Original Article

An improved algorithm for cleaning Ultra High-Frequency data

Received (in revised form): 29th June 2009

Thanos Verousis

is a lecturer in Accounting and Finance at Swansea University. He holds a PhD from the University of Wales. His research interests are in the fields of market microstructure and options trading. He is a review editor for market microstructure at the New Economic Papers network and a member of the British Accounting Association.

Owain ap Gwilym

is a professor of Finance at Bangor Business School, which was ranked first for research in accounting and finance in the 2008 UK university research assessment exercise. He has published over 50 research articles, including many on derivatives and market microstructure. He is a co-editor of credit ratings publications by InteractiveData.

Correspondence: Thanos Verousis, School of Business & Economics, Swansea University, Singleton Park, Swansea SA2 8PP, UK.
E-mail: T.Verousis@swansea.ac.uk

ABSTRACT We develop a multiple-stage algorithm for detecting outliers in Ultra High-Frequency financial market data. We show that an efficient data filter needs to address four effects: the minimum tick size, the price level, the volatility of prices and the distribution of returns. We argue that previous studies tend to address only the distribution of returns, and may tend to ‘overscrub’ a data set. In this study, we address these issues in the market microstructure element of the algorithm. In the statistical element, we implement the robust median absolute deviation method to take into account the statistical properties of financial time series. The data filter is then tested against previous data-cleaning techniques and validated using a rich individual equity options transactions data set from the London International Financial Futures and Options Exchange. The paper has many relevant insights for any practitioner who uses high frequency derivatives data, for example, for market analysis or for developing trading strategies.

Journal of Derivatives & Hedge Funds (2010) **15**, 323–340. doi:10.1057/jdhf.2009.16

Keywords: Ultra High Frequency; data mining and cleaning; equity options; LIFFE

INTRODUCTION

Ultra High-Frequency Data (UHFD) refers to a financial market data set in which all transactions

are recorded.¹ A number of studies highlight the importance of detecting outliers in UHFD,^{2–4} but there is a general lack of published literature

on data-cleaning filters for implementation in historical UHFD series.

This article surveys the existing literature on data-cleaning filters and proposes a new algorithm for detecting outliers in UHFD. To our knowledge, this is the first study that develops a data filter that encompasses the data-cleaning arrangements proposed by historical data providers (Olsen and Associates and Tick Data Inc.). The algorithm is compared with a previous data filter (Huang and Stoll,⁵ henceforth HS), and its validity is confirmed by applying the filter to options market data.

An outlier or a data error is defined as an observation that does not reflect the trading process, and hence there is no genuine connection between the market participants and the recorded observation. Muller⁶ argues that there are two types of errors: human errors that can be caused unintentionally (for example, typing errors) or intentionally, for example producing dummy quotes for technical testing⁷; and computer errors (technical failures), making it even more difficult to detect the origins of outlying observations.⁸ On this basis, Falkenberry⁴ remarks that 'the most difficult aspect of cleaning data is the inability to universally define what is unclean'. The problem lies in the trade-off between applying too strict ('overscrubbing', Falkenberry⁴) or too loose outlier detection models, and in the fact that it is very difficult to systematically identify causes of data errors.

HS, Chung *et al*⁹ and Chung *et al*¹⁰ develop and implement different versions of a data-cleaning algorithm that is based on the assumption that excess returns (positive or negative) are in principle caused by the presence of outlying data. Returns that are found to lie outside the prescribed return window are dropped from the sample as outliers. In contrast,

historical data providers stress the importance of accounting for the time effect in data filtering.^{4,6} The latter models, however, tend to be too complex to be implemented in specific data samples, and the specifications of the filters are not disclosed by the data providers. The problem is particularly severe where exchanges have no (reliable) in-house data filtering process.

In this article, we identify four distinctive effects that should be accounted for in detecting outlying observations in UHFD. In particular, we support the proposition that while HS focus on the application of a 10 per cent return criterion, the latter may lead to labelling an excessive number of observations as outliers.¹¹ This study implements the following four data selection criteria:

- *The minimum tick size effect*: we document how low-priced securities are affected by a relatively large minimum tick size;
- *The price level effect*: we assert that the uniform application of a return criterion may lead to 'overscrubbing' the lower-priced observations of a data set;
- *The daily price range effect*: a method of selecting observations that fall within the average daily price range is proposed that controls for large price differences across trading days, and that can also be used as a robustness test; and
- *The return effect*: finally, similar to HS, we apply a return criterion; however, ours also controls for the effect of differences in the price level of assets.

A statistical algorithm is established to implement these concepts. The results are tested on an UHF transactions data set for 28 individual equity options contracts traded on the London International Futures and Options Exchange (LIFFE) during 2005. The latter data set is used

as it appropriately encompasses all the issues discussed above. The results are compared with an existing data filter, and the consistency of the filters is analysed.

The remainder of this article is organised as follows. The next section discusses the issues that arise with regard to data filtering. The subsequent sections present the steps for detecting outliers in UHFD and discuss data selection criteria and the returns' calculation method, respectively. The next section presents the algorithm for detecting outliers in UHFD. The penultimate section presents the results and analysis and the last section offers the conclusions.

EXISTING STUDIES ON UHFD CLEANING

Olsen and Associates and Tick Data Inc. develop and apply data filters in historical price data sets. These filters share some common traits.^{4,6} Bad (outlying) ticks are compared with a moving threshold so that the effect of time is addressed.¹² Ticks that exceed the threshold are identified as outliers. Finally, a procedure is in place to either replace the outliers with 'corrected' values (Tick Data Inc.) or to delete the outliers (as used by Olsen and Associates).

Although the outlier detection algorithms developed by private firms and exchanges can have wide applications, data-cleaning techniques applied in finance are mostly data-specific. Yet, articles on market microstructure tend to share some common characteristics that are mainly dictated by the nature of financial data. Values with the following characteristics are commonly omitted:

- Recorded trades and quotes occurring before the market open and after the market close

(widely applied in the market microstructure literature).

