THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

Date October 12, 2017

WP # 0004MSS-253-2017

Piecewise Solutions to Big Data

Henry Chacon
University of Texas at San Antonio

Anuradha Roy
University of Texas at San Antonio

UTSA

ONE UTSA CIRCLE
SAN ANTONIO, TEXAS 78249-0631

Piecewise Solutions to Big Data

Henry Chacón* Anuradha Roy†

Abstract

Outliers in the financial market data often carry important information, which requires attention and investigation. Many outlier detection techniques, including both parametric and nonparametric, have been developed over the years which are specific to certain application domains. Nonetheless, outlier detection is not an easy task, because sometimes the occurrence of them is pretty easy and evident, but in some other times, it may be extremely cumbersome. Financial series, which are not only pretty sensitive in reflecting the world market conditions due to the interactions of a very large number of participants in its operation, but also influenced by other stock markets that operate in other parts of the world, produce a non-synchronous process. In this research, we detect the presence of outliers in financial time series over the S&P 500 during the year 2016. We detect the beginning of some shocks (outliers) such as the Brexit referendum and the United States Presidential election held in the year 2016. Generally, the impacts of these events were not drastic.

Histogram time series was implemented over a daily closing price on intervals of five minutes for the S&P 500 index during 2015 and 2016. In this case, the linear dependency between days of atypical returns were analyzed on quantiles $[0-40]\%$ and $[60-100]\%$, while Wassertein distance and an approximation of entropy were used to quantify the presence of instant shocks in the index.

Key Words: Big data, Outlier detection, Financial market, Histogram time series, Entropy

Mathematics Subject Classifcation (2010) 62M10, 62P20, 94A17

1. Introduction

Detecting outliers is not an easy task since this behaves irregularly compared to the normal performance. Histogram time series has the possibility to focus on specific sections of a histogram, with potential applications in financial metrics such as Value at Risk (VAR). Wassertein distance is widely used as a metric to compare the distance between two histograms, however we found that an approximation of the entropy has an equivalent but conservative results in explaining the variability of a financial series.

In this article, a commonly used market reference such as the S&P 500 during the years 2015 and 2016 for inter-daily records of five minutes index is analyzed using different quantile sections of the auto correlation function (ACF) in a similar fashion implemented by González-Rivera and Arroyo (2012). In their paper, González-Rivera and Arroyo (2012) used the S&P 500 during 2007 and 2008, and showed how to use histogram time series with smoothing and $k-$NN methods for modeling the autocorrelation and the data generating process (DGP).

We focus on two important events that happened in 2016, the Brexit (British exit) and the United States (US) presidential election using the ACF on quantiles $[40-60]\%$ to understand or quantify their effects on the market in order to determine if their effects were expected by the market or they can be considered as an atypical event.

As a histogram comparative metric, Verde and Irpino (2008) demonstrated the use of the Mahalanobis-Wassertein distance, also known as Kantorovich metric, to quantify the

*Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA

†Department of Management Science and Statistics, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA

difference between two histograms and used this metric in clustering data over 60 meteorological stations of the People's Republic of China. The same Wassertein metric is applied to our research; this time to quantify how different two consecutive histograms for two consecutive days are with the intention to detect outliers behavior (Irpino and Verde, 2015) in the financial market. We have also implemented an approximation of the entropy on histograms as suggested by Wallis (2006). In this case, we notice that the performance of the entropy is conservative, but equivalent to the Wassertein distance in the detection of outliers in the financial market. In this research[1] we detect the permanent effect of outliers in financial time series over the S&P 500 in 2016 due to the Brexit referendum and the US Presidential election, and compare it to the effect of some shocks happened in 2015.

2. New Ways to See Today's Big Data

The advances in computer processing speed and most space efficient way to store massive data set in the present day have brought the possibility of analyzing the stock market beyond the daily closing price, but in much smaller time periods such as five minutes or less, commonly described as *high-frequency* data. We use these aspects of computer technology to our advantage to get valuable information that produce abnormal behaviors in the market, the reasons behind the abnormal behaviors in the market, that is not possible on a daily reference price, like daily closing price of the market. Two approaches are commonly found in the literature besides the traditional time series analysis with only one reference price per day:

- Intra-daily return prices (hourly, thirty, ten, five or fewer minutes) that produce an inter-daily histogram indexed by day, generate a *histogram time series*, similar to a traditional time series, but in this case, the datum is the complex object, histogram.
- If only the minimum (low) and the maximum (high) daily prices are processed, an interval object indexed by day is appropriate; in this case, an *interval time series* representation is suitable.

