High Frequency Data and Volatility in Foreign Exchange Rates

Bin Zhou [1]
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

September 1992

Abstract:

Exchange rates, like many other financial time series, display substantial heteroscedasticity. This poses obstacles in detecting trends and changes. Understanding volatility becomes extremely important in studying financial time series. Unfortunately, estimating volatility from low frequency data, such as daily, weekly, or monthly observations, is very difficult. The recent availability of ultra-high frequency observations, such as tick-by-tick data, to large financial institutions creates a new possibility for the analysis of volatile time series. This article uses tick-by-tick Deutsche Mark and US Dollar (DM/$) exchange rates to explore this new type of data. Unlike low frequency data, high frequency data have extremely high negative first order autocorrelation in their return. A model explaining the negative autocorrelation and volatility estimators using the high frequency data are proposed. Daily and hourly volatility of the DM/$ exchange rates are estimated and the behaviors of the volatility are discussed.

KEY WORDS: Financial time series; tick-by-tick data; heteroscedasticity.

[1]The author thanks Professor David Donoho for his advice and helpful comments. The author also thanks Morgan Guaranty Trust Company for providing the data for this research.

1 INTRODUCTION

There is considerable literature analyzing the behavior of exchange rates. However, structural exchange rate modeling has not been very successful. By studying monthly data, Meese and Rogoff (1983a,b) have shown that a random walk model fits at least as well as more complicated structural models.

Empirical studies such as those by Hsieh (1988) and Diebold (1988) have shown that daily returns are approximately symmetric and leptokurtic (i.e., heavy tailed). The autocorrelations are weak but not independent and identically distributed (i.i.d.). One explanation for the heavy tailed distribution is the hypothesis that data are independently distributed as a normal distribution whose mean and variance change over time (Friedman and Vandersteel 1982, Hsieh 1988 and Diebold 1988). Since market volatility depends information flow and the amount of "information flow" is not constant over time. There is no reason to believe that the variance of the price changes is constant over time. Clark (1973) and many others (Mandelbrot and Taylor 1969, Praetz 1972) have argued that observed returns come from a mixture of normal distributions. If the random variable X_t denotes the daily return of the price, the conditional distribution of X_t given information is:

$$X_t|\omega_t \sim N(\mu, f(\omega_t)) \tag{1}$$

where ω_t is all the information available at time t. The quantity ω_t could be the number of transactions (Mandelbrot and Taylor 1969), or trading volume (Clark 1973).

One parametrization of this conditional heteroscedasticity was first studied by Engle (1982). Engle approximates the volatility by

$$f(\omega_t) = \alpha_0 + \sum_{i=1}^{p} \alpha_i (X_{t-i} - \mu)^2, \quad \alpha_0 > 0 \quad \alpha_i \geq 0, \quad i = 1, ..., p \tag{2}$$

It is called the autoregressive conditional heteroscedasticity (ARCH) model since the heteroscedasticity is represented in an autoregressive fashion.

Because the ARCH model exhibits the conditional heteroscedasticity present in financial time series and is mathematically easy to manipulate, it has been used to analyze many financial time series. Previous studies found that the ARCH model provides a close approximation of many financial time series. Since then many other parametrizations, such as the generalized autoregressive conditional heteroscedasticity (GARCH) model (Bollerslev 1986), have been proposed. They capture some characteristics of the volatility such as

volatility clustering. However, like other parametric models, the parameters need to be estimated. The volatility estimates from ARCH, GARCH models are often lagged because the historical data are used.

The availability of high-frequency data has opened up new possibilities in estimating volatility. Tick-by-tick data provide us with a near continuous observation of the process. It gives us the chance to study volatility in great detail. Understanding volatility is the key issue in the conditional heteroscedasticity model (1) and any other financial time series models. This paper explores tick-by-tick data and uses the data to estimate and study volatility.

2 HIGH-FREQUENCY DATA

Because of fast growing computer power, gathering financial data is easier than ever. Data are no longer recorded daily or weekly. Many large institutions began to collect so called *tick-by-tick* exchange rate in the early eighties.

In contrast to stock markets, foreign exchange markets have no geographical location, and no "business-hour" limitations. Traders negotiate deals and make exchanges over the telephone. The transaction prices and trading volume are not known to the public. The exchange rates used for most research are the quotes from large data suppliers such as Reuters, Telerate, or Knight Ridder. Any market-maker can submit new quotes to the data suppliers. The quotes are then conveyed to data subscribers' screens. The data suppliers cover the market information worldwide and twenty-four hours a day. The quotes are intended to be used by market participants as a general indication of where exchange rates stands, but does not necessarily represent the actual rate at which transactions are being conducted. It is possible for some participants to manipulate indicative prices occasionally and create a favorable market movement. However, since a bank's reputation and credibility as a market-maker emerges from favorable relations with other market participants, it is generally felt that these indicative prices closely match the true prices experienced in the market.[2] Goodhart and Figliuoli (1991) studied minute-by-minute exchange rates (the closing tick of a minute) from Reuters. They found that the series exhibited (time varying) leptokurtosis, unit roots, and first-order negative correlation. The paper used only three days' data from Reuters.

In this study, tick-by-tick data for the entire year of 1990 are used. The data are provided by Morgan Guaranty Trust Company (J. P. Morgan). They

[2]Reader who is unfamiliar with this type of data in foreign exchange markets may want to read Goodhart and Figliuoli's (1991) paper for details.

Table 1: A Sample of Tick-by-tick Exchange Rates

PUB.	UNIX TIME	BANK	LOC.	RATE
263	632672082	BERGEN BK	OSL	1.6980 -90
WRLD	632672083	SBZX		1.6980/90
263	632672083	COCO	COP	1.6985 -95
263	632672085	AKTIVBANK	VEJ	1.6988 -95
263	632672088	CHEMICAL	N Y	1.6985 -90
263	632672089	CHEMICAL	LDN	1.6987 -92
263	632672091	SWISS BANK	BAS	1.6980 -90
263	632672092	SE BANKEN	MAL	1.6985 -90
263	632672094	MIDLAND BK	LDN	1.6983 -93
WRLD	632672095	DBNY		1.6985/90
263	632672098	SOC GEN	PAR	1.6985 -90

contain spot rate quotes from Reuters and Telerate. To save computation time, this study concentrates on the Deutsche Mark and US Dollar (DM/$) exchange rate, the most active in the market. Other exchange rates have similar results. One year of DM/$ yield more than two million observations. Each record of data contains the following information: data publishers, UNIX time stamp, originator of the data, location, bid and ask prices. A sample record is listed in Table 1. The rate quoted in the form x.xxxx and 0.0001 is called one basis point. The spread between bid and ask is most often 10 points or less. Since banks are not obligated to trade at the price they quoted, I find quite a few keying errors in the data. To illustrate the problem, a small portion of the raw data is plotted in Figure 1. There are quite a few "outliers" in the data. Before the data can be used for statistical analysis, a validation procedure is necessary to remove the "outliers". Our validation procedure removes any sudden jumps with a reversal. The detailed program is listed in appendix B. The starred values in Figure 1 are identified as the "outliers" by our validation program. The validated data are plotted in Figure 2. A sample of removed data has been manually checked. In most cases, the cause of the errors has been identified as a keying mistake.