- Quotes or trades with negative or zero prices.^{9,10,15,16}
- Trades with non-positive volume,^{9,10,16,17} and
- Trades that are cancelled or identified by the exchange as errors.^{15,16,18}

HS develop a set of codes that is widely used in the relevant data-cleaning literature. The most important criterion within these codes is that not only are cancelled and before-open/after-close trades deleted, but outliers are also identified with respect to returns. In particular, trades (quotes) are classified as outliers when returns on trades (quotes) are greater than 10 per cent. In addition, quotes are deleted when spreads are negative or greater than US\$4 (zero spreads are possible, for example, on NASDAQ).¹⁹ Further criteria applied by HS entail deleting observations whose prices are not multiples of the minimum tick (see also Bessembinder¹⁵) and a market open condition based on the first-day return.

However, one point to consider from HS is the subjectivity of the 10 per cent return, signifying that data selection rules in UHFD are always prone to somewhat arbitrary data selection rules. This is demonstrated in the study by Chung *et al*,¹⁰ in which a 50 per cent return rule is applied, and in that by Bessembinder,¹⁵ in which prices that involve a price change of 25 per cent are omitted. In addition, Chung *et al*¹⁶ and Chung *et al*⁹ raise the issue of selecting only positive returns, and hence they expand on HS by selecting observations with less than 10 per cent absolute returns.²⁰

Outlier data-cleaning methods that rely on the statistical properties of the data offer the advantage of uniformity in data selection. Leung *et al*²¹ develop a two-phase outlier detection

system wherein the phase of data identification is followed by the second phase of detecting short-lived price changes based on the statistical properties of the data.

As an alternative to the outlier detection systems proposed, Brownlees and Gallo²² suggest a procedure that relies more on the deviation of observations from neighbouring prices. Observations are omitted when the absolute difference between the current price and the average neighbouring price is outside three standard deviations plus a parameter that controls for the minimum price variation. However, the authors conclude that the judgement of the validity of the parameters selected (the number of neighbouring prices and the minimum price parameter) can only be achieved by graphical inspection.

Finally, some studies rely on bid-ask spread criteria to eliminate outlying observations. Chordia *et al*²³ remove the following observations sampled from the NYSE: (1) those that lie outside a \$5 quoted spread and (2) those in which the fraction of the effective spread²⁴

over the quoted spread is greater than \$4.

In contrast, Benston and Harland¹⁷ use an effective spread of 20 per cent as their cut-off point, combined with the value of price per share for stocks traded on NASDAQ.

STEPS FOR DETECTING OUTLIERS IN UHFD

The common element of previous studies on deleting outliers in UHFD is the assumption that excess returns are the product of outlying data being present in the data set (see HS and Chung *et al*⁹). Hence, the objective of these studies is to appropriately define excess returns. In contrast, commercial data providers also focus on the effect of time in the calculation of returns.^{4,6} Below, we address these issues and discuss the appropriate steps that would need to be considered for an efficient data filter for UHFD (see also Figure 1).

The minimum tick size effect

In view of the fact that assets are often low-priced, the effect of a large minimum tick size can lead to an overly restrictive data-cleaning

Figure 1: Data filter steps.

technique that distorts valid data. For example, with a minimum tick of 0.5 pence, an asset that is priced at 3p with a previous price of 2.5p will be classified as an outlier with HS's 10 per cent return criterion solely owing to the minimum tick. Thus, data would be rejected even at one-tick movements, leading to excessive deletions and a clear bias in favour of retaining more data for higher-priced securities.

The price level effect

HS's and subsequent studies (see Bessembinder,¹⁵ Chung *et al.*,¹⁶ Chung *et al.*,⁹ and Chung *et al.*¹⁰) that uniformly apply a return criterion (10 per cent or 5 per cent) face the risk of 'overscrubbing' the lower end of the sample. As the price level of assets may vary widely, a uniform return criterion may not have the desired effects for low-priced assets. For example, a one-penny increase in two assets priced at 2p and 20p will generate returns of 50 and 5 per cent, respectively. Hence, the 'clean' data set would be skewed, as there is a higher probability of low-priced assets being classified as potential outliers. Clearly, the price level effect is also found in the calculation of returns, and thus the above discussion also applies to returns calculations.

In addition, although subsequent to that by HS, the studies of Chung *et al.*,¹⁶ Chung *et al.*⁹ and Chung *et al.*¹⁰ have remedied the problem of selecting only positive returns by defining outliers by using absolute return, another issue still remains. That is, even though the latter definition solves the problem of defining outliers as only those prices that are abnormally (more than 10 per cent) above the preceding price, it might also lead to removing observations that

are actually 'corrections' to an outlying price. For example, if $T=3$ and at t_1 , $p_1=5p$; t_2 , $p_2=20p$ and t_3 ; $p_3=5p$, then even though HS's model will classify p_2 as an outlier, the absolute returns model will delete both p_2 and p_3 on the basis of classifying the 'correct' p_3 price as an outlier.²⁵

The daily price range effect

A problem arises with applying a uniform return (absolute or not) criterion to the whole data set; the price range is not identified, which might lead to classifying an excessively large number of observations for deletion. The latter means that volatile assets will always generate high numbers of observations classified as outliers, even though the average price is close to the observed prices. For example, an asset priced at 3p will be classified as an outlier if the previous price is 2p and the minimum tick is 0.5p. Thus, a two-tick movement will actually be sufficient to lead to 'overscrubbing' of the sample.

Statistical data mining and robustness

Barnett and Lewis²⁶ note that real-time analytical data are often long tailed, containing a disproportionate (compared with the normal distribution) number of observations further away from the mean, and tend to contain erratic observations (that is, outliers). Hence, a statistical algorithm that will act as a robustness check for the data mining algorithm will have to take into account this specific characteristic of UHFD.

