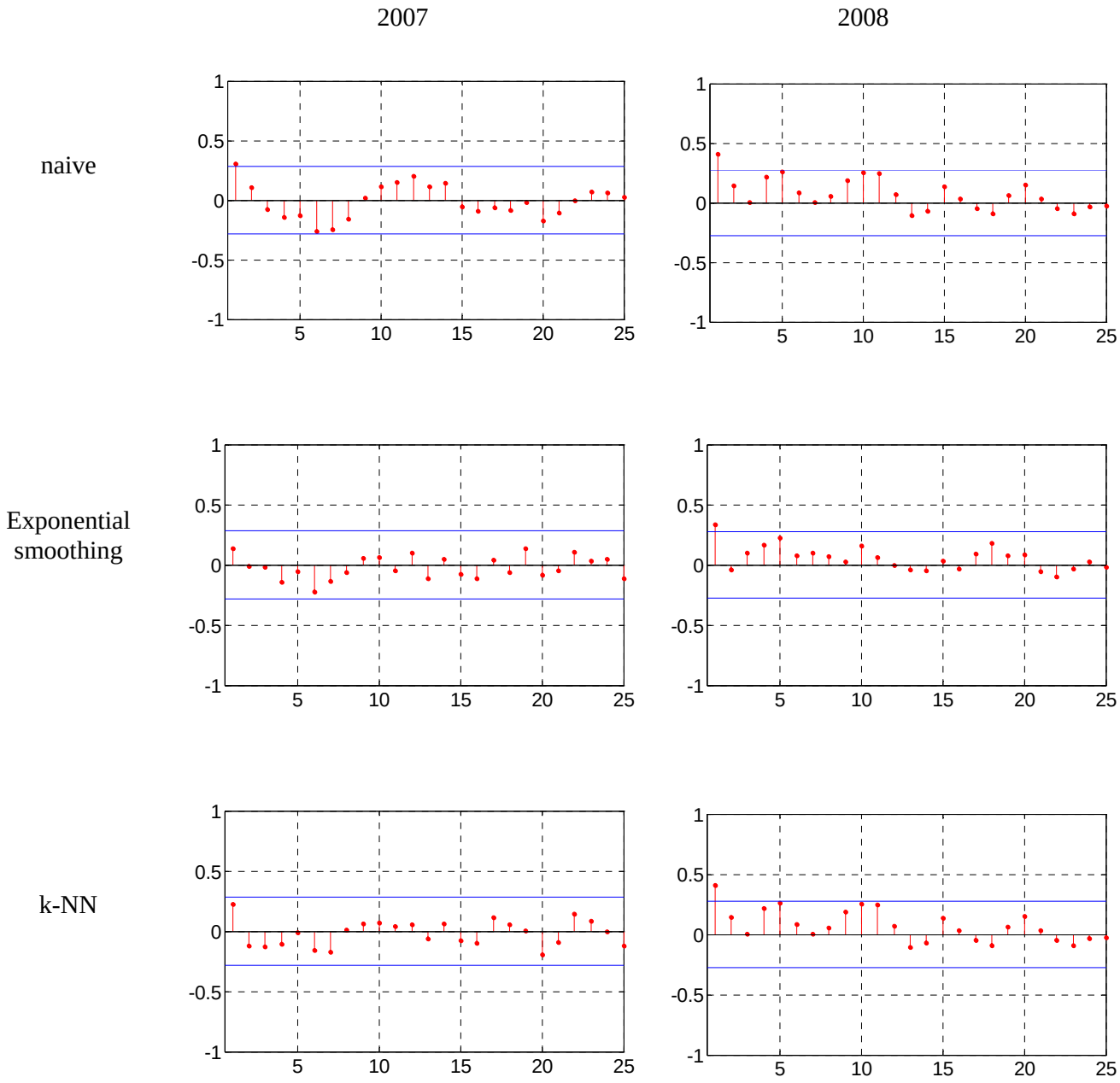


Figure 6. Empirical ACFs of the one-step-ahead forecast distance error from the prediction with naïve, exponential smoothing, and k-NN methods in 2007 and 2008

