Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2006/03

Financial Econometric Analysis at Ultra–High Frequency: Data Handling Concerns

Christian T. Brownlees,
Giampiero M. Gallo

Università degli Studi
di Firenze

Financial Econometric Analysis at Ultra–High Frequency: Data Handling Concerns

Christian T. Brownlees * Giampiero M. Gallo *

Abstract

The financial econometrics literature on Ultra High-Frequency Data (UHFD) has been growing steadily in recent years. However, it is not always straightforward to construct time series of interest from the raw data and the consequences of data handling procedures on the subsequent statistical analysis are not fully understood. Some results could be sample or asset specific and in this paper we address some of these issues focussing on the data produced by the New York Stock Exchange, summarizing the structure of their TAQ ultra high-frequency dataset. We review and present a number of methods for the handling of UHFD, and explain the rationale and implications of using such algorithms. We then propose procedures to construct the time series of interest from the raw data. Finally, we examine the impact of data handling on statistical modeling within the context of financial durations ACD models.

Keywords: Ultra-high Frequency Data, ACD models, Outliers, New York Stock Exchange.

*Dipartimento di Statistica "G. Parenti", Viale G.B. Morgagni 59, I-50134 Firenze, Italy, e-mail: ctb@ds.unifi.it, gallog@ds.unifi.it. The software implementation of the methods used here can be found in the MATNEM package for MATLAB at http://www.ds.unifi.it/ctb. We thank participants in the 3^{rd} IASC World Conference on CSDA, Limassol (Cyprus), Oct. 2005 and in the Robust Classification and Discrimination with High Dimensional Data Conference, Florence, Jan. 2006, for many useful comments. Finally, we would like to acknowledge the contribution SKYPE™ gave to this project allowing long free conversations through the last bits of revision between the two authors across nine time zones. Financial support from the Italian MIUR (Grants PRIN and FISR) is gratefully acknowledged.

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 over the last years (for a survey, cf. Engle and Russell (2006)). The expression "financial high-frequency data" refers to the datasets containing detailed reports of all the financial markets activity information which is available. Some authors like Engle (2000) emphasize these new sources of information by using the expression financial "ultra" high-frequency data (UHFD) in order to stress that it is not possible to access more information than that contained in these data sets.

The atomic unit of information which builds up ultra high-frequency data is the *"tick"*. The word tick comes from the practitioners' jargon. 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 analyze, 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,
  - might not be time ordered,
  - might have to be reconstructed from other tick sequences and
  - might sometimes exhibit some anomalous behavior as a result of particular market conditions (e.g. opening, closing, trading halts, etc);
- 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 only without distorting its content.

There are many financial markets which produce tick data. The most relevant financial markets analyzed in an ultra high-frequency perspective are the exchange

rate and the equity markets. The exchange rate market is an over-the-counter (OTC) market while the equity market is organized in exchange and OTC markets. The data produced by the exchange rate market and the OTC equity market is usually collected by the data providers which electronically disseminate their information, like Reuters or Bloomberg, while the information produced by the exchanges is collected by the exchanges themselves. A very important difference between these two different marketplaces is that exchanges are usually regulated by special government laws. This implies that exchanges are usually much more monitored than the OTC markets and thus the information collected by the exchanges is much more extensive than the one collected by the OTC market.

As observed by Bollerslev (2001) in a survey on the state of the art in this area, the analysis of these data is not particularly simple, also because regulatory changes and technological advances make this field a rapidly changing one (cf., for example, the minimum price change set at USD 0.01 cent in Jan. 2001, the automated way of updating quotes introduced in May 2003, and so on). This notwithstanding, the investigation of financial markets at an intra-daily scale requires a different approach when compared with what can be done at the daily level. Some of the main causes of the difficulties involved are the need to understand the specific market mechanisms which have an impact in the dynamics of a series and the operational complications which arise when manipulating these data. Furthermore, the literature is not always too clear about how the time series of interest was constructed from the raw data, nor whether specific choices of preliminary data handling may bear consequences on subsequent statistical analysis. Notable exceptions are Bauwens and Giot (2001) who describe the trading mechanisms at work at the NYSE and provide some data handling suggestions related to the structure of the TAQ database and Vergote (2005) who is concerned by the widespread use of the 5–second rule suggested by Lee and Ready (1991) to synchronize trades and quotes without checking its robustness for the period and the asset at hand.

In this paper we examine the case of ultra high-frequency data in the Trades and Quotes (TAQ) database available from the New York Stock Exchange (NYSE) and

1. summarize the structure of the high frequency datasets and highlight some important aspects of the market procedures;

2. propose a simple method for outlier detection;

3. propose procedures to construct time series of interest and

4. present some insights on the impact UHFD handling on econometric analysis in the context of financial duration modeling (Engle and Russell, 1998).

The paper is structured as follows. Section 2 reviews the market rules and procedures of the NYSE and will describe its UHF database, the TAQ. Section 3 shows the patterns which emerge from the raw data and presents methods for data handling. Section 4 presents some descriptive statistics on financial durations and highlights the impact of data handling procedures on subsequent econometric analysis. Concluding remarks follow in Section 5.

2 Exchange Market Rules and Procedures

2.1 How does an exchange market work?

Trading securities on the various stock exchanges follows very complex rules and procedures which are subject to modifications in the course of time to adapt to technological advances and evolving regulatory needs. Ultra high-frequency data contain information regarding *all* market activity: as a result, it is not an easy task to synthesize the institutional features which may have relevant consequences on ultra high-frequency data collection and analysis. While we will try to highlight the main problems arising from a practical perspective, we are in no position to be comprehensive on these issues. The interested reader can find further details on general aspects of exchange structures in classic texts such as Hasbrouck (1992), while reference to specific updated information in technical reports produced by the exchanges themselves cannot be eschewed prior to any analysis with real data. Bauwens and Giot (2001) also provide an accurate description of the rules and procedures of the NYSE as well as other exchanges.