A popular approach to detecting outliers is the process of Windsorisation: instead of deleting the outlying value, replacing them with the closest 'clean' values, which, however, distorts the distribution of prices. Instead, trimming techniques are more appropriate. The Grubbs

test (Grubbs cited by Barnett and Lewis²⁶) is used to measure the largest absolute deviation of a price from the mean, standardised in units of standard deviation. A test statistic that follows a t -distribution is used to test the hypothesis of an observation being an outlier. However, as this test assumes normality, which cannot be directly inferred in UHFD (for example, ap Gwilym and Sutcliffe²⁷), and also can only be applied successively for one observation at a time, it is rejected on data-specific and computational grounds.

In contrast, the median absolute deviation (MAD) test relies on the fact that the median value of a data set is more resistant to outliers than the mean value. In addition, if normality cannot be inferred, the median value is more efficient than the mean value. The latter is true, as the mean can be affected by the presence of extreme values, whereas the median is less sensitive to the presence of non-normal distributions. MAD gives the median value of the absolute deviation around the median (see Fox²⁸).

$$\text{MAD} = \text{median}\{|p_1 - \mu|\}$$

where p_1 is price at $t=1$ and μ is the daily median value. MAD is not normally distributed; however, for a normal distribution one standard deviation from the mean is $1.4826 \times \text{MAD}$ (see Hellerstein²⁹ and Hubert *et al*³⁰). Hence, for the appropriate measure of two standard deviations from the mean, it is hypothesised that a value is an outlier if its standardised value is greater than $2.9652 \times \text{MAD}$.^{28,29,31}

DATA AND RETURNS CALCULATION

One market that demonstrates a number of difficulties in detecting outliers is the options

market. Options contracts are often low-priced and the minimum tick size can be large.

Computational difficulties arise because of the nature of options data and the complexity in the calculation of returns. In order to address these issues and demonstrate the appropriateness of the data-cleaning filter, the data sample is comprised of individual equity options contracts trading at LIFFE. The data set consists of all trades and quotes posted on the exchange during 2005.

In order to control for stale and non-synchronous pricing problems, we select the most heavily traded assets.^{27,32} Specifically, we select option contracts that report more than 1500 trades during 2005,³³ leading to a sample based on 28 equity options.

In general, the calculation of volatility follows the procedure introduced by Sheikh and Ronn.³⁴ Returns are calculated only for the at-the-money, nearest-to-mature contracts. As the calculation of the spread, even for the highly traded options, may lead to the use of stale prices, only ask prices are used.^{35,36} At each time interval, the first ask price is obtained. For the closing return calculation, the last ask price of the day is obtained. The closing ask price and the first ask quote of the next day are used for the computation of the opening returns. Different strike prices can meet the criteria for a given contract at consecutive intervals. The procedure adopted is as follows: at every hourly interval i the first ask price is obtained. Then, at the next hourly time interval $i+1$, the ask price with the same strike price is obtained. The logarithmic return is calculated from these two prices. If, however, there is no ask with the same strike price at the next interval $i+1$, we search for the next available ask price at interval i that satisfies that criterion. When the returns for intervals i

Figure 2: Stages in the proposed outlier detection process. Price (Pr) denotes the price of the asset after the data are defined into categories based on each option type, trade type, delivery date and strike price. μ denotes the average daily price. R is the simple return and SP denotes the standardised price. Finally, $NMAD$ is the normalised Median Absolute Deviation.

and $i + 1$ are calculated, the same procedure is repeated for the next interval, $i + 2$.

AN ALGORITHM FOR DETECTING OUTLIERS IN INDIVIDUAL EQUITY OPTIONS

In the interests of data homogeneity,⁶ the data selection method would be applied to the finest market structure available. That is, UHFD are employed and there is no aggregation of data in, for example, strike price or maturity date clusters. Hence, option contracts are classified at the following levels of variability: option types (call/put); trade types (trades, asks and bids); delivery dates and strike prices. It is worth mentioning that when the data are classified according to the above classification structure, the number of groupings found in the sample of 28 equity options for 2005 is 17 076.³⁷

Cancelled, block and outside the market open and close trades and quotes are deleted. Observations that show zero or non-positive volume are also dropped. Finally, three trading days are discarded from the data set as missing data is found on these dates.^{38,39}

Consistent with the above analysis, in order to capture the effect of the minimum tick size, we distinguish between low- and high-priced assets. In addition, we account for a large price movement for all options and for a large deviation of the observed price from the daily mean price. The algorithm also has a statistical property by applying the MAD criterion for the observations that are identified as potential outliers. The algorithm is presented in Figure 2. Below, we demonstrate how we controlled for the effects identified in the earlier section.

In order to capture the minimum tick size effect, assets with price change (price less lagged

price at previous transaction time) less than or equal to 0.5p (minimum tick) are immediately retained in the final sample. Figure 2 shows that options with prices less than or equal to 20p are treated differently than options with higher prices. For the first category of options, the algorithm identifies those observations with absolute return greater than 20 per cent. If the price of these stocks is outside a 20 per cent window around the mean daily price, the observation is classified as a possible outlier. The above avoids the problem of deleting low-priced options, captures the effect of the tick size and is able to take into account the daily range of prices; thus, price jumps (volatility) are also accounted for. For example, options priced at 3p with lagged price of 2.5p will not be deleted. Even if the lagged price is 2p, the observation will not be deleted as long as the price is within the 20 per cent of the mean daily price window.

For options priced at more than 20p, the algorithm identifies observations with price spread greater than 0.5, price outside the price range of 10 per cent around the daily mean price and absolute return greater than 10 per cent. Hence, the high-priced securities are treated differently, for which the code is more similar to HS.

