Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

Computational Statistics & Data Analysis 51 (2006) 2232–2245

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Financial econometric analysis at ultra-high frequency: Data handling concerns

C.T. Brownlees*, G.M. Gallo

Dipartimento di Statistica “G. Parenti”, Università di Firenze, Viale G.B. Morgagni 59, I-50134 Firenze, Italy

Available online 16 October 2006

Abstract

Data collection at ultra high-frequency on financial markets requires the manipulation of complex databases, and possibly the correction of errors present in the data. The New York Stock Exchange is chosen to provide evidence of problems affecting ultra high-frequency data sets. Standard filters can be applied to remove bad records from the trades and quotes data. A method for outlier detection is proposed to remove data which do not correspond to plausible market activity. Several methods of aggregation of the data are suggested, according to which corresponding time series of interest for econometric analysis can be constructed. As an example of the relevance of the procedure, the autoregressive conditional duration model is estimated on price durations. Failure to purge the data from “wrong” ticks is likely to shorten the financial durations between substantial price movements and to alter the autocorrelation profile of the series. The estimated coefficients and overall model diagnostics are considerably altered in the absence of appropriate steps in data cleaning. Overall the difference in the coefficients is bigger between the dirty series and the clean series than among series filtered with different algorithms.

Keywords: Financial ultra high-frequency data; Outliers; ACD modeling; Trades and quotes (TAQ)

1. Introduction

The increasing availability of data at the highest frequency possible (tick-by-tick) has allowed for many advances in the field of the quantitative analysis of financial markets (for a recent survey, cf. Engle and Russell, 2006). Different data sets allow for different types of economic or econometric analysis, spanning from time series analysis (volatility, duration, etc.) to market microstructure studies (O’Hara, 1997) on price formation, asymmetric information diffusion, impact of news, and so on.

In what follows we will concentrate on transaction prices and quotes. In so doing, admittedly, we do not discuss the possible problems affecting the recently available data on order books which have attracted a lot of attention in the literature related to market microstructure. The available information (the “*tick*”) refers to specific aspects of market activity recorded at a given time (the “*time stamp*”). High frequency data are commonly encountered in the data sets collected by Olsen Associates (cf. Dacorogna et al., 2001 for a general overview, Muller, 2001 for a description of the filter adopted to correct the data, Breymann et al., 2003 for an application). Stock exchange data are available extensively

* Corresponding author. Tel.: +39 55 4237211; fax: +39 55 4223560.
E-mail addresses: ctb@ds.unifi.it (C.T. Brownlees), gallog@ds.unifi.it (G.M. Gallo).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.09.030

for the US exchanges (NYSE, NASDAQ, AMEX), and also, among the major Stock Exchanges, for London, Paris, Frankfurt and Tokio.

The sequence and the structure of the ticks strongly depends on the rules and procedures of the institutions that produce and collect the information (e.g., electronic versus hybrid markets). Furthermore, regulatory changes and technological advances may make the data structurally different across sample periods (e.g. the January 2001 NYSE move to a minimum tick size of $0.01). The latter aspect is particularly delicate since some stylized facts established for some sample period may not be valid for others. In turn, the sequence of ticks might contain some *wrong* records (beyond what is classified as such by the data provider), might not be time ordered, and exhibits some anomalous behavior as a result of particular market conditions (e.g. opening, closing, trading halts, etc.).

It should be stressed that in the steps required to transform raw data into a time series for subsequent analysis, some choices are involved and these may have consequences on the results one may obtain. There are many contributions in the literature that address the issues regarding market rules and procedures, data management and data cleaning (cf. many publications by the New York Stock Exchange available from the www.nyse.com site). Hasbrouck et al. (1993) is often quoted as a reference for NYSE rules and procedures, but attention must be paid to the modifications adopted ever since. Some authors, such as Madhavan and Sofianos (1998) and Sofianos and Werner (2000), analyze various empirical aspects of the NYSE trading mechanisms related to the way orders are handled and executed. Bauwens and Giot (2001) provides an extensive treatment of such issues, more oriented towards econometric analysis, with specific reference to the NYSE Trades and Quotes (TAQ) data. Some authors describe some relevant issues in UHFD management: e.g. Blume and Goldstein (1997) details how they eliminated NYSE quote data which appeared to be affected by errors; Oomen (2006) suggests a filter for detecting outliers for the IBM stock; Vergote (2005) raises concerns about the widespread use of the 5-s rule suggested by Lee and Ready (1991) to synchronize trades and quotes without checking its robustness for the sample period and the asset at hand; Boehmer et al. (2006) discusses the problems in matching information from TAQ and the order book records in the system order data of the NYSE.

In this paper, we first describe the content of TAQ high frequency data sets (Section 2) and pinpoint handling needs in specific reference to some practical examples (Section 3). Given the nature of these data, we propose a simple method for identifying single records as outliers (3.1). We then detail (3.2) the methods to translate the clean tick data into time series of interest for subsequent analysis (duration, realized volatility, realized range, volume accumulation, and so on). As a leading empirical example of the impact that UHFD handling may have on econometric analysis, we take the autoregressive conditional duration (ACD) model by Engle and Russell (1998) (Section 4). The results show that there will be a difference in estimated coefficients of a simple ACD model as a consequence of the choice of inclusion/exclusion of possible outliers. Since the appropriate parameters of the filtering procedures may be stock-dependent, our feeling is that empirical work in the field would be enriched by showing the robustness of the results to the adopted convention in defining records to be deleted.

The purpose of the filter we propose is to eliminate observations which seem incompatible with the prevailing market activity: there is a trade-off between possible information contained in the discarded data and the noise that outliers bring along. At times that information is relevant, when the actual interaction among agents or the diffusion of information is of main interest. One may note that in studying ultra high-frequency based measures of volatility (cf. Andersen et al., 2006) the problem of outliers is less severe when data are sampled at a fixed interval (e.g. 5 min), but it becomes more worrisome when alternative sampling schemes such as the one suggested by Aït-Sahalia et al. (2005) are adopted where all data available are used. In deriving duration data (e.g. price duration, as in our application) the problem is the most relevant because outliers will signal movements above a certain threshold while none has happened in practice.

2. The TAQ database

The categories of data collected within the TAQ are *quotations* and *transactions*. The NYSE is probably the first exchange which has been distributing its ultra high-frequency data sets since the early 1990s. In 1993 the trades, orders and quotes (TORQ) database was released (Hasbrouck, 1992) which contained a 3 month sample of data. Since 1993, the NYSE has started marketing the trades and quotes (TAQ) database. It has undergone some minor modifications through the years leading to 3 different TAQ versions (1, 2 and 3). Since 2002, order book data has also been separately available for research purposes, but will not be discussed here. Although through time many improvements have been added to the quality of the data, the TAQ data are raw, in that the NYSE does not guarantee the degree of accuracy of the data, so that further manipulations are needed for using them in research.