Agents interact with the market of a given asset through a sell or buy transaction proposal which is called *order*. It is not possible to submit orders continuously on the exchange but only in a specific period of time devoted to transactions, the *trading day*. Orders can be classified into two broad categories: *market* orders and *limit* orders. A market order represent a buy or sell order of a certain number of assets of a stock at the current standing (bid or ask) price. The relevant feature of these orders is that there is certainty about transaction but uncertainty as to the actual price of the transaction. On the other hand, a limit order specifies a limit buy or sell price of a certain number of assets of a stock at which the transaction has to be executed. A buy limit order specifies the maximum price at which a trader is willing to buy while a sell limit order specifies the minimum price at

which a trader is willing to sell. The important feature of limit orders is the uncertainty regarding the execution of the order (whether it will be executed and when) but the certainty as to the fact that if the order were executed, it would be executed at the requested price or better.

The set of all buy and sell limit orders of given stock forms the so called *book*, which lists the orders on the basis their price and time of submission. The book provides a very detailed picture of the market for the asset, in fact market participants usually access information regarding the portion of the book which is "near–the–market", i.e. the list of sell and buy limit orders near the current market price, and hence more likely to be executed soon.

The current best sell and buy limit order form the current *bid* and *ask*, that is the *quote*. The quote provides the upper and lower bound within which transactions will occur.

The market orders reaching the exchange will be executed with the current quote generating a *trade*. The priority rules used to match orders on an exchange can be very complex and specific. Special mention should be made of large orders (block transactions), which may generate many separate transactions. Generally speaking, the exchange regulations try and minimize the number of such transactions, while maintaining a time priority principle.

2.2 Trading on the NYSE

The NYSE is a hybrid market, in that it is both an *agency* and an *auction* market. It is an agency market since the buy and sell orders are executed through an agent, the market maker, who is called the *specialist* at the NYSE. The NYSE is also an auction market, since on the exchanges *floor* the *brokers* participate actively in the negotiation and thus contribute to the determination of the transaction price. Also note that despite the global trend towards computer automated trading systems, the "human" element has a very crucial role in the trading mechanisms, and this fact as a deep impact on the data. Rules and procedures of the NYSE are described in detail in Hasbrouck (1992), Hasbrouck et al. (1993), Madhavan and Sofianos (1998), O'Hara (1997).

Trading takes place on the *floor* Monday to Friday, from 9:30AM to 4:00PM. The trading floor of the NYSE is composed of a series of contiguous rooms, where the *trading posts* are located. All the activity linked to a given stock is done in proximity of the *panels* located on the trading posts. At each panel the specialist of the assigned stock and the brokers interested in trading form the so called *trading crowd*. The trading crowd's function is to determine the transaction price of a

stock through the negotiation. As a market maker, the specialist is obliged to ensure that the market is liquid enough and stable, that is s/he has to *make the market* at her/his own risk with the aim of making it as easy as possible for the brokers to execute their transactions. In different words, the specialist behaves against the market buying when the price goes up and selling when the price goes down. During the trading day, the specialist remains close to the trading post and the panel relative to the stock s/he is assigned to together with his clerks. The brokers on the other hand can trade any stock on the floor, thus they move across the various trading posts, even if they specialize just on a few stocks. Brokers can participate in the trading either in an active mode, interacting within the trading crowd, or in a passive mode, leaving her/his orders to the specialist.

The order handling mechanisms of the exchange floor are constructed around the figure of the specialist. Orders reach the specialist through either a broker or through the NYSE network. The market orders that reach the specialist wait until they are executed by the specialist. The specialist can execute them with another order, on her/his inventory, against a broker in the trading crowd or in an another linked exchange. All orders are always executed within the current bid and ask quotations, that is to say the highest buying price and lowest selling price set by the specialist. It is important to stress that the current highest buy limit order and the lowest sell limit order in the book do not automatically become the current bid and ask. An order becomes the current quote when the specialist communicates this to the trading crowd (there are cases when the specialist is forced to do so). It is also important to emphasize that transactions are almost never automatically executed. Thus in practice, all relevant transactions are executed under the specialist's supervision.

Special regulations by the NYSE apply to important events in the trading activity, namely market opening, closing and the trading interruptions. At the opening of the trading day, the specialist is obliged to present a quotation that is as close as possible to the closing price of the previous day. The specialist is helped in this operation by an automated program developed by the NYSE IT infrastructure which matches the outstanding sell and buy orders and presents the balance to the specialist. On this basis s/he will then decide how many buy or sell orders will be presented.

The closing price is set on the basis of the balance of *market-on-close* (MOC) orders, which are buy or sell orders that are executed entirely at the closing price or are canceled. In case there is a non null balance of the MOC orders, the difference is executed against the current bid or ask at the closing determining the closing price. The other orders will be executed at that price. If the size of the buy and

sell MOC order is equal, the closing price is the price of the last transaction.

Under some particular circumstances the specialist can declare a delay of the opening of the transactions for some stocks, an *opening delay*, or a temporary interruption of the transactions, a *trading halt*. The possible causes for these types of delays or halts are news pending, news dissemination, or big order balances needing to be executed. After the market breaks of October 1987 and 1989, the NYSE also introduced some market-wide "circuit-breakers" the function of which is to slow down or even stop transactions in high volatility phases. Finally the Board of Directors of the NYSE can declare in particular conditions other types of interruption of the normal trading day (snow storms, network problems, commemorations and so on).

2.3 Comments on the NYSE Trading Mechanisms

Even if a large part of the procedures of the NYSE are automated, its dynamics are clearly strongly characterized by the interaction on the floor between the specialists and the brokers. From the point of view of ultra high-frequency data sensitivity to extra noise and dynamics that could be originated by the human component bit of the black box should be adequately taken into consideration.

