IMPROVED ALGORITHM FOR CLEANING HIGH FREQUENCY DATA: AN ANALYSIS OF FOREIGN CURRENCY

Nirodha I. Jayawardena*, Jason West, Bin Li, Neda Todorova

Abstract

High-frequency data are notorious for their noise and asynchrony, which may bias or contaminate the empirical analysis of prices and returns. In this study, we develop a novel data filtering approach that simultaneously addresses volatility clustering and irregular spacing, which are inherent characteristics of high-frequency data. Using high frequency currency data collected at five-minute intervals, we find the presence of vast microstructure noise coupled with random volatility clusters, and observe an extremely non-Gaussian distribution of returns. To process non-Gaussian high-frequency data for time series modelling, we propose two efficient and robust standardisation methods that cater for volatility clusters, which clean the data and achieve near-normal distributions. We show that the filtering process efficiently cleans high-frequency data for use in empirical settings while retaining the underlying distributional properties.

Keywords: High Frequency, Market Microstructure, Realized Volatility, Volatility Clusters

** Department of Accounting, Finance and Economics, Griffith Business School, Griffith University, Nathan, Queensland 4111, Australia*

1. Introduction

High frequency financial data (or tick-by-tick data) are observations measured at a fine time scale. The analysis of high frequency data has gained great traction in recent years. The rapid development of computing technology has made it possible to capture, transfer, store and process huge volumes of high frequency data efficiently for use in algorithmic trading strategies. As such, the demand for high-quality high frequency data (intra-day data, intra-hour data or intra-minute data) is soaring. When prices are sampled at much finer intervals, the microstructure issues become more pronounced with increasing data volume and frequency. Finely sampled data can thus become unusable (Falkenberry, 2002). A clean data set is critical for empirical data manipulation and research. This demands an efficient and robust, yet simple data filter to properly recognise erroneous and misleading data points.

On the contrary, the most difficult aspect of cleaning intraday data is the inability to universally define what is “unclean” (Dacarogna et al., 2001). Moreover, filtering of outliers always involves a tradeoff: as the data are filtered too tightly there is a possibility of ‘over scrubbing’ data, leading changes to its statistical properties and taking reality out of the data; on the other hand, if the data is too loosely filtered there is a possibility of ‘under scrubbing’ making the data still unusable for testing. Hence, the prime objective in developing an efficient algorithm should be to manage over scrubbing/under scrubbing and produce a time series that removes false information without distorting the real dynamics of the data set.

In this paper, we propose a novel data filtering approach that simultaneously addresses several inherent characteristics of high frequency data such as volatility clustering and irregular spacing. The treatment of these features has not received extensive treatment in the existing literature, so through this analysis we propose a simple yet robust filtering solution. We find the presence of vast microstructure noise and large deviations from the Gaussian assumption in the original data set. We then present two standardisation methods that can be applied to achieve virtual normality in the filtered data. Specifically, we employ a univariate standardisation approach that uses realised volatilities and a GARCH-based standardisation, which both account for volatility clustering and irregular observation spacing. We show that these standardisation methods yield almost Gaussian-distributional outcomes that are representative of the underlying dynamics for use in further empirical analysis. Our results also allow for the further development of data filters to cater for other unique characteristics associated with high frequency data such as the information embedded in traded volumes, bid-offer spreads and other intra-day factors.

The paper is organised as follows. The next section presents the methodology of the study. Subsequent sections present and discuss our empirical results. The final section concludes.

2. Methodology

2.1. GARCH Framework

We obtain foreign exchange data for AUD/USD, EUR/USD and JPY /USD currency pairs at 5-minute intervals from 2 January to 30 April 2013 from the Securities Industry Research Centre of Australia (SIRCA) database. In the first part of the analysis, we deseasonalize the series using the approach proposed by Taylor and Xu (1997).

