96-15

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

THE ECONOMETRICS OF ULTRA-HIGH FREQUENCY DATA

BY

ROBERT F. ENGLE

DISCUSSION PAPER 96-15
MAY 1996

THE ECONOMETRICS OF ULTRA-HIGH FREQUENCY DATA

Robert F. Engle[1]
Chancellor's Associates Chair
Department of Economics, UCSD.

March 1996

Prepared for the Frank Paish Lecture
Royal Economic Society
Swansea, U.K.
April 2,1996

and
The Fisher Schultz Lecture
Econometric Society
Istanbul, Turkey
August, 1996

ABSTRACT

Ultra-high frequency data are complete transactions data which inherently arrive at random times. Marked point processes provide a theoretical framework for analysis of such data sets. The ACD model developed by Engle and Russell(1995) is then applied to IBM transactions data to develop semi-parametric hazard estimates and measures of instantaneous conditional variances. The variances are negatively influenced by surprisingly long durations as suggested by some of the market micro-structure literature.

[1] The author is very grateful to Jeff Russell whose collaboration on two previous papers forms the basis for this paper. He carried out the computer analysis and continually helped to extend the models presented here. Financial support is gratefully acknowledged from NSF Grant SES 94-22575. Many offered insights into this work including David Brillinger, Sir David Cox, Clive Granger, Jim Hamilton, Ken Wallis, Mark Watson, Hal White, and participants at Princeton's Econometric Workshop.

THE ECONOMETRICS OF ULTRA HIGH FREQUENCY DATA

Robert F. Engle
Department of Economics
University of California, San Diego
La Jolla, Ca. 92093-0508

619-534-4553
619-454-3985
fax:619-534-0877
email rengle@ucsd.edu

THE ECONOMETRICS OF ULTRA-HIGH FREQUENCY DATA

I. Introduction

One measure of progress in empirical econometrics is the frequency of data used. Upon entering Graduate School, I learned that my thesis supervisor, T.C. Liu, had just broken the sound barrier by estimating the first quarterly model of the U.S. economy in Liu(1964). Shortly thereafter, he doubled the feat by publishing a monthly macro model, Liu(1969) I've not yet seen a weekly model of the macroeconomy but I suspect there may be some in both government and private sector research groups. In finance a similar transition has lead from the analysis of annual data to monthly data, to weekly data, to daily data and now there is great interest in intradaily models.

In each case, much of the movement to higher frequency econometrics was a consequence of the availablility of higher frequency measurements of the economy. It is natural to suppose that this will continue and we will have ever increasing frequencies of observations. However, a moment's reflection will reveal that this is not the case. The limit in nearly all cases, is achieved when *all transactions are recorded*. These may be transactions which occur in the supermarket, on the internet or in financial markets. It is difficult to think of economic variables which really are measurable at arbitrarily high frequencies.

I will call this limiting frequency "*ultra-high frequency*" and spend the time in this paper discussing econometric methods for the analysis of ultra-high frequency data.

The salient feature of such ultra-high frequency data is that it is fundamentally irregularly spaced. Of course, one can aggregate this data up to fixed intervals of time, but at that point one might argue that it is no longer ultra-high frequency data. There is naturally a loss of information in such aggregates. This loss occurs partly because the features within the interval may be lost if the interval is too large, and partly because irregular spacing of the observations in an interval makes econometric analysis very complex if the intervals are small.

The thrust of this paper will be to develop methods which are directly tailored to the irregular spacing of the data rather than to adapt fixed interval econometrics to this new situation. The statistics literature is replete with models for data of this form. These models treat events as arriving in time according to some probability law. Famous stochastic processes such as the Poisson Process and its Doubly Stochastic, Cluster and Self-Exciting forms, Birth and Death Processes and many other continuous time discrete state processes have been developed to solve problems in science and engineering. Many of these processes can be used or extended to address economic problems. The basic model to be presented and extended in this paper is the Autoregressive Conditional Duration model developed by Engle and Russell(1995,a,b) which is a type of dependent Poisson Process. In order to motivate the statistical methodology it is necessary to examine the types of economic questions which may be asked of the data.

Section II of the paper will formulate the economic questions in a statistical framework. Sections III and IV develop the econometric models to be applied. Section V gives results from IBM stock transactions, while VI presents the IBM price distribution.

II. Formulating the Economic Questions Statistically

Transactions data can be described by two types of random variables. The first is the time of the transaction, and the second is a vector observed at the time of the transaction. In the literature of point processes, these latter variables are called marks as they identify or further describe the event that occured. In the type of financial data to be examined here, the point of time is the time at which a contract to trade some number of shares of IBM stock is agreed upon, and the marks are the volume of the contract, the price of the contract, and the posted bid and asked prices at the time. Additional marks which could be used or observed would be the counter-parties to the trade , the posted bid and asked prices for other stocks, the order mechanism and many other features of a trade which are of interest in studying market microstructure.