Since the consecutive prices are nonstationary, it is appropriate to study changes in prices. Like many other researchers, I prefer to study the compound return (which is defined as the difference in the logarithmic value of the bid prices). Table 2 shows the summary statistics of the tick-by-tick returns.

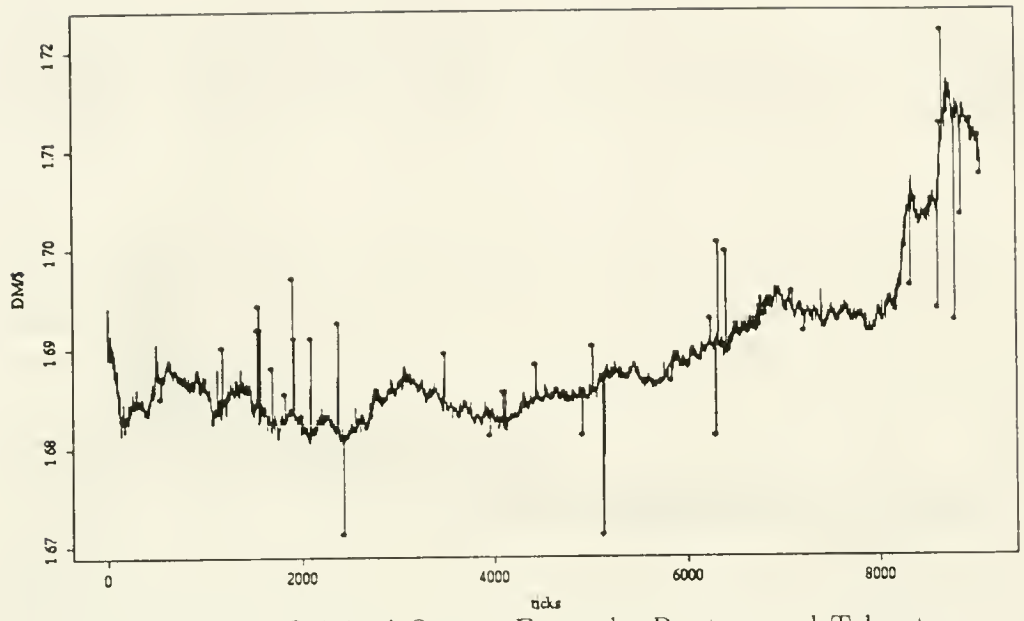

Figure 1: Original Quotes From the Reuters and Telerate.

Figure 2: Validated Quotes.

Table 2: Summary Statistics of Tick-by-tick Returns

n	Min.	Max.	Mean	sd	Skew.	Kurt.	Autocorr.
2129364	-.00662	.00752	-5.55e-8	2.40e-4	-.0399	10.75	-.464

The number of ticks in each minute varies greatly. It ranges from zero to several hundreds. The average return of the tick-by-tick data is negligible in comparison to its standard deviation. Positive and negative moves are equally likely. The returns are skewed slightly to the left. The kurtosis is much higher than 3, the kurtosis of a normal distribution.

Whistler (1988) indicated that kurtosis rises as periodicities become shorter and frequencies become higher by comparing moments of hourly, daily, weekly, and monthly returns. However, Goodhart and Figliuoli (1991) disputed his finding and found that kurtosis for minute-by-minute returns is less than those for hourly and daily returns. To investigate this phenomenon, a subsequence of tick-by-tick data at every n-th tick is sampled. Then the sample kurtosis of return of the subsequence is calculated. For $n = 1, 2, ..., 1000$, the sample kurtoses are shown in Figure 3. This figure shows that as the frequency increases (n decreases), the sample kurtosis passes three stages: rises, becomes unstable and then decreases. The last stage, which contradicts to Whistler's finding may be due to the increasing negative autocorrelation in the data (see Figure 4).

Although I expected the slightly negative first lag autocorrelation that had been reported in other exchange rate studies, a -47% negative autocorrelation in tick-by-tick returns is a surprising. To be cautious, I also calculated the autocorrelations in four subgroups according to four quarters. The autocorrelation coefficients are -0.4718, -0.4691, -0.4665 and -0.4632. The negative autocorrelation is consistent. Obviously, high frequency data do not follow a Brownian motion as is assumed for low frequency data. Our question is if there is any fundamental difference between the high and low frequency data.

After further studying the data, I have found that the difference in high and low frequency data is the level of noise. The noise is negligible in low frequency data, but becomes very significant in high frequency data. The noise may come from many different sources. For example, it could be round-off error. All financial data are quoted in finite digits. If a Brownian motion has been rounded off, there will be negative autocorrelation in its return.

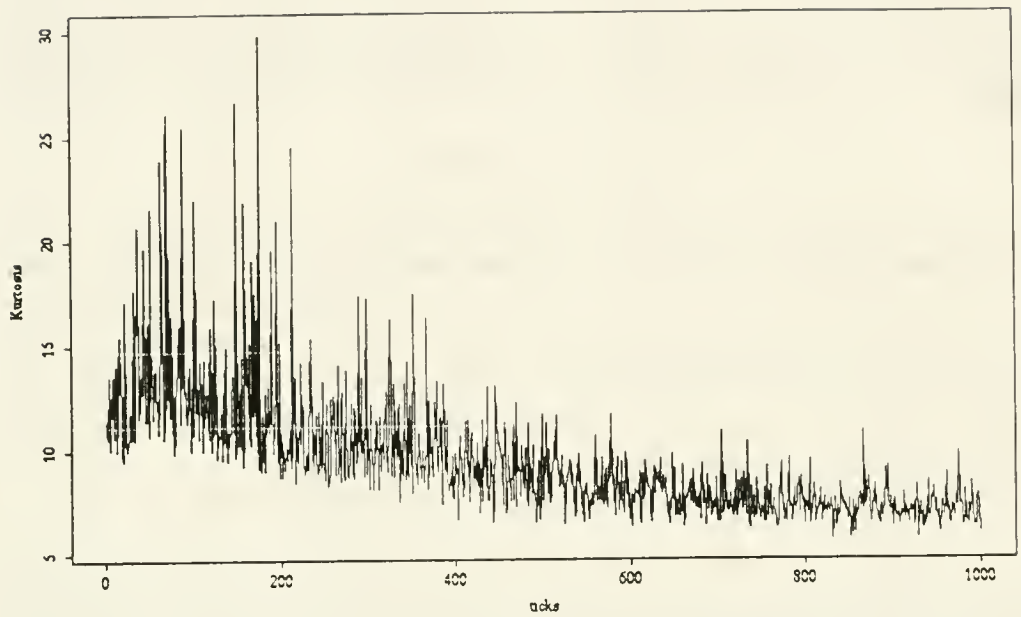

Figure 3: Sample Kurtoses of Return of Different Frequencies.