A note of caution should be made regarding the minimum tick size that is found in the data set. Option contracts selected for this study are traded either at the minimum tick of 0.25p or at the minimum tick of 0.50p, and thus for those assets that are traded at multiples of 0.25, the minimum tick restriction employed is also applicable, as the selection criterion of 0.5 is only twice the minimum tick size. The latter implies that securities whose prices differ from the lagged price by less than or equal to 0.5 are automatically retained, irrespective of the two minimum tick sizes found in this data set.

Table 1: The evolution of the data filter (trades only)

1. Firm	2. Raw data	3. Huang and Stoll (HS)			4. HS plus minimum tick (HSMT)			5. HSMT plus price level (HSMTPL)			6. HSMTPL plus volatility (no MAD)			7. Final data set		
		Obs. retained	% Outliers		Obs. retained	% Outliers		Obs. retained	% Outliers		Obs. retained	% Outliers		Obs. retained	% Outliers	
OAAM	2388	1807	24.33	1813	24.08	1837	23.07	2340	2.01	2382	0.25					
OAWS	1733	1382	20.25	1405	18.93	1479	14.66	1723	0.58	1728	0.29					
OAZA	7904	6463	18.23	6486	17.94	6566	16.93	7705	2.52	7873	0.39					
OBBL	5211	4359	16.35	4422	15.14	4602	11.69	5169	0.81	5191	0.38					
OBLT	3380	2764	18.22	2776	17.87	2838	16.04	3350	0.89	3371	0.27					
OBOT	2222	1867	15.98	1889	14.99	1964	11.61	2200	0.99	2216	0.27					
OBP	6883	5663	17.72	5711	17.03	5878	14.60	6816	0.97	6869	0.20					
OBSK	2724	2269	16.70	2297	15.68	2383	12.52	2702	0.81	2716	0.29					
OBTG	4044	3384	16.32	3571	11.70	3735	7.64	4025	0.47	4035	0.22					
OCPG	1588	1269	20.09	1276	19.65	1329	16.31	1568	1.26	1584	0.25					
OCUA	3174	2596	18.21	2622	17.39	2737	13.77	3145	0.91	3169	0.16					
OEMG	2566	2038	20.58	2042	20.42	2060	19.72	2529	1.44	2558	0.31					
OGNS	3669	3091	15.75	3138	14.47	3227	12.05	3628	1.12	3656	0.35					
OGXO	9551	7835	17.97	7870	17.60	8076	15.44	9351	2.09	9516	0.37					
OHSB	5797	4996	13.82	5082	12.33	5262	9.23	5776	0.36	5780	0.29					
OKGF	2437	2072	14.98	2087	14.36	2145	11.98	2421	0.66	2434	0.12					
OLS	2000	1588	20.60	1594	20.30	1623	18.85	1971	1.45	1993	0.35					
OPRU	2841	2302	18.97	2322	18.27	2381	16.19	2808	1.16	2833	0.28					
ORBS	8196	6874	16.13	6933	15.41	7074	13.69	8048	1.81	8166	0.37					
ORTZ	5085	3911	23.09	3918	22.95	3961	22.10	4941	2.83	5069	0.31					
ORUT	2153	1776	17.51	1784	17.14	1832	14.91	2136	0.79	2151	0.09					
OSAN	2084	1759	15.60	1810	13.15	1904	8.64	2068	0.77	2076	0.38					

Table 1 *continued*

1. Firm	2. Raw data	3. Huang and Stoll (HS)		4. HS plus minimum tick (HSMT)		5. HSMT plus price level (HSMTPL)		6. HSMTPL plus volatility (no MAD)		7. Final data set	
		Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers
OSCB	2777	2204	20.63	2212	20.35	2258	18.69	2728	1.76	2765	0.43
OSPW	1952	1639	16.03	1663	14.81	1724	11.68	1927	1.28	1938	0.72
OTAB	2600	2058	20.85	2069	20.42	2121	18.42	2577	0.88	2600	0.00
OTCO	2006	1706	14.96	1737	13.41	1818	9.37	1998	0.40	2001	0.25
OTSB	7259	6092	16.08	6182	14.84	6402	11.81	7175	1.16	7224	0.48
OVOD	5136	4266	16.94	4567	11.08	4739	7.73	5108	0.55	5125	0.21

However, for any implementations of the data filter in future research, the minimum tick size criterion would have to be more flexible in order to capture any drastic differences in the tick size. For example, if the minimum tick ranges between whole integers and 0.01, it is clear that every tick would need its own category. The above demonstrates that the tick rule is not arbitrary, yet prudence is required for future implementations of the algorithm in other settings.

Finally, we compare the normalised MAD (NMAD) value with the standardised price (see previous section) of the potential outliers, adopting a conservative approach in outlier detection. The latter is consistent with the findings of Barnett and Lewis,²⁶ hence capturing data that are long-tailed. Only those observations that are identified as outliers from both techniques are eventually discarded from the sample.

RESULTS AND ANALYSIS

One problem with UHFD filtering is that the actual 'clean' data set is not observable, and hence it is difficult to evaluate the efficacy of any filter. The method used here is to compare the results with those using the HS algorithm and also with the established level of outliers reported in the relevant literature.

For this reason, we apply the HS method to our data set. As two-way quotes in LIFFE equity options are not continuous, the second part of the algorithm cannot be applied directly; however, we replicate the HS method for trades. The results are presented in Table 1, Column 3. In Table 1 we also demonstrate the appropriateness of the data-cleaning steps identified in Figure 2. Thus, columns 4–6 show

Table 2: The evolution of the data filter (bids only)