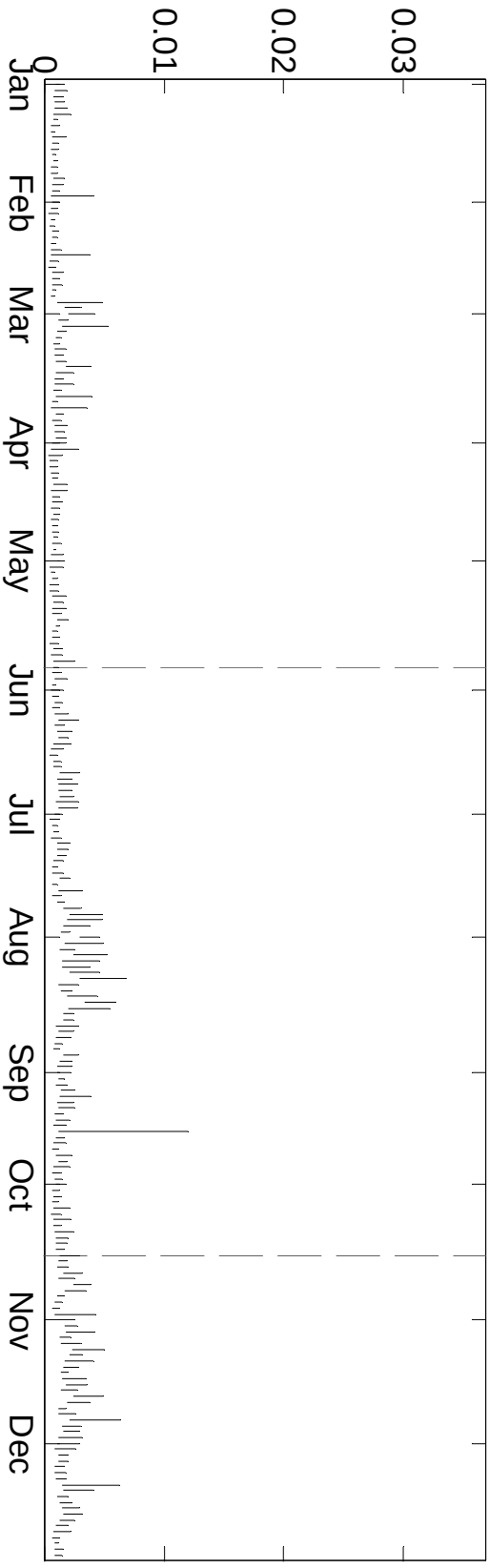


Figure 7. ITS of the [90,100] quantile interval in 2007

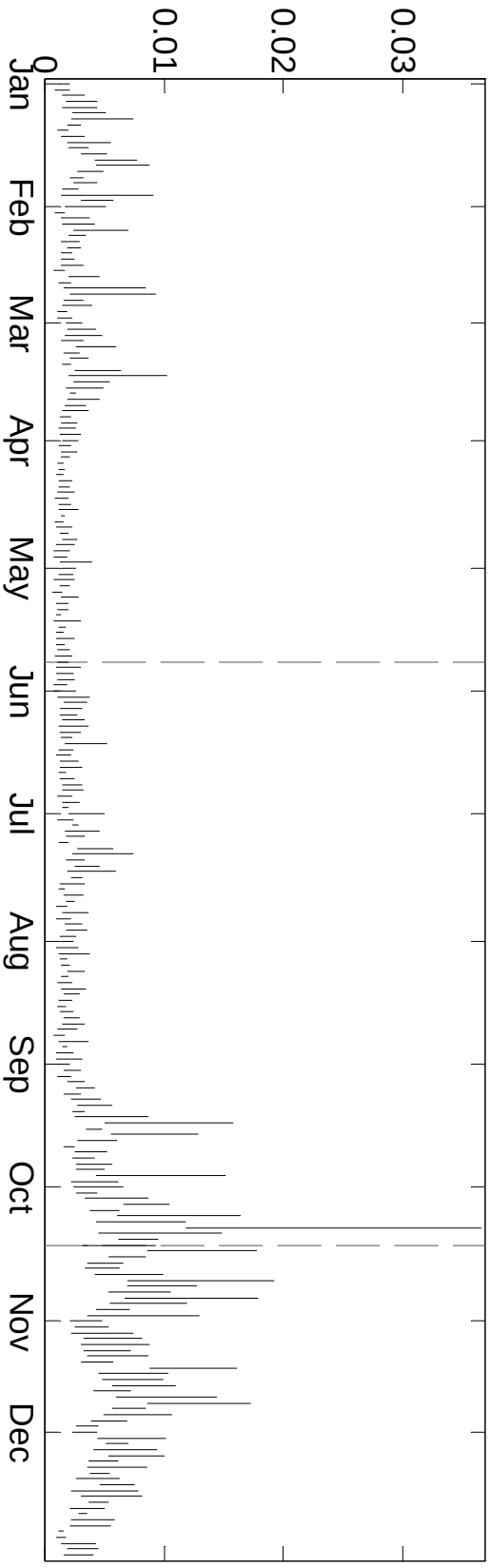


Figure 8. ITS of the [90,100] quantile interval in 2008