Table 1
TAQ quote records

SYMBOL	EX	QDATE	QTIM	BID	OFR	BIDSIZ	OFRSIZ	QSEQ	MODE	MMID
GE	B	020321	35603	37.550000	37.900000	2	7	0	12	
GE	T	020321	35606	37.690000	75.400000	4	1	0	12	ARCA
GE	T	020321	35606	37.690000	37.900000	4	7	0	12	CAES
GE	N	020321	35606	37.690000	37.710000	1	1	2190411	6	
GE	N	020321	35606	37.680000	37.710000	1	1	2190412	6	
GE	X	020321	35607	37.640000	37.850000	1	1	0	12	

Note: SYMBOL is the Stock symbol; EX is the exchange on which the quote occurred; QDATE is the quote date; QTIM is the quote time expressed as cumulative number of seconds since 00:00 AM; BID is the bid price; OFR is the offer price; BIDSIZ is the bid size in number of round lots (100 share units); OFRSIZ is the offer size in number of round lots (100 share units); QSEQ is the Market Data Systems (MDS) sequence number; MODE is the quote condition; MMID is the NASD Market Maker.

Table 2
TAQ trade records

SYMBOL	EX	TDATE	TTIM	PRICE	SIZ	CORR	TSEQ	COND	G127
GE	N	020321	35605	37.700000	20,000	0	2190410		40
GE	B	020321	35605	37.690000	100	0	0		0
GE	T	020321	35605	37.700000	200	0	0		0
GE	B	020321	35605	37.690000	800	0	0		0
GE	T	020321	35606	37.690000	100	0	0		0
GE	M	020321	35606	37.700000	600	0	0		0
GE	B	020321	35608	37.700000	2000	0	0		0

Note: SYMBOL is the Stock symbol; EX is the exchange on which the trade occurred; TDATE is the trade date; TTIM is the trade time expressed as cumulative number of seconds since 00:00 AM; PRICE is the trade price per share; SIZ is the number of shares traded; CORR is the Correction Indicator (see text); TSEQ is the Market Data System (MDS) sequence number; COND is the Sale Condition (see text); G127 is a field indicating simultaneously: a G trade (a sell or buy transaction made by a NYSE member on his own behalf); a rule 127 transaction, i.e. a transaction executed as a block position; a stopped stock (which *should* also be used to identify the closing price).

Quote data contain information regarding the best trading conditions available on the exchange. Table 1 displays a few sample records from the quote database with an explanations of the various fields. The quote table fields unfortunately do not include any information on the quality of the reported data. However, the MODE field (quote condition) contains many useful information which can be used to reconstruct accurately the trading day events and some specific market conditions. Some values of this field indicate various types of trading halts that can occur during the trading day. Furthermore, the field also contains values indicating the opening and closing quotes.

Trade data contain information regarding the orders which have been executed on the exchange. Table 2 displays few sample records from the trade database. Some fields of the database containing information on the quality of the recorded ticks, allowing for the removal of wrong or inaccurate ticks from subsequent use: e.g. the CORR field (correction indicator) signals whether a tick is correct or not, and the “Z” and “G” value of COND field (sale conditions) indicate a trade reported at a later time.

3. Ultra high-frequency data handling

The preliminary steps needed before starting the econometric analysis of the time series from UHFD are:

(1) *Data cleaning*, i.e. detecting and removing wrong observations from the raw UHFD;
(2) *Data management*, i.e. constructing the time series of interest for the objectives of the analysis.

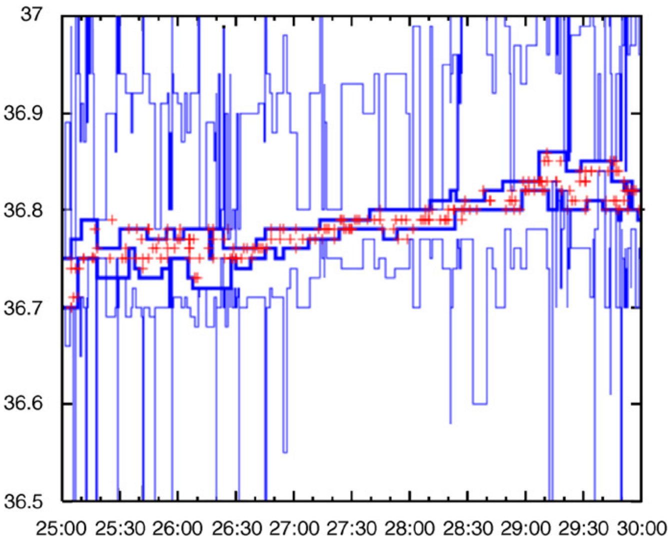

Fig. 1. Trade and quote data for the General Electric stock on April 1, 2002 from 10:25:00 to 10:30:00 AM. Tick-by-tick transaction prices (crosses); NYSE tick-by-tick quotes (thick line), non NYSE tick-by-tick quotes (thin line). The vertical axis has been truncated as non NYSE data contain zeroes and large values.

3.1. Data cleaning

To obtain a clean sample we identify and discard the records which are not of interest using available information. Falkenberry (2002) reports that errors are present both in fully automated and partly automated trading systems, such as the NYSE, and attributes the main determinant of errors to trading intensity. The higher the velocity in trading, the higher the probability that some error will be committed in reporting trading information.

For the quote data, all quotes with a primary NYSE listing that did not occur on the NYSE (EX field different from N) should be eliminated. Fig. 1 illustrates this point: non-NYSE quotes have a very large spread and there are often extremely large quotes or suspicious zeros. For NYSE-listed stocks this pattern is probably due to the fact the NYSE is recognized as the leader market and thus other exchanges do not post competing quotes. Non-NYSE spreads are hence bound to be much larger than the NYSE ones. It is important to note that while discarding incorrect and delayed trade records implies removing a usually very small fractions of observations, removing non-NYSE quote records can have a dramatic impact on the reduction of sample size (cf. below for an example). As some authors suggest (e.g. Vergote, 2005; Boehmer et al., 2006) we also remove from quote data those quotations which were generated by non normal market activity, as indicated by the MODE field values 1, 2, 3, 6 or 18.

For trade data all transactions that are not correct (CORR field different from 0) and delayed (COND field equal to Z) should be eliminated from the sample. Contrary to quote data, we prefer not to discard transaction prices that did not occur on the NYSE, though in some cases this is not advisable (cf. Dufour and Engle, 2000). Fig. 2 compares a 1 h sample of NYSE and non-NYSE transaction data: the two series clearly exhibit the same dynamics, although non-NYSE transaction prices seem to contain a higher share of outliers.

After this TAQ based filtration of the data of interest, the remaining tick-by-tick price series still show observations that are not consistent with the market activity. Fig. 3 shows 30 min of raw trade and quote prices for the General Electric stock. Graphical inspection of the graphs reveals clearly that both series present some suspicious observations, especially in the transaction price series.