Let $r_{d,n}$ be the n^{th} intraday return on day d, and suppose that we have D days and N intraday periods. Then the seasonal variance estimate is given by:

$$S_n^2 = \frac{1}{D}\sum_{d=1}^{D} r_{d,n}^2 \quad (n = 1, \dots\dots\dots N)(1)$$

Using the seasonal terms we filter the returns:

$$X_t = \frac{r_{d,n}}{s_{d,n}}, \qquad (2)$$

and estimate the conditional volatilities of the filtered intraday data using a GARCH (1,1) process as the following system of equations[1]:

$$X_t = \mu + \varepsilon_t \qquad (3)$$

$$\sigma_t^2 = \alpha + \lambda_i \varepsilon_{t-1}^2 + \beta_j \sigma_{t-1}^2 \qquad (4)$$

where X_t denotes filtered index returns, μ denotes the average return, σ_t^2 denotes the conditional variance at time t, and $\varepsilon_t \sim N\ (0, \sigma_t^2), \alpha > 0, \beta_j \geq 0$ and $\lambda_i \geq 0$. Our approach is not unlike the standard GARCH approach (Engle, 1982; Bollerslev, 1986).

2.2. Cluster Detection Algorithm

We introduce a cluster detection algorithm, originally proposed by Laurini (2004). The algorithm advances a two-threshold detection method and defines an independent extreme return cluster in terms of movements in returns and volatility. This approach seems to be appealing as it is motivated by the empirical findings exhibited by financial returns. For instance, when a cluster of extreme observations is ended, the volatility drops down to a suitable level. This marks the end of a period of influence of a certain type of independent information and causes a cluster to terminate. In their study of two-threshold method, a time series of returns and a corresponding time-series of volatility for each index are needed. The two-threshold method of detecting clusters is as follows.

Let $\{X_t\}$ denote a return series and $\{\sigma_t\}$ denote the conditional volatility series. Let u denote the threshold for the $\{X_t\}$ process and c denote the threshold for the $\{\sigma_t\}$ process and m denote the length of a run (cluster size). The following conditions yield the cut off point t* that defines the endpoint of an extreme return cluster as:

if $X_1 > u$,
$Max(X_2 \dots\dots, X_m) \leq u$,
$Min(\sigma_2 \dots\dots, \sigma_{m-1}) > c$, and $\sigma_m \leq c$,
then $t^* = m$.

2.3. Standardisation of Returns

It is well known that high frequency returns are generally leptokurtic (fat-tailed) relative to a Gaussian distribution. The next step in the construction of the filter focuses on standardisation methods that normalise the returns observed in the empirical data by using standard stochastic volatility models. We introduce the two main standardisation methods as a univariate method which refer to as the Sigma (RV) standardisation and a GARCH-based method which we refer to as the Sigma(GARCH) standardisation.

The main purpose of employing standardization is to reduce the impact of '*overscrubbing*' of data when filtering the data. As depicted in the Q-Q plots in Figure 3 and the summary statistic tables, the impact of '*fat tails*' is significantly reduced after standardization, making it clear to distinguish extreme observations without distorting the statistical properties. Furthermore, it is evident in Table 1 that after employing the standardization, the proportion of outliers decreases considerably.

2.4 Data Filtering Framework

We develop an efficient data filter to capture the unique characteristic of leptokurtic return distributions of foreign currency pairs after normalisation.

[1] For a justification of using a GARCH(1,1) model, the reader is referred to Hansen and Lunde (2005) who use the DM-USD exchange rate data and compare the forecasting ability of 330 GARCH-type models. They find that none of the models can consistently outperform GARCH (1,1).

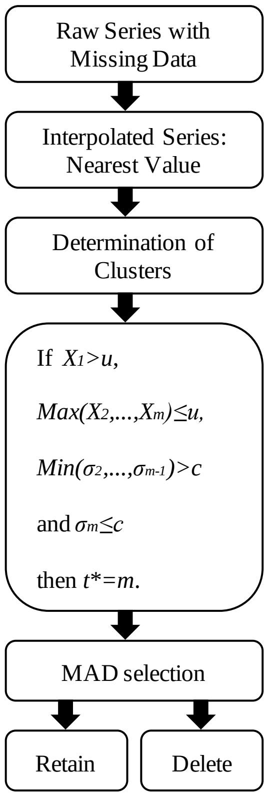

Raw Series with Missing Data
Interpolated Series: Nearest Value
Determination of Clusters
If $X_1>u$,
$Max(X_2,...,X_m)\leq u$,
$Min(\sigma_2,...,\sigma_{m-1})>c$
and $\sigma_m\leq c$
then $t^*=m$.
MAD selection
Retain
Delete

Figure 1. Selection and filtering process applied to currency pairs using Sigma(RV) and Sigma(GARCH) standardisation methods.