Let t_i be the time at which the i^{th} trade occurs and let $x_i=t_i-t_{i-1}$ be the duration between trades. At the i^{th} event the marks are observed and let these be denoted y_i which is a kx1 vector from a sample space Ξ. The data therefore can be viewed as:

(1) $$\{(x_i, y_i), i = 1,\dots N\}$$

where the i^{th} observation has joint density conditional on the past filtration of (x,y) given by:

(2) $$(x_i , y_i)\big|F_{i-1} \sim f(x_i, y_i\big|\breve{x}_{i-1}, \breve{y}_{i-1};q_i)$$

where $\breve{z}_i = \{z_i, z_{i-1},\dots, z_1\}$ denotes the past of z and q's are parameters which are potentially different from observation to observation.

Economic hypotheses or measures of interest can now be expressed in terms of this density function. The analysis of quantity data in traditional fixed interval econometrics typically involves estimating the expected number of transactions or volume of transactions in a particular time interval. The data measures the realization of a random variable but generally it is the mean of this distribution and its determinants which is of interest. For transactions data the realized transactions are zero at almost every point in time while the probability of an event at each point of time is well defined.

The expected probability of a transaction in any instant of time can easily be derived from (2). This is the instantaneous expected transaction rate and might be a measure desired for various economic purposes. For any $t>t_{i-1}$, the probability of an event must be conditioned not only on all past events but also on the fact that there has not been an event since t_{i-1}. The hazard function at a value $t>t_{i-1}$ is exactly the probability of an event at time $t+\Delta t$ given that there has not been an event since t_{i-1}. This is simply expressed as the density of $t-t_i$ (or x) divided by the survival probability which is the probability that the next event will be at a time greater than t. Since y is irrelevant to this calculation it must be integrated out giving:

(3) $$l_i(t) = \frac{\int\limits_{y\in\Xi} f(t - t_{i-1}, u\big|\breve{x}_{i-1}, \breve{y}_{i-1};q_i)du}{\iint\limits_{s\geq t, y\in\Xi} f(s - t_{i-1}, u\big|\breve{x}_{i-1}, \breve{y}_{i-1};q_i)duds},\ \text{for } t_{i-1} \leq t \leq t_i$$

A simpler expression can easily be obtained. Without loss of generality, the joint density can be written as the product of the marginal density of the duration times the conditional density of the marks given the duration, all conditioned upon the past transactions.

(4) $$f(x_i, y_i|\breve{x}_{i-1}, \breve{y}_{i-1}; q_i) = g\ (x_i,|\breve{x}_{i-1}, \breve{y}_{i-1}; q_{1i})q\ (y_i|\breve{x}_i, \breve{y}_{i-1}; q_{2i})$$

Substituting (4) into (3) gives an extension of the standard formula for the hazard which now allows for past influences of both durations and marks (See for example Kalbfleisch and Prentice(1980))

(5) $$l_i(t) = \frac{g\ (t - t_{i-1}|\breve{x}_{i-1}, \breve{y}_{i-1}; q_{1i})}{\int\limits_{s \geq t} g\ (s - t_{i-1}|\breve{x}_{i-1}, \breve{y}_{i-1}; q_{1i})ds}, \text{for } t_{i-1} \leq t \leq t_i$$

Notice in particular that it is not necessary to assume weak exogeneity or any other restriction on the joint distribution to get this expression for the hazard.

More elaborate economic hypotheses are associated with the distribution of the marks. Two such questions are the distribution of the next mark, regardless of when it occurs, or the distribution of the next mark if it occurs at time t. To calculate the distribution of the next mark conditional on t, one simply needs the function q defined in equation (4).

To calculate the distribution of the next mark regardless of when it occurs, requires calculating the marginal density of the mark.

(6) $$r\ (y_i|\breve{x}_{i-1}, \breve{y}_{i-1}; q_i)|F_{i-1} = \int\limits_{s \geq 0} f\ (s, y_i|\breve{x}_{i-1}, \breve{y}_{i-1}; q_i)ds$$

If x and y are independent conditional on the past, then q will not depend upon x and consequently, $r \equiv q$. However, in the more common case where transaction rates are related to the distribution of the marks, a numerical integral of some sort will be required.

Corresponding to each of these questions are prediction questions. What is the hazard rate expected to be at some specified time in the future or after a certain number of trades? Similarly, what is the distribution of the marks at some fixed time in the future or after a certain number of transactions have occurred. Each of these questions can also be answered by manipulation of the densities in (4) , although in most cases, closed form solutions cannot be obtained. Instead, simulations can be used to generate answers. These simulations are precisely defined by the joint density functions in (4) which are conditional on past observations.

In the examples to be discussed below, the prices and times are modelled jointly. This allows measurement not only of the transaction rate but its interaction with volatility. The relation between volume and volatility has received vast attention from both a theoretical and empirical point of view. It has been addressed as simply a data analysis issue, or as a component of theories of market microstructure. Sometimes the empirical models are developed for transactions data directly, but more often they are based on fixed interval aggregates.

A popular approach to this analysis is through models of time deformation where the relevant time scale is “economic time” rather than “calendar time.” Intuitively, economic time measures the arrival rate of new information which influences both volume and volatility. The joint analysis of transaction times and prices generalizes the standard time deformation models by obtaining a direct measure of the arrival rate of new information and then measuring exactly how this influences the distribution of the other observables in the market.

III. Econometric Issues

The econometric issues in applying these techniques are specifying and testing the parameterizations of the functions g and q in equation (4) since the relevant economic questions can all be determined from these functions.