Little help comes from information which is not the tick-by-tick price sequence. Volumes, for example, cannot be used as a general guideline, since it is not straightforward to assess the plausibility of a certain volume beyond the plausibility of the corresponding price. One procedure would be to mix trades and quotes data and, after matching the two series, cross check the plausibility of either series. This approach stumbles upon the difficulty of matching trades and quotes given the non-synchronicity of the recording of either observations; the five-second rule by Lee and Ready (1991) is widely used but as Vergote (2005) points out, there is some doubt about its robustness across stocks and

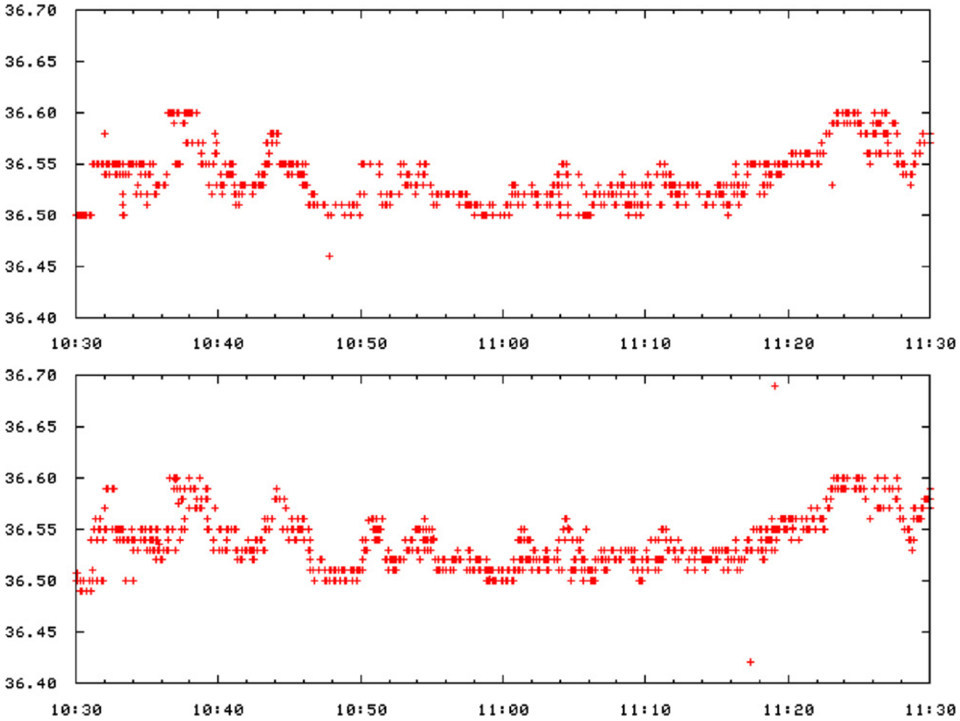

Fig. 2. Trade data for the General Electric stock on the April 8, 2002 from 10:30 to 11:30 AM. Tick-by-tick NYSE transaction prices (top); Tick-by-tick non NYSE transaction prices (bottom).

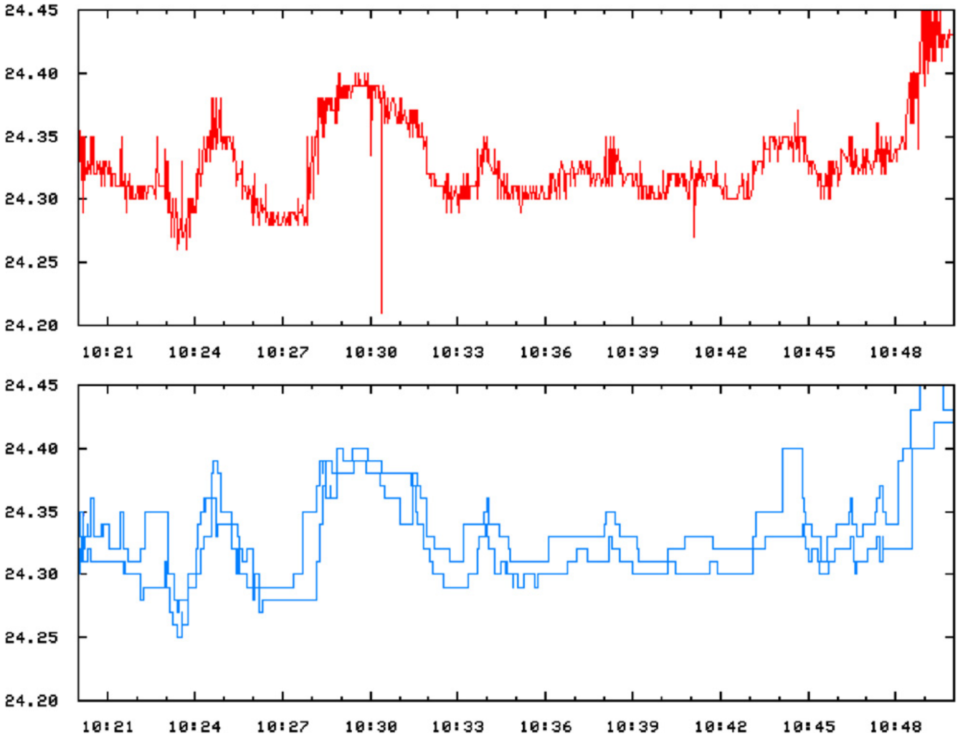

Fig. 3. Trade and Quote Data for the General Electric stock on November 11, 2002 from 10:20:00 to 10:50:00 AM. Tick-by-tick transaction prices (top); tick-by-tick bid and ask prices (bottom).