We first identify outliers as data points that are three standard deviations away from the mean calculated within *k*-equal size blocks. However, the number of outliers detected is quite sensitive to the parameter *k* chosen and such an algorithm may lead to *overscrubbing* of data[2]. In order to eliminate these limitations, we introduce Laurini's (2004) approach, as depicted in Figure 1 to detect cluster size (*m*) together with median absolute deviation (MAD) test and standardisation approaches to reduce the impact of fat tails[3].

2.4 Univariate Standardisation by Realised Volatility – Sigma (RV)

The univariate return series are naturally decomposed as $r_t = \sigma_t \varepsilon_t$, where $\varepsilon_t \sim (0,1)$ and σ_t is the time *t* conditional standard deviation. By rearranging this decomposition, we obtain the σ-standardized return,

$$\varepsilon_t = \frac{r_t}{\sigma_t}. \qquad (3)$$

However, the σ_t is unknown and must be estimated. As shown in Anderson et al. (2000) we estimate the *ex-post* volatility over a day by summing up the high frequency returns within a day. We refer to this as the Sigma(RV) approach. Considering the 24-hour trading day, the daily variances can be estimated by simply summing the 288 squared returns within each day. That is:

$$\sigma^2{}_{AUD_t}(RV) = \sum\nolimits_{j=1,\ldots\ldots288}(r^2_{AUD})$$
(4)

$$\sigma^2{}_{EUR_t}(RV) = \sum\nolimits_{j=1,\ldots\ldots288}(r^2_{EUR})$$
(5)

$$\sigma^2{}_{JPY_t}(RV) = \sum\nolimits_{j=1,\ldots\ldots288}(r^2_{JPY})$$
(6)

where $t = 1,2,\ldots\ldots288$, and RV represents realized volatility[4].

[2] For the sake of clarity, we do not present the results of parameter sensitivity of equal cluster sizes in this paper. However, as the parameter *k* varies between 5 and 40 the number of identified outliers range from 16% to 23%.. Refer to Brownlees (2006) for similar study in high frequency data cleaning using equal cluster sizes.

[3] 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 (Fox, 2008).

[4] The choice of 5-minute returns is motivated by examining the market structure noise graph in Figure 1, which indicates that the RV is a function of sampling frequency. The benefit of using this approach is that the observed output remains stable as the sampling frequency increases up to 5-minute returns.

2.5 Standardisation by GARCH (1,1) – Sigma(GARCH)

In order to capture the conditional temporal dependencies in σ_t, the most commonly used specification is the simple univariate GARCH (1,1) model:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2. \qquad (7)$$

Both the ARCH and GARCH coefficients are significant indicating the persistence in volatility for both of the series. This is the Sigma(GARCH) approach and its main benefits are that it can efficiently account for volatility clustering and the time varying effects of volatility.

3. Results

3.1. Distributional Properties and Stylised Factors

First, we summarise the distributions of the raw daily AUD/USD, EUR/USD and JPY/USD exchange rate return series. These currency pairs are chosen because they represent three out of the top five highest volume traded currencies in the global foreign exchange market (BIS, 2013; The Economist, 2013).

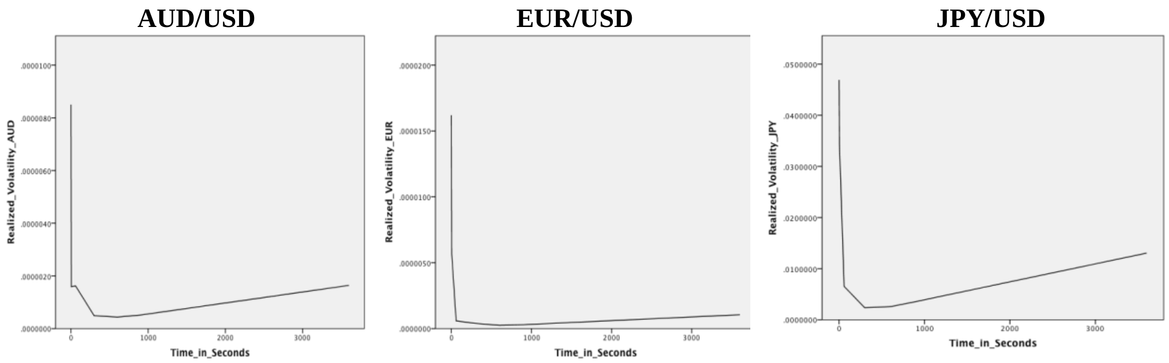


Figure 2. Plot of realised volatility as a function of the sampling interval for AUD/USD, EUR/USD and JPY/USD currency pairs at 5-minute intervals, 2 Jan-30 Apr 2013