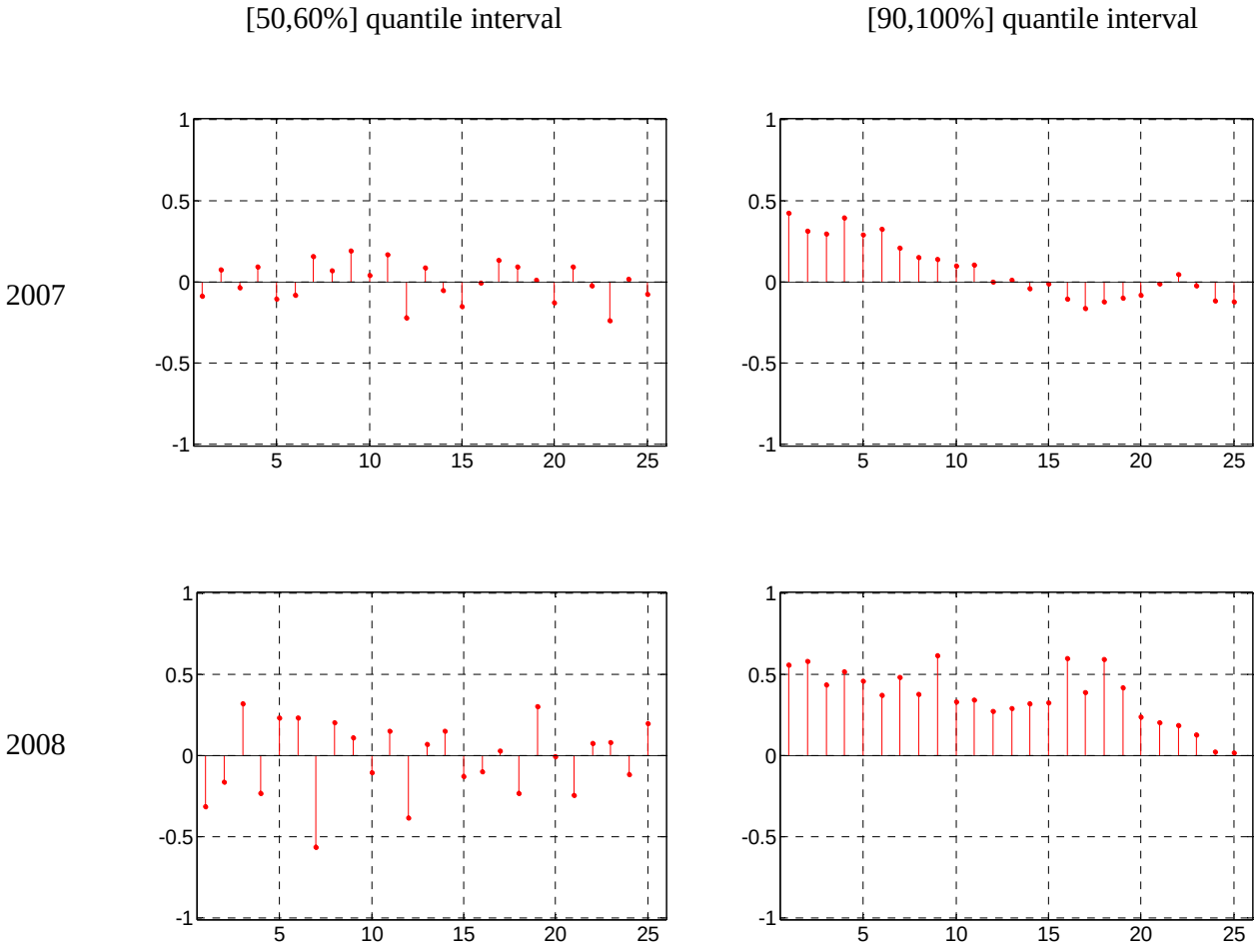


Figure 9. Empirical ACFs of the ITS corresponding to the [50,60] and [90,100] quantile intervals in the estimation periods in 2007 and 2008