The approach developed here is maximum likelihood with all its associated parametric inference and testing procedures. In the subsequent section, a semiparametric approach to hazard estimation will be presented.

In order to estimate parameters by maximum likelihood, it is necessary to formulate the process so that there are a finite number of parameters θ which are invarient over events. The log likelihood function is simply the sum of the logs of all the N individual joint densities conditional on the past and can therefore be written as:

$$\mathcal{L}(X,Y;q)=\sum_{i=1}^{N}\log f(x_i,y_i|\breve{x}_{i-1},\breve{y}_{i-1};q) \tag{7}$$

where X and Y are all the data and θ∈Θ is the set of parameters.

In principle such a likelihood can be derived from a model where there are unobserved latent processes. For example, some parameters may actually be unobserved stochastic processes. Initially suppose that there is a series {ϕ} which is unobserved by the econometrician but which is assumed to follow a probability law with conditional density given by p with unknown parameters which are included in θ. The conditional density of the observables can be expressed in terms of the different filtrations as follows:

$$\begin{aligned} & f_i \left| \mathrm{F}_{i-1}^{(x,y)} \cup \mathrm{F}_{i-1}^{f} \right. \sim p(f_i|\breve{x}_{i-1},\breve{y}_{i-1},\breve{f}_{i-1};q) \\ & (x_i,y_i)\left| \mathrm{F}_{i-1}^{(x,y)} \cup \mathrm{F}_{i-1}^{f} \right. \sim f^{*}(x_i,y_i,f_i|\breve{x}_{i-1},\breve{y}_{i-1},\breve{f}_{i-1};q)\,p(f_i|\breve{x}_{i-1},\breve{y}_{i-1},\breve{f}_{i-1};q) \end{aligned} \tag{8}$$

To obtain the density of (x,y) conditional only on observables, the density in (8) must be integrated with respect to $\{f_i, i=1,\ldots,N\}$ leaving a likelihood as in (7) with only the fixed parameters θ. In practice, the multidimensional integral is very difficult to evaluate and requires sophisticated Monte Carlo methods. See for example Poulson, Jacquier and Rossi(1994) for a volatility process and Shephard(1993) for more general problems. Furthermore, it is not clear that it is easier to specify the processes in (8) than the processes in (2) from *a priori* considerations.

The specification of the conditional density of the durations given covariates is a familiar problem in statistics and biostatistics. Much of this literature is focussed on the treatment of censored durations which are endemic in survival analysis. In the transactions analyses here, there are no censored durations, so the range of specifications which can be considered is more generous. However, since the focus is on the temporal dependence of the durations, the covariates are typically going to be lagged dependent variables and functions of lagged dependent variables.

Engle and Russell(1995a,b) propose a specification of the conditional density which requires only a mean function. They define ψ as the conditional duration given by

(9) $$y_i \equiv y(\breve{x}_{i-1}, \breve{y}_{i-1}; q) = E_{i-1}(x_i | \breve{x}_{i-1}, \breve{y}_{i-1}; q) = \int s g(s | \breve{x}_{i-1}, \breve{y}_{i-1}; q) ds$$

and then assume that

(10) $$\frac{x_i}{y_i} \equiv \tilde{x}_i \sim i.i.d.$$

This assumption insures that all the temporal dependence in the durations is captured by the mean function. Such an assumption is testable in the sense that the standardized durations can be checked for various forms of deviation from independence or identical distribution. Under this assumption,

(11) $$g(x_i | \breve{x}_{i-1}, \breve{y}_{i-1}; q) = g(x_i | y_i; q)$$

where possibly there are no remaining unknown parameters in g which are not already in ψ. Because the x's are now conditionally independent, the likelihood is easily evaluated. The assumption in (10) is powerful and is incorporated in some but not all familiar specifications for hazard models.

For example, the familiar log linear regression models or accelerated failure time models can be specified in terms of covariates z as:

(12) $$\log x_i = z_i b + w_i$$

where the error density is independent and identically distributed over observations and does not depend upon b. Thus

(13) $$x_i = v_i \exp z_i b$$

where v_i is the exponential of w and is therefore positive. The expected value of x conditional on the covariates will be proportional to exp(zβ) so that x/ψ will be independent and identically distributed as assumed in (10). If the hazard for v is l_0, then the hazard for x is simply

(14) $$l(t, z) = l_0(t e^{-zb}) e^{-zb}$$

indicating that the covariates not only multiply the hazard, but also adjust the rate at which the individual passes through the period. The name “accelerated failure time” has intuitive appeal in medical applications where the patient is viewed as progressing through the disease at a faster or slower pace depending upon the covariates. A similar interpretation has been often used in finance where it is hypothesized that “economic time” sometimes moves faster or slower than calendar time. Such models are described as time deformation models and will be compared with the ACD model.

An alternative popular specification is the proportional hazard model (Cox(1972)) in which the baseline hazard is simply multiplied by the function of the covariates giving rise to the specification:

$$l(t,z) = l_0(t)e^{-zb} \tag{15}$$

This model does not in general satisfy assusmption (10) although it does when the baseline hazard is constant as in the exponential or when it is proportional to t^r as in the Weibull as shown by Kalbfleisch and Prentice(1980,p34).