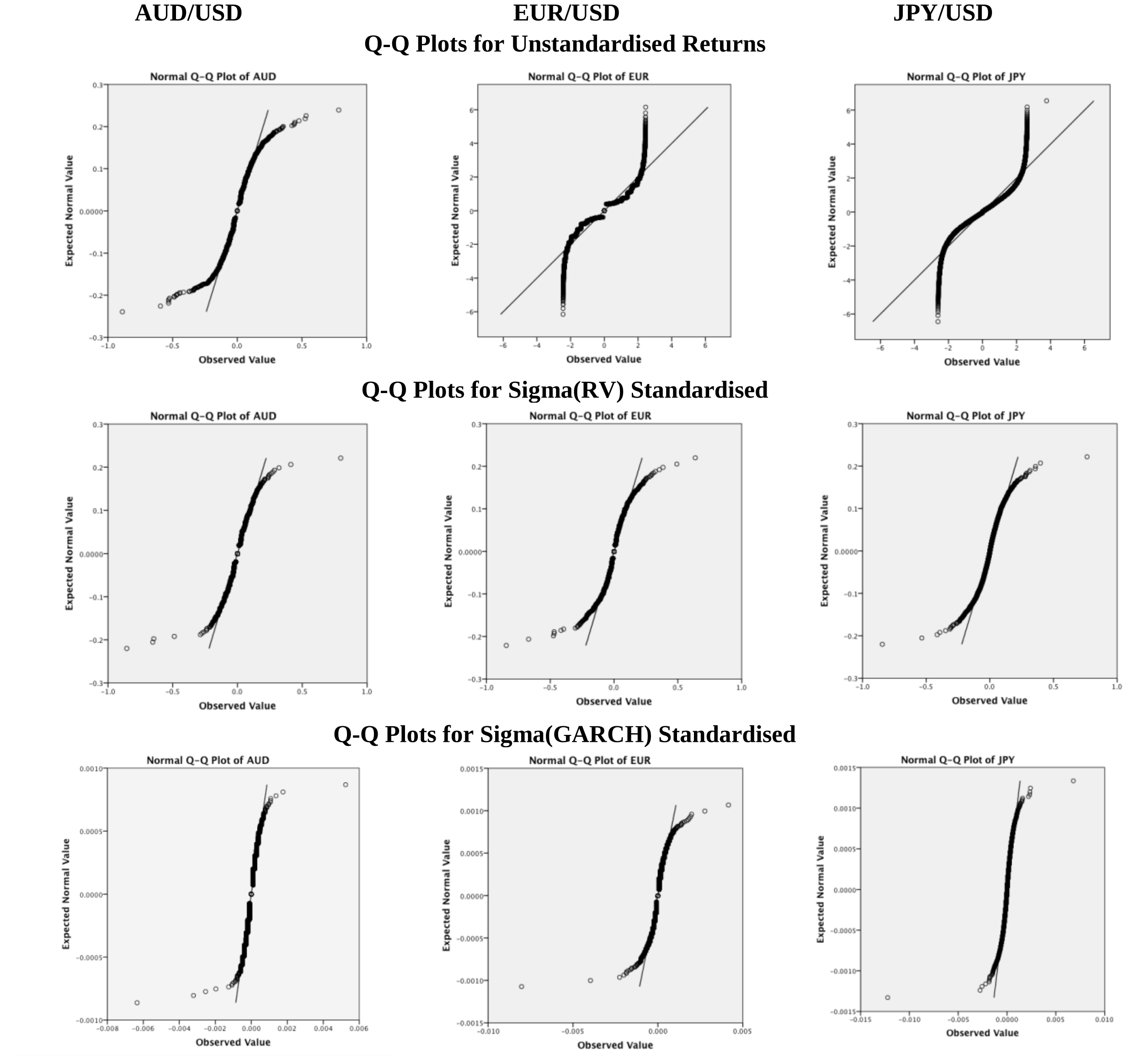
A shorter sampling time interval is equivalent to a higher the sampling frequency. Figure 2 illustrates that when the sampling interval decreases, the realised volatility (RV) increases, which suggests the presence of the microstructure noise. Hansen and Lunde (2006) argue that if there is no noise in the data, the RV should be to stabilise as sampling frequency increases. The presence of noise in the observed exchange rate data provides an explanation for the 'explosion' of RV near the origin.