Financial data in an object format such as histograms or intervals provide "a more complete picture" and the "dynamics" of the data. There are many other areas such as marketing, environmental sciences, quality control and biomedical sciences in which the object of analysis is not a single-valued variable, but an object variable including histogram-valued or interval-valued variable.

2.1 Symbolic Data

Interval-valued and the histogram-valued data are classified as symbolic data as opposed to classical data. According to Bock and Diday (2000), "*A histogram variable is a modal variable that associates a histogram with each observation*". Symbolic data analysis (Billard and Diday, 2006; Bock and Diday, 2000), especially the histogram analysis has come up with a novel technique to reduce the complexity of the information, discovering the aspects of the data that are not possible under traditional techniques. In this article, histogram analysis is used to detect the presence of outliers in financial time series over the S&P 500 per day during 2016.

Symbolic data can deal with massive information contained in nowadays massive data (Big data) sets found across many disciplines. Development of new methodologies to deal with the massive data sets is very much needed in this decade; it is moving, but at a much

[1] This research is based on the work of González-Rivera and Arroyo (2012).

slower pace. Symbolic data analysis is a new and developing area of statistics, the new frontier of the analysis of the big data.

Symbolic data analyses bring massive size to a manageable size of the data by retaining as much of their original information as possible. The new frontier in the search for patterns and complex structures on massive data, collected over a short period of time, in particular the high-frequency data in the financial market, has promising use of symbolic data analysis, not only in a new symbolic way, but also in its capabilities to reduce the complexity of the vast amount of information in today's big data.

3. Histogram Time Series

Consider a random variable of interest $X \in \mathbb{R}$, partitioned in continuous intervals of the form $[x]_j = [x_{Lj}, x_{Uj})$, such as $-\infty < x_{Lj} \leq x_{Uj} < \infty$ and $x_{Uj-1} < x_{Lj}, \forall j$, for $j \geq 2$ (González-Rivera and Arroyo, 2012). Then, a histogram is defined as *datum* of the form:

$$h_x = \{([x]_1, \pi_1), \cdots, ([x]_n, \pi_n)\},$$

where $\pi_j, j = 1, \cdots, n$ is a probability for the j^{th} interval.

A collection of histograms $\{h_{x_i}, i = 1, \cdots, m\}$ is defined by the probability space $(\Omega, \mathcal{F}, P)$, where Ω is the set of elementary events for the random variables, $\mathcal{F}$ is the $\sigma-$field of events and $P : \mathcal{F} \to [0, 1]$ is the probability measure. Define a partition of Ω into sets $A_X(x)$, such that $A_X(x) = \{\omega \in \Omega | X(\omega) = x\}$, where $x \in \{h_{X_i}, i = 1, \cdots, m\}$.

According to González-Rivera and Arroyo (2012), a mapping $h_X : \mathcal{F} \to \{h_{X_i}\}$, such that there is a set $A_X(x) \in \mathcal{F}, \forall x \in \{h_{X_i}, i = 1, \cdots, m\}$, called a histogram random variable. The collection of histogram random variables indexed by time $\{h_{X_t}\}$ for $t \in T \subset \mathbb{R}$ is a stochastic process.

As in González-Rivera and Arroyo (2012), the histogram stochastic process is assumed weakly stationary and ergodic. Then, a barycentric histogram is introduced as a first moment and given by the following expression:

$$\widehat{h}_c = \arg\min_{h_c} \left(\sum_{t=1}^{T} D(h_{X_t}, h_c)^2 \right)^{1/2},$$

where the Wasserstein distance $D(h_{X_t}, h_c)$ is a measure of dissimilarity between two histograms, and is defined as:

$$D(h_1, h_2) = \sqrt{\int_0^1 \left(H_1^{-1}(r) - H_2^{-1}(r) \right)^2 dr},$$

with $H^{-1}(r)$ as the inverse of the distribution function H of h_X.