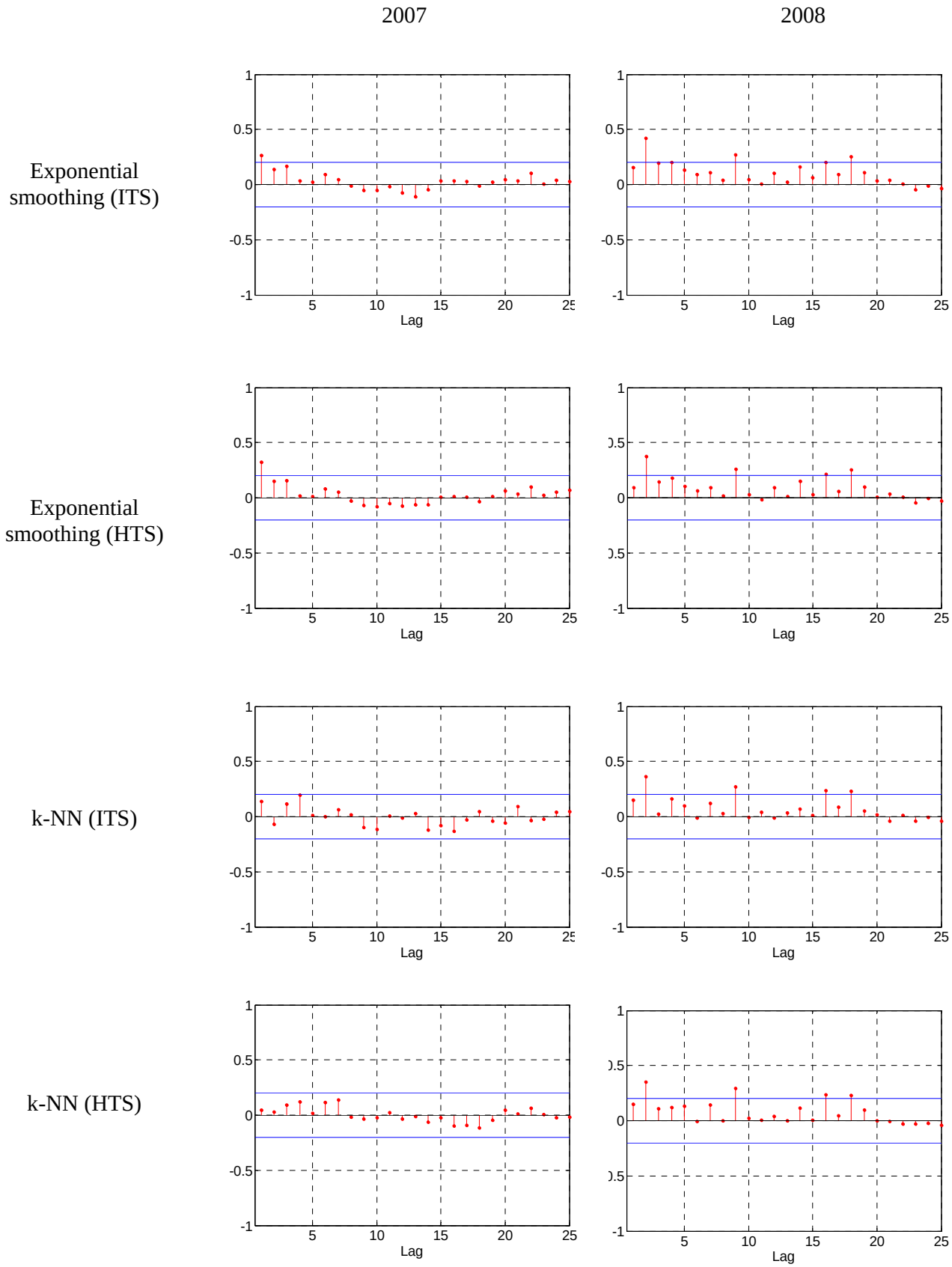


Figure 10. Empirical ACFs of the residuals from the estimation of exponential smoothing (TSI and TSH), and k-NN (TSI and ISH) for the [90,100%] quantile interval in 2007 and 2008