As far as the extra noise is concerned, the NYSE trading mechanisms seemingly do not ensure that the timing of the data be necessarily reliable and significant. Although since 1993 the NYSE has established some standards regarding the time of execution (most transaction are matched by the specialist after an interaction with the trading crowd) the execution of market orders is by no means always immediate. Trade reporting can be affected by delays as well. In case the specialist matches two orders which have reached the floor via the NYSE network, the execution and the report are simultaneous. In the case of an order execution involving a floor member, the transaction will be reported with some delay given the time needed to generate the report. Also, since the reporting mechanisms on the floor have undergone dramatic changes through the years the order of magnitude for delays varies through time.

Trading by brokers on the floor are likely to complicate the price process dynamics. Although almost all transactions which are executed on the NYSE come from the exchange network system, floor brokers execute a very large share of the total volume of transactions executed. There are some interesting empirical studies which have tried to further characterize the quality of brokers trading. Some estimates made by Sofianos and Werner (2000) found that the average orders of the broker (from 10000 to 49000 stocks) are on average 5 times bigger than the or-

ders reaching the floor through the network. Brokers will then divide an order into smaller trades, which will be executed in a time range which on average ranges from 11 to 29 minutes. This brings to mind the fact that the *on floor* information set of brokers contains lots of important information which is not available *off floor*.

Thus, these examples seem to suggest that the higher the frequency the higher attention should be devoted to the analysis of the data.

2.4 UHFD Resources at NYSE

The categories of data collected by the NYSE are: *orders details*, *quotations* and *transactions*. The NYSE is probably the first exchange which has been distributing its ultra high-frequency data sets since the early '90s. In 1993 the Trades, Orders and Quotes (TORQ) database, which contains a 3 month sample of data was released. Since 1994 the NYSE has started marketing the Trades And Quotes (TAQ) database. The TAQ database has undergone some minor modifications through the years leading to 3 different TAQ versions (0, 1 and 2). Finally, since 2002 book data has also been available for research purposes.

The high frequency data of the NYSE is raw, in that the NYSE does not guarantee the degree of accuracy of the data, so that further manipulations are needed for using the data in research.

2.4.1 Quote Data

Quote data contains information regarding the best trading conditions available on the exchange. Table 1 describes the fields which the TAQ database contains. Table 2 displays a few sample records from the quote database.

The basic piece of information of interest that a quote tick contains are

- the *quote time stamp* which is the date and approximate time of the order execution,
- the *bid price* which is the bid price of a single share of the asset exchanged,
- the *bid volume* which is the number of round lots (100 share units) which are offered,
- the *ask price* which is the ask price of a single share of the asset exchanged,

Field	Description
SYMBOL	Stock symbol
EX	Exchange on which the quote occurred
QDATE	Quote date
QTIM	Quote time (expressed as cumulative number of seconds since 00:00am)
BID	Bid price
OFR	Offer price
BIDSIZ	Bid size in number of round lots (100 share units)
OFRSIZ	Offer size in number of round lots (100 share units)
QSEQ	Market Data Systems (MDS) sequence number
MODE	Quote condition
MMID	NASD Market Maker

Table 1: TAQ Quote data description.

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	

Table 2: TAQ Quote records.

Field	Description
SYMBOL	Stock symbol
EX	Exchange on which the trade occurred
TDATE	Trade date
TTIM	Trade time (expressed as cumulative number of seconds since 00:00am)
PRICE	trade price per share
SIZ	Number of shares traded
CORR	Correction Indicator
TSEQ	Market Data System (MDS) sequence number
COND	Sale conditions
G127	this field simultaneously indicates: G trade, a sell or buy transaction made by a NYSE member on his behalf; rule 127 transaction, i.e. a transaction executed as a block position; stopped stock. It is important to recall that a stopped stock *should* be used to identify the closing price as well

Table 3: TAQ Trade data description.

- the *ask volume* which is the number of round lots (100 share units) which are asked.

The quote table fields unfortunately do not include any information on the quality of the reported data. On the other hand, the MODE field (quote condition) contains lots of useful information which can be used to reconstruct accurately the trading day events. More specifically some of this field values indicate various types of trading halts that can occur during the trading day. Furthermore, the field also contains values to indicate the opening and closing quotes.

2.4.2 Trade data

Trade data contains information regarding the orders which have been executed on the exchange. Table 3 describes the fields which the TAQ database contains. Table 4 displays few sample records from the trade database.

The basic piece of information of interest that a transaction tick contains are

SYMBOL	EX	TDATE	TTIM	PRICE	SIZ	CORR	TSEQ	COND	G127
GE	N	020321	35605	37.700000	20000	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

Table 4: TAQ Trade records

- the *transaction time stamp* which is the date and approximate time of the order execution,
- the *transaction price* which is the price of a single share of the asset exchanged,
- the *transaction volume* which is the number of shares that were exchanged.

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

- the CORR field (correction indicator) signals whether a tick is correct or not,
- the "Z" and "G" value of COND field (sale conditions) indicate a trade reported at a later time.

3 Ultra High-Frequency Data Handling

Let us start looking at the data. The starting point of the research is that a sample of the trade and quote information of interest has been extracted from the TAQ. In the case of trades, incorrect and delayed transactions are eliminated from the sample. As it will be suitably justified, the preliminary steps needed before starting the econometric analysis of UHFD are:

1. **UHFD Cleaning**, i.e. detecting and removing wrong observations from the raw UHFD;

2. **UHFD Management**, i.e. constructing the time series of interest for the objectives of the analysis.

Unfortunately, these operations are far less trivial than one could expect. Furthermore, there are very few references in the econometric/quantitative finance literature which deal specifically with these issues, a notable exception being the works produced by Olsen & Associates (e.g Dacorogna et al. (2001)) who concentrated mainly on foreign exchange data.