González-Rivero and Arroyo (2012) showed that *the minimization with the Mallows (Wasserstein) distance reduces to a least squares problem for which the barycentric solution involves a collection of averages. The r^{th} quantile $\widehat{H}_c^{-1}(r)$ of the barycentric histogram $\widehat{h}_c$ is the mean of the r^{th} quantiles $H_{X_t}^{-1}(r)$ of all histograms h_{X_t} in the set*. Therefore, the barycentric histogram is computed as:

$$\widehat{H}_c^{-1}(r) = \frac{1}{T} \sum_{t=1}^{T} H_{X_t}^{-1}(r). \tag{1}$$

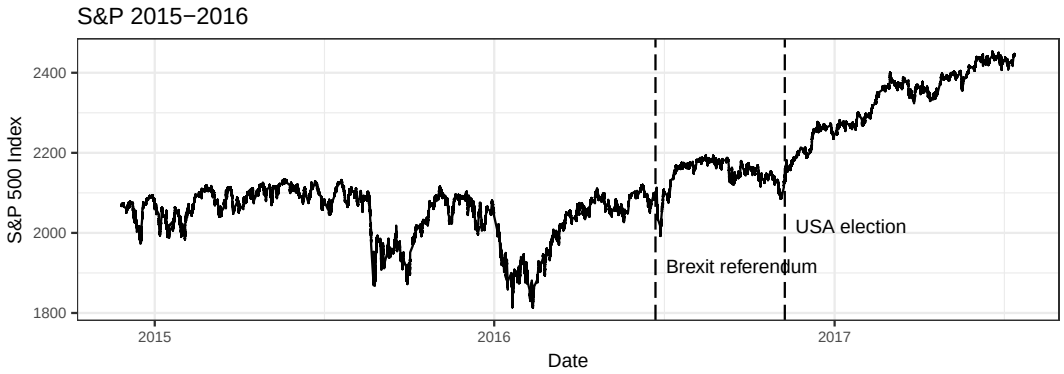

Figure 1: S&P 500 daily - 2015 and 2016.

Finally, the sample autocovariance of order k is defined as:

$$\gamma_k = \text{Cov}\Big(h_{X_t}, h_{X_t-k}; \widehat{h}_c\Big) = \frac{1}{T}\sum_{t=1}^{T}\int_0^1 \Big(H_{X_t}^{-1}(r) - \widehat{H}_c^{-1}(r)\Big)\Big(H_{X_{t-k}}^{-1}(r) - \widehat{H}_c^{-1}(r)\Big)\, dr,$$

and the empirical autocorrelation function (ACF) is defined and computed as:

$$\widehat{\rho}_k = \frac{\gamma_k}{\gamma_0}. \tag{2}$$

The large sample distribution of the ACF (Shumway and Stoffer, 2011), for a series of white noise with large T, is approximately normally distributed with zero mean and standard deviation

$$\sigma_{\widehat{\rho}_k} = \frac{1}{\sqrt{T}}.$$

Hence, a hypothesis test for the linear dependence with lag k provided by the ACF has the rejection region $\{\widehat{\rho}_k > \frac{1.96}{\sqrt{T}}\}$ at 5% level of significance.

4. Entropy for Histograms

In a note on the calculation of entropy from histograms, Wallis (2006) presented the entropy $\big($given by Harris (2006)$\big)$ for discrete distribution as follows

$$E = -\sum_{k=1}^{n} p_k \log\left(\frac{p_k}{w_k}\right),$$

where n is the number of bins for a given histogram. In order to compare histograms with different bin sizes, the width ($w_k = u_k - l_k$) of the bins should be included. However, for $w_k < p_k$, the entropy $E < 0$. For the discrete case, Wallis (2006) expressed the entropy as follows

$$E = -\sum_{k=1}^{n}\int_{l_k}^{u_k} f(x)\log f(x)dx,$$

which can be approximated by

$$E = -\sum_{k=1}^{n} w_k f(x_k)\log f(x_k).$$

This metric is used to compare the amount of information found in each histogram.