1. Firm	2. Raw data	3. Huang and Stoll (HS)		4. HS plus Minimum tick (HSMT)		5. HSMT plus price level (HSMTPL)		6. HSMTPL plus volatility (no MAD)		7. Final data set	
		Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers
OAMM	1 721 053	1 709 307	0.68	1 713 512	0.44	1 715 654	0.31	1 719 698	0.08	1 720 662	0.02
OAWS	886 596	880 502	0.69	884 207	0.27	885 677	0.10	886 414	0.02	886 473	0.01
OAZA	7 471 164	7 357 649	1.52	7 372 151	1.33	7 380 560	1.21	7 451 886	0.26	7 469 087	0.03
OBBL	4 660 639	4 626 376	0.74	4 645 320	0.33	4 647 362	0.28	4 659 253	0.03	4 659 754	0.02
OBLT	1 355 383	1 347 188	0.60	1 352 822	0.19	1 353 285	0.15	1 354 878	0.04	1 355 181	0.01
OBOT	744 089	732 185	1.60	740 743	0.45	742 522	0.21	743 672	0.06	743 850	0.03
OBP	6 014 104	5 963 291	0.84	5 986 292	0.46	5 990 054	0.40	6 009 328	0.08	6 012 117	0.03
OBSK	876 706	865 696	1.26	872 846	0.44	874 641	0.24	876 118	0.07	876 349	0.04
OBTG	1 747 487	1 710 922	2.09	1 735 662	0.68	1 738 069	0.54	1 745 865	0.09	1 746 517	0.06
OCPG	152 946	149 475	2.27	151 796	0.75	152 191	0.49	152 740	0.13	152 840	0.07
OCUA	2 527 120	2 490 906	1.43	2 506 050	0.83	2 508 111	0.75	2 525 047	0.08	2 526 538	0.02
OEMG	958 206	952 253	0.62	954 898	0.35	955 810	0.25	957 542	0.07	958 043	0.02
OGNS	2 615 968	2 576 539	1.51	2 596 717	0.74	2 601 449	0.56	2 613 681	0.09	2 615 341	0.02
OGXO	4 030 726	3 984 811	1.14	4 003 264	0.68	4 008 008	0.56	4 025 745	0.12	4 029 677	0.03
OHSB	2 182 076	2 153 499	1.31	2 170 768	0.52	2 173 186	0.41	2 180 053	0.09	2 181 354	0.03
OKGF	360 296	354 546	1.60	358 529	0.49	359 184	0.31	359 909	0.11	360 093	0.06
OLS	1 695 452	1 678 717	0.99	1 684 444	0.65	1 688 748	0.40	1 693 866	0.09	1 695 104	0.02
OPRU	3 043 850	3 005 650	1.25	3 021 586	0.73	3 024 835	0.62	3 040 647	0.11	3 042 754	0.04
ORBS	7 732 452	7 672 142	0.78	7 698 984	0.43	7 705 610	0.35	7 728 165	0.06	7 730 868	0.02
ORTZ	3 136 347	3 115 887	0.65	3 124 102	0.39	3 127 436	0.28	3 133 585	0.09	3 135 722	0.02
ORUT	1 540 332	1 529 007	0.74	1 535 851	0.29	1 537 508	0.18	1 539 601	0.05	1 540 076	0.02
OSAN	1 112 881	1 104 577	0.75	1 111 750	0.10	1 112 257	0.06	1 112 651	0.02	1 112 737	0.01

Table 2 continued

1. Firm	2. Raw data	3. Huang and Stoll (HS)		4. HS plus minimum tick (HSMT)		5. HSMT plus price level (HSMTPL)		6. HSMTPL plus volatility (no MAD)		7. Final data set	
		Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers
OSCB	2030023	2015651	0.71	2020297	0.48	2024306	0.28	2028485	0.08	2029404	0.03
OSPW	367927	357495	2.84	367007	0.25	367337	0.16	367669	0.07	367818	0.03
OTAB	2282656	2259677	1.01	2271516	0.49	2275591	0.31	2280067	0.11	2281960	0.03
OTCO	802936	796684	0.78	801757	0.15	802219	0.09	802833	0.01	802862	0.01
OTSB	2127955	2101962	1.22	2115508	0.58	2117528	0.49	2126293	0.08	2127248	0.03
OVOD	1319193	1273073	3.50	1300781	1.40	1305440	1.04	1317544	0.13	1318383	0.06

the evolution of the data-cleaning filter when adding the minimum tick, the price level and the daily price level criteria, respectively. Column 7 shows the final 'clean' data set. Results are presented for bids (Table 2) and asks (Table 3) for comparison.

Table 1 strongly suggests that the HS algorithm would lead to 'overscrubbing' for equity options trades UHFD. According to HS, data identified as outliers range from 13.82 to 24.33 per cent, with an average of 18 per cent. The latter implies that the HS algorithm is overly conservative for high-priced assets. Hence, Figure 3 shows that as price level increases, the percentage of data classed by the HS algorithm as outliers also tends to increase. Further analysis in Table 4 reveals that the correlation coefficient between price level and the percentage of outliers from the HS algorithm across the data set is 64 per cent.

Columns 4-7 in Table 1 demonstrate the evolution of the data-cleaning filter.⁴⁰ It is shown that with the inclusion of the minimum tick effect, the overall proportion defined as outliers falls. The same applies for the price level effect. Column 6 shows that adjusting for the daily volatility of prices may have substantial effects on the distribution of outliers. The latter is an expected and well-documented finding in the literature.⁴¹ Finally, Column 7 shows that by adopting the robust MAD criterion, the percentage of data defined as outliers falls significantly. The latter is a desirable end result, as it demonstrates a high level of consistency with previous research (see below).

