# Different Methods to Clean Up Ultra High-Frequency Data(⋆)

*Differenti metodi per la pulizia di dati ad altissima frequenza*

Angelo M. Mineo, Fiorella Romito
Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo
e-mail: romito@dssm.unipa.it

**Keywords:** ultra high-frequency data, stock exchange, outliers, ACD model

## 1. Introduction

Ultra high-frequency data contain detailed information of all the financial market activity, by considering all the transactions tick-by-tick. In this data set the available information is constituted by the times at which all the market events take place and by their associate characteristics. The two main features of such data are that the number of observations is huge and the observations are irregularly time spaced. However, it is well known that almost every ultra high-frequency data set contains bad observations; problems may arise from human input errors or from computer system errors, for example. If we do not take into account these data problems, we can have misleading results in subsequent statistical analyses. Therefore, data cleaning is a necessary step, but only few methods specific for this kind of data are proposed in literature. In this paper we compare the performance of three data cleaning methods briefly described in the following section. The comparison is done by considering the impact of each data cleaning method on autoregressive conditional duration (ACD) models (Engle and Russell, 1998).

## 2. The considered data cleaning methods

For ultra high-frequency data cleaning it is necessary to implement automatic procedures based on some criteria in order to decide on the possible elimination of each observation. We consider the method proposed by Dunis *et al.* (1998), based on the median of the last three ticks, and the method proposed by Brownless and Gallo (2006), that verifies the validity of an observation on the basis of its relative distance from a neighborhood of the closest valid observations. Let $\{p_i\}_{i=1}^N$ be an ordered tick-by-tick price series. If

$$(|p_i - \bar{p}_i(k)| < 3s_i(k) + \varphi) = \begin{cases} \text{TRUE} & \text{observation } i \text{ is kept} \\ \text{FALSE} & \text{observation } i \text{ is removed} \end{cases} \quad (1)$$

where $\bar{p}_i(k)$ and $s_i(k)$ denote respectively the $\alpha$-trimmed mean and the standard deviation of a neighborhood of $k$ observations around *i* and $\varphi$ is a parameter called granularity parameter. The granularity parameter is considered because the ultra high-frequency series often contains sequences of equal prices which would lead to a zero variance; thus, it is useful to introduce a lower positive bound on price variation in order to have always

(⋆) Work supported by grants of the University of Palermo.

admissible solutions. In Mineo and Romito (2007) it is proposed a slight modification of the method proposed by Brownless and Gallo, based on the following rule:

$$(|p_i - \bar{p}_{-i}(k)| < 3s_{-i}(k) + \varphi) = \begin{cases} \text{TRUE} & \text{observation } i \text{ is kept} \\ \text{FALSE} & \text{observation } i \text{ is removed} \end{cases} \quad (2)$$

where $\bar{p}_{-i}(k)$ and $s_{-i}(k)$ denote respectively the mean and the standard deviation of a neighborhood of $k$ observations around *i* without the *i*-th observation and $\varphi$ is the granularity parameter. In this way, we do not take into account the value of the observation under investigation, which can influence the value of the computed mean and standard deviation. Then, even if the two procedures are quite similar, this one is more restrictive. Moreover, we not have the necessity to set the value of $\alpha$ to compute the $\alpha$-trimmed mean.

## 3. Comparison and conclusions

In order to compare the three methods, we have applied them to trade and quote data extracted from the TAQ database of the New York Stock Exchange (NYSE). The TAQ database is a collection of intraday transaction data (trades and quotes) for all the securities listed on the NYSE. Some involved variables are the transaction prices, the number of traded shares, the date and the time of the trade and of the quote, the bid prices and size, the ask prices and size. This database has also some fields containing information on the quality of the recording of the ticks (for more information on the TAQ database see the user's guide at the site `http://nyse.com`). In this work, we use TAQ data from 01.05.1997 to 15.05.1997 for two stocks: General Electric and Microsoft.

The obtained results show that the method proposed by Dunis *et al.* (1998) has the worst performance. By comparing the two methods proposed by Brownless and Gallo (2006) and by Mineo and Romito (2007), it seems that the performance are quite similar, but the latter method can be considered better in terms of autocorrelation of the model residuals. In fact, the autoregressive structure of the ACD model implies that small durations are more likely followed by small durations and long durations are more likely followed by long durations. This effect is well known in literature as "clustering" of the durations. A possible outlier stops this clustering effect. A good data cleaning allows to avoid this situation. Anyway, the considered methods have a sure merit, i.e., their implementation is very simple and effective and the obtained results seem to respect the structure of the original raw data.

## References


Brownless C., Gallo G. (2006) Financial econometric analysis at ultra-high frequency: data handling concerns, *Computational Statistics & Data Analysis*, 51, 2232–2245.

Dunis C., Gavridis M., Harris A., Leong S., Nacaskul P. (1998) An Application of genetic algorithms to high frequency trading models: a case study, in: *Nonlinear Modelling of High Frequency Financial Time Series*, Dunis C. & Zhou B. (Eds.), Wiley, 247-278.

Engle R., Russell J.R. (1990) Autoregressive conditional duration: a new model for irregularly spaced transaction data, *Econometrica*, 66, 1127–1162.

Mineo A.M., Romito F. (2007) A method to "clean up" ultra high-frequency data, *Statistica & Applicazioni*, 5, 167–186.