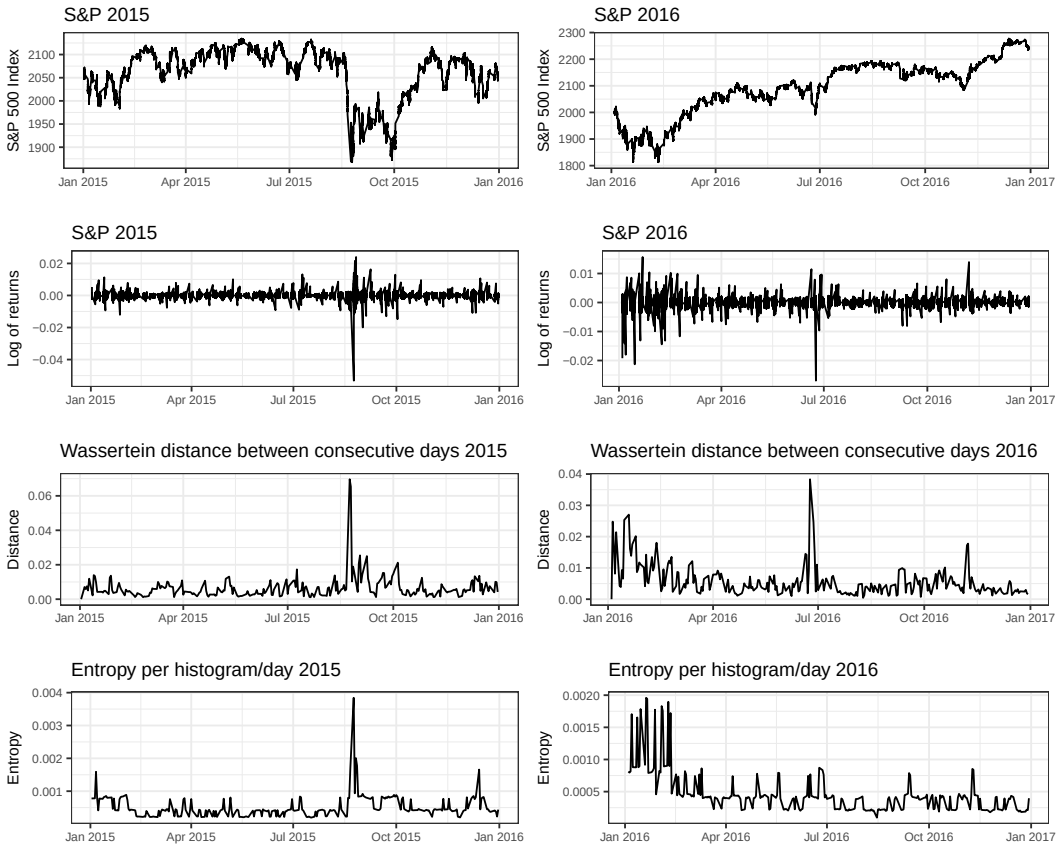

Figure 2: Analyses of S&P 500 index - 2015 and 2016.

5. An Application of Symbolic Data in Shock Detection

The S&P 500 is one of the most representative indicators of the stock market. It is composed of the 500 leading companies, and captures approximately 80% of the available market capitalization (S&P Dow Jones Report, 2017). Since it covers a wide portion of the market, it is used as a sensor of context information and its impact.

As mentioned before, two important events had happened during the year 2016, first one is the *Brexit* referendum (June 23) and the second one is the US presidential election (November 8). A big source of uncertainty was observed in the market prior to these two events, as spotted and perceived in Figure 1. The daily index value during the years 2015 and 2016 for an interval of five minutes is downloaded from the search engine 'Bloomberg': a total of 20,344 observations in 2015 and a total of 20,372 observations in 2016 are obtained. Since the period of stock prices is enough small, the continuously compounded return is calculated for this data. A commonly used metric in financial statistics is given by the natural logarithmic transformation of the simple gross return, and is calculated and showed as *log of returns* in the following equation

$$r_t = \log(P_t) - \log(P_{t-1}), \qquad \text{where } P_t \text{ is the index price at time } t.$$

The biggest variability of returns is observed during the final quarter of 2015 due to the intent of a liberalization of the economy and implementation of some financial reforms of Chinese government (Perkowski, 2015), while the Brexit referendum and the US election results have a conservative variation in the inter-daily prices with respect to the beginning of the year 2016. Following González-Rivera and Arroyo (2012), the histogram analysis in