Table 1. Descriptive statistics for AUD/USD, EUR/USD and JPY /USD currency pairs at 5-minute intervals, 2 Jan-30 Apr 2013. * represents statistical significance at the 5% level.

	Mean	**Std. Dev**	**Skewness**	**Kurtosis**	**JB statistic**
Unstandardised Returns					
AUD/USD	0.0000	0.0002	-2.0010	141.9341	5.0390e+06*
EUR/USD	0.0000	0.0003	-3.4031	118.2775	3.4772e+06*
JPY/USD	0.0000	0.0004	-5.4710	243.7400	1.5073e+07*
Sigma(RV) Standardised Returns					
AUD/USD	0.0006	0.0593	-0.5500	21.6350	2.4180e+03*
EUR/USD	-0.0006	0.0593	-0.7220	20.4140	6.5968e+03*
JPY/USD	0.0009	0.0594	-0.3900	18.6240	4.8601e+03*
Sigma(GARCH) Standardised Returns					
AUD/USD	-0.0014	1.4228	-0.0357	11.8822	1.2540e+03*
EUR/USD	0.0014	1.5151	-0.0052	13.4515	2.5126e+03*
JPY/USD	0.0466	1.6010	-0.0239	13.1023	1.7542e+03*

The summary statistics in Table 1 are consistent with results from existing literature. The statistics confirm that the sample skewness is near zero implying symmetry but the sample kurtosis is well in excess of that expected under a Gaussian assumption. Non-normality is confirmed by the Jarque-Bera goodness of fit statistics rejecting the null hypothesis of normality. It is important to note that even after implementing the standardisation method, the skewness is dampened but the excess kurtosis remains high.

Figure 3 presents the Quantile-Quantile (Q-Q) plots for the raw marginal distributions and the two standardised marginal distributions. The s-shaped profile for each of the raw currency pairs visually indicates that these returns are symmetric but fat-tailed, relative to the normal distribution. When the data is standardised using the Sigma (RV) approach and the Sigma(GARCH) approach, the results become significantly less skewed and leptokurtic. The Q-Q plots in the middle panel of Figure 3 representing the Sigma(RV) method look radically different from those in the top panel. In particular, they appear almost linear indicating that a Gaussian distribution affords a close approximation to each of the marginal distributions. The diagnostic statistics in the second panel of Table 1 also confirms this result. Further improvements are observed using the Sigma(GARCH) approach. The distributions of the standardised daily returns are remarkably close to a standard normal distribution (near-zero mean, standard deviations close to unity, skewness coefficients close to zero and kurtosis closer to three).



AUD/USD EUR/USD JPY/USD

Q-Q Plots for Unstandardised Returns

Q-Q Plots for Sigma(RV) Standardised

Q-Q Plots for Sigma(GARCH) Standardised

Figure 3. Quantile plots for AUD/USD, EUR/USD and JPY /USD currency pairs at 5-minute intervals, 2 Jan-30 Apr 2013.