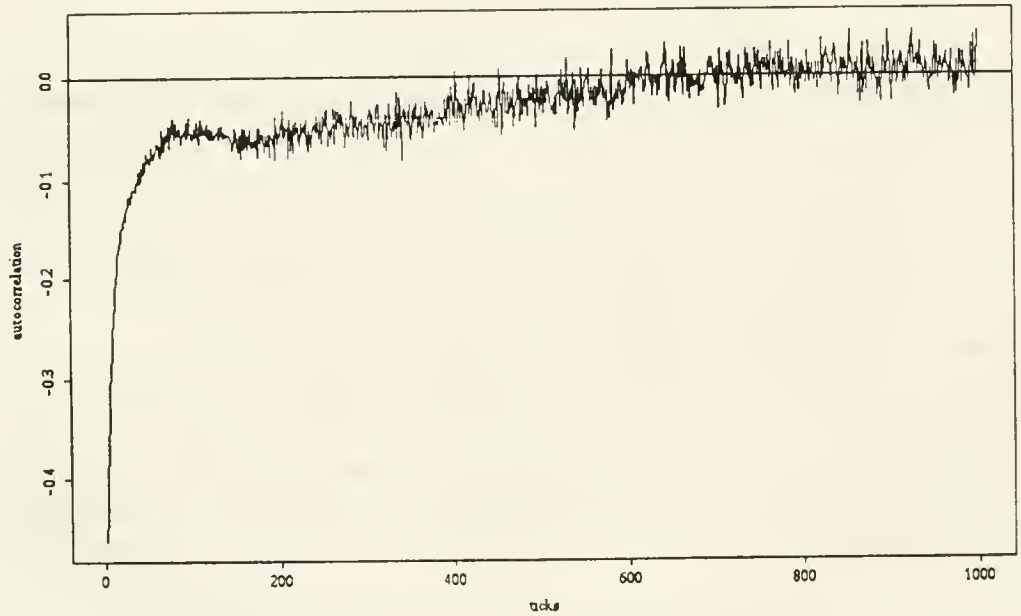

Figure 4: First-order Autocorrelations of Return of Different Frequencies.

There are also updating noises in quotes. To be visible in the market, traders keep updating their quotes. The new update is often slightly different from the previous quotes even if the market price remains the same. Typographical errors are another source of noise. Summarizing these arguments, I assume the following process for the exchange rate:

$$S(t) = d(t) + B(\tau(t)) + \epsilon_t \tag{3}$$

where $S(t)$ is the logarithm of the exchange rate, $B(\cdot)$ is the standard Brownian motion, both $d(\cdot)$ and $\tau(\cdot)$ are assumed deterministic functions, $\tau(\cdot)$ has positive increments, and ϵ_t is the mean zero random noise independent to the Brownian motion $B(.)$. The noise, ϵ_t, is a combination of several sources that were mentioned above. No distribution assumptions have been made for this noise. It could be a very general stochastic process. This extra noise causes most of the negative autocorrelation in high frequency data.

Let $X(s,t) = S(t) - S(s)$, the return in interval $[s,t]$. Then

$$X(s,t) = \mu(s,t) + \sigma(s,t)Z_t + \epsilon_t - \epsilon_s \tag{4}$$

where Z_t is a standard normal random variable, $\sigma^2(s,t) = \tau(t) - \tau(s)$ and $\mu(s,t) = d(t) - d(s)$. The variance of the return is:

$$\mathrm{Var}(X(s,t)) = \sigma^2(s,t) + \eta_t^2 + \eta_s^2 - 2c(s,t)$$

where $\eta_t^2 = \mathrm{Var}(\epsilon_t)$ and $c(s,t) = \mathrm{Cov}(\epsilon_s, \epsilon_t)$. When $|s-t|$ increases, $\sigma^2(s,t)$ increases as well. For large $|s-t|$, the variances of noises become negligible, so does the noise. $X(s,t)$ behaves just like a random walk. When $|s-t|$ decreases to near zero, $\sigma^2(s,t)$ diminishes. The return, $X(s,t)$, is the difference of two noises. The sample first order autocorrelation of such series is about -50%. When we study high-frequency data, the noise is no longer negligible. An autocorrelation of -47% for the DM/$ exchange rate indicates that the level of noises is very high in tick-by-tick data.

There are several difficulties in analyzing the process (3). One of the difficulties is lack of information about $\tau(t)$, which I call the *cumulative volatility*. $\sigma^2(t-\delta, t) = \tau(t) - \tau(t-\delta)$ is called the δ-increment of volatility or simply *δ-volatility*. In the next section, I will devote my attention to estimating the volatility increment.

3 VOLATILITY ESTIMATION

In this section, I will concentrate on estimating the volatility of a given time $[a,b]$, $\tau(b) - \tau(a)$. The function $\tau(t)$ can be estimated increment by increment.

I first derive an optimal estimator of the volatility based on the assumption of constant variance and zero mean. For simplicity, I also add normal assumptions on the noise component. A more generalized result will be given in later of the section. Proofs of the theorems are listed in appendix A.

Theorem 1 *Assume that* $\{S(t_i), i = 0, 1, ..., n\}$ *is a series of observations from the process*

$$S(t) = B(\tau(t)) + \epsilon_t \tag{5}$$

where $\epsilon_{t_i}, i = 1, .., n$, *are independent and identically distributed with normal distribution and* $\tau(t) = \sigma^2 t + b$. *Let* $X_i = S(t_i) - S(t_{i-1})$. *Then the maximum likelihood estimator of* σ^2 *is*

$$\hat{\sigma^2}_{MLE} = (1/n)\sum_{i=1}^{n}(X_i^2 + 2X_iX_{i-1}\frac{\rho}{\rho'})\frac{(1-\rho\rho')}{(1-\rho^2)} \tag{6}$$

where

$$\rho = \frac{\sum_{i=1}^{n} X_iX_{i-1}}{\sum_{i=1}^{n} X_{i-1}^2} \qquad \textit{and} \qquad \rho' = \frac{\sum_{i=1}^{n} X_iX_{i-1}}{\sum_{i=1}^{n} X_i^2}$$

This MLE is not unbiased. However, noticed that ρ and ρ' are very close for large n, I can have an unbiased estimator by eliminating the factors $(1 - \rho\rho')/(1 - \rho^2)$ and ρ/ρ':

$$\hat{\sigma^2}_U = (1/n)\sum_{i=1}^{n}(X_i^2 + 2X_iX_{i-1}). \tag{7}$$

Theorem 2 *Under the assumptions of Theorem 1, the mean and variance of the estimator (7) are:*

$$\mathbf{E}\hat{\sigma^2}_U = \sigma^2 \tag{8}$$

and

$$\mathbf{Var}(\hat{\sigma^2}_U) = (1/n)\sigma^4(6 + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{\sigma^4}) + 4\eta^4/n^2, \tag{9}$$

where η^2 *is variance of* ϵ_{t_i}.