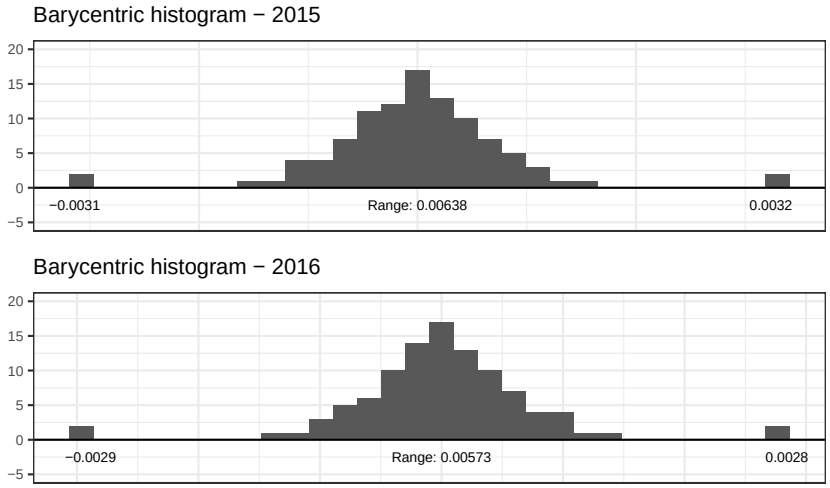

Figure 3: Barycentric histograms per year.

this article is divided per year and analyses are made independently of each other. Figure 2 shows the daily price evolution for the index, the log of returns, the Wasertein distance and the entropy results during 2015 as well as during 2016 for comparison purpose. Notice how the Wassertein distance and the entropy are sensible to abrupt changes in the index. During the biggest volatility observed between August-October in 2015, the entropy seems to be more stable after the drastic loss of value reported in the middle of August 2015, nevertheless a different performance is observed in the Wassertein distance, basically due to the dissimilarities on two consecutive daily histograms. In the same fashion, it is important to point out how both methods reflect the variability presented at the end of the year: the Wassertein distance is almost negligible for that period, but the entropy obtained during that period is not.

In our opinion uncertainty for the year 2016 has two important events, the Brexit decision and the US presidential election. Due to the impact of the outcomes of these decisions over the worldwide economy, volatility increases in the market at the beginning of the year 2016 followed by a stabilized period prior to these two events, as observed in Figure 1. The continuous compound return or the log of returns also reflects the same phenomenon, see Figure 2. At the beginning of the year, a negative perception was observed in the market; probably due to these elections and the uncertainty of the outcome in each case. Wassertein distance accounts the volatility periods in the same fashion as that of the Entropy. However, in the former, the biggest fluctuation happens at the Brexit referendum, while in the last one, at the beginning of the year. As mentioned before, the Wassertein distance stresses the difference between two consecutive histograms, whereas the entropy stresses on the amount of information in each histogram. Days before the Brexit referendum, some brokers could manage to increase the volatility in the market in order to take advantage of the market based on the result of the Brexit referendum: a temporary erratic behavior increases the range of the log of the returns per day and then its variability. On the other hand, in the Entropy the impact of the Brexit and the US presidential election is more conservative, significantly less than the impact at the beginning of the year. From the information point of view, the uncertainty at the beginning of the year was bigger than the two studied events, mainly due to the influence of public opinion polls prior to each event and the knowledge of the decision after each one. This is the reason we presume the difference in the performance of both metrics for the year 2016, however, it is just a presumption.

In the same fashion, the daily histogram for the log of returns is computed per year