Table 4 shows the effect of each data-cleaning step in relation to each firm's price level.⁴² We show that when we control for the minimum tick size and price level differences, the

Table 3: The evolution of the data filter (asks only)

1. Firm	2. Raw data	3. Huang and Sroll (HS)		4. HS plus minimum tick (HSMT)		5. HSMT plus price level (HSMTPL)		6. HSMTPL plus volatility (no MAD)		7. Final data set	
		Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers
OAAM	1 562 899	1 553 738	0.59	1 555 669	0.46	1 557 954	0.32	1 560 888	0.13	1 561 675	0.08
OAWS	1 012 847	1 005 730	0.70	1 007 025	0.57	1 008 624	0.42	1 010 934	0.19	1 011 371	0.15
OAZA	7 528 893	7 448 668	1.07	7 453 081	1.01	7 459 620	0.92	7 486 443	0.56	7 512 104	0.22
OBBL	4 965 868	4 940 774	0.51	4 948 601	0.35	4 951 156	0.30	4 953 916	0.24	4 954 652	0.23
OBLT	1 353 797	1 344 590	0.68	1 346 896	0.51	1 347 505	0.46	1 348 797	0.37	1 351 372	0.18
OBOT	734 019	724 005	1.36	728 635	0.73	730 054	0.54	731 242	0.38	732 755	0.17
OBP	6 244 652	6 189 744	0.88	6 209 291	0.57	6 212 987	0.51	6 222 324	0.36	6 228 060	0.27
OBSK	918 345	907 581	1.17	913 473	0.53	915 361	0.32	916 689	0.18	916 911	0.16
OBTG	1 921 538	1 882 393	2.04	1 906 545	0.78	1 909 513	0.63	1 911 512	0.52	1 912 029	0.49
OCPG	152 387	151 094	0.85	151 363	0.67	151 662	0.48	152 218	0.11	152 296	0.06
OCUA	2 649 015	2 616 227	1.24	2 625 398	0.89	2 627 563	0.81	2 630 953	0.68	2 632 962	0.61
OEMG	1 250 976	1 246 414	0.36	1 246 928	0.32	1 248 007	0.24	1 249 922	0.08	1 250 549	0.03
OGNS	2 663 394	2 622 631	1.53	2 639 650	0.89	2 645 503	0.67	2 649 724	0.51	2 651 931	0.43
OGXO	4 045 342	4 010 125	0.87	4 019 012	0.65	4 022 210	0.57	4 028 465	0.42	4 037 343	0.20
OHSB	2 370 726	2 335 916	1.47	2 353 683	0.72	2 356 081	0.62	2 359 401	0.48	2 361 826	0.38
OKGF	357 682	354 305	0.94	355 914	0.49	356 497	0.33	357 189	0.14	357 374	0.09
OLS	1 833 009	1 820 526	0.68	1 821 887	0.61	1 826 037	0.38	1 829 063	0.22	1 831 678	0.07
OPRU	3 325 794	3 294 275	0.95	3 302 686	0.69	3 305 674	0.60	3 310 887	0.45	3 313 730	0.36
ORBS	7 905 659	7 843 259	0.79	7 859 907	0.58	7 868 473	0.47	7 881 166	0.31	7 889 313	0.21
ORTZ	3 053 503	3 033 212	0.66	3 037 250	0.53	3 041 245	0.40	3 047 483	0.20	3 049 984	0.12
ORUT	1 848 108	1 837 700	0.56	1 843 268	0.26	1 844 702	0.18	1 845 907	0.12	1 846 390	0.09
OSAN	1 071 040	1 063 588	0.70	1 066 644	0.41	1 067 785	0.30	1 068 579	0.23	1 068 917	0.20

Table 3 continued

1. Firm	2. Raw data	3. Huang and Stoll (HS)		4. HS plus minimum tick (HSMT)		5. HSMT plus price level (HSMTPL)		6. HSMTPL plus volatility (no MAD)		7. Final data set	
		Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers	Obs. retained	% Outliers
OSCB	2073 844	2063 642	0.49	2065 188	0.42	2068 374	0.26	2070 805	0.15	2072 397	0.07
OSPW	373 189	366 167	1.88	371 659	0.41	372 128	0.28	372 653	0.14	372 862	0.09
OTAB	2240 305	2223 979	0.73	2228 431	0.53	2231 619	0.39	2235 788	0.20	2238 626	0.07
OTCO	863 079	857 281	0.67	860 449	0.30	861 100	0.23	861 429	0.19	861 669	0.16
OTSB	2139 699	2117 253	1.05	2125 656	0.66	2127 591	0.57	2130 099	0.45	2131 850	0.37
OVOD	1496 422	1442 081	3.63	1472 662	1.59	1478 850	1.17	1482 909	0.90	1484 743	0.78

correlation coefficients of the price level and the proportion of outliers fall to -4 and 1 per cent, respectively. We view the latter as a significant finding, as it demonstrates a desirable property of the data filter. Finally, when the MAD criterion is applied, the correlation coefficient is 6 per cent.

Tables 2 and 3 show the application of the data filter for bids and asks, respectively. It is clear that as the frequency of quotes is relatively higher, the HS algorithm is much less conservative. The percentage of outliers from the HS algorithm applied to quotes ranges from 0.60 to 3.50 per cent. In the last columns of Table 2 and 3, the percentage of outliers for our data filter ranges between 0.01 and 0.07 per cent, which is consistent with prior literature (see below).

Dacorogna *et al*² note that for foreign exchange data, the percentage of outliers is between 0.11 and 0.81 per cent. Dacorogna *et al*³ report the outlier rates for a number of different financial markets. It is worth noting that the data filter employed for the data cleaning in the above two articles is implemented by Olsen and Associates. In the latter article, from eight data samples six are found to have a percentage of outliers between 0.07 and 0.24 per cent. However, for the remaining two thinly traded assets, the percentage outlier rates are 1.14 and 7.59 per cent, signifying the possible downsides of ‘overscrubbing’.

Chordia *et al*²³ apply a bid-ask spread data selection model in US equities, effectively eliminating 0.02 per cent of the data. However, such an algorithm is less useful for securities traded in order-driven markets, as the bid-ask spread is not as appropriate for use in outlier detection.⁴³ Finally, Bessembinder¹⁵ applies an algorithm to NYSE and NASDAQ stock data

Figure 3: Average price level and the HS algorithm. The left scale refers to the average price level per asset. The right scale refers to the percentage of observations that are classed as outliers by the HS algorithm.