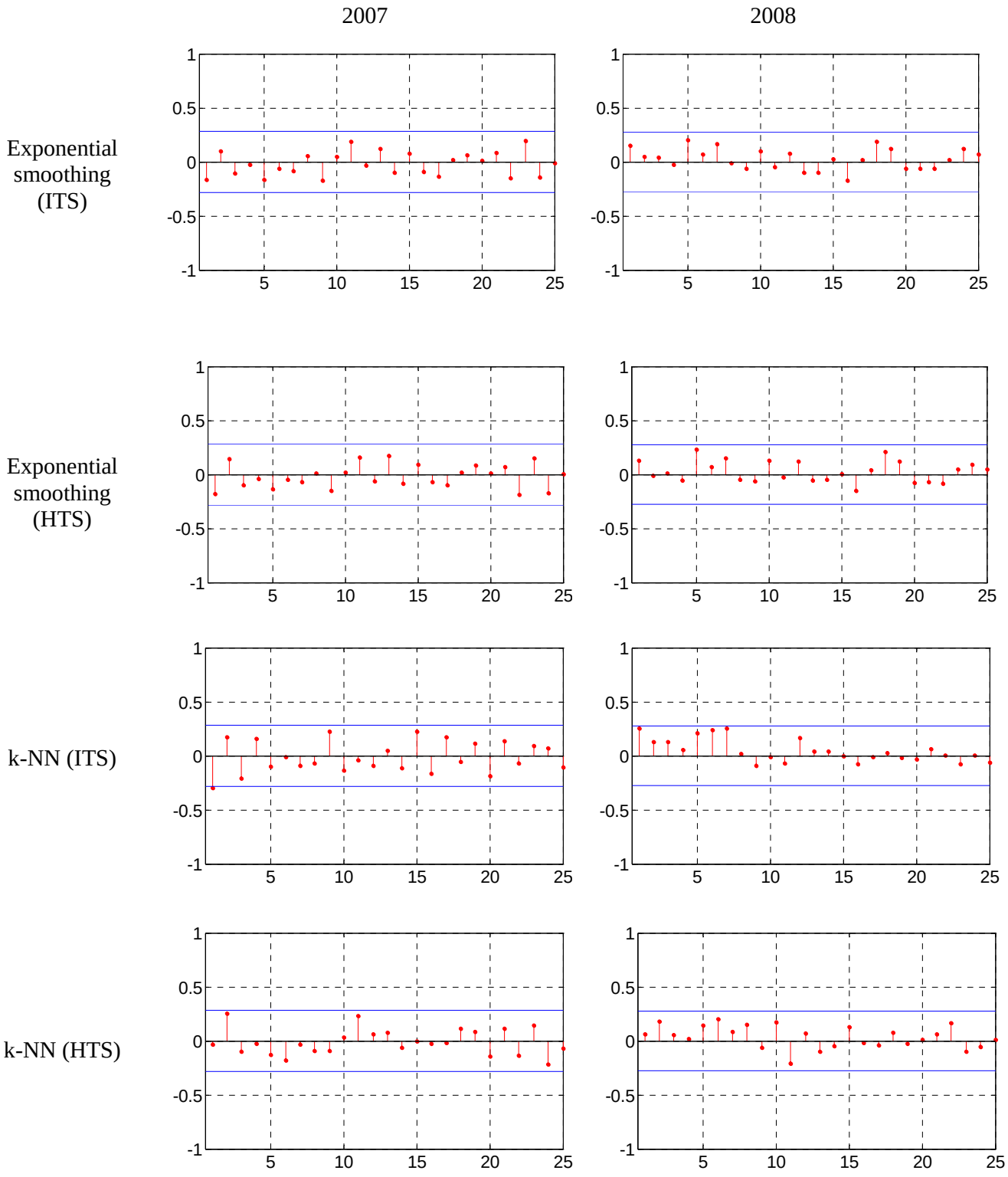
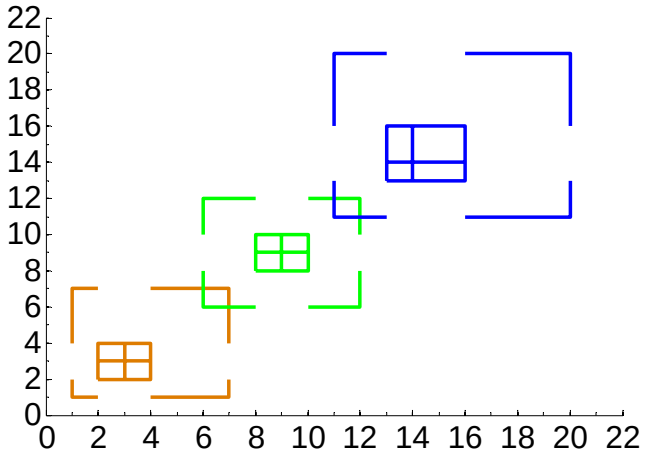
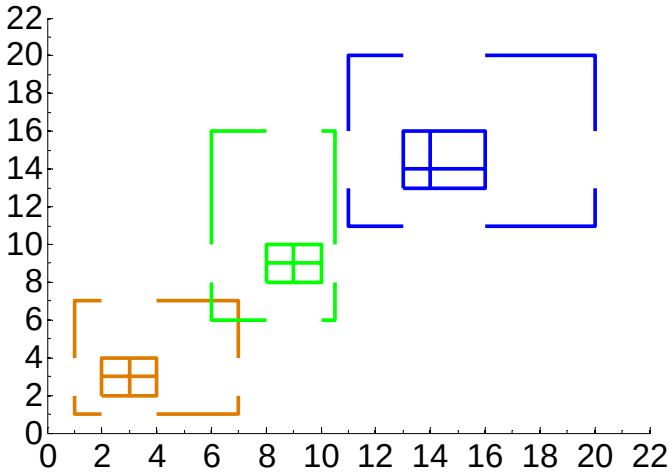