From (9), I find that variance of $\hat{\sigma^2}_U$ can be optimized by properly adjusting the variance ratio η^2/σ^2. Since aggregation increases the variance σ^2, I apply estimator (7) to $X_{i,k} = S(i) - S(i-k), i = k, 2k, ..., n$ where I assume that n is multiple of k. Let

$$\hat{\sigma^2}_{U,k} = \frac{1}{n}\sum_{i=k,2k,...,n}(X_{i,k}^2 + 2X_{i,k}X_{i-k,k}). \tag{10}$$

Estimator (10) is unbiased and the variance is given in following theorem:

Theorem 3 *Under the assumptions of Theorem 1*

$$\mathbf{Var}(\hat{\sigma^2}_{U,k}) = (1/n)\sigma^4(6k + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{k\sigma^4}) + 4\eta^4/n^2 \tag{11}$$

The variance is minimized at $k = [(2\eta^2)/(\sqrt{3}\sigma^2)]$ *or* $k = [(2\eta^2)/(\sqrt{3}\sigma^2)] + 1$, *where* $[x]$ *rounds* x *down to the next integer.*

Since the noise ϵ_t is independent, the variance of the estimator can be further reduced by averaging the estimator (7) using overlapping returns:

$$\hat{\sigma^2} = \frac{1}{kn}\sum_{i=1}^{n}(X_{i,k}^2 + 2X_{i,k}X_{i-k,k}) \tag{12}$$

In fact, it can be proved that:

$$\mathbf{Var}(\hat{\sigma^2}) \leq \frac{1}{n}\sigma^4(6k + 16\frac{\eta^2}{k\sigma^2} + 8\frac{\eta^4}{k^2\sigma^4}) + 4\eta^4/n^2 \tag{13}$$

The above three theorems assumed i.i.d. noises and constant variances. However, if I relax all these assumptions, I can still have a nearly unbiased estimator. Suppose that we have observations $\{S(t_i), i = -2k, -2k+1, ..., n\}$ from process (3). The variances of the returns are not necessarily constant. The noises can be nonstationary. Under minimal assumptions, I propose to estimate the volatility $\tau(t_n) - \tau(t_0)$ by:

$$V(t_0, t_n) = \frac{1}{k}\sum_{i=1}^{n}(X_{i,k}^2 + 2X_{i,k}X_{i-k,k}) \tag{14}$$

Theorem 4 *Assume* $Cov(\epsilon_{t_i}, \epsilon_{t_{i-k}})$*=0 for all* i. *Then*

$$\begin{aligned}
\mathbf{E}V(t_0, t_n) &= \tau(t_n) - \tau(t_0) \\
&\quad + \sum_{i=1}^{k-1}(i/k)[\sigma^2(t_{i-1-k}, t_{i-k}) - \sigma^2(t_{n-i+1}, t_{n-i})] \\
&\quad + (1/k)\sum_{i=0}^{k-1}[\eta^2_{t_{i-k}} - \eta^2_{t_{n-i}}] \\
&\quad + (1/k)\sum_{i=0}^{n}[\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})]
\end{aligned} \tag{15}$$

where $\eta^2(t) = Var(\epsilon_t)$.

The only assumption I have made in this theorem is that of uncorrelated noises. Since

$$\sum_{i=k}^{n}[\mu^2(t_i,t_{i-k})+2\mu(t_i,t_{i-k})\mu(t_{i-k},t_{i-2k})]\le 3\max\{|\mu(t_i,t_{i-k})|\}[d(n)-d(0)]$$

the last term in (15) is negligible in high frequency data if the drift $d(t)$ is smooth in interval $[t_0,t_n]$. Therefore, for large n, the estimator (14) is approximately unbiased if no jumps occurred in the time interval $[t_0,t_n]$. This estimator is also easy to update when new data become available. It will allow us to estimate the volatility $\tau(t)$ dynamically.

4 ESTIMATING VOLATILITY OF EXCHANGE RATES

In this section, I estimate the volatilities of DM/$ exchange rates at different frequencies. By examining these estimates, I want to evaluate my assumptions about the exchange rates as well as the volatility estimator. Again, 1990 DM/$ exchange rates are used. The data have one discontinuity: in the week of August 13, the data base was shutdown due to a power outage in lower Manhattan. The return for that week is set at zero, as is the volatility estimate.

To choose the parameter k, I estimated the variance ratio η^2/σ^2 to be approximately 6. Minimizing the upper bond of variance (13), I have $k=6$. Using all available tick-by-tick data, I first estimate the volatility of the DM/$ exchange rate in entire 1990. The estimate is .010349. To verify this estimate, I compared it to the estimate under the Brownian motion assumption. If the data had no noise and followed a Brownian motion, the quadratic variation

$$Q_k=\sum_{i=k}^{(n/k)}X(t_{ik-k},t_{ik})^2$$

would be a standard estimator of the volatility. When the quadratic variation is used on data with noise, it overestimates the volatility. As the sample frequency decreases, so does the bias. The expectation of the quadratic variation is:

$$\begin{aligned}\mathrm{E}Q_k &= \sum_{i=1}^{(n/k)}[\tau(t_{ik})-\tau(t_{ik-k})+\eta^2_{t_{ik}}+\eta^2_{t_{ik-k}}-c(t_{ik},t_{ik-k})+\mu^2(t_{ik},t_{ik-k})]\\ &= \tau(t_n)-\tau(t_0)+\sum_{i=1}^{(n/k)}[\eta^2_{t_{ik}}+\eta^2_{t_{ik-k}}-c(t_{ik},t_{ik-k})+\mu^2(t_{ik},t_{ik-k})]. \quad (16)\end{aligned}$$