Table 4: Price level, minimum tick size and the evolution of the data filter

Name	Tick size	Price level	HS (%)	HSMT (%)	HSMTPL (%)	HSMTPL plus volatility (no MAD) (%)	Final (%)
OTCO	0.25	11.36	14.96	24.08	23.07	2.01	0.25
OSAN	0.25	10.86	15.60	18.93	14.66	0.58	0.38
OBTG	0.25	7.04	16.32	17.94	16.93	2.52	0.22
OBBL	0.25	16.36	16.35	15.14	11.69	0.81	0.38
OVOD	0.25	3.67	16.94	17.87	16.04	0.89	0.21
OAWS	0.25	13.85	20.25	14.99	11.61	0.99	0.29
OHSB	0.5	21.43	13.82	17.03	14.60	0.97	0.29
OKGF	0.5	20.32	14.98	15.68	12.52	0.81	0.12
OGNS	0.5	19.55	15.75	11.70	7.64	0.47	0.35
OBOT	0.5	22.82	15.98	19.65	16.31	1.26	0.27
OSPW	0.5	14.93	16.03	17.39	13.77	0.91	0.72
OTSB	0.5	24.08	16.08	20.42	19.72	1.44	0.48
ORBS	0.5	43.39	16.13	14.47	12.05	1.12	0.37
OBSK	0.5	18.87	16.70	17.60	15.44	2.09	0.29
ORUT	0.5	29.59	17.51	12.33	9.23	0.36	0.09
OBP	0.5	32.06	17.72	14.36	11.98	0.66	0.20

Table 4 *continued*

Name	Tick size	Price level	HS (%)	HSMT (%)	HSMTPL (%)	HSMTPL plus volatility (no MAD) (%)	Final (%)
OGXO	0.5	38.02	17.97	20.30	18.85	1.45	0.37
OCUA	0.5	18.60	18.21	18.27	16.19	1.16	0.16
OBLT	0.5	31.91	18.22	15.41	13.69	1.81	0.27
OAZA	0.5	72.89	18.23	22.95	22.10	2.83	0.39
OPRU	0.5	28.86	18.97	17.14	14.91	0.79	0.28
OCPG	0.5	18.35	20.09	13.15	8.64	0.77	0.25
OEMG	0.5	55.97	20.58	20.35	18.69	1.76	0.31
OLS	0.5	42.68	20.60	14.81	11.68	1.28	0.35
OSCB	0.5	38.22	20.63	20.42	18.42	0.88	0.43
OTAB	0.5	33.30	20.85	13.41	9.37	0.40	0.00
ORTZ	0.5	70.97	23.09	14.84	11.81	1.16	0.31
OAAM	0.5	60.15	24.33	11.08	7.73	0.55	0.25
Correlation coefficient			0.64	−0.04	0.01	0.18	0.06

similar to the selection model originated by HS, and reports that 4.1 per cent of trades and 1.1 per cent of quotes were classified as outliers.

This prior evidence suggests that data selection models typically should not reject more than 1 per cent of the overall number of trades and quotes, which indicates that the algorithm developed here is operating within sensible bounds for options contracts.

CONCLUSION

This article develops a new algorithm for data cleaning in UHFD. Although there is substantial published research on market microstructure issues, we identify a gap in the literature on data cleaning and filtering for UHFD. The main objective of this study is to discuss relevant data

filters with an intention to evaluate the validity of the filters. We also find that the most popular method of outlier selection in the literature⁵ is rather inappropriate for contracts with inbuilt time characteristics or very low prices, such as equity options.

We develop a data filtering technique that takes full consideration of a wider range of issues than discussed in prior studies. This new data-cleaning method is an amalgam of the structural characteristics of options contracts and of the statistical properties of the sample. A multiple-stage algorithm is developed and implemented in UHFD with the robust MAD method to validate the first (market microstructure) part of the algorithm.

The validity of the model is justified not only on statistical grounds (*ex-ante*), but also, *ex-post*, the model is found to perform in a

manner consistent with many strands of previous literature. As this is a unique study on the case of options, the comparability of the results of this algorithm with earlier studies uses other asset classes.

The findings suggest that the algorithms developed can also be applied to other types of derivative contracts with very few alterations, subject to controlling for the effect of the minimum tick size. To our knowledge, this is the first study that offers a data filter that can be implemented in a range of asset classes taking full account of the characteristics of the data.