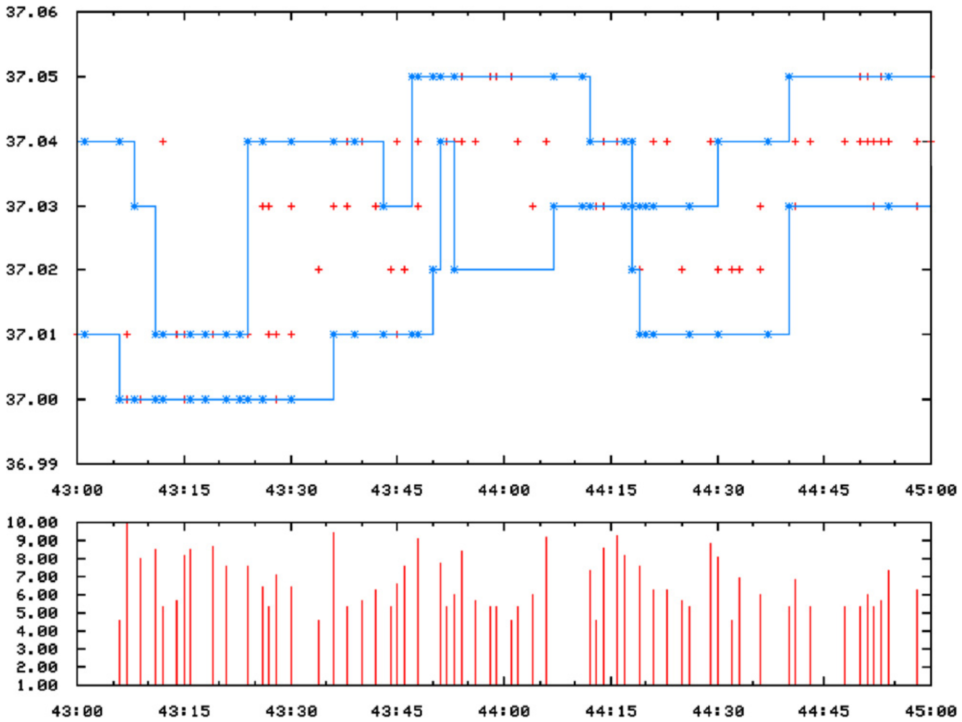

Fig. 4. Two minutes of trading activity for the General Electric stock on March 21, 2002 from 12:43 to 12:45 PM. Transaction price within the bid-ask prices (top panel). Each cross on the transaction price line marks a different transaction. Log-volumes associated with the transaction (bottom panel).

sample periods. Moreover, apart from the possibility that errors may occur in both series, one is left with the choice of which record to delete, when an inconsistency is detected.

Let us consider $\{p_i\}_{i=1}^N$ be an ordered tick-by-tick price series. Our proposed procedure to remove outliers is

$$(|p_i - \bar{p}_i(k)| < 3s_i(k) + \gamma) = \begin{cases} \text{true observation } i \text{ is kept,} \\ \text{false observation } i \text{ is removed,} \end{cases}$$

where $\bar{p}_i(k)$ and $s_i(k)$ denote respectively the δ-trimmed sample mean and sample standard deviation of a neighborhood of k observations around i and γ is a *granularity* parameter. The neighborhood of observations is always chosen so that a given observation is compared with observations belonging to the same trading day. That is, the neighborhood of the first observation of day are the first k ticks of the day, the neighborhood of the last observation of the day are the last k ticks of the day, the neighborhood of a generic transaction in the middle of the day is made by approximately the first preceding $k/2$ ticks and the following $k/2$ ones, and so on. The idea behind the algorithm is to assess the validity of an observation on the basis of its relative distance from a neighborhood of the closest valid observations. The role of the γ parameter is to avoid zero variances produced by sequences of k equal prices.

The percentage of trimming δ should be chosen on the basis of the frequency of outliers, the higher the frequency, the higher the δ. The parameter k should be chosen on the basis of the level of trading intensity. If the trading is not very active k should be "reasonably small", so that the window of observations does not contain too distant prices (the contrary is true if the trading is very active). The choice of γ should be a multiple of the minimum price variation allowed for the specific stock. The procedure is inevitably heuristic but it has the virtue of simplicity and effectiveness: we will show the sensitivity of parameter estimation and model diagnostics to the choice of such parameters.

3.2. Data management

Simultaneous observations: Fig. 4 displays 2 min of ultra high-frequency transaction prices, transaction log-volumes and the two series of bid and ask prices. As it can be noted, 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. There are different explanation for this phenomenon. First of all, note that the trading of NYSE securities can also be performed on other exchanges, and thus simultaneous trades at different prices are normal. Also, the execution on one exchange of market orders will in some cases produce more than one transaction report. Finally, even non simultaneous trades/quotes could be all reported as simultaneous due to trade reporting approximations.

As ultra high-frequency models for the modeling of tick-by-tick data usually require one observation per time stamp, some form of aggregation has to be performed. 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 will be performed, the method of aggregation choice becomes progressively less relevant as the difference between prices will be negligible, and simpler methods such as the last or first price of the sequence should not cause any problems. 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 and the number of simultaneous transactions.

Irregularly spaced data: The most striking feature of the data displayed in Fig. 4 is that the plotted time series are *irregular*, with a random time separating two subsequent observations. To turn it into a time series with discrete, equally spaced time intervals, let us consider an irregular time series $\{(t_i, y_i)\}_{i=1}^{N}$, where t_i and y_i indicate, respectively, the time and value of the ith observation, and let $\{(t_j^*, y_j^*)\}_{j=1}^{N^*}$ be the lower frequency time series that we intend to construct using an appropriate aggregation function such as

$$y_j^* = f(\{(t_i, y_i) \mid t_i \in (t_{j-1}^*, t_j^*]\}).$$

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

First: $y_j^* = y_f$ where $t_f = \min\{t_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
Minimum: $y_j^* = \min\{y_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
Maximum: $y_j^* = \max\{y_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
Last: $y_j^* = y_l$ where $t_l = \max\{t_i | t_i \in (t_{j-1}^*, t_j^*]\}$;
Sum: $y_j^* = \sum_{t_i \in (t_{j-1}^*, t_j^*]} y_i$;
Count: $y_j^* = \#\{(y_i, t_i) | t_i \in (t_{j-1}^*, t_j^*]\}$.

In the first four methods if the set $\{t_i | t_i \in [t_j^*, t_{j+1}^*)\}$ is empty the jth observation will be considered missing. The "First", "Minimum", "Maximum" and "Last" methods can be useful for the treatment of price series (e.g. the "Maximum" and the "Minimum" are the base for the realized range; "Last" can be used for the computation of realized variance, and so on). The "Sum" method is appropriate for aggregating volumes and "Count" can be used to obtain the number of trades and/or quotes in a given interval.

As far as the construction of regular price series is concerned, 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:

Previous point interpolation: $y_j^* = y_p$ where $t_p = \max\{t_i | t_i < t_j^*\}$;
Next point interpolation: $y_j^* = y_n$ where $t_n = \min\{t_i | t_i > t_j^*\}$;
Linear point interpolation: $y_j^* = \left(1 - \frac{t_j^* - t_p}{t_n - t_p}\right) y_p + \frac{t_j^* - t_p}{t_n - t_p} y_n$.

The problem in using these methods, however, is that they might employ information not available at t_j^*. For liquid stocks, the choice of the interpolation schemes does not seem to be particularly relevant as the neighborhood of t_j^* will be very dense of observations, and the different interpolation schemes will deliver approximately the same results of the "Last" method. On the other hand, results may be unsatisfactory for infrequently traded stocks since at certain frequencies the interval $(t_{j-1}^*, t_j^*]$ will not contain any observation and the interpolation may refer to prices recorded at some remote time. In these cases we think it more appropriate to treat the observation as missing, in order to avoid long sequences of zero or identical returns.

Bid–Ask bounce: A common pattern which can often be observed in tick-by-tick transaction price series is the so-called bid–ask bounce: since transactions are not necessarily generated by the arrival of news, in the absence of any significant event, market orders will tend to be executed at the current bid and ask, displaying the "bouncing" pattern. Insights into this microstructure based mechanism have been provided by Roll (1984).