3.1 Data Cleaning

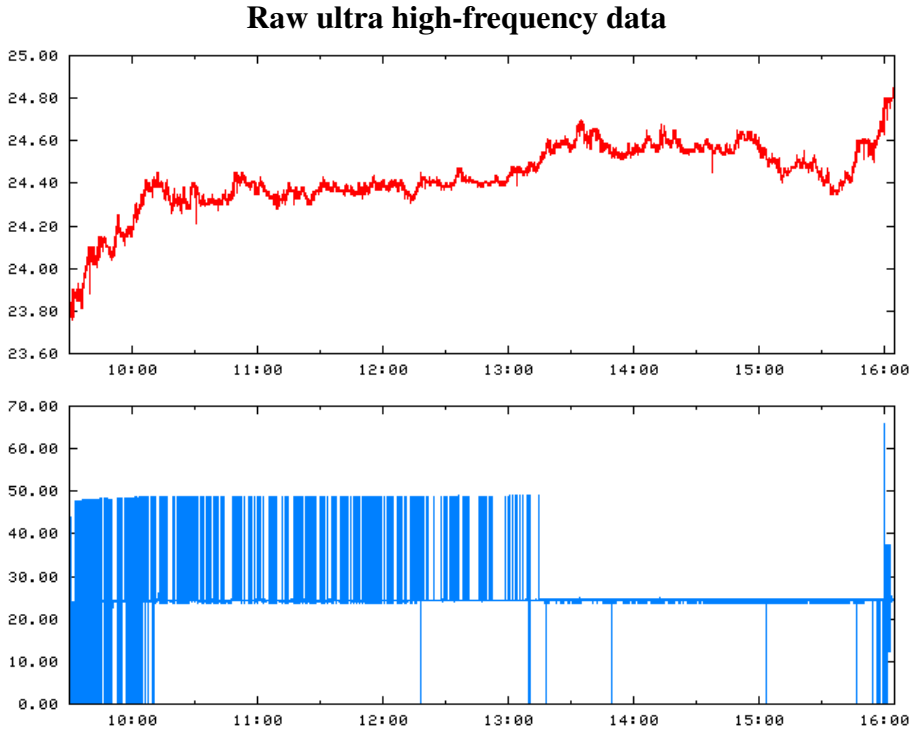


Figure 1: Trade and quote data for the General Electric stock on the 11^{st} of November 2002 from 9:30:00AM to 4:05:00PM. Tick–by–tick transaction prices (top); Tick–by–tick bid and ask prices (bottom).

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 the time series

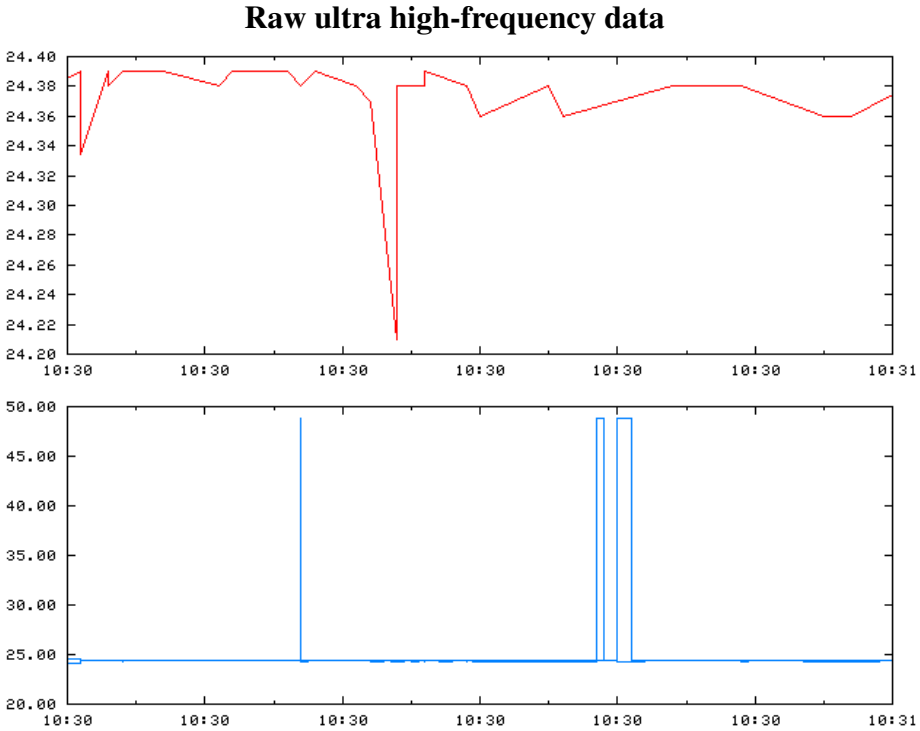


Figure 2: Trade and Quote Data for the General Electric stock on Nov. 11, 2002 from 10:30:00AM to 10:31:00AM. Tick–by–tick transaction prices (top); tick–by–tick bid and ask prices (bottom).

analysis. This kind of preliminary data manipulation is often referred as filtering or *data cleaning*. Its aim is to eliminate any observation which does not reflect the market activity from the ultra high frequency time series; in practice, however, it seems that these methods cannot go beyond detecting outliers.

It is not really clear where errors come from. Falkenberry (2002) reports that errors are present both in fully automated and partly automated trading systems such as the NYSE. According to the author, trading intensity is the main determinant of errors: the higher the velocity in trading, the higher the probability that some error will be committed in reporting trading information.

Most data cleaning procedures deal only with tick–by–tick price time series. This is probably due to the fact that while it is "easy" to detect a price which is coherent with the current market conditions, it is much harder to assess the plausi-

bility of a certain volume. Thus, for volume data there are no special data cleaning recipes other from 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 and 2 show respectively one day and one minute of trade and quote prices for the General Electric stock. Trade data seems to be much more accurate than quotation data. Coherently with the intuition of Falkenberry (2002), a possible explanation for this fact is that there are many more quotes than trades during a trading day. Another explanation is that trade data reports contracts which have been executed rather simple exchange proposals, and thus market agents are much more careful in reporting the former kind of information in comparison to the latter. An interesting aspect of the two images is that they seem to suggest that at least some of the series errors are indeed very evident.