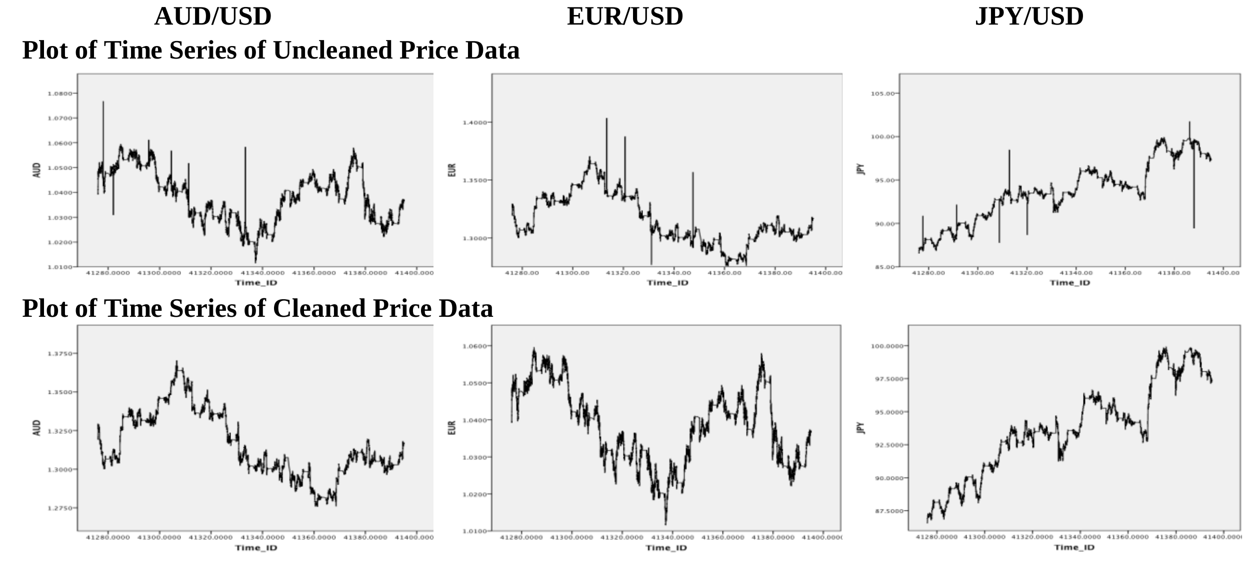
Table 2. Filtered data results using the Sigma(RV) and Sigma(GARCH) filters for AUD/USD, EUR/USD and JPY /USD currency pairs at 5-minute intervals, 2 Jan-30 Apr 2013.

	Raw Return Series	**Standardised Sigma(RV)**	**Standardised Sigma(GARCH)**
AUD/USD Series			
Original Sample	25329	25221	25329
Retained Observations	24417	24687	24911
Discarded Observations	912	534	418
% of Discarded Observations	3.6	2.1	1.6
% of Retained Observations	96.4	97.9	98.4
EUR/USD Series			
Original Sample	25329	25218	25329
Retained Observations	24113	24626	23898
Discarded Observations	1216	592	1431
% of Discarded Observations	4.8	2.3	5.6
% of Retained Observations	95.2	97.6	94.4
JPY/USD Series			
Original Sample	25329	25223	25329
Retained Observations	24341	24659	23773
Discarded Observations	988	564	1556
% of Discarded Observations	3.9	2.2	6.1
% of Retained Observations	96.1	97.7	93.9

Table 2 illustrates the results from the proposed data filter where the parameter *k* is determined by the algorithm suggested by Laurini and the observations are evaluated within each cluster based on their MAD[5]. This process combines both standardisation and volatility clustering properties to derive an efficient data-filtering algorithm. It is worth noting that the number of outliers identified is significantly reduced after standardisation. This reduction may be due to the reduction in the presence of fat tails (Table 1). However, even in the standardised series we observe the presence of volatility clusters.

To confirm the robustness of the filter, we examine the consequences of using the original (unclean) data when removing outliers. To illustrate this process, Figure 4 plots the series with and without data cleaning. In the first panel, we observe the original series. The presence of spikes and other anomalous observations is clear. In panel 2 of the same figure, we demonstrate the effectiveness of the Sigma (GARCH) filtering process. The observed series is not only smoothed but erroneous spikes are eliminated from the data.

[5] The initial parameters for return and volatility series (*u* and *c*, respectively) are chosen by way of observing the data structures and using a graphical analysis, similar to Brownlees (2006.