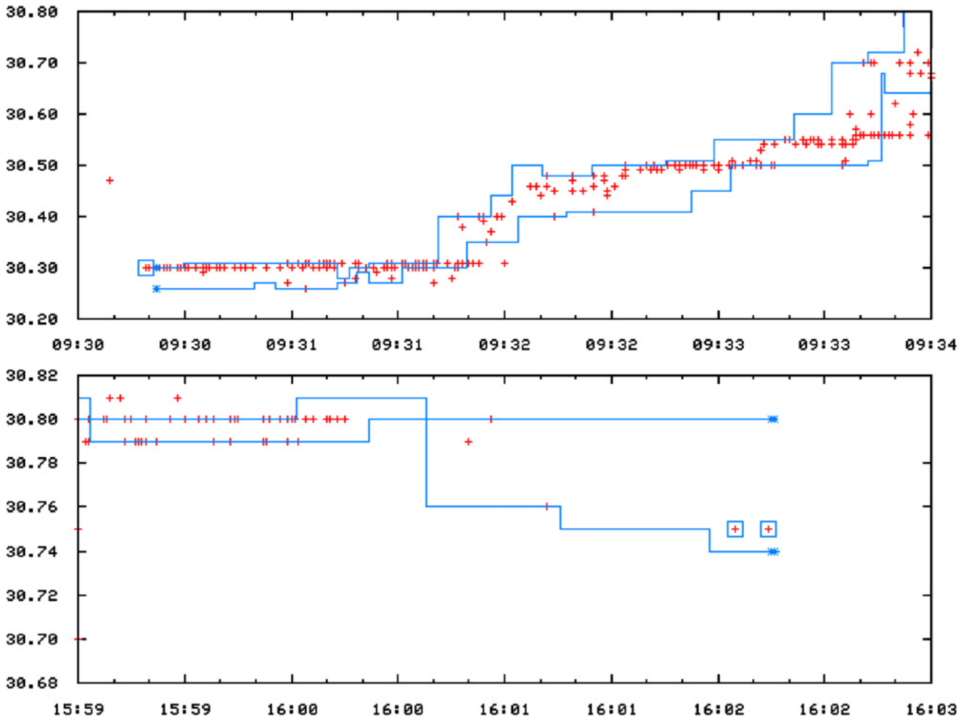

Fig. 5. A few minutes of trading activity for the General Electric stock on August 7, 2002 around 9:30 AM and 4:00 PM. Each cross marks a different transaction. The square boxes in the top panel indicate the NYSE official opening transaction, while the square boxes on the bottom panel indicate the NYSE official closing transaction. The stars on the bid–ask price lines indicate the NYSE opening and closing quote.

As observed in Fig. 4, the transaction price tends to "bounce", but this phenomenon is far less accentuated than in the early pre 2001 NYSE series. This is mainly because of the fine price "granularity". In fact, the minimum price variation on the NYSE since January 2001 is USD 0.01.

Traditionally, bid–ask bounces can be considered not containing useful information and they can lead to undesirable problems in that they can show price movements where none occurred in practice. It is thus of interest to find ways to remove or reduce the impact of this component on the data. There are several ways to achieve this result. The first consists of not using the transaction price series to compute returns but rather use the quote mid-price series. This is possible within TAQ but there is the above mentioned problem of non synchronicity between the two sets of data. Yet another solution is to construct an algorithm which eliminates all price movements which do not move the price above a certain threshold (larger than the bid–ask spread) from the last selected price.

Opening and closing: It would be natural to expect the first (respectively, last) trade/quote of the day to be the recorded opening (respectively, closing) trade/quote. From Fig. 5 showing the closing and the opening of the trading day considered, we can see that the first trade of the day is not the official NYSE opening trade. As expected, the opening quote is also reported after the opening trade. At the closing of the NYSE of the same day the last trade reported before 4:05 PM is actually the closing trade, but it was reported 2 min later the official closing. Thus, the detection of the opening and the closing of the NYSE is far less trivial than one could think.

The official NYSE trading day begins at 9:30 AM and finishes at 4:00 PM, but in practice trading will start some time after the official opening and will go some time beyond the closing, which implies that de facto the actual times of the opening and closing are indeed random. In addition to this, it is common to find reports of trades and quotes in the data which clearly do not belong to the NYSE trading day, that is, transactions before 9:30 AM and after 4:00 PM. These trade and quotes records may either come from other exchanges or from the off-hours (crossing) sessions of the NYSE and are therefore discarded. Lastly, to complicate matters, it is not uncommon that the actual trading of a stock will begin later than the opening time because of opening delays; also on some special days (like for example those preceding a holiday) the stock exchange may close at an earlier time.

It is thus not always possible to exactly identify the opening and the closing data. The MODE field in the quote data (Table 1) and the G127 and the COND fields of the trade data (Table 2) 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 practice, the difference between the first and last transaction prices of the day should not significantly differ from the true opening and closing and can be used as a proxy. However, this is not the case for transaction volume.

In order to adequately capture the closing price of the day, we adopt the convention that the trading day hours span between 9:30 AM and 4:05 PM, which ensures (to a large degree) that closing prices possibly recorded with a delay are accounted for. When using fixed-time intervals such as when building 10-min returns series, the last interval will span a nominally longer period (in the example, 15 min between 3:50 and 4:05 PM). This will give the return computed as the log-difference between the closing price and the price recorded at 3:50 PM as the last observation of the day.

4. An econometric application

We finally turn to an econometric analysis with UHFD. The goal is to show the consequences of using the original (dirty) data and of removing outliers from the data. We focus on financial durations, defined as the time in-between transaction price movements of a size above a given threshold (e.g. Giot, 2000). The application highlights the details of the time series construction and the impact of the data cleaning on the modeling exercise. The sample of observations used for the application is the transaction data for the GE stock recorded during the month of April 2002.

4.1. From the raw transaction data to the clean series

The number of raw transactions for the GE stock in April 2002 is 362 028 (22 trading days), 1499 of which were immediately discarded in that they did not lie within the (extended) NYSE trading day time defined as the period 9:30 AM–4:05 PM.

The plot of the raw tick-by-tick transaction price series of the first day of the sample in Fig. 6 clearly contains anomalous observations. The data was cleaned using the procedure described in the Section 3.1 above. For illustrative purposes, the data cleaning algorithm was run several times for a grid of different values of its parameters (k, γ); δ was kept fixed at 10%. Given that the GE stock is very liquid, that is frequently traded, the size of the window parameter k was set to reasonably large values. The bar diagram in Fig. 7, which displays the relative frequencies of the price variations between USD -0.06 and 0.06, guided the choice of the granularity parameter γ. In Table 3 we report the results of the number of excluded observations from the dirty data series according to values of k ranging from 40 to 80 and of γ from USD 0.02 to 0.06.

The cleaning procedure turns out to be more sensitive to the choice of γ, than it is to the choice of k, at least for this stock and sample period. Table 3 shows that with a strict choice of $\gamma = 0.02$ the number of outliers found is more than double the ones in the looser setting where $\gamma = 0.06$. The judgment on the quality of the cleaning can be had only by a visual inspection of the clean tick-by-tick price series graphs. In our view, a choice of $k = 60$, $\gamma = 0.02$ provides the most satisfactory results. See, for instance, Fig. 6 depicting the differences between the dirty and the clean series for different values of γ.