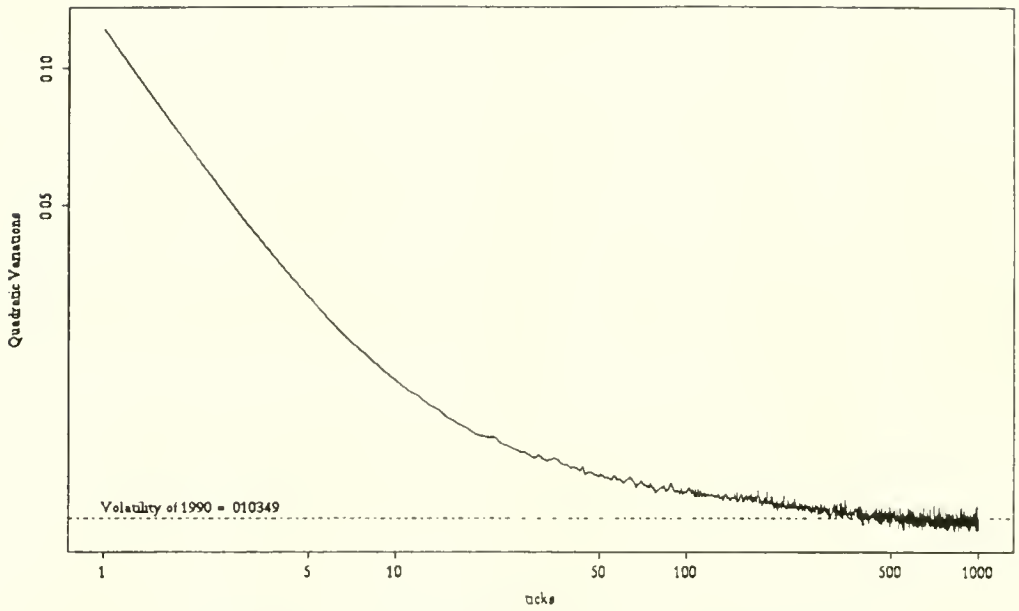

Figure 5: Quadratic Variations Using n-tick Returns

where $c(t_{ik}, t_{ik-k}) = \mathrm{Cov}(\epsilon_{t_{ik}}, \epsilon_{t_{ik-k}})$. When ϵ_t's are uncorrelated and μ's are negligible,

$$\mathrm{E}Q_k \approx \tau(t_n) - \tau(t_0) + 2 \sum_{i=1}^{(n/k)} \eta_{t_{ik}}^2 \tag{17}$$

which decreases as k increases. I plotted Q_k against k in Figure 5. Both axes have a logarithmic scale. When the frequency is low, the quadratic variation is about the same as our estimate except for a high variation due to a small sample size. At a high frequency, the bias is tremendous. When $k = 1$, the quadratic variation is about thirteen times the size of our estimate. Therefore, from (17), the total variance of the noise is about six times the total volatility which confirmed our early estimate of the ratio.

To estimate daily volatilities, I need to define the start and the end of a day since the foreign exchange market is a twenty-four hour international market. I choose 24 hours from 0:00 Greenwich Mean Time (GMT) as a day because that 0:00 GMT is 9:00am Tokyo time and 24:00 GMT is 7:00pm New York time. This twenty-four hour period covers most activities of the world market. There are average more than 7,000 ticks per day in weekdays. There are many fewer ticks on weekends and holidays. For small n, the volatility estimate from (14) could be negative. If such is the case, I let the estimate go to zero to avoid negative volatility. The daily volatility estimates of DM/$ are

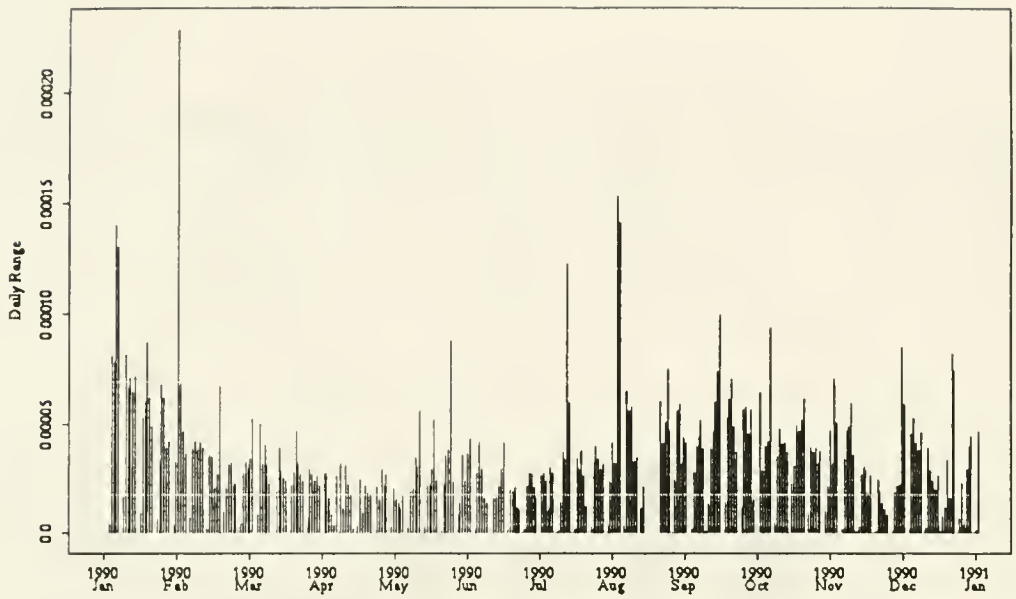

Figure 6: Daily Volatility Estimates of 1990 DM/$

plotted in Figure 6.

A good volatility estimator should be able to catch market turbulences. High volatility estimates should indicate unusual activities in the market. In examining the daily volatility estimates, I found such correlation. There are six volatility estimates above .0001. There are on January 4, 5, 30, July 12 and August 2, 3, 1990. On all these six days, there is significant news released. On January 4 and 5, the German central bank surprisingly intervened in the foreign exchange market and pushed the dollar lower. On January 30, a wild market followed a rumor that Mr. Gorbachev was considering resigning as secretary of the former Soviet Communist Party. On July 12, the dollar tumbled because possible lower interest rate by the US Federal Reserve. On August 2 and 3, the Dollar had another wild ride as the news of Iraq's invasion of Kuwait spread around the world. However, large volatility does not always mean a large change in price. The daily changes for above six days were -0.0537, 0.0162, 0.0212, -0.0178, 0.0058 and -0.0074 respectively. On August 2 and 3, the exchange rate only changed 58 and 74 points that are about the average. The price changes prior to August 2 were not large either. Therefore if an ARCH model is used, these activities will be missed.