Under the specification (11), the log likelihood can be expressed as:

$$\ell(X,Y;q) = \sum_{i=1}^{N}[\log g(x_i|y_i;q_1) + \log q(y_i|x_i,\breve{x}_{i-1},\breve{y}_{i-1};q_2)] \tag{16}$$

which can be maximized with respect to the unknown parameters (θ_1,θ_2). If in addition one can assume that the marks are weakly exogenous for the parameters of interest which might be taken to be θ_1, then the joint estimation is not required and the parameters can be estimated with no loss of efficiency simply by maximizing the first term. Similarly, if the parameters of interest are in θ_2, then weak exogeneity of x justifies maximizing simply the second term.

IV. Semiparametric Hazard Estimation

With simply the assumption (10) and correct specification of (9), it is desirable to find an estimate of the hazard function for x. This is called a semiparametric estimate of the hazard because it does not require parameterizing the density of x but does require specifying the mean of x. It is proposed to maximize the quasi-likelihood function

$$Q(X,q) = -\sum_{i=1}^{N}[\log y_i + x_i / y_i] = -\sum_{i=1}^{N}\ell_i \tag{17}$$

which would be the true log likelihood function if g were the exponential density. Define the score, hessian and expected hessian of **l** as s, h and a respectively:

$$
\begin{aligned}
& s_i = \frac{\partial l_i}{\partial \theta} = -\frac{(\psi_i - x_i)}{\psi_i^2}\frac{\partial \psi_i}{\partial \theta} \\
(18) \qquad & h_i = \frac{\partial^2 l_i}{\partial \theta \partial \theta'} \\
& a_i = E_{i-1}(h_i) = -\frac{1}{\psi_i^2}\frac{\partial \psi_i}{\partial \theta}\frac{\partial \psi_i}{\partial \theta'}
\end{aligned}
$$

A QMLE estimator can then be established which is consistent for the parameters and which has a well defined asymptotic covariance matrix. This result was first pointed out by Gourieroux, Monfort, and Trognon(1984) for a closely related problem, and follows the QMLE results for ARCH models given in Bollerslev and Wooldridge(1992).

THEOREM 1: Under the following conditions:

1) For some $\theta_0 \in \text{int}\, \Theta$ a compact parameter space

$$E(x_i | F_{i-1}; \theta_0) = \psi_i(\theta_0)$$

2) The random variable $\dfrac{x_i}{\psi_i(\theta_0)}$ is a Martingale difference sequence

3) A set of regularity conditions nearly identical to those in Bollerslev and Wooldridge

Then if $\hat{\theta}_N$ maximizes (17) over Θ, then

$$\left[A_N^{o-1} B_N^o A_N^{o-1}\right]^{-1/2} \sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, I)$$

where

$$A_N^o = N^{-1}\sum_{i=1}^{N} E[a_i(\theta_0)], \quad B_N^o = N^{-1}\sum_{i=1}^{N} E[s_i(\theta_0) s_i(\theta_0)'].$$

Furthermore:

$$\hat{A}_N - A_N^o \xrightarrow{p} 0, \quad and \quad \hat{B}_N - B_N^o \xrightarrow{p} 0$$

where

$$\hat{A}_N = N^{-1}\sum_{i=1}^{N} [a_i(\hat{\theta}_N)], \quad \hat{B}_N = N^{-1}\sum_{i=1}^{N} [s_i(\hat{\theta}_N) s_i(\hat{\theta}_N)']$$

This theorem supports estimation and inference by QMLE assuming an exponential density. The result of such an estimate is a set of conditional durations for each observation in the sample. By assumtion (10), the ratio of the realized duration to its expectation is i.i.d. following some density g(). From this density, one can empirically estimate a hazard function $\hat{\lambda}_0()$ called the baseline hazard and then compute the hazard for x from

$$(19) \qquad \hat{\lambda}_{x_i}(t) = \hat{\lambda}_0(t/\hat{\psi}_i)/\hat{\psi}_i$$

There are many ways to empirically estimate the hazard for the standardized durations. As there is no censoring or truncation in these data sets, various approaches are available. One can estimate the density non-parametrically, compute the survivor function from it and then take a ratio, or one can calculate the sample hazard function and smooth it. Here we employ a version of the latter estimator which is essentially a k-nearest neighbor estimator. When k is chosen as 1, this is a Kaplan Meier type estimate. Since the durations are continuous and have few mass points, it is necessary to smooth this estimator by choosing a wider bandwidth.

Consider the failure rate of the smallest 2k standardized durations. The time interval for this set of failures is therefore $(0,t_{2k})$ and the estimate of the hazard is the number of failures divided by the time interval times the number at risk. In general, let n_i be the number of individuals surviving at time t_i, then the 2k nearest neighbor estimate of the hazard rate is computed from

$$\hat{l}(t_i) = \frac{2k}{n_i(t_{i+k} - t_{i-k})} \tag{20}$$

The nearest neighbor estimator gives a variable bandwidth as the number of individuals at risk varies over the time axis. This is exactly the estimate which would be obtained if a nearest neighbor approach to estimating the density of the durations was then used to define the survival function and hazard. If the true density is exponential, then the hazard is constant and the hazard estimates should be unbiased.

