Ultra High-Frequency Data Management

Christian T. Brownlees¹ and Giampiero M. Gallo¹

Dipartimento di Statistica G. Parenti, Viale G.B. Morgagni 59, I-50134 Firenze, Italy, e-mail: ctb@ds.unifi.it, gallog@ds.unifi.it.

ABSTRACT. The Financial Econometrics literature on Ultra High-Frequency Data (UHFD) has been growing steadily in recent years. However no standard procedures for the data management have been established yet. The present work reviews and presents a number of methods for the treatment of UHFD, and explain the ratio and implications of using such procedures. In this paper we shall concentrate on the data produced by the New York Stock Exchange (NYSE).

1 INTRODUCTION

The advent of financial high-frequency data has been one of the most relevant innovations in the field of the quantitative analysis of the financial markets of the last years. Unfortunately, as observed by Bollerslev (2001) in a survey on the state of this subject, the analysis of this data is not particularly simple. The intra-daily analysis of the financial markets requires a different approach in comparison with the more familiar daily-frequency studies. One of the problems which arises in Ultra High-Frequency Data (UHFD) analysis regards the ways raw data can be managed in order to obtain the time series of interest. This contribution addresses this issue in that it will present and motivate simple and general methods for the data management and manipulation.

The atomic unit of information which builds up ultra high-frequency data is the "tick". Broadly speaking, a tick is made up of a time stamp and a set of information which refers to some specific aspect of the market activity. The ultra high-frequency databases containing tick-by-tick information are very complex to analyse, in that:

- the number of ticks is usually huge;
- the time interval which separates two following ticks is random since the time which separates two following markets events is indeed stochastic;
- the sequence of ticks could contain some wrong ticks
- a tick can contain additional information which is not of interest for analysis purposes;
- the sequence and the structure of the ticks strongly depends on the rules and procedures of the institution which produces and collects the information.

It is therefore fundamental to understand the structure of this data carefully in order to efficiently extract the information of interest without distorting its content.

Although there are many financial markets which produce tick data, the focus of this paper is however on the data produced by the New York Stock Exchange (NYSE).
Raw ultra high-frequency data

Figure 1. Trade and quote data for the General Electric stock on the 11\textsuperscript{th} of November 2002 from 9:30:00 to 16:05:00. The top figure shows the tick–by–tick transaction prices while the bottom figure shows the tick–by–tick bid and ask prices.

2 Ultra High-Frequency Data Handling

The preliminary steps that have to be undertaken before starting the analysis of UHFD are:

- detecting and removing wrong observations from the raw UHFD,
- constructing the time series of interest for the objectives of the analysis.

Unfortunately these operations are far less trivial than one could expect.

2.1 Data Cleaning

Raw UHFD is well known for being prone to errors and some method for the detection and elimination of such values has to be used prior to the time series analysis. This kind of preliminary data manipulation is often referred to as filtering or data cleaning. The aim of data cleaning is to eliminate from the ultra high frequency time series any observation which does not reflect the market activity, in practise however it seems that these methods cannot go beyond detecting outliers.

Most data cleaning procedures deal only with tick–by–tick price time series (transaction, bid, ask, quote mid-price). For volume data there are no special data cleaning recipes other than assessing the plausibility of a certain volume on the basis of the plausibility of the corresponding price and resorting to classical statistical outlier detection methods.
Figure 1 shows one day trade and quote prices for the General Electric stock in the year 2002. The figure shows that presence of outliers is very evident, and it is much pronounced in quote rather than trade data.

There are some algorithms which have been proposed in the literature for washing away wrong observations. Dacorogna, Gencay, Muller, Olsen & Pictet (2001) for instance outline the algorithm used at the Olsen & Associates for performing this kind of task. Their algorithm seems however to be much more complex than needed at least for the NYSE data. In our experience a method which has showed to work well in practise consists in constructing a heuristic $z$-score and than eliminating observations if the absolute value of the score is beyond a certain threshold. The $z$-score for each price observation is computed using a centred moving average, centred moving variance and a parameter $g$ which induces a positive lower bound on the heuristic variance which is useful for the handling of the sequences of identical prices:

$$z_i = \frac{p_i - \bar{p}_{[i-k,i+k]}}{\sqrt{s^2_{[i-k,i+k]} + g}}$$

2.2 DATA MANAGEMENT

Even when the UHFD has been “cleaned”, the construction of the appropriate time series for the purposes of the analysis can still be quite problematic in that

- the data has a complex structure and
- there are not always unique and optimal ways to aggregate information.

In the following paragraphs we will present some of the peculiar patterns that emerge in the data and the methods for their management.

Simultaneous Observations Figure 2 displays 1 minute of ultra high-frequency transaction prices, transaction log-volumes and the two series of bid and ask prices. Note that each cross on the transaction price line marks a transaction. As can be observed, there are several transactions reported at the same time which were executed at different price levels. Simultaneous prices at different levels are also present in quote data. Note that the trading of NYSE securities can also be performed on other regional exchanges, and thus simultaneous trade and quotes at different prices are not unusual.