Another way to check the assumption of our model for exchange rates (3) and the accuracy of our volatility estimates is to test normality of scaled daily returns. When a model (3) provides a good approximation and the volatility

Figure 7: Q-Q Plot of Daily Returns

Table 3: Basic Statistics of Daily Returns

	n	Mean	Var.	Skew.	Kurt.	KS-Test
X	320	-3.1e-4	3.62e-5	-.429	6.25	1.275(p=.00)
Y=X/σ	320	-6.0e-2	1.17e+0	-.329	3.20	0.788(p=.13)
SE		(.060)	(.093)	(.137)	(.274)	

estimate is accurate, the rescaled daily return $Y_i = X_i/\sigma_i$ should be close to a standard normal random variable. Noise is negligible since I am studying daily returns here. Excluding zero volatility estimates (all on Saturdays), I plot Q-Q normal plots for both return X_i and rescaled return Y_i in Figure 7. The basic statistics of both X_i and Y_i are shown in Table 3. The standard errors are also given in the parentheses. Table 3 also shows the Kolmogorov-Smirnov goodness-of-fit test for normality. If we compare the statistics in column X_i and column Y_i, we see that Y_i is much closer to having a normal distribution. It indicates that the volatility estimates are reasonably good and the process (3) well describes the high frequency observations of the exchange rates.

The daily volatility estimates can be used in many other ways. For example, we can use them to check calendar effects on daily volatilities. Seven average daily volatilities are calculated. The results are listed in Table 4 as

Table 1: Average Daily Volatility

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Ave. Vol.	7.54e-6	3.25e-5	4.13e-5	3.63e-5	4.45e-5	4.12e-5	7.07e-8
SE	1.11e-6	2.15e-6	4.42e-6	2.43e-6	4.39e-6	3.92e-6	3.19e-8
n	50	51	50	51	51	51	37

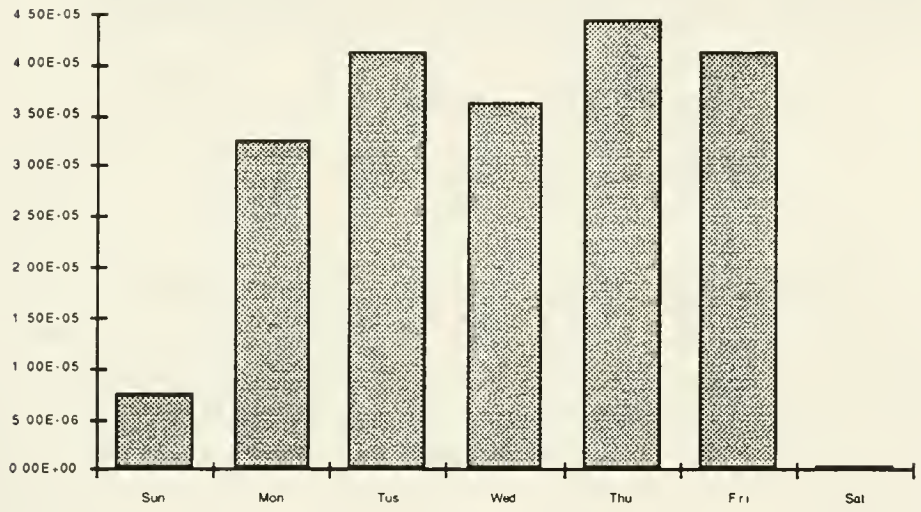

Figure 8: Average Daily Volatilities

well as plotted in Figure 8. Sunday's nonzero volatility is due to the New Zealand market, which opens 18:00 GST Sunday night. The volatility is low on Monday and high on Thursday and Friday. However, F-statistic testing of the null hypothesis of equal means shows that the difference is not significant during weekdays.

Because of over two million data points, I can repeat the above procedure for hourly volatilities. For 1990, I estimated a total of 6,354 hourly volatilities. The average number of ticks in an hour is about 335. Using all weekday hourly volatility estimates, I calculated average hourly volatilities and plot them in Figure 9. The figure shows that the volatility is high when both the American and European markets are open. A very low volatility around 10:00pm EST(12:00 noon Tokyo time) is corresponds to the lunch hour in

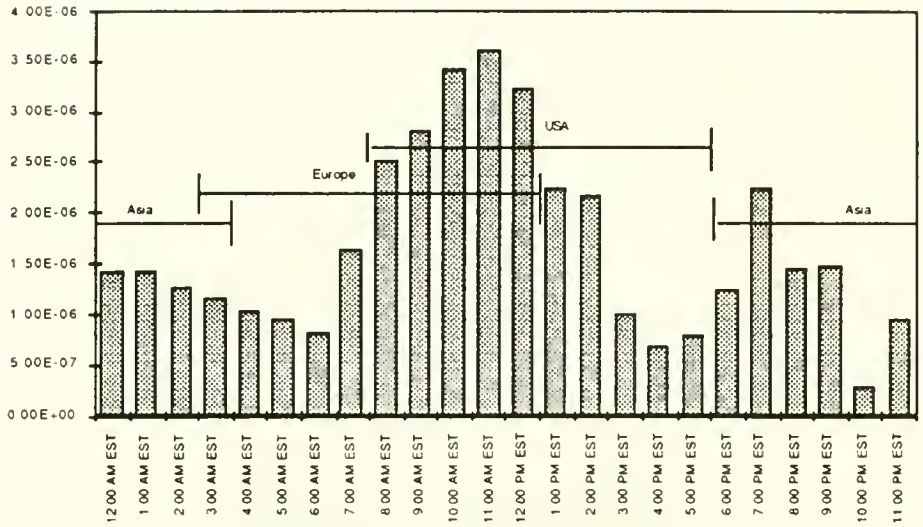

Figure 9: Average Hourly Volatilities

Tokyo.

Again, the Q-Q normal plots of hourly return X_i and rescaled hourly return $Y_i = X_i/\sigma_i$ for all $\sigma^2 > 5\epsilon - 7$ are given in Figure 10. The basic statistics of both X_i and Y_i are shown in Table 5. The sample kurtosis of rescaled returns is about 3. This time the contrast is more significant. The Q-Q plot of hourly returns is curved indicated heavy tails. However, the Q-Q plot of the rescaled returns is almost a straight line. The normality is rejected by the classic Kolmogorov-Smirnov (KS) test. However, the normality is not rejected by Shapiro-Wilk test (Shapiro and Wilk 1965, Royston 1982a and 1982b), which is considered a more powerful test for normality. Since Y_i is not a simulated data, it can't be exactly normal. With size of more than 6,000 data points, it is almost certain that some tests would reject the normality of this time series. However, many other statistics indicate that Y_i is very close to having an i.i.d. normal distribution. This implies that I have reasonably good estimates of the volatilities.