V. Estimating the Hazard for IBM Trades

Data on all trades for a random collection of stocks on the NYSE is available from the exchange on the TORQ database which stands for Trades, Orders and Quotes. The sample period runs from November 1990 through January 1991. For IBM there are approximately 60,000 trades during this period although in this paper only 15,000 observations will be examined. Engle and Russell(1995a) analyzed both the first and second set of 15,000 observations obtaining quite similar estimates even though the first set included the Friday after Thanksgiving which was not only extraordinarily slow but which also included a computer failure. Here only the second data set is examined.

Following E&R, the data is first “seasonally adjusted” to take out the typical time of day effect. This is accomplished by regressing the durations on the time of day using a piecewise linear spline specification and then taking ratios to get “seasonally adjusted” durations which are expressed as fractions above or below normal. While this could be done in one step, there is little to be gained in such a large data set. The range of the adjusted data is from .025 to 19.9 or 1/40 of normal to 20 times normal. The standard deviation is 1.33.

The adjusted data show striking evidence of autocorrelation with a Ljung-Box statistic with 15 degrees of freedom of 1209 and all 15 autocorrelations between .055 and .1 indicating a small but persistent signal. The model estimated is the ACD(1,1) expressed as

(21) $$y_i = w + a x_{i-1} + b y_{i-1}$$

where the density is assumed to be exponential. The estimates are therefore interpreted as QMLE estimates as they are consistent regardless of the true density. The estimates are presented below with the robust standard errors computed as in Theorem 1.

Table 1
Estimates for the ACD(1,1) for IBM Trades

	Coeff.	St. Error	T-Stat	Robust St.Err	Robust T-Stat
ω	.011	.0011	9.66	.0017	6.46
α	.056	.0023	24.2	.0034	16.5
β	.933	.0028	325	.0043	216

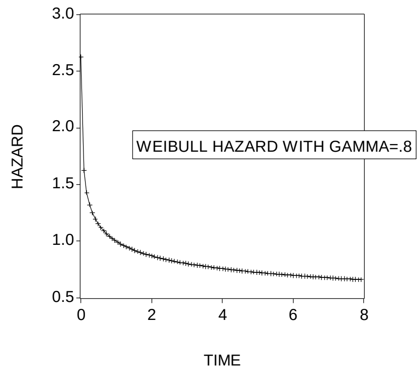
The standardized durations from this model, given by

(22) $$\tilde{x}_i = x_i / y_i$$

show very little evidence of autocorrelation with a Ljung Box statistic of 17.7 which is well below the 5% point of 25 and six of 15 autocorrelations positive with a maximum absolute value of .014. There is no evidence of time varying heteroskedasticity; the Ljung Box statistic on the squares is only 20.6! There is very little evidence against serial independence for this set of 15,000 observations. In E&R, further examination of this result indicated that economic variables could improve the fit of this model but this line is not pursued here.

The standardized durations are used to compute a density and a semiparametric hazard following the approach of section IV. The empirical density with 50 nearest neighbors on each side is plotted in Figure 1 with the hazard in Figure 2. There is strong evidence of a sharp drop in the hazard for very small durations after which it only gradually declines. Such a picture is broadly consistent with a Weibull density with parameter less than unity as was found in E&R. The hazard in this case is proportional to $1/x^{.2}$ which is consistent with the sharp drop for very small durations.

The actual shape of the Weibull hazard when γ=.8 is however not as abrupt as in the figures. This can be seen in the following plot.



VI. Estimating Price Volatility with Transaction Data

The most important mark which is available with each trade is the price at which it occurred. These prices convey information about the volatility of the market and about a variety of market micro structure hypotheses. In this section, a preliminary analysis of price data corresponding to the IBM trades will be undertaken. The analysis focusses on the relation between the timing of trades and the volatility of prices.

Since the time between trades is the reciprocal of the transaction rate which is very highly correlated with volume, this study can draw on both the vast theoretical and empirical literature on the relation between volatility and volume. Much of the empirical literature is based on aggregated data which shows that there is a strong contemporaneous relation between volume and volatility, e.g. see Lamoureux and Lastrapes(1990), Gallant, Rossi and Tauchen(1992), and a survey by Karpoff(1987), although the predictive information in volume is much less clear.

A theoretical construct which is often used in modelling both volume and volatility is time deformation. Following the original ideas of Clark(1973) and Tauchen and Pitts(1983), the relation between economic time and calendar time is specified either as a latent process or as a function of observables. For example, Ghysels and Jasiak(1994) propose having time pass as a function of quote arrival rates while Müller et al.(1990) use absolute quote changes and geographical information on market closings.

With transactions data, models are typically estimated in transaction time without explicit account of the calendar time. For example see Hasbrouck(1988,1991) and Harris(1986). Hausman Lo and MacKinlay(1992) introduce the duration of the last trade as an exogenous explanatory variable but do not discuss its implications. Pai and Polasek(1995) treat time as exogenous but allow the parameters of the process to depend upon durations in simple ways.