Figure 11. Empirical ACFs of the one-step-ahead forecast distance error from the prediction of the [90,100%] quantile interval with exponential smoothing (ITS and HTS) and k-NN (ITS and HTS) methods in 2007 and 2008

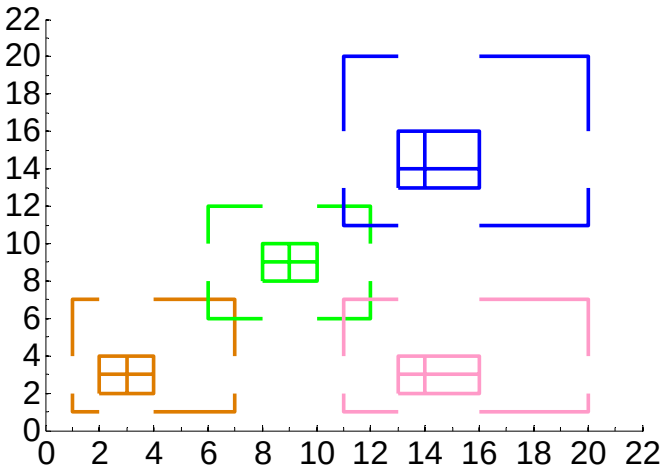
CASE 1 $\rho = 1$



CASE 2 $\rho = .9872$



CASE 3 $\rho = 0.4559$



CASE 4 $\rho = 0$

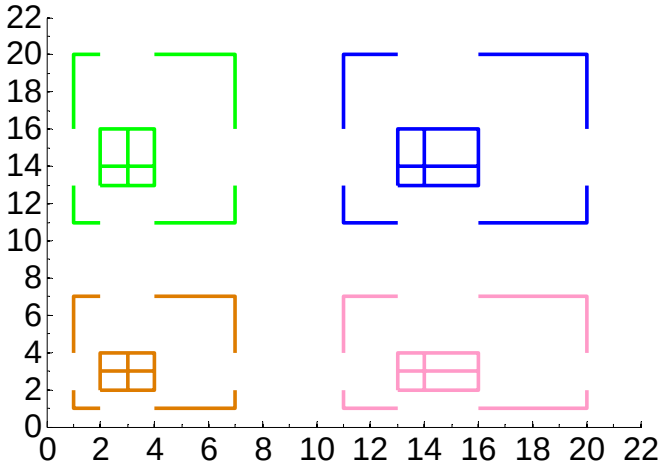


Figure 12. Illustrative examples of correlation between histogram random variables