5 CONCLUSION AND DISCUSSION

High frequency data can be described as a Brownian motion with noise. This noise brings a strong first lag negative autocorrelation in high frequency data. The autocorrelation decreases as frequency decreases since the role of the noise

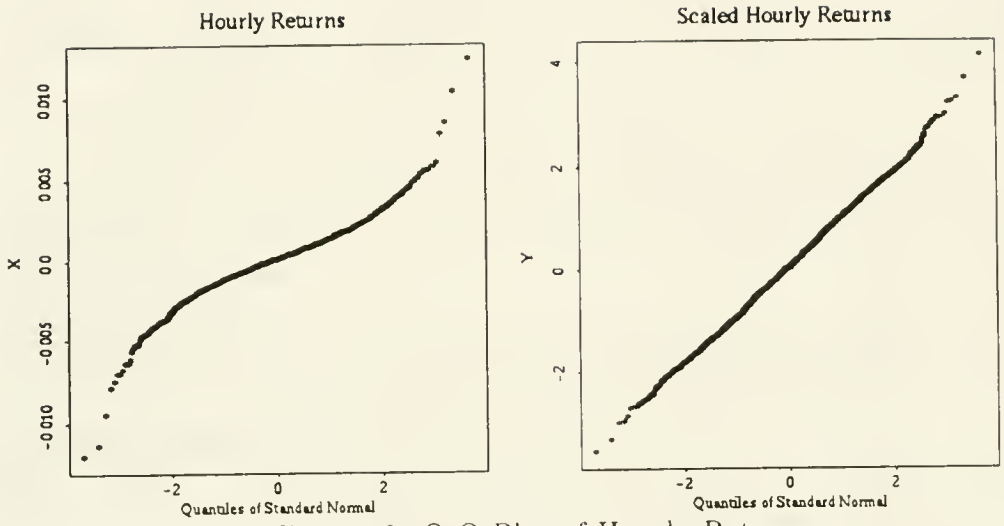

Figure 10: Q-Q Plot of Hourly Returns

Table 5: Basic Statistics of Hourly Returns

	n	Mean	Var.	Skew.	Kurt.	KS-Test
X	4377	-4.3e-5	2.28e-6	-.255	8.32	3.847(p=.00)
Y=X/σ	4377	-6.4e-3	9.13e-1	.038	2.95	1.604(p=.00)
SE		(.014)	(.020)	(.037)	(.074)	

reduces. High frequency data can be used to estimate the volatility in high frequency with reasonable precision. In contrast to other volatility estimators, our volatility estimator mainly uses the data within the period we are interested in instead of historical data. This allows us to capture the market volatility quickly without delay. The estimate is nearly unbiased when prices have no jumps. The Q-Q normal plot of rescaled daily or hourly returns by the volatility estimates shows an almost straight line, which other volatility estimators can not achieve.

If we consider volatility as a stochastic process, we can have the same discussions as we did in this paper by condition on volatility. After we obtained the volatility estimates, we can further study the volatility process itself. For example, we can study the distribution of the daily volatility, the dynamic structure of the daily process. Our study indicates that daily volatility can be approximated by a lognormal distribution. Therefore, we can apply regular ARIMA techniques on the logarithm of volatility for modeling and forecasting the volatility. Unlike estimating ARCH coefficients, where we often run into divergence in MLE iterations, estimating ARIMA coefficients for volatilities are much easier. The volatility forecasting is also improved. Besides, these volatility estimates can be used in many other ways. Since the sample mean and the sample variance of i.i.d. normal variables are independent, it is easy to image that the daily volatility estimator is little dependent on daily return. This property enables us to do research on relations between the market volatility and market price movement.

APPENDIX A: PROOF OF THEOREMS

Proof of Theorem 1:

Under assumptions of the theorem,

$$X_t = \sigma Z_t + \epsilon_{t_t} - \epsilon_{t_{t-1}}$$

is a normal random variable with mean zero and variance $v^2 = \sigma^2 + 2\eta^2$, where η^2 is the variance of ϵ_{t_t}. X_t has the first lag autocorrelation $\rho = -\eta^2/v^2$. The likelihood function of X_t is

$$L(s^2, \rho; X_1, \ldots, X_n) = f(X_n|X_{n-1})f(X_{n-1}|X_{n-2})\ldots f(X_1|X_0)$$

Notice that $X_t|X_{t-1}$ is also a normal variable with mean ρX_{t-1} and variance $s^2 = v^2(1-\rho^2)$, I have

$$L(s^2, \rho; X_1, \ldots, X_n) = (2\pi s^2)^{-n/2} \exp(-\sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})^2}{2s^2}).$$

The log-likelihood function is

$$\ell(s^2, \rho; X_1, \ldots, X_n) = -\frac{n}{2}\log(2\pi s^2) - \sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})^2}{2s^2}$$

The partial derivative of the likelihood function with respect to ρ is

$$\frac{\partial \ell}{\partial \rho} = \sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})X_{t-1}}{s^2}$$

Setting the derivative equal to zero and solve for ρ, I have

$$\hat{\rho} = \frac{\sum_{t=1}^{n} X_t X_{t-1}}{\sum_{t=1}^{n} X_{t-1}^2}.$$

Similarly, for s^2, I have

$$\frac{\partial \ell}{\partial s^2} = -\frac{n}{2s^2} + \sum_{t=1}^{n} \frac{(X_t - \rho X_{t-1})^2}{2s^4} = 0$$

or

$$\begin{aligned} \hat{s^2} &= \frac{1}{n}\sum_{t=1}^{n}(X_t - \rho X_{t-1})^2 \\ &= \frac{1}{n}(\sum_{t=1}^{n} X_t^2 + \rho^2 \sum_{t=1}^{n} X_{t-1}^2 - 2\rho \sum_{t=1}^{n} X_t X_{t-1}) \end{aligned}$$

Substituting ρ by $\hat{\rho}$ in above formula. I have

$$\begin{aligned}\hat{s^2} &= \frac{1}{n}(\sum_{i=1}^{n} X_i^2 - \hat{\rho}^2 \sum_{i=1}^{n} X_{i-1}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2(1 - \hat{\rho}\hat{\rho}')\end{aligned}$$

where

$$\hat{\rho}' = \frac{\sum_{i=1}^{n} X_i X_{i-1}}{\sum_{i=1}^{n} X_i^2}.$$

It is easy to show that

$$\sigma^2 = v^2(1 + 2\rho) = s^2(1 + 2\rho)/(1 - \rho^2)$$

Therefore, the maximum likelihood estimator of σ^2 is

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2(1 + 2\hat{\rho})\frac{(1 - \hat{\rho}\hat{\rho}')}{(1 - \hat{\rho}^2)} \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 + 2X_i X_{i-1}\frac{\hat{\rho}}{\hat{\rho}'})\frac{(1 - \hat{\rho}\hat{\rho}')}{(1 - \hat{\rho}^2)}\end{aligned}$$