REFERENCES AND NOTES

- 1 Engle, R.F. (2000) The econometrics of ultra-high-frequency data. *Econometrica* 68(1): 1–22.
- 2 Dacorogna, M.M., Müller, U.A., Jost, C., Pictet, O.V. and Ward, J.R. (1995) Heterogeneous real-time trading strategies in the foreign exchange market. *European Journal of Finance* 1: 383–403.
- 3 Dacorogna, M.M., Gencay, R., Müller, U., Olsen, R.B. and Pictet, O.V. (2001) *An Introduction To High-Frequency Finance*. San Diego, CA: Academic Press.
- 4 Falkenberry, T.N. (2002) High Frequency Data Filtering. Technical Report, Tick Data.
- 5 Huang, R.D. and Stoll, H.R. (1996) Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE. *Journal of Financial Economics* 41(3): 313–357.
- 6 Muller, U. (2001) The Olsen Filter for Data in Finance. Zurich, Switzerland. Olsen and Associates Working Paper UAM, 27 April 1999.
- 7 In the latter case, these entries always appear in the data file.
- 8 It is worth noting, however, that only computer errors that are caused by human intervention (for example, typing errors) affect outliers.
- 9 Chung, K., Van Ness, B. and Van Ness, R. (2004) Trading costs and quote clustering on the NYSE and NASDAQ after decimalization. *Journal of Financial Research* 27(3): 309–328.
- 10 Chung, K.H., Chuwongnant, C. and McCormick, T.D. (2004) Order preferencing and market quality on NASDAQ before and after decimalization. *Journal of Financial Economics* 71(3): 581–612.
- 11 HS delete observations with spreads that are negative or larger than \$4. However, while the spread criterion can be applied in continuous quote markets like NASDAQ, it will lead to stale pricing and non-synchronous data problems in markets with no obligation for continuous quotes.
- 12 Uniquely in high frequency finance, there is a departure from using fixed-interval data to using unequally spaced data. This implies that the event is now of more importance than the time interval during which it occurred, dictating the recording of an observation (see Goodhart and O'Hara¹³ and Engle and Russell¹⁴).
- 13 Goodhart, C.A.E. and O'Hara, M. (1997) High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance* 4(2–3): 73–114.
- 14 Engle, R.F. and Russell, J.R. (2004) Analysis of High Frequency Financial Data. Chicago, USA. University of Chicago Working Paper.
- 15 Bessembinder, H. (1997) The degree of price resolution and equity trading costs. *Journal of Financial Economics* 45(1): 9–34.
- 16 Chung, K.H., Van Ness, B.F. and Van Ness, R.A. (2002) Spreads, depths, and quote clustering on the NYSE and NASDAQ: Evidence after the 1997 securities and exchange commission rule changes. *Financial Review* 37(4): 481–505.
- 17 Benston, G.J. and Harland, J.H. (2007) Did NASDAQ market makers successfully collude to increase spreads? A re-examination of evidence from stocks that moved from NASDAQ to the New York or American Stock Exchanges. London, UK. Financial Markets Group, FMG Special Papers. sp170.
- 18 Cooney, J., Van Ness, B.F. and Van Ness, R.A. (2003) Do investors prefer even-eighth prices? Evidence from NYSE limit orders. *Journal of Banking & Finance* 27(4): 719–748.
- 19 See also Bessembinder,¹⁵ Chung *et al.*,⁹ Chung *et al.*¹⁰ and Chung *et al.*¹⁶ The selection of \$4 as a spread measure is not justified by the authors. In addition, subsequent studies use a selection of different benchmark spreads (for example, Chung *et al.*,⁹ use \$5). The latter reflects the subjectivity of this criterion.
- 20 It is very surprising that HS do not mention an absolute-returns measure, thus it is plausible that this point has been unintentionally omitted from their published article. Some literature has also made the supposition that HS failed to model absolute returns (see Chung *et al.*,⁹ Chung *et al.*¹⁰ and Chung *et al.*¹⁶).
- 21 Leung, C.K.-S., Thulasiram, R.K. and Bondarenko, D.A. (2006) An efficient system for detecting outliers from financial time series. In: D. Bell and J. Hong (eds.) *Flexible and Efficient Information Handling*. Heidelberg, Germany: Springer, p. 4026/2006.

- 22 Brownlees, C.T. and Gallo, G.M. (2006) Financial Econometric Analysis at Ultra-High Frequency: Data Handling Concerns. Università degli Studi di Firenze Dipartimento di Statistica 'Giuseppe Parenti'. Working Papers (2006/03).
- 23 Chordia, T., Roll, R. and Subrahmanyam, A. (2001) Market liquidity and trading activity. *Journal of Finance* 56(2): 501–530.
- 24 Defined as the difference between the execution price and the quote midpoint.
- 25 This is true unless the algorithm makes two passes through the data. HS give no indication that their algorithm has multiple iterations.
- 26 Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*. Chichester, UK: John Wiley & Sons.
- 27 ap Gwilym, O. and Sutcliffe, C. (2001) Problems encountered when using high frequency financial market data: Suggested solutions. *Journal of Financial Management & Analysis* 14(1): 38–51.
- 28 Fox, J. (2008) *A Mathematical Primer for Social Statistics*. Los Angeles, CA: Sage Publications.
- 29 Hellerstein, J.M. (2008) *Quantitative Data Cleaning for Large Databases. Report for United Nations Economic Commission for Europe*. Berkley, CA: EECS Computer Science Division.
- 30 Hubert, M., Pison, G., Struyf, A. and Aelst, S.V. (2004) *Theory and Applications of Recent Robust Methods*. Basel, Switzerland: Birkhauser Verlag AG.
- 31 This technique is also referred to as Hampel X84 (see Hellerstein²⁹). A value is standardised when we deduct the mean value and divide by the standard deviation. A standardised value follows a normal distribution.
- 32 ap Gwilym, O. and Sutcliffe, C. (1999) *High Frequency Financial Market Data: Sources, Applications and Market Microstructure*. London: Risk Books.
- 33 Thirty-one assets were identified; however, three assets were dropped from the sample owing to price distortions.
- 34 Sheikh, A.M. and Ronn (1994) A characterization of the daily and intraday behaviour of returns on options. *Journal of Finance* 49(3): 557–579.
- 35 ap Gwilym, O., Clare, A. and Thomas, S. (1998) The bid-ask spread on stock index options: An ordered probit analysis. *Journal of Futures Markets* 18(4): 467–485.
- 36 Bollerslev, T. and Melvin, M. (1994) Bid-ask spread and volatility in the foreign exchange market: An empirical analysis. *Journal of International Economics* 36(3–4): 355–372.
- 37 This number reflects the number of combinations found in the data and not the potential number, which is much higher.
- 38 Hameed, A. and Terry, E. (1998) The effect of tick size on price clustering and trading volume. *Journal of Business, Finance & Accounting* 25(7–8): 849–867.
- 39 The following dates are discarded: 13/01/05, 09/08/05 and 22/09/05.
- 40 In column 4, we apply the price level algorithm accounting for differences in returns. In Column 6 we further enhance the algorithm by applying the average daily range of prices (volatility measure).
- 41 Gutierrez, J.M.P. and Gregori, J.F. (2008) Clustering Techniques Applied to Outlier Detection of Financial Market Series Using a Moving Window Filtering Algorithm. European Central Bank Working Paper No. 948.
- 42 In order to conserve space, we present the results for trades only.
- 43 In quote-driven markets there are always active bid and ask quotes.