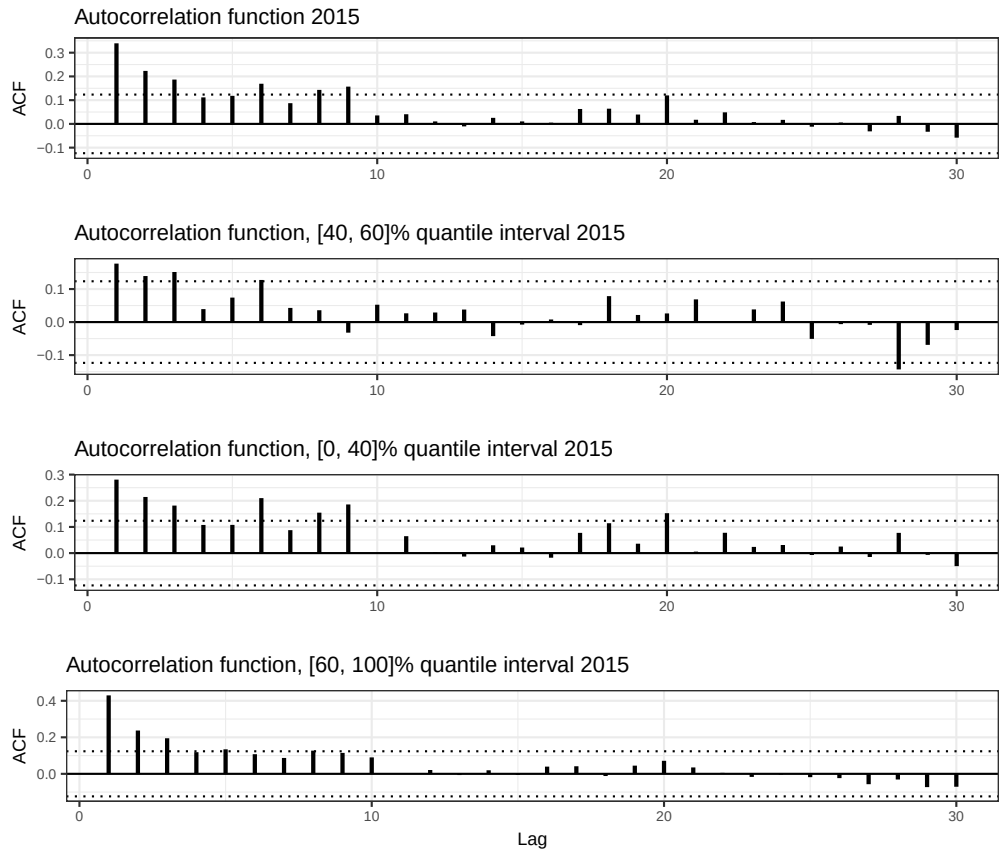

Figure 4: ACF per quantile segments - 2015.

and then, the barycentric histogram is obtained using the Equation (1) and based on the stationary and ergodic assumptions stated in Section 3. Barycentric histograms for the years 2015 and 2016 are displayed in Figure 3. These barycentric histograms look pretty similar, however the range of the log of returns for 2015 is $6.38E-3$, while it is $5.73E-3$ for 2016. This difference is mainly related with the variability observed in the former with respect to the later as illustrated in Figure 1. As pointed out by González-Rivera and Arroyo (2012), one of the advantages of using histogram analysis is the possibility to focus on some specific sections or quantiles in each histogram. So, it is reasonable to reckon that the regular behavior in the log of returns is between $[40-60]\%$ and extreme values or outliers are outside this interval. Figure 4 presents the ACF $\big($Equation (2)$\big)$ for 2015 using all the quantiles, and subsequently the ACF for quantiles $[40-60]\%$, $[0-40]\%$ and $[60-100]\%$ for each daily histogram of the year. Our main focus is on the quantiles $[40-60]\%$, since we deem them as the expected behavior of returns. In this case there is a linear dependency in the first three days in their regular behavior in 2015. A different phenomenon is observed in the quantiles $[0-40]\%$, the linear dependency is extended to nine days, while in the upper range $[60-100]\%$ the dependency is similar to the center. A sign that negative events have a more permanent impact than those of positive or in the regular performance, a kind of temporary snowball effects for the first range of quantiles in the year 2015.

Contrastingly in the year 2016, the ACF for quantiles $[40-60]\%$ (Figure 5) on the log of returns, the linear dependency is present in all the previously mentioned sections of the quantiles: around ten days for the regular behavior and more than twenty days for the lower and upper segments of the quantiles in each histogram. These statistics justify the fact that