As ultra high-frequency models for tick–by–tick data usually require one observation per time stamp, some form of aggregation has to be performed. For tick–by–tick prices the time series do not seem to suggest that the sequence of simultaneous observations reported is meaningful. Taking the median price could be a reasonable solution given the discrete nature of the tick–by–tick data. In case further aggregations at lower frequencies are performed the method of aggregation becomes progressively less relevant (as the difference between the chosen prices will be negligible). Simpler methods such as the last or first price of the sequence could hence be used. For tick–by–tick volumes or transaction counts the natural way to aggregate observations is to substitute the simultaneous observations with the sum of the simultaneous volumes or the number of simultaneous transactions.
Irregularly Spaced Data The most striking feature of the data displayed in Figure 2 is that the plotted time series are irregular, that is the time which separates two subsequent observations is random. Rather than working with ultra high-frequency irregularly spaced observations for several kinds of studies one might be interested in analysing a regular time series, i.e. a time series in which the time separating two observations is deterministic.

Let \( \{(t_i, x_i)\}_{i=1}^N \) be an irregular time series, where \( t_i \) and \( x_i \) indicate respectively the time and value of the \( i^{th} \) observation, and let \( \{(t_j^*, x_j^*)\}_{j=1}^{N^*} \) be the lower frequency time series that we intend to construct. As we are aggregating higher frequency information at a lower frequency, it seem coherent to use aggregation methods of the form:

\[
x_j^* = f( \{ x_i \mid t_i \in (t_{j-1}^*, t_j^*) \} )
\]

which basically imply that the aggregated observation value \( x_j^* \) is constructed using all the information available from the previous observation at \( t_{j-1}^* \) until \( t_j^* \).

Some simple but useful aggregation methods which are coherent with this scheme are:

**First:** \( x_j^* = x_j \) where \( t_j = \min \{ t_i \mid t_i \in (t_{j-1}^*, t_j^*) \} \)

**Minimum:** \( x_j^* = \min \{ x_i \mid t_i \in (t_{j-1}^*, t_j^*) \} \)

**Maximum:** \( x_j^* = \max \{ x_i \mid t_i \in (t_{j-1}^*, t_j^*) \} \)
\[
\text{Last: } x_j = x_t \text{ where } t_t = \max\{t|t_t \in (t^*_j, t^-_j]\}
\]
\[
\text{Sum: } x_j = \sum_{t \in [t^*_j, t^-_j]} x_t
\]
\[
\text{Count: } x_j = \#\{ (x_t, t)| t_t \in (t^*_j, t^-_j]\}
\]

In the first four methods if the set \( \{ t_t | t_t \in [t^*_j, t^-_j+1] \} \) is empty the \( j \)th observation will be considered missing. The “First”, “Minimum”, “Maximum” and “Last” methods can be useful for the treatment of price series. The “Sum” method is appropriate for aggregating volumes and “Count” can be used to obtain the number of trade and quotes in an interval.

For price data, Dacorogna et al. (2001) proposed some methods which are based on the interpolation at \( t^*_j \) of the previous and the next observation in series. For liquid stocks however the choice of the interpolation schemes will deliver approximately the same results of the “Last” aggregation method. On the other hand results may be very different for infrequently traded stocks, where the interpolation schemes proposed by the authors will lead to long sequences of approximately equal returns which will increase serial correlation and are likely to generate problems in numerical optimisation procedures.

**Bid–Ask Bounce** A pattern which can often be observed in tick–by–tick transaction price series is the so called bid-ask bounce, i.e. the series of transaction prices does not significantly change, and new transactions are executed at the current bid or ask. As can be observed in figure 2 the transaction price tends to “bounce”, but this phenomenon is not too strong because of the fine price granularity and because of the possibility on the NYSE of executing transactions withing the current quote.

The economic explanation of bid–ask bounce is the fact that transactions are not necessarily always generated by the arrival of news. Thus, if no significant event has occurred market orders will tend to be executed at the current bid or ask, thus generating the “bounce” pattern.

The bounce can generate problems when dealing with the construction of a regular time series of high frequency returns. In such cases, the higher the frequency of the returns series, the higher the chance that returns will not reflect the arrival of news. There are two methods to eliminate this problem. The first consists in not using the transaction price series to compute returns but rather to use the quote mid-price series. Another solution is to construct an algorithm which eliminates all price movements which do not move the price away a certain threshold from the last selected price.

**Opening and Closing** Figure 3 shows the closing and the opening of the trading day. As can be observed, the first trades and quotes of the day are not the NYSE opening trades and quotes. At the closing of the NYSE of the same day the last trade reported before 16:05 was actually the closing trade, which was reported 125 seconds after the official closing, while the last reported quote of the day was not. Some specific fields of the NYSE UHF database contain some flags that can be used to identify the exact opening/closing trades and quotes, but unfortunately this piece of information is not always accurately reported.

In practise the difference between the the first and last transaction, bid or ask prices of the day should not significantly differ from the true opening and closing and can be used as a proxy. On the other hand, however, for transaction, bid and ask volumes this is not the case. Furthermore, in order to capture the closing price we suggest considering the NYSE trading day as the time period 9:30 16:05 rather than the official 9:30 16:00. This extension should ensure that the closing price is included in the data.
3 CONCLUSIONS

In this paper we have shown a number of patterns that emerge from NYSE UHFD and some methods for the UHFD management. We believe that the awareness of this ultra high-frequency patterns is important for conscious modelling of this rich and complex source of data.

REFERENCES