The market microstructure approach to models of time and volume are more useful as starting points. O'Hara(1995) points out that "if market participants can learn from watching the timing of trades, then the adjustment of prices to information will also depend on time." It is useful to formulate the price adjusment process in terms of the asymmetric information models introduced originally by Glosten and Milgrom(1985). In these models, the specialist is faced with setting bid and asked prices to use when trading with individuals who may be better informed than he is. In the original structure, there is new information which is revealed to a fraction of the agents, while the rest will buy or sell with a probability which is unrelated to the new information. When the specialist faces a seller he offers a price which incorporates the probability that this trader is informed and the probability that the information is bad. This inference is by Bayesian updating. The more informed traders there are the lower the bid, and the higher the fraction of sellers, the lower the bid. Eventually, the specialist discovers the private information and sets bids and asks at the true value. If the informed traders act distinctively, such as selling large volumes, then the specialist will rapidly reduce his bid.

Several extensions of this model are relevant for this analysis. Diamond and Verrecchia(1987) note that sellers who do not own the stock must therefore short sell. If some fraction of the informed traders are prohibited from short selling, then they cannot profit from their information and when offered a chance to trade will not trade at the existing prices. Thus non-trades are evidence that the news may be bad so the specialist learns from the time between trades and lowers his prices and increases his spread. This model can be summarized as *no trade means bad news*.

An alternative model is in Easley and O'Hara(1992) where there is also uncertainty as to whether there even was an information event. Again a fraction of the agents are informed and thus know whether there was news or not. When it is their turn to trade, they will generally decline so that long intervals between trades is interpreted by the specialist as evidence that there is no relevant news. The specialist therefore keeps prices relatively stable if the trading intervals are long and reduces the bid asked spread. This model can be interpreted as *no trade means no news.*

The goal of the analysis at this point is to determine a measure of the instantaneous volatility using transaction data and discover how the timing of trades influences this volatility. The prices are assumed to be best measured by the midquote which is the average of the bid and asked price at the time of the transaction. This choice of price measure reduces the econometric issues of bid asked bounce.

The model is formulated with the durations depending on past information following an ACD type of model, and with prices following a GARCH type of process. However, the volatility specification is now conditional not only on past information, but also on the current duration. Such a measure of conditional volatility is no longer a prediction but only becomes a prediction when integrated over durations using the density estimated as part of the ACD.

To derive the variance process, let σ_i^2 be the conditional variance of the process per second and let h_i be the conditional variance of the i^{th} transaction. Both are condional on the current duration as well as past information. These two variances are related by the definition:

(23) $$h_i = x_i s_i^2$$

so that a pair of prices separated by zero time will have zero variance. This is in fact correct in the data since at any point in time only one set of quotes can be current. The longer the time, the bigger the variance of the transaction if the underlying variance remains constant. If for example, σ depends only on the past, then the expected variance over the next trade is simply given by

(24) $$E_{i-1}(r_i^2) = E_{i-1}(h_i) = E_{i-1}(x_i s_i^2) = y_i s_i^2$$

where r_i is the log change in the midquote from transaction i-1 to transaction i. A simple GARCH specification where current durations are not informative assumes:

$$(25) \qquad s_i^2 = w + a\frac{e_{i-1}^2}{x_{i-1}} + bs_{i-1}^2$$

which is a GARCH(1,1) modified to take account of the irregular trading intervals. This model can be rewritten as

$$(26) \qquad h_i = x_i\left[w + a\frac{e_{i-1}^2}{x_{i-1}} + b\frac{h_{i-1}}{x_{i-1}}\right]$$

which is the expression which appears in the likelihood function.

It is clear that this model does not yet recognize the possibility that variations in x and variations in σ could be related to the same news events. A more realistic and a more interesting specification which also happens to have the highest likelihood of all the models tried, is

$$(27) \qquad s_i^2 = w + a\frac{e_{i-1}^2}{x_{i-1}} + bs_{i-1}^2 + g_1\frac{x_{i-1}}{y_{i-1}} + g_2\frac{x_i}{y_i} + g_3 x_i + g_4 y_i^{-1}$$

where ξ_i is the long run volatility computed by exponentially smoothing r^2 with a parameter .995. That is:

$$(28) \qquad x_i = .005 r_{i-1}^2 + .995 x_{i-1}$$

Presumably, optimizing this parameter could give even better results although that was not done here. The half-life of this smoother is 138 trades. This exponentially smoothed estimate of volatility is computed in transaction time so it is here divided by ψ in order to convert it to a per second volatility.

Because there is some serial correlation in the returns, an autoregressive moving average process of order (1,1) was estimated. In addition, the trade interval was included in the mean to allow a drift in returns or to allow the bad news effect of long durations to reduce prices.

The impact of durations on volatility is incorporated in three coefficients which measure the effects of surpises in durations, actual durations, and expected durations respectively. These allow a rather interesting pattern of responses.

The results are presented in Table 2 along with robust standard errors as discussed by Bollerslev and Wooldridge(1992). The first two rows correspond to the variables which appear in the mean while the rest are in the variance.