There are some algorithms which have been proposed in the literature for washing away wrong observations (cf. Dacorogna et al. (2001) for the algorithm used at Olsen & Associates for performing this kind of task on exchange rate data which seems to be much more complex than needed for the NYSE data). Zhou (1996) mentions a simple method to validate foreign exchange quote data by comparing each quote to the medians of the three preceding and the three following observations and removing it if it is outside a fixed distance from those medians.

As an alternative, we suggest a procedure which has given satisfactory results, as argued below. Let $\{p_i\}_{i=1}^N$ be an ordered tick-by-tick price series. Our proposed procedure to remove outliers is

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

where $\bar{p}_i(k)$ and $s_i(k)$ denote respectively the 10% 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 most close valid observations. The role of the γ parameter is particularly important. Ultra-high frequency series often contain sequences of equal prices which would

lead to a zero variance, thus it is useful to introduce a lower positive bound on price variations which are always considered admissible.

It is important to make some remarks on the choice of the parameters of the algorithm: k and γ. 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. On the the other hand, if the trading is very active k should be "reasonably large" so that the window contains enough observations to obtain precise estimates of the price local characteristics. 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 an illustrative example to the choice of such parameters.

3.2 Data Management

Even once 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 suggested methods for their management.

Simultaneous Observations Figure 3 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 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 trade and quotes at different prices are normal. Another explanation of this phenomenon for trade data is that the execution on one exchange of market orders will in some cases produce more than one transaction report. A third and final explanation is that even non simultaneous trades/quotes could be all reported as simultaneous due to trade/quote reporting approximations. Thus in practice is

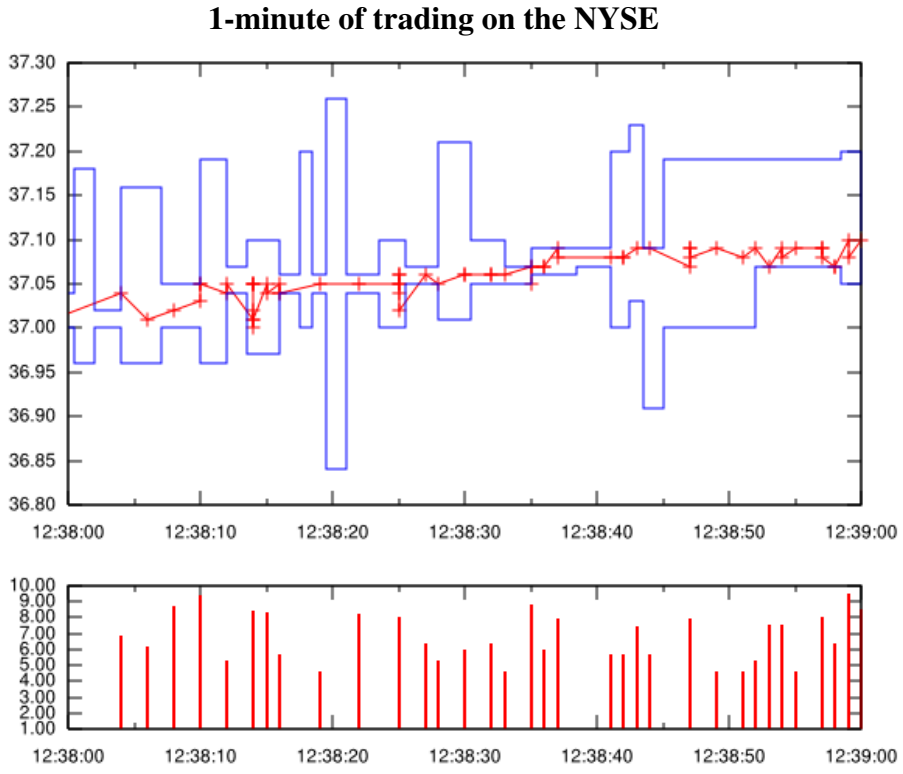


Figure 3: A minute of trading activity for the General Electric stock on the 21^{st} of March 2002 from 12:38 to 12:39PM. Transaction price within the bid-ask prices (top panel). Each cross on the transaction price line marks a different transaction. Log–volumes associated with the transaction (bottom panel).

common to find that more than one trade and quote have be reported at the same time.

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. For tick–by–tick prices the time series do not seem to suggest that the sequence of simultaneous observations reported is particularly 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 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 Figure 3 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 analyzing a time series with discrete, equally spaced time intervals.

Let $\{(t_i, y_i)\}_{i=1}^{N}$ be an irregular time series, where t_i and y_i indicate respectively the time and value of the i^{th} observation, and let $\{(t_j^*, y_j^*)\}_{j=1}^{N^*}$ be the lower frequency time series that we intend to construct. The problem then consist in using an appropriate aggregation function which employs the high frequency information to obtain the series of observations $\{y_j^*\}_{j=1}^{N^*}$ of the regular series. As we are aggregating higher frequency information at lower frequency, it seems coherent to use aggregation functions like:

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

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

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

First: $y_j^* = x_f$ where $t_f = \min\{t_i | t_i \in (t_{j-1}^*, t_j^*]\}$

[1]Note that alternatively f could have been defined as a function of $\{(t_i, y_i) \mid t_i \in [t_j^*, t_{j+1}^*),\ i = 1, ..., N\}$. It is more of a notation convention.

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: $y_j^* = x_l$ 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 the first four methods if the set $\{t_i | t_i \in [t_j^*, t_{j+1}^*)\}$ 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.

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: $y_j^* = y_p + \frac{t_j^* - t_p}{t_n - t_p}(y_n - y_p)$

The problem in using these methods, however, is that they might employ information which is 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^* 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 very different for infrequently traded stocks. In case of infrequently traded stocks we believe that interpolation schemes are not completely satisfactory. Say that we are constructing a high frequency series of regularly spaced prices, say 5 minutes. What often happens for non frequently traded stocks and during halts is that there will be some intervals, say $(t_{j-1}^*, t_j^*]$ which will not contain any observation and thus in case of previous point interpolation the interpolated price at t_j^* could be a price belonging to some time before, say $(t_{j-k-1}^*, t_{j-k}^*]$ for some $k > 1$.