AUD/USD EUR/USD JPY/USD

Plot of Time Series of Uncleaned Price Data

Plot of Time Series of Cleaned Price Data

Figure 4. Comparison of clean data and unclean data for AUD/USD, EUR/USD and JPY /USD currency pairs at 5-minute intervals, 2 Jan - 30 Apr 2013

At first glance, it may appear that it would be possible for both the Sigma (RV) and Sigma (GARCH) filtering processes to eliminate important informational data by mistaking a genuine data point for an erroneous price spike. Some price spikes may actually contain useful information that, when eliminated, reduces the distributional quality of the data. But the double-threshold filter coupled with the volatility-normalised price method avoids removing genuine price data because such data is only eliminated if neighbouring normalised prices lie in isolation from the erroneous data point. However, if neighbouring volatility-normalised price data contributes to the informational context then the price spike will remain. Figure 4 shows that the filtering process is robust because volatility clusters and some price spikes are retained while erroneous outliers are eliminated.

Concluding Remarks

We develop a new algorithm for data cleaning high frequency time series with specific reference to commonly traded exchange rates. We introduce a novel technique to deal with volatility clustering and variable return intervals, which have not been adequately addressed in prior studies. The Sigma(RV) and Sigma(GARCH) standardised filter processes ‘clean’ the high frequency data without adversely altering the underlying dynamics of the distributions. The stylised factors inherent to those types of data still hold after filtering as the high kurtosis and tail dependencies remain. Maintaining the underlying distributional properties that eliminates spikes due to artefacts or other factors is a significant benefit of using these filtering techniques. Importantly, when the returns are standardised by volatilities, the presence of fat tails in the distributions are greatly reduced but not eliminated. This provides a critical informational advantage relative to other filtering methods. While the potential to over scrub the data and lose important information exists, our standardisation approaches avoid this possibility because we maintain the key features of clustered price volatility and other higher-order moments in the distribution.

Acknowledgment:

We thank the SIRCA for providing us with the tick data for this analysis.

References:

1. Andersen, T.G., Bollerslev, T., Diebold, F.X., and Labys, P., 2000, “Exchange Rate Returns Standardized by Realized Volatility are (Nearly) Gaussian”, *Multinational Finance Journal*, 4, 159-179
2. Bank for International Settlements (BIS), 2013, “Triennial Central Bank Survey Foreign Exchange Turnover in April 2013: Preliminary Global Results”, Bank for International Settlements Report
3. Bollerslev, T., 1986. “Generalized Autoregressive Conditional Heteroskedasticity”, *Journal of Econometrics*, 31, 307-327
4. Brownlees, C.T. and Gallo, G., 2006, “Financial Econometric Analysis at Ultra High Frequency: Data Handling Concerns”, *Computational Statistics and Data Analysis*, 51, 2232-2245
5. Engle, R.F., 1982. “Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation”, *Econometrica*, 50, 987-1008
6. Hansen, P. and Lunde, A., 2005, “A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH (1,1)? ”, *Journal of Applied Econometrics*, 20, 873–889.

7. Hansen, P. and Lunde, A., 2006, "Realized Variance and Market Microstructure Noise ", *Journal of Business and Economic Statistics*, 24, 127-161
8. Laurini, F. and Tawn, J.A., 2003, "New Estimators for the Extremal Index and Other Cluster C.haracteristics", *Studies in Nonlinear Dynamics & Econometrics*, 6, 189-211
9. Laurini, F., 2004, "Clusters of Extreme Observations and Extremal Index Estimate in GARCH Processes", *Studies in Nonlinear Dynamics & Econometrics*, 8, 1-21
10. Martens, M., Chang, Y. C., & Taylor, S. J.,2002, "A Comparison of Seasonal Adjustment Methods When Forecasting Intraday Volatility", *Journal of Financial Research*, *25*, 283-299
11. Meinl, T. and Sun, E.W., 2012, "A Nonlinear Filtering Algorithm based on Wavelet Transforms for High-Frequency Financial Data Analysis", *Studies in Nonlinear Dynamics & Econometrics*, 16, 1-24
12. Taylor, S. J., & Xu, X.,1997, "The incremental volatility information in one million foreign exchange quotations", *Journal of Empirical Finance, 4*, 317-340
13. The Economist, 2013, "*The Foreign-Exchange Market: Special FX*", 21-27, 69.
14. Verousis, T. and Gwilym, O., 2009, "An Improved Algorithm for Cleaning Ultra High-Frequency Data", *Journal of Derivative and Hedge Funds*, 15, 323-340.