TABLE 2
ESTIMATES OF VOLATILITY σ^2 FOR (27)
Log Likelihood=-18852

	coef.	st.err	t-stat	rob err	rob t-stat
			MEAN		
AR(1)	.51	.009	54	.06	8.2
MA(1)	-.66	.008	-83	.054	-12
x	-.005	.007	-7.3	.003	-1.6
			VARIANCE		
const	.60	.008	71	.12	4.9
e_{i-1}^2 / x_{i-1}	.38	.007	52	.045	8.2
s_{i-1}^2	.38	.006	59	.041	9.3
x_{i-1} / y_{i-1}	.40	.011	38	.052	7.8
x_i / y_i	-.07	.002	-33	.007	-10.7
x_i	-.009	.002	-4.5	.007	-1.4
$1/y_i$	-.22	.008	-26	.082	-2.7

This model reveals strong autocorrelation in the mean through the highly significant AR and MA coefficients. This is a familiar result in transaction data sets because of bid asked bounce in transaction prices. It is also familiar in midquote models such as Hasbrouck(1991) which is probably due to the discreteness in the quoted prices as well as the occasional erroneous quote. The Ljung Box statistic for the original data with 15 lags is 1339, while the residuals reduce it to 658 and the standardized residuals fall to 88. Nevertheless this has a very small p value. The autocorrelations are still primarily negative.

The mean is expressed as a function of the duration of the trade. This should be the drift in returns but also provides evidence of the bad news effect of long durations. Interestingly this is negative supporting the Diamond and Verrecchia model but at least the robust t-statistic is not very significant.

More interesting results are found in the variance equation. There is strong evidence of time varying volatility as the Ljung Box on the squared returns is 1253 and on the squared residuals is 1122 so that the fitting of the mean coefficients has little impact on the volatility. The final squared standardized residuals have a Ljung Box of 34 which has a p value of .006 which may not be too small for a sample of 15,000 observations. There is however no reason why these should be serially uncorrelated since they are conditioned on current durations as well as on the past. Only when conditioned on a filtration of the observables should this be white noise.

The coefficients of the conventional GARCH model α and β sum to just over .75 which is very small for such a high frequency data set. This result is also familiar in the papers of Kroner and Lee(1995) and Andersen and Bollerslev(1995) where it is found that the persistence of a GARCH model drops dramatically with intradaily data. As shown in the Appendix, the plain vanilla GARCH(1,1) shows similar features.

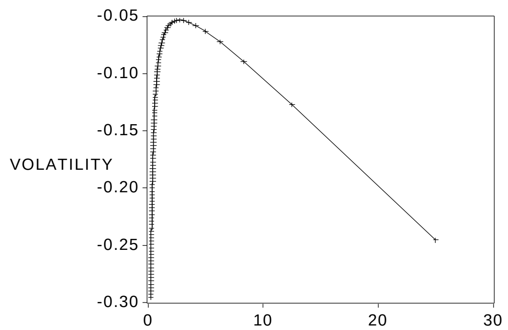
The remaining four variables illuminate the nature of high frequency volatility. The first reveals that there is some substantial persistence in volatility since the slowly decaying exponential smoother is an important explanatory variable.

Duration enters by itself and divided by its expectation. In both cases, longer durations lead to lower volatilities. This finding supports the Easley and O'Hara formulation in which no trade is interpreted as no news so that volatility is reduced.

The two terms have rather different impacts for forecasting. Taking expectation of equation (27) conditional on the past gives:

$$(29) \qquad E_{i-1}(\sigma_i^2) = \omega + \alpha \frac{\varepsilon_{i-1}^2}{x_{i-1}} + \beta \sigma_{i-1}^2 + \gamma_1 \frac{x_{i-1}}{\psi_{i-1}} + \gamma_2 + \gamma_3 \psi_i + \gamma_4 \psi_i^{-1}$$

so that γ_2 which is highly significant has no persistence. On the other hand, γ_3 has a long term impact on future volatility since durations are themselves persistent. In this form it is now clear that there is a non-linear response to ψ where both coefficients are negative. Plotting this relation shows that high expected duration leads to lower volatility while very low expected durations also promise low volatilities.



VOLATILITY AGAINST EXPECTED DURATION

The response to high volatility merely supports the intuition that information in transaction rates will help to predict future volatilities. When durations are very short, it is possible that many of the transactions are actually trades which have been broken up and thus do not represent separate events. This could explain the finding that when the market is expected to be very active, volatility is lower than would otherwise be expected.

Several other simpler models have been estimated for this data set. These are given in Table 4. In each case there is an ARMA(1,1) in the mean and an intercept. The first model is a plain GARCH(1,1) in transaction time while the second is a similar plain GARCH(1,1) but translated into calendar time. The latter has a somewhat better fit but the differences are not very large. The second is theoretically more interesting so it is retained in the preferred specifications given above.

TABLE 3
MORE MODELS FOR VARIANCES
(each model has ARMA(1,1) in mean)

Coef.	*EQN.1*	*EQN.2*	*EQN.3*
x_i^{-1}	.16		
	.019		
e_{i-1}^2 / x_i	.14		
	.013		
s_{i-1}^{2} / x_i	.67		
	.030		
const		.56	.34
		.060	.039
e_{i-1}^2 / x_{i-1}		.38	.37
		.042	.045
s_{i-1}^{2}		.47	.41
		.037	.038
x_{i-1} / y_{i-1}			.39
			.058
x_i / y_i			-.081
			.0049
Log Lik	-19184	-19172	-18871

VII. Conclusions

This paper has introduced a framework to estimate models when the data arrive at random intervals when these intervals themselves may carry information. The basic procedure is to model the associated variables called marks conditional on the times, and to separately model the times. Under some assumptions, there is no loss of information by this two step procedure.