Proof of Theorem 2:

$$\begin{aligned}\mathbf{E}(1/n)\sum_{i=1}^{n}(X_i^2 + 2X_i X_{i-1}) &= (1/n)\sum_{i=1}^{n}(\mathbf{Var}(X_i) + 2\mathbf{E}X_i X_{i-1}) \\ &= (1/n)\sum_{i=1}^{n}(\sigma^2 + 2\eta^2 - 2\eta^2) = \sigma^2.\end{aligned}$$

and

$$\begin{aligned}\mathbf{Var}(\hat{\sigma^2}_U) &= (1/n)^2\mathbf{Var}\sum_{i=1}^{n}(X_i^2 + 2X_i X_{i-1}) \\ &= (1/n)^2\mathbf{Var}[\sum_{i=1}^{n}(\sigma^2 Z_i^2 + 2\sigma Z_i\epsilon_{t_i} + 2\sigma^2 Z_i Z_{i-1} - 2\sigma Z_i\epsilon_{t_{i-2}} \\ &\quad +2\sigma Z_{i-1}\epsilon_{t_i} - 2\sigma Z_{i-1}\epsilon_{t_{i-1}} - 2\epsilon_{t_i}\epsilon_{t_{i-2}} + 2\epsilon_{t_{i-1}}\epsilon_{t_{i-2}}) + \epsilon_{t_n}^2 - \epsilon_{t_0}^2] \\ &= \frac{1}{n}(6\sigma^4 + 16\sigma^2\eta^2 + 8\eta^4) + \frac{2}{n^2}\mathbf{Var}(\epsilon_{t_0}^2) \\ &= \frac{1}{n}\sigma^4(6 + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{\sigma^4}) + \frac{4}{n^2}\eta^4.\end{aligned}$$

Proof of Theorem 3:

Since $\mathbf{Var}(X_{i,k}) = k\sigma^2$, (9) implies that

$$\mathbf{Var}(k\hat{\sigma}^2_{U,k}) = \frac{k}{n}(k\sigma^2)^2(6 + 16\frac{\eta^2}{k\sigma^2} + 8\frac{\eta^4}{k^2\sigma^4}) + \frac{4\eta^4}{(n/k)^2}$$

or

$$\mathbf{Var}(\hat{\sigma}^2_{U,k}) = \frac{1}{n}(\sigma^2)^2(6k + 16\frac{\eta^2}{\sigma^2} + 8\frac{\eta^4}{k\sigma^4}) + \frac{4\eta^4}{n^2}$$

The variance reaches minimum when

$$k = [\frac{2}{3}\frac{\eta^2}{\sigma^2}] \quad \text{or} \quad k = [\frac{2}{3}\frac{\eta^2}{\sigma^2}] + 1$$

Proof of Theorem 4:

$$\begin{aligned}
EV(t_0, t_n) &= (1/k)\sum_{i=1}^{n}[\sigma^2(t_{i-k}, t_i) + \eta^2_{t_i} - \eta^2_{t_{i-k}} \\
&\quad +\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})] \\
&= [\tau(t_n) - \tau(t_0)] \\
&\quad + \sum_{i=1}^{k-1}(i/k)[\sigma^2(t_{i-1-k}, t_{i-k}) - \sigma^2(t_{n-i+1}, t_{n-i})] \\
&\quad +(1/k)\sum_{i=0}^{k}[\eta^2_{t_{i-k}} - \eta^2_{t_{n-i}}] \\
&\quad +(1/k)\sum_{i=0}^{n}[\mu^2(t_i, t_{i-k}) + 2\mu(t_i, t_{i-k})\mu(t_{i-k}, t_{i-2k})]
\end{aligned}$$

APPENDIX B: VALIDATION PROGRAM

There are many reasons to have outliers in the original data set shown in Figure 1. Most quotes are typed in by humans, there are unavoidable keying errors. Most outliers I found are these types of errors. Outlier also could be caused by large bid and ask spreads. In such a case, at least one of bid or ask prices does not reflect the true market price and becomes an outlier. Occasionally, electronic errors also make outliers. The following program is designed to remove above three types of outliers:

A quote is considered as an outlier and removed from the time series if

i) a rate is more than 5 or less than 1, or

ii) bid and ask spread is more than 50 points, or

iii) a rate is above or below its neighbor prices by more than a certain threshold.

For (iii), I carry out two regressions using ten bid prices on each side of the data. If the current bid price is higher or lower than c-point from both regressions, it is considered an outlier, where c is range from 15 to 30 points dependent on the variances of the neighbor points. Whenever an outlier is detected, I go back ten steps and repeat above procedure.

References

[1] Baillie, Richard T. and Tim Bollerslev (1989), "Intra-day and inter-market volatility in exchange rates." *Review of Economic Studies* **58**, 565-585.

[2] Bollerslev, T (1986), "Generalised autoregressive conditional heteroskedasticity." *Journal of Econometrics*, **31**, 307-28.

[3] Calderon-Rossel, Jorge and Moshe Ben-Horim (1982), "The behavior of foreign exchange rates, " *J. of Inter. Business Studies*, **13**, 99-111.

[4] Clark, P. K. (1973), "A subordinate stochastic process model with finite variance for speculative price." *Econometrica*, **41**, 135-155.

[5] Diebold, Francis X. (1988), *Empirical modeling of exchange rate dynamics*, Springer-Verlag, New York.

[6] Engle, R.F. (1982), "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation." *Econometrica*, **50**, 987-1008.

[7] Friedman, Daniel and Stoddard Vandersteel (1982), "Short-run fluctuation in foreign exchange rates." *J. Intern. Econ.*, **13**, 171-186.

[8] Goodhart, C.A.E. and L. Figliuoli (1991), "Every minute counts in finacial markets." *J. of Inter. Money and Finance*, **10**, 23-52.

[9] Hsieh, David A. (1988), "The statistical properties of daily foreign exchange rates: 1974-1983." *J. of Inter. Econ.*, **24**, 129-145.

[10] Mandelbrot, B. and H. Taylor (1969), "On the distribution of stock price differences." *Operations Research*, **15**, 1057-1062.

[11] Meese, R. A. and K. Rogoff (1983a), "Empirical exchange rate models of the seventies: do they fit out of sample?." *J. of Inter. Econ.*, **14**, 3-24.

[12] Meese, R. A. and K. Rogoff (1983b), "The out of sample failure of empirical exchange rate models: sampling error or misspecification?." in J. Frenkel, ed., *Exchange Rates And International Microeconomics*, Chicago: University of Chicago Press.

[13] Praetz, P.D. (1972), "The distribution of share price changes." *Journal of Business*, **45**, 49-55.

[14] Royston, J. P. (1982a). "An extension of Shapiro and Wilk's W test for normality to large samples." *Applied Statistics* **31**, 115-124.

[15] Royston, J. P. (1982b), "The W test for normality." *Applied Statistics* **31**, 176-180.

[16] Shapiro, S.S. and M.B. Wilk (1965). "An analysis of variance test for normality." *Biometrika* **52**, 591-611.

[17] Taylar, Stephen (1988). *Modeling financial time series.* John Willey & Sons.