In these cases we think that rather than setting an interpolated price is more appropriate to treat the observation as missing. If these price series are used to generate high frequency returns, this avoids introducing in the regularized series

long sequences of zero returns (if previous or next point interpolation are used) or identical returns (linear interpolation) which will increase serial correlation and are likely to generate problems in numerical nonlinear estimation procedures.

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, that is the tendency of the transaction price to bounce between the current bid and ask price. Insights into this microstructure founded mechanism have been provided by Roll (1984).

From an economic perspective, the reason behind 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 refer to purchases and sales and will tend to be executed at the current bid and ask, displaying the "bounce" pattern.

As it can be observed in figure 3 the transaction price tends to "bounce", but this phenomenon is far less accentuated than in other series and, especially, than early NYSE series. This is because of

- the fine price "granularity", that is the minimum price variation on the NYSE since January 2000 is 0.01 USD;
- price improvement, i.e. on the contrary of other exchanges on the NYSE it is possible to trade at better conditions than the current quote, and thus new market orders will not necessarily execute at the bid or the ask, reducing the extent of the "bounce" phenomenon.

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 it did not occur 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 in 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 a problem with non synchronicity between the two sets of data, so that, for example, the quotes data for 3:00PM are not comparable with the transaction price recorded at the same time, and cannot really be used to eliminate the problem at hand unless one resorts to the 5-second rule suggested by Lee and Ready (1991). Recently Vergote (2005) has reported that the delay between quotes and trades is time-varying and stock–specific and has proposed a new version of the algorithm proposed by Lee and Ready (1991). 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.

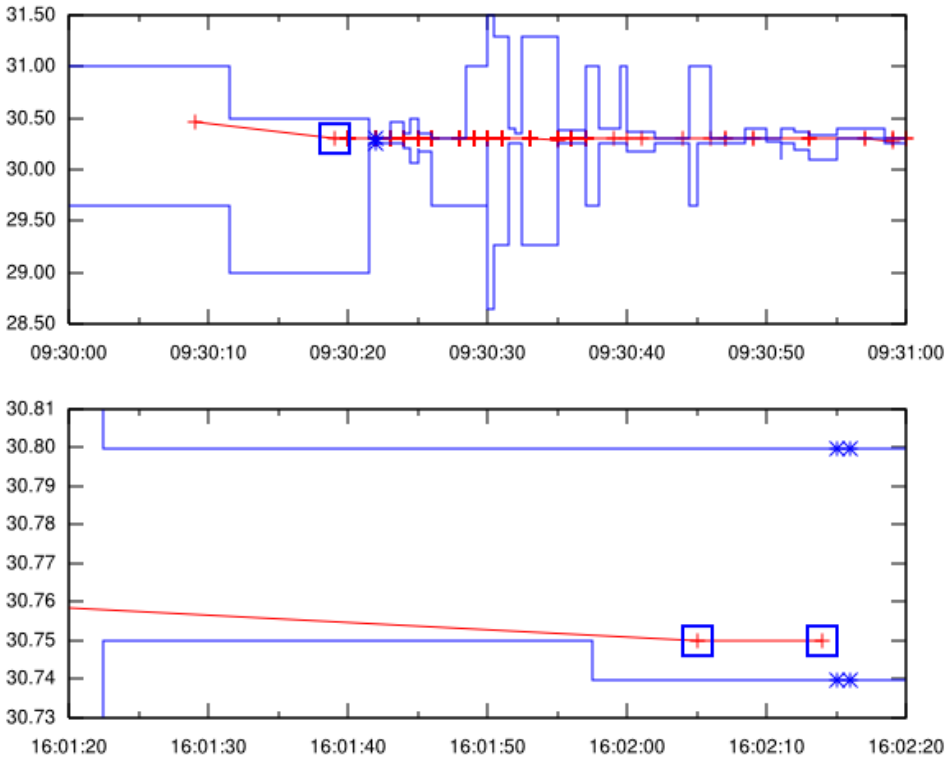


Figure 4: A first and last minute of trading activity for the General Electric stock on the 7^{st} of August 2002. Each cross on the transaction price line marks a different transaction. The square boxes in the top panel indicate the NYSE opening transaction, while the square boxes on the bottom panel indicate the NYSE closing transaction. The stars on the bid–ask price lines indicate the NYSE opening and closing quote.

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 Figure 4 showing the closing and the opening of the trading day considered, we can see that the first trades and quotes of the day are not the official NYSE opening trades and quotes. Interestingly, 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 4:05PM is actually the closing trade but the last reported quotes of the day were not. 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 9:30AM and finishes at 4:00PM, 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 9:30AM and after 4:00PM. 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 unfortunately 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 3) 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, bid or ask 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 and quote volumes.

In order to adequately capture the closing price of the day, we adopt the convention that the trading day hours span between 9:30AM and 4:05PM, 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-minute returns series, the last interval will span a nominally longer period (in the example, 15 minutes between 3:50PM and 4:05PM). This will give the return computed as the log-difference between the closing price and the price recorded at 3:50PM 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. 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 Data to the Time Series

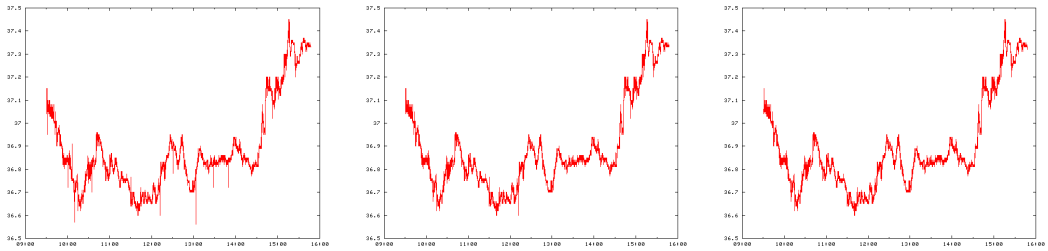
The number of raw transactions for the GE stock in April 2002 is 362028 (22 trading days), 1499 of which were immediately discarded in that they did not lie within the (extended) NYSE trading day time defined as the period 9:30AM–4:05PM.