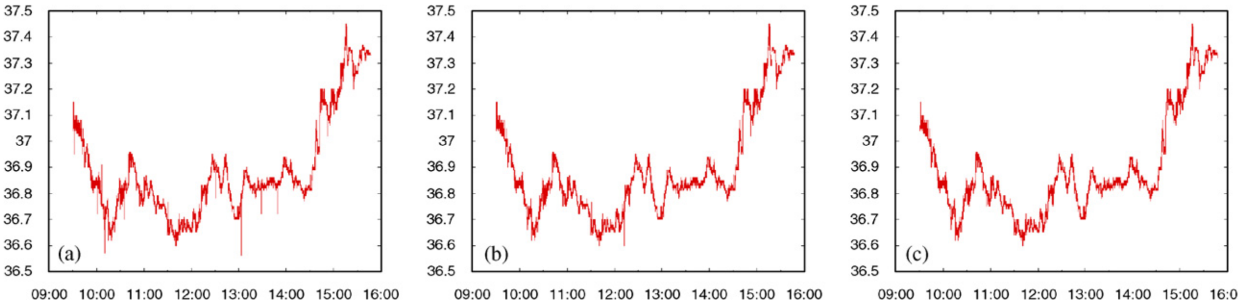

Fig. 6. Tick-by-tick transaction price series of the first day of the sample: (a) dirty time series, (b) clean time series with $k = 60$, $\gamma = 0.06$, (c) clean time series with $k = 60$, $\gamma = 0.02$.

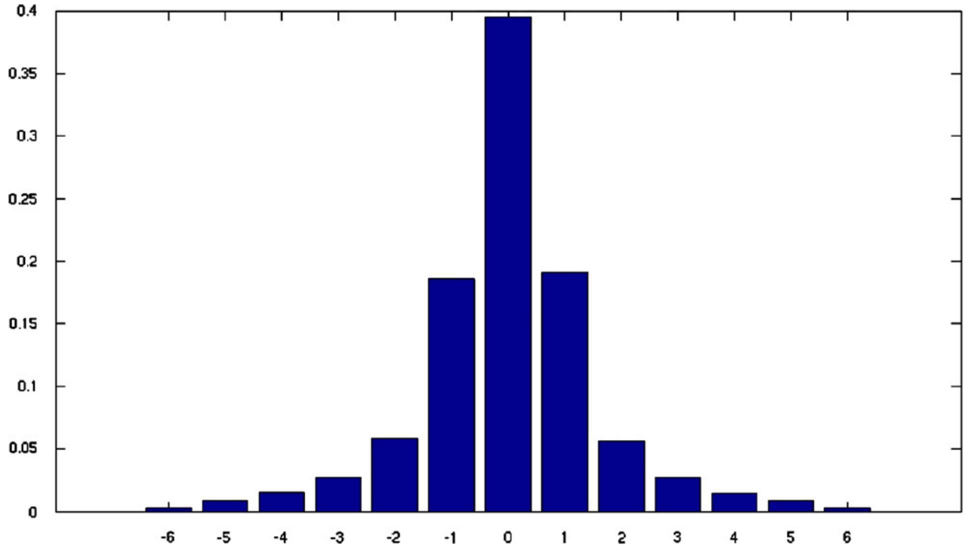

Fig. 7. Relative frequencies of transaction price changes between USD −0.06 and +0.06.

Table 3
Outlier detection. Results from the data cleaning algorithm as a function of the window length for the average, k, and of the granularity parameter γ

(k, γ)	Outliers	Average outlier per day
(40, 0.02)	659	30
(40, 0.04)	344	16
(40, 0.06)	248	11
(60, 0.02)	647	29
(60, 0.04)	359	16
(60, 0.06)	302	14
(80, 0.02)	638	29
(80, 0.04)	352	16
(80, 0.06)	255	11

Simultaneous prices were substituted with one observation at the median value of the simultaneous observations. A share-weighted average could be chosen instead, with little impact on the overall subsequent results, at least for the stock at hand. There were approximately 73, 000 sequences of simultaneous observations in both the dirty and clean series (corresponding to approximately 149, 000 observations removed). Furthermore, as the first half hour of the trading day is affected by a peculiar price formation mechanism, the corresponding observations have also been removed.

4.2. Transforming data into durations

Finally, the time series of price changes and its associated price duration series were constructed. To be clear, the resulting series will be dependent on the choice of the threshold giving rise to the corresponding durations $x_i = t_i - t_{i-1}$ be they derived from the raw data or from the clean data according to a specific choice of k, γ. For the illustration at hand, we considered price changes recorded at time t_i at or above a certain threshold chosen at USD 0.10 for all series.

Removing outliers in the data has the effect of decreasing the number of durations (there are fewer moves above the chosen threshold): this dramatically reduced the sample size. Table 4 gives an example of the impact the data handling has on the sample size of one clean data series, relative to the original dirty series.

Table 4
Data handling steps for the dirty and clean ($k = 60$, $\gamma = 0.2$) series

Operation	Dirty		Clean	
Number of raw observations minus	363,527	100%	363,527	100%
Ticks out of time scale	1499	0.41%	1499	0.41%
Filtered ticks	0	0.0%	647	0.18%
Simultaneous ticks	149,153	41.03%	148,672	40.90%
Within threshold	211,753	58.25%	211,863	58.28%
Final sample size	1122	0.31%	843	0.23%

Table 5
Descriptive statistics on the number of observations N of durations for transaction price changes above USD 0.10

Series type	(k, γ)	N	Mean	Min	Max	Std. Dev.
Dirty		1121	423	1	5375	631
Clean	(40, 0.02)	831	575	1	7606	838
Clean	(40, 0.04)	908	526	1	7606	794
Clean	(40, 0.06)	944	508	1	7606	762
Clean	(60, 0.02)	843	566	1	7606	841
Clean	(60, 0.04)	905	530	1	7606	802
Clean	(60, 0.06)	934	513	1	7606	764
Clean	(80, 0.02)	837	570	1	7606	844
Clean	(80, 0.04)	908	527	1	7606	801
Clean	(80, 0.06)	939	507	1	7606	751

In Table 5 we report the descriptive statistics about the durations series with data cleaning and without. The time series constructed from the dirty data series contains 1121 irregularly spaced observations, while, for example, the time series obtained from the clean data series ($k = 60$, $\gamma = 0.02$) contains 843 observations. Correspondingly, the clean series should also exhibit longer durations: in fact, the means are higher for the clean series (almost 10 min versus 7 min in the dirty series), but also the maximum value (approximately, 2 h after the cleaning as opposed to 1 h and a half before) and the standard deviations.

A way to visualize the inter-daily dynamics of the series is to count the daily number of price changes. Fig. 8 displays the plot of such series. Across days, the series exhibit the same dynamics but the dirty series overestimates the number of true price changes.