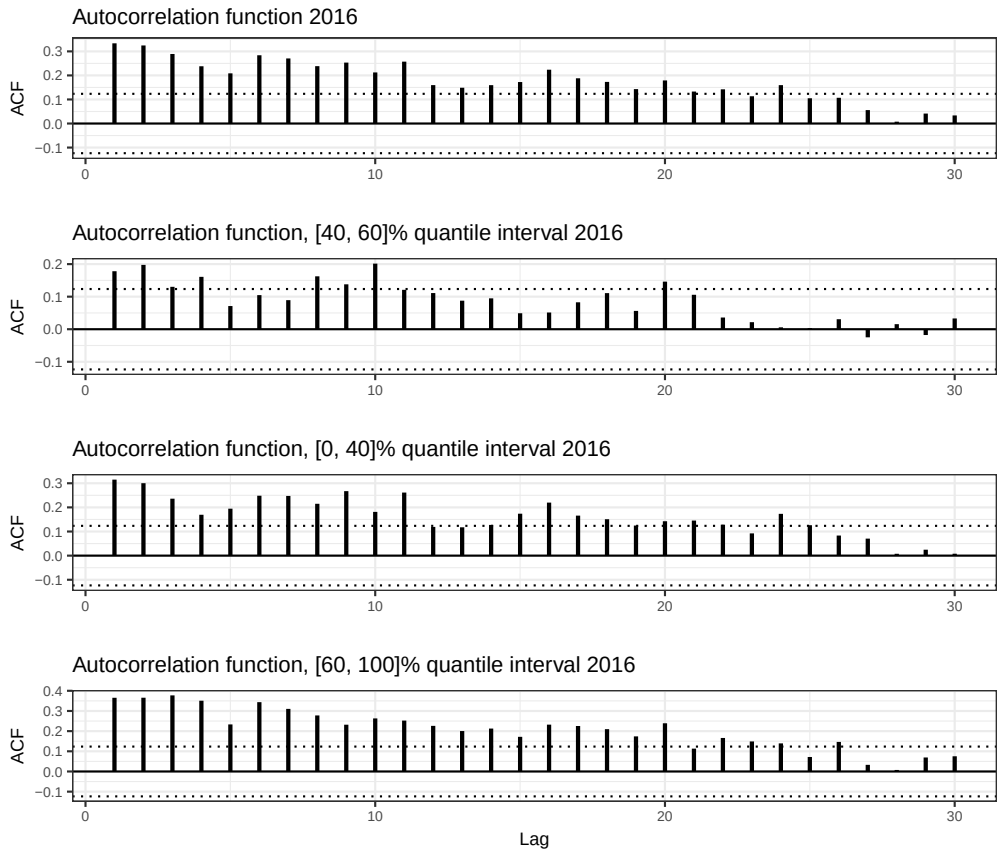

Figure 5: ACF per quantile segments - 2016.

the variability stimulated by the Brexit and the US presidential election or other events in the year 2016, has an overall permanent impact in the financial market. In plain English, the variability in all the quantiles is conditioned to the Brexit and the US presidential election; the impacts or results of the Brexit and the US presidential election are not random, and the market reactions to all this information stabilize themselves.

6. Concluding Remarks

Histogram time series has a tremendous potential to detect outliers as the analysis can be performed on segregated quantiles of the data. In this article, we present three different approaches to determine the impact of announced events in the S&P 500 index and other unexpected situations during the years 2015 and 2016. Focusing on the $[40-60]\%$ quantiles ACF, we notice that sudden variations observed in 2015 has a short permanent dependency over the log of returns during the year. On the contrary, the Brexit and the US presidential election seem to have a longer permanent impact over the market.

The two metrics, Wassertein distance and entropy approximation are appropriate to detect instant variations on the daily financial series; however, the former exaggerates much to the abrupt changes, that is Wassertein distance is too sensitive to abrupt changes. We are working on the quantification of the performance of both the metrics and will report it in a future correspondence.

REFERENCES

Billard, L. and Diday, E. (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley & Sons Ltd.

Bock, H. H., and Diday, E. (2000), *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer.

González-Rivera, G., and Arroyo, J. (2012), "Time series modeling of histogram-value data: The daily histogram time series of S&P 500 intradaily returns," *International Journal of Forecasting*, 28

Harris, B. (2006), "Entropy," in N. Balakrishnan, C.B. Read and B. Vidakovic (eds), Encyclopedia of Statistical Sciences (2nd edn, vol.3), New York: Wiley, pp.1992-1996.

Irpino, A., and Verde, R. (2006), "A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data," in *Data science and classification, proceedings of the IFCS* Berln: Springer, pp. 185192.

Perkowski, J. (2015), "Putting China's Stock Market Into Perspective," Forbes Asia (online edition).

S&P Dow Jones Indices. (2017), http://us.spindices.com/indices/equity/sp-500. Technical report.

Shumway, R. H. and Stoffer, D. S. (2011), *Time Series Analysis and Its Applications Applications with R Examples* (3rd ed.), Springer.

Verde, R., and Irpino, A. (2008), "Comparing histogram data using a MahalanobisWasserstein distance," in *COMPSTAT 2008. Proceedings in computational statistics*, Porto: Springer, pp. 7789.

Wallis, K. (2006), "A note on the calculation of entropy from histograms," Department of Economics, University of Warwick.