The plot of the tick–by–tick transaction price series of the first day of the sample in Figure 5 clearly contains anomalous observations. Before starting the analysis, the data was cleaned using the procedure described in the subsection 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). 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 Figure 6, which displays the relative frequencies of the price variations between USD -0.06 and 0.06, guided the choice of the granularity parameter γ which regulates the threshold above the moving average which identifies an observation as an outlier to be excluded. In Table 5 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 sensitive to the choice of γ, more than it is to the choice of k, at least for this stock and time period. Table 5 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, Figure 5 depicting the differences between the dirty and the clean series for different values of γ.

4.2 Durations Data

Finally, the time series of price changes and its associated duration series were constructed. Simultaneous prices were substituted with one observation at the



a b c

Figure 5: 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 $k = 60, \gamma = 0.02$.

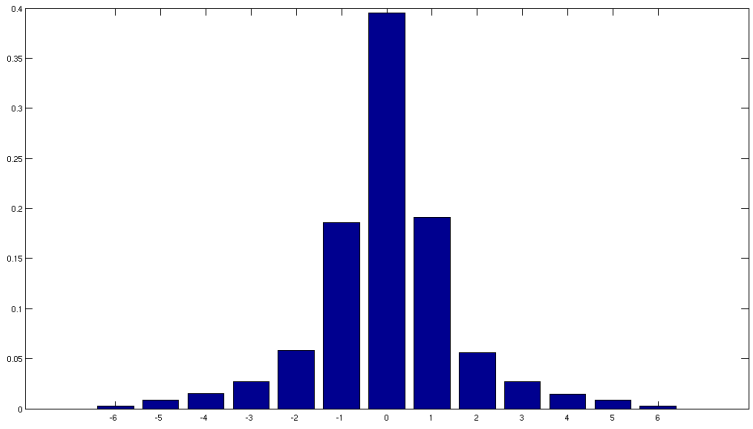


Figure 6: Relative frequencies of transaction price changes between USD -0.06 and $+0.06$.

(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

Table 5: Results from the data cleaning algorithm as a function of the window length for the average, k, and of the granularity parameter γ.

median value of the simultaneous observations. There were approximately 73000 sequences of simultaneous observations in both the dirty and clean series. We considered only price changes recorded at time t_i at or above a certain threshold set to USD 0.10 as giving rise to the corresponding durations $x_i = t_i - t_{i-1}$. 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. This dramatically reduced the sample size. Table 6 summarizes the impact of the data handling on the sample size for the dirty and one clean data series.

Operation	Dirty		Clean	
number of raw observations	363527	100%	363527	100%
minus				
ticks out of time scale	1499	0.41%	1499	0.41%
filtered ticks	0	0.0%	647	0.18%
simultaneous ticks	149153	41.03%	148672	40.90%
within threshold	211753	58.25%	211863	58.28%
final sample size	1122	0.31%	844	0.23%

Table 6: Data handling steps for the dirty and clean ($k = 60, \gamma = 0.2$) series.

In Table 7 we report the descriptive statistics about durations data series with data cleaning and without. The presence of outliers in the data has the effect

of splitting one "true" duration into at least two smaller ones. 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 minutes versus 7 minutes in the dirty series), but also the maximum value (approximately 2 hours after the cleaning as opposed to 1 hour and a half before) and the standard deviations.

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

Table 7: Descriptive statistics on the number of observations N of durations for price changes above USD 0.10.

A way to visualize the inter-daily dynamics of the series is to count the daily number of price changes. Figure 7 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 Autoregressive Conditional Duration (ACD) model. 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,

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

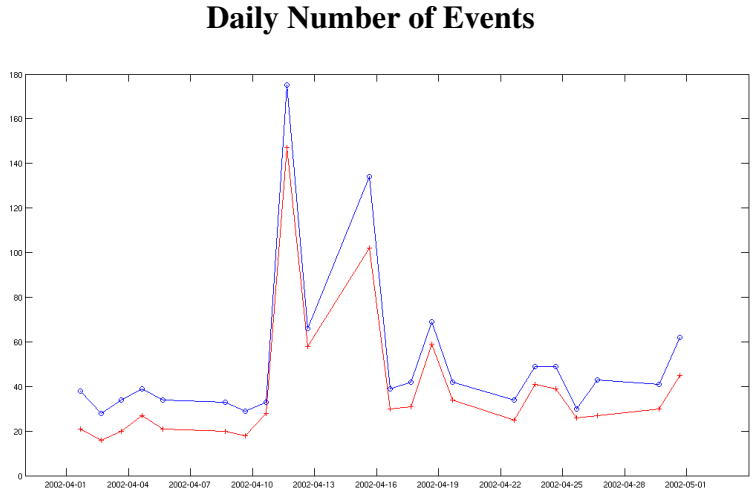


Figure 7: Daily number of transaction price changes at or above 0.10 USD. Dirty series above, clean series ($k = 60, \gamma = 0.02$) below.