4.3. Financial duration modeling

The model introduced in Engle and Russell (1998) for the modeling of financial durations is the ACD model (cf. Bauwens et al., 2004 for an empirical application of several ACD models). Let $\{x_i\}_{i=1}^{N}$ be the series of financial durations. The standard ACD model decomposes the series in the product of a diurnal component ϕ_i, a conditionally autoregressive component ψ_i and an iid innovation term ε_i,

$$x_i = \phi_i \psi_i \varepsilon_i, \tag{1}$$

where ε_i can be assumed to be distributed as a Gamma r.v. with parameters $(\lambda, 1/\lambda)$ so that $E(\varepsilon_i) = 1, \ \forall i$. The seasonal factor ϕ_i is modeled using a cubic spline with knots set at each hour starting from 10:00 AM. Fig. 9 shows the intra-daily seasonality patterns emerging from the durations series. The patterns are practically the same, with the only difference that the clean series is shifted up as its durations are longer.

Table 6 displays the ACF of the seasonally adjusted dirty and clean duration series, together with the Ljung–Box test statistic for 15 lags. The persistence is strong in both series, but the clean series exhibits a stronger persistence: this

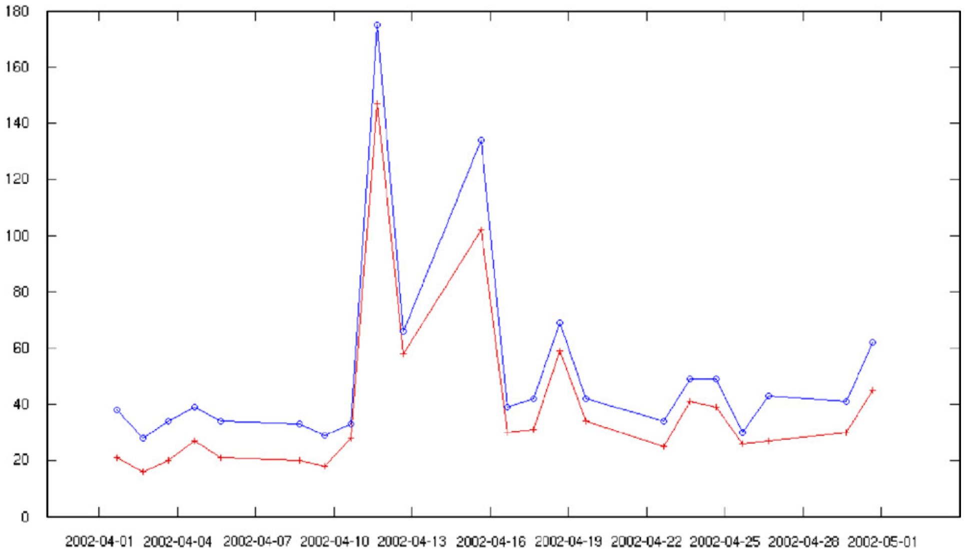

Fig. 8. Daily number of transaction price changes at or above 0.10 USD. Dirty series above, clean series ($k = 60$, $\gamma = 0.02$) below.

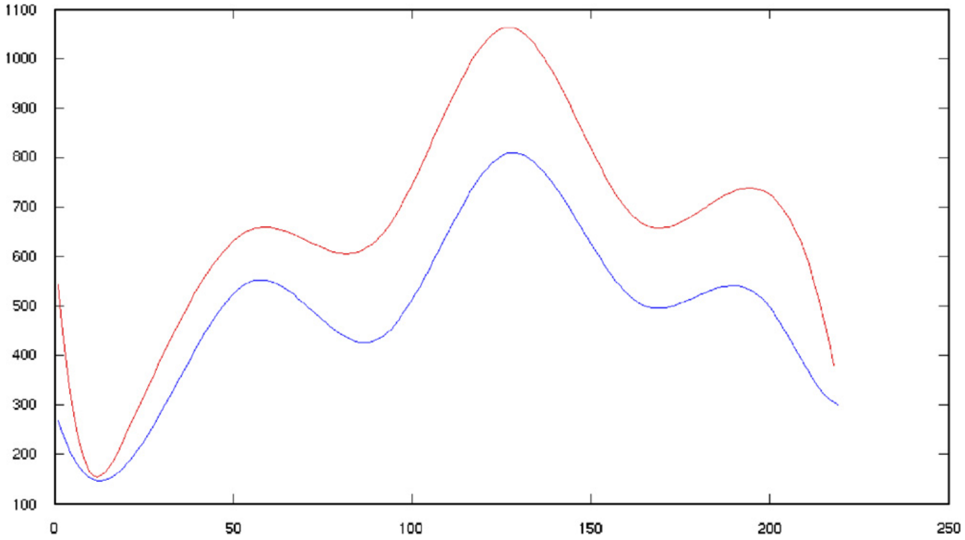

Fig. 9. Intra-daily seasonality pattern. Clean series above, dirty series below.

confirms the idea that the outliers interrupt the sequence of “true” durations at random times, hence interfering with duration clustering.

Clustering suggests the specification of ψ_i as

$$\psi_i = \omega + \alpha x_{i-1} + \beta\psi_{i-1}.$$

The model 1 was fitted to both the dirty and clean duration series. Estimation results and diagnostics on the residuals are presented in Table 7. The overall characterization of the duration dynamics is similar (and no major autocorrelation in the residuals is detected), but, as expected, the coefficient estimates on the various sets of data are quite different. In particular it seems that the main effect of the data cleaning is on the α and β coefficients in the simple ACD(1,1) adopted. The former are generally higher and they are more sensitive to the pair k, γ chosen, while the β's are smaller and less sensitive to the parameters used in the data cleaning procedure.

Table 6
Empirical ACF of the dirty and clean duration series

Series type	(k, γ)	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	LB(15)
Dirty		0.1389	0.1166	0.0827	0.1364	0.1559	227.3616
Clean	(40, 0.02)	0.2339	0.1467	0.1016	0.1413	0.2200	292.6566
Clean	(40, 0.04)	0.2244	0.1616	0.0970	0.1063	0.1625	347.4113
Clean	(40, 0.06)	0.2194	0.1805	0.1245	0.0995	0.1703	357.9523
Clean	(60, 0.02)	0.2395	0.1563	0.0956	0.1537	0.2327	331.8018
Clean	(60, 0.04)	0.2495	0.1715	0.1141	0.1503	0.2150	343.6282
Clean	(60, 0.06)	0.2165	0.1816	0.1090	0.1113	0.1666	361.2818
Clean	(80, 0.02)	0.2407	0.1626	0.0977	0.1415	0.2245	327.8541
Clean	(80, 0.04)	0.2510	0.1706	0.1091	0.1459	0.2027	332.2779
Clean	(80, 0.06)	0.1871	0.1892	0.1165	0.1476	0.2043	321.0136

Table 7
Estimation results and diagnostics on the estimated ACD(1,1) model with Gamma innovations: dirty and clean durations