In this example, 15,000 IBM stock transactions are analyzed to find a model of the timing of trades and then to measure the impact of this timing on the price volatility. The ACD model introduced by Engle and Russell(1995) is used to estimate the dependent point process for the arrival rates. A semiparametric approach to estimating the hazard function is introduced and applied.

Finally, the price quotes are examined to obtain models of volatility conditional on transaction times. Here it is found that volatility has a short and a long run component and that longer durations are associated with lower volatilities as predicted by the Easley and O'Hara model. In a predictive sense, very long expected durations have a negative impact on expected volatility, but short expected durations also have a negative volatility impact.

REFERENCES

Clark, P.K.,(1973), " A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices", *Econometrica,*41, pp135-156

Cox, D.R.,(1972), "Regression Models and Life Tables (with discussion)", *J.R. Stat.Soc. B. 34, pp 187-220*

Diamond, D.W. and R.E.Verrecchia,(1987), "Constraints on Short-selling and Asset Price Adjustments to Private Information", *Journal of Financial Economics*,18,pp277-311

Easley, D. and M. O'Hara,(1992)," Time and the Process of Security Price Adjustment", *Journal of Finance*,47,pp905-927

Engle, R.F. and J.E. Russell,(1995a) "Forecasting Transaction Rates: The Autoregressive Conditional Duration Model," UCSD Discussion Paper

_________(1995b) Forecasting the Frequency of Changes in Quoted Foreign Exchange Prices with the ACD Model*, in High Frequency Data in Finance, Proceedings,* Olsen and Associates, Zurich Switzerland.

Gallant, R., P.E. Rossi, and G. Tauchen(1992) "Stock Prices and Volume", *Review of Financial Studies,*5,pp871-908

Ghysels, E. and J. Jasiak(1995), "Trading Patterns, Time Deformation and Stochastic Volatility in Foreign Exchange Markets", *in High Frequency Data in Finance, Proceedings,* Olsen and Associates, Zurich Switzerland.

Glosten, L., and P. Milgrom,(1985)," Bid, Ask, and Transaction Prices in a Specialist Market with Hetrogeneously Informed Traders," *Journal of Financial Economics*,13, pp 71-100

Gourieroux, C., Monfort,A., and Trognon,A.,(1984), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models", *Econometrica*; 52, pages 701-20.

Harris, L.(1986), " A transaction Data Study of Weekly and Intradaily Patterns in Stock Returns," *Journal of Financial Economics*, 16, pp99-117

Hasbrouck, J.(1988), "Trades, Quotes, Inventories and Information", *Journal of Financial Economics*, 22, pp229-252

___________(1991), " Measuring the Information Content of Stock Trades," *Journal of Finance*,46,pp179-207

Hausman, J.A., A.W.Lo, and A.C. MacKinlely,(1992), “An Ordered Probit Analysis of Transaction Stock Prices”, *Journal of Financial Economics*, 31, pp319-330

Jacquier,E., Polson,N.G., and Rossi, P.E.,(1994), “Bayesian Analysis of Stochastic Volatility Models”, *Journal of Business and Economic Statistics*; 12(4), pages 371-89.

Kalbfleish, J.D. and R.L Prentice,(1980), *The Statistical Analysis of Failure Time Data*, Wiley, New York

Karpoff,J.(1987), “The Relation Between Price Change and Trading Volume: A Survey”, *Journal of Financial and Quantitative Analysis*, 22, March, 109-126

Lamoureux, C. and W. Lastrapes(1990), “Heteroskedasticity in Stock Return Data: Volume Versus GARCH Effect”, *Journal of Finance*,45, pp221-229

Liu, T.C.,(1963), “An Exploratory Quarterly Econometric Model of Effective Demand in the Postwar U.S. Economy”, *Econometrica*, 31,pp.301-348

________(1969), “A Monthly Recursive Model of the U.S. Economy”, *Review of Economics and Statistics*

Müller, U.A., M.M. Dacorogna, R.B.Olsen, O.V. Pictet, M. Schwarz and C. Morgenegg(1990), “Statistical Study of Foreign Exchange Rates. Empirical Evidence of a Price Change Scaling Law and Intraday Analysis,” *Journal of Banking and Finance,*14,pp1189-1208

O’Hara, M.(1995), *Market Microstructure Theory*, Basil Blackwell, Oxford, England

Pai, J.S. and W. Polasek,(1995), “Irregularly Spaced AR and ARCH (ISAR-ARCH) Models”, *in High Frequency Data in Finance, Proceedings*, Olsen and Associates, Zurich Switzerland.

Shephard, N.E. (1993), “ Fitting Nonlinear Time-Series Models with Applications to Stochastic Variance Models”, *Journal of Applied Econometrics*; 8, Suppl., pages 135-52.

Tauchen, G.E. and M. Pitts,(1983), “The Price Variability Volume Relationship on Speculative Markets”, *Econometrica*, 51, pp485-505