The seasonal factor ϕ_i is modeled using a cubic spline with knots set at each hour starting from 10:00AM. Figure 8 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 8 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 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

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

Lastly, the specification of the iid innovation term ϵ_i is a version of a Gamma r.v. appropriately restricted as to have unit mean:

$$\epsilon_t \sim Gamma(\varphi, 1/\varphi) \quad f_{\text{gamma}}(x, \varphi) = \frac{1}{\Gamma(\varphi)} x^{\varphi - 1} \exp(-\varphi x) \quad \varphi > 0.$$

The model 1 was fitted to both the dirty and clean duration series. Estimation results and diagnostics on the residuals are presented in Table 9. The overall

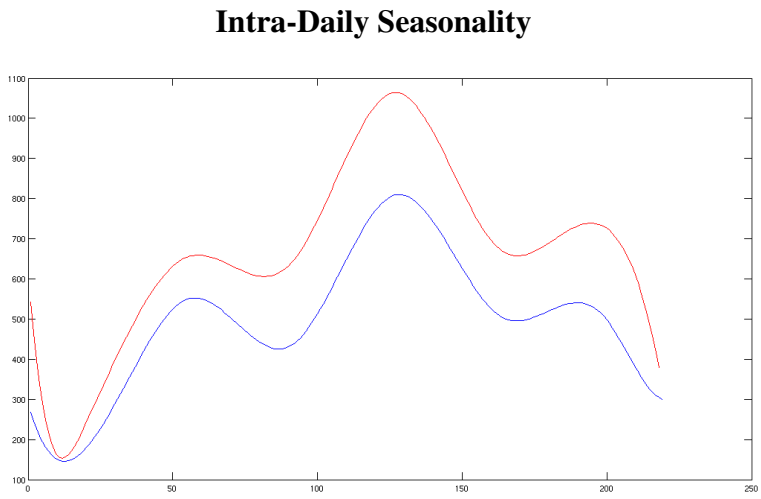


Figure 8: Intra-daily seasonality pattern. Clean series above, dirty series below.

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.

5 Conclusions

In this paper we have taken a long view on many issues surrounding the collection and distribution of ultra high frequency data in specific relationship to the NYSE and its commercially available TAQ database. We point out how errors are present in the data and how they can be handled by applying some basic outlier detection procedure (easier done for prices than it is for volumes, however).

A clean dataset is a preliminary necessary condition for moving into the second step of data manipulation involved in building a time series (durations, five-minute returns, realized volatility, realized range, and so on). Also for this step we document a framework within which elementary clean data can be aggregated

		ACF					
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 8: Empirical ACF of the dirty and clean duration series.

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 4:05PM since minor delays in recording the closing price past 4:00PM are likely.

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 considerably altered when appropriate steps such as the ones we suggest are not taken.

References

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

Bollerslev, T. (2001), 'Financial econometrics: Past developments and future challenges', *Journal of Econometrics* **100**, 45–51.

Dacorogna, M. M., Gencay, R., Muller, U. A., Olsen, R. and Pictet, O. V. (2001), *An introduction to high frequency finance*, Academic Press, London.

ACD(1,1) Estimation Results

Series Type	(k, γ)	ω	α	β	φ	LogLik	LB(15)
Dirty		0.008	0.091	0.903	0.499	-733.3	14.886
		0.005	0.017	0.016	0.019		
Clean	(40,0.02)	0.0228	0.1739	0.8107	0.6337	-628.6	12.7417
		0.0070	0.0287	0.0281	0.0282		
Clean	(40,0.04)	0.0123	0.1315	0.8599	0.6169	-644.1	14.6100
		0.0047	0.0213	0.0217	0.026		
Clean	(40,0.06)	0.0130	0.1388	0.8530	0.5939	-656.1	8.8255
		0.0049	0.0219	0.0211	0.0248		
Clean	(60,0.02)	0.0177	0.1403	0.8456	0.6208	-614.4	11.6391
		0.0061	0.0229	0.0230	0.0274		
Clean	(60,0.04)	0.0139	0.1321	0.8577	0.6057	-638.5	12.8831
		0.0050	0.0218	0.0217	0.0255		
Clean	(60,0.06)	0.0132	0.1255	0.8643	0.5914	-654.8	9.2771
		0.0053	0.0215	0.0222	0.0248		
Clean	(80,0.02)	0.0174	0.1388	0.8473	0.6239	-615.0	11.5708
		0.0063	0.0228	0.0229	0.0276		
Clean	(80,0.04)	0.0147	0.1332	0.8557	0.6021	-639.9	15.7645
		0.0053	0.0220	0.0220	0.0253		
Clean	(80,0.06)	0.0118	0.1158	0.8751	0.5771	-654.4	11.5803
		0.0052	0.0197	0.0208	0.0241		

Table 9: Estimation results and diagnostics on the estimated ACD(1,1) model with Gamma innovations: dirty and clean durations.

Engle, R. F. (2000), 'The economics of ultra high frequency data', *Econometrica* **68**, 1–22.

Engle, R. F. and Russell, J. R. (1998), 'Autoregressive conditional duration: A new model for irregularly spaced transaction data', *Econometrica* **66**, 987–1162.

Engle, R. F. and Russell, J. R. (2006), Analysis of high frequency data, *in* Y. Ait Sahalia and L. P. Hansen, eds, 'Handbook of Financial Econometrics', Elsevier.

Falkenberry, T. N. (2002), High frequency data filtering, Technical report, Tick Data.

Hasbrouck, J. (1992), Using the torq database, Nyse working paper #92-05, New York Stock Exchange.

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

Lee, C. M. C. and Ready, M. J. (1991), 'Inferring trade direction from intraday data', *Journal of Finance* **46**, 733–746.

Madhavan, A. and Sofianos, G. (1998), 'An empirical analysis of the nyse specialist trading', *Journal of Financial Economics* **48**, 189–210.

O'Hara, M. (1997), *Market Microstructure Theory*, Blackwell.

Roll, R. (1984), 'A simple implicit measure of the effective bid-ask spread in an efficient market', *Journal of Finance* **39**, 1127–1139.

Sofianos, G. and Werner, I. M. (2000), 'The trades of nyse floor brokers', *Journal of Financial Markets* **3**, 139–176.

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

Zhou, B. (1996), 'High-frequency data and volatility in foreign-exchange rates', *Journal of Business and Economic Statistics* **14**, 45–52.