Series type	(k, γ)	ω	α	β	φ	LogLik	LB(15)
Dirty		0.008 (0.005)	0.091 (0.0017)	0.903 (0.016)	0.499 (0.019)	−733.3	14.886
Clean	(40, 0.02)	0.0228 (0.0070)	0.1739 (0.0287)	0.8107 (0.0281)	0.6337 (0.0282)	−628.6	12.7417
Clean	(40, 0.04)	0.0123 (0.0047)	0.1315 (0.0213)	0.8599 (0.0217)	0.6169 (0.0260)	−644.1	14.6100
Clean	(40, 0.06)	0.0130 (0.0049)	0.1388 (0.0219)	0.8530 (0.0211)	0.5939 (0.0248)	−656.1	8.8255
Clean	(60, 0.02)	0.0177 (0.0061)	0.1403 (0.0229)	0.8456 (0.0230)	0.6208 (0.0274)	−614.4	11.6391
Clean	(60, 0.04)	0.0139 (0.0050)	0.1321 (0.0218)	0.8577 (0.0217)	0.6057 (0.0255)	−638.5	12.8831
Clean	(60, 0.06)	0.0132 (0.0053)	0.1255 (0.0215)	0.8643 (0.0222)	0.5914 (0.0248)	−654.8	9.2771
Clean	(80, 0.02)	0.0174 (0.0063)	0.1388 (0.0228)	0.8473 (0.0229)	0.6239 (0.0276)	−615.0	11.5708
Clean	(80, 0.04)	0.0147 (0.0053)	0.1332 (0.0220)	0.8557 (0.0220)	0.6021 (0.0253)	−639.9	15.7645
Clean	(80, 0.06)	0.0118 (0.0052)	0.1158 (0.0197)	0.8751 (0.0208)	0.5771 (0.0241)	−654.4	11.5803

5. Conclusions

In this paper we have discussed some issues surrounding the collection and distribution of ultra high frequency data in specific relationship to the NYSE and its commercially available TAQ database. We illustrate common problems related to errors present in the data and how they can be handled by applying some basic outlier detection procedure.

A clean data set is a preliminary necessary condition for moving into the second step of data manipulation involved in building a time series (durations, 5-min returns, realized volatility, realized range, and so on). Also for this step we document a framework within which elementary clean data can be aggregated to form the relevant time series to be analyzed. Special attention is devoted to the discussion of data at opening and closing time: for the latter, we suggest the extension of the trading day time to 4:05 PM since minor delays in recording the closing price past 4:00 PM are likely to occur within the NYSE.

The whole procedure is illustrated with reference to an estimation exercise of the ACD model proposed by Engle and Russell (1998). We show that failure to purge the data from “wrong” ticks is likely to shorten the financial durations between substantial price movements and to alter the autocorrelation profile of the series. The estimated coefficients and overall model diagnostics are altered when appropriate steps such as the ones we suggest are not taken. Overall the difference in the coefficients is bigger between the dirty series and the clean series than among series filtered with different values of the parameters.

Acknowledgments

We thank the Guest Editor Alessandra Amendola for her guidance and encouragement and two anonymous referees who helped us in focussing on the contribution of the paper. We thank participants in the third IASC World Conference on Computational Statistics and Data Analysis, Limassol (Cyprus), October 2005 and in the ROBCLA Conference, Florence, January 2006, for many useful comments. All the numerical procedures used in this paper as well as other methods useful for UHFD analysis are implemented in the MATNEM library for MATLAB available at http://www.ds.unifi.it/ctb.

References

Aït-Sahalia, Y., Mykland, P., Zhang, L., 2005. How often to sample a continuous-time process in the presence of market microstructure noise. Rev. Financial Studies 28, 351–416.

Andersen, T., Bollerslev, T., Christoffersen, P., Diebold, F., 2006. Volatility and correlation forecasting. In: Elliott, G., Granger, C., Timmermann, A. (Eds.), Handbook of Economic Forecasting. North-Holland, Amsterdam.

Bauwens, L., Giot, P., 2001. Econometric Modelling of Stock Market Intraday Activity. Kluwer, Dordrecht.

Bauwens, L., Giot, P., Grammig, J., Veredas, D., 2004. A comparison of financial duration models via density forecasts. Internat. J. Forecasting 20, 589–609.

Blume, M.E., Goldstein, M., 1997. Quotes, order flow, and price discovery. J. Finance 52, 221–244.

Boehmer, E., Grammig, J., Theissen, E., 2006. Estimating the probability of informed trading—does trade misclassification matter?. J. Financial Markets, in press, doi:10.1016/j.finmar.2006.07.002.

Breymann, W., Dias, A., Embrechts, P., 2003. Dependence structures for multivariate high-frequency data in finance. Quantitative Finance 3, 1–14.

Dacorogna, M.M., Gencay, R., Muller, U.A., Olsen, R., Pictet, O.V., 2001. An Introduction to High Frequency Finance. Academic Press, London.

Dufour, A., Engle, R., 2000. Time and the price impact of a trade. J. Finance 555, 2467–2498.

Engle, R.F., Russell, J.R., 1998. Autoregressive conditional duration: a new model for irregularly spaced transaction data. Econometrica 66, 987–1162.

Engle, R.F., Russell, J.R., 2006. Analysis of high frequency data. In: Ait Sahalia, Y., Hansen, L.P. (Eds.), Handbook of Financial Econometrics.

Falkenberry, T.N., 2002. High frequency data filtering. Technical Report, Tick Data.

Giot, P., 2000. Time transformations, intraday data and volatility models. J. Comput. Finance 4, 31–62.

Hasbrouck, J., 1992. Using the torq database. Nyse Working Paper #92-05, New York Stock Exchange.

Hasbrouck, J., Sofianos, G., Sosebee, D., 1993. New york stock exchange system and trading procedures. Nyse Working Paper #93-01, New York Stock Exchange.

Lee, C.M.C., Ready, M.J., 1991. Inferring trade direction from intraday data. J. Finance 46, 733–746.

Madhavan, A., Sofianos, G., 1998. An empirical analysis of the nyse specialist trading. J. Financial Econom. 48, 189–210.

Muller, U.A., 2001. The olsen filter for data in finance, Uam.1999.04.27. Olsen & Associates, Seefeldstrasse 233, 8008 Zurich, Switzerland.

O’Hara, M., 1997. Market Microstructure Theory. Blackwell, London.

Oomen, R.C.A., 2006. Properties of realized variance under alternative sampling schemes. J. Business Econom. Statist. 24, 219–237.

Roll, R., 1984. A simple implicit measure of the effective bid–ask spread in an efficient market. J. Finance 39, 1127–1139.

Sofianos, G., Werner, I.M., 2000. The trades of nyse floor brokers. J. Financial Markets 3, 139–176.

Vergote, O., 2005. How to match trades and quotes for nyse stocks?. Ku wp, Katholieke Universiteit Leuven.