Modeling High-Frequency FX Data Dynamics*

Òscar Jordà
University of California, Davis
ojorda@ucdavis.edu

Massimiliano Marcellino†
Università Bocconi
massimiliano.marcellino@uni-bocconi.it

February, 2002

Abstract

This paper shows that high-frequency, irregularly-spaced, FX data can generate non-normality, conditional heteroskedasticity, and leptokurtosis when aggregated into fixed-interval, calendar time even when these features are absent in the original D.G.P. Furthermore, we introduce a new approach to modeling these high-frequency irregularly spaced data based on the Poisson regression model. The new model is called the autoregressive conditional intensity (ACI) model and it has the advantage of being simple and of maintaining the calendar time scale. To illustrate the virtues of this approach, we examine a classical issue in FX microstructure: the variation in information content as a function of fluctuations in the intensity of activity levels.

- *JEL Classification Codes: C43, C22, F31*

*The paper has benefited from the valuable comments of an anonymous Associate Editor. We also thank Rob Engle, Tom Rothenberg, Jim Stock and seminar participants at the European University Institute, Harvard University, the Midwestern Econometrics Group, U.C. Berkeley, U.C. Davis and U.C. Riverside for useful comments. All errors remain our responsibility.

†Corresponding author: Massimiliano Marcellino, IGIER - Università Bocconi, Via Salasco 5, 20136, Milano, Italy. Phone: +39-02-5836-3327. Fax: +39-02-5836-3302.

- *Keywords:* time aggregation, irregularly-spaced high-frequency data, dependent point process.

1. Introduction

The spot foreign exchange (FX) market is an around-the-clock, around-the-globe, decentralized, multiple-dealer market characterized by an enormous trading volume (an average of $1.5 trillion of FX traded according to Lyons, 2001). Interdealer trading accounts for roughly two-thirds of this volume and unlike other, more traditional financial markets (such as the NYSE's hybrid auction/single-dealer or the NASDAQ's centralized multiple-dealer markets), FX trade data is not generally observable because there are no disclosure regulatory requirements.

Arguably, the sheer volume and peculiar organizational features of the spot FX market makes its study one of the most exciting topics of investigation in theoretical and empirical macroeconomics and finance. Thus, this paper examines the unconventional temporal properties of FX data and the effect that these properties have on typical econometric investigations of microstructure effects.

Specifically, this paper addresses two important issues: (1) to what extent are the conventional stylized facts of these high-frequency financial data (such as non-normality, leptokurtosis and conditional heteroskedasticity) attributable to the stochastic arrival in time of tick-by-tick observations, and (2) the introduction of new modeling approaches for high-frequency FX data and in particular of a new dynamic count data model, the autoregressive conditional intensity (ACI) model. To be sure, we believe that many of the observations we make concerning the FX market are not limited to these data and are generally applicable in other contexts as well.

The first of these two issues is intimately related with the extensive literature on time aggregation and time deformation and has to do with the irregular nature in which FX events arrive over time. There are a number of physical phenomena, such as temperature, pressure, volume, and so on for which sampling at finer and finer intervals would be desirable since in the limit, the sampling frequency would deliver

continuous measurements in line with the stochastic differential equations that usually describe these phenomena in continuous-time. Thus, there exists an extensive literature on sampling theory and aliasing designed to establish the minimum sampling rates necessary to identify the model. This is a classical issue in the literature on fixed-interval, time aggregation.

Nevertheless, high-frequency financial data are, in a sense, already sampled at their limiting frequency since sampling at finer intervals would not deliver any further information: there is no new activity between two contiguous observations. This unusual characteristic makes econometric modeling problematic. On one hand, continuous-time formulations stand in the face of the irregular manner in which new observations are generated. On the other hand, discrete-time models overlook the information content enclosed in the varying time intervals elapsing between observations.

The second of the issues we investigate relates more generally to modern analysis of high-frequency, tick-by-tick data. Early studies estimated models in event time, without explicit account of calendar time (see Hasbrouck, 1988, 1991 and Harris, 1986). Hausman, Lo and MacKinlay (1992) and Pai and Polasek (1995) treated time as an exogenous explanatory variable. The introduction of the autoregressive conditional duration (ACD) model by Engle and Russell (1998) represents the first direct attempt at jointly modeling the process of interest and the intervals of time between observations in a dynamic system. By contrast, we propose conducting the analysis in the usual calendar time scale and instead extract the information contained in the random intensity of event arrival per unit of calendar time – that is, the count process that represents the dual of the duration process in event time.

As an illustration of the considerable advantages of our approach, we investigate a classical issue in the microstructure literature: whether trades arriving when trade intensity is high contain more information than when this intensity is low. Inventory based models of information flow (see Lyons 2001) suggest that low intensity trades are

more informative because inventory management generates a flurry of activity designed to rebalance dealer positions. Alternatively, Easly and O'Hara (1992) argue that if there exists private information in the market, the arrival of new trades raises the probability that dealers attach to those trades containing new information. As we shall see, our investigation with quote data suggests that the story is somewhat more complicated and lies somewhere between these two explanations.

2. Temporal Properties of the FX Market

This section investigates in what manner does the temporal pattern of the FX data affect the salient statistical properties of these data. More specifically, we will suggest that many of the properties to be discussed below can be explained as artifacts of time aggregation of data that is inherently irregularly spaced in time. Thus, we begin by summarizing these stylized facts (for a survey and summary see Guillaume *et al.*, 1997), which largely motivate the typical econometric techniques used in the literature. Hence, denote price at time t as

$$\mathsf{x}_t \equiv \log(fx_t), \tag{2.1}$$

where fx denotes exchange rate quotes or prices (as the data may be), and t refers to a calendar-time interval during which k_t observations (or ticks) of the variable fx have elapsed. Then, if we denote τ as the operational time scale in which observations arrive, we have that the correspondence between calendar-time t and operational-time τ is given by

$$\tau = \varphi(t) = \varphi(k_t) = \sum_{j=1}^{t} k_j \qquad for \quad k = \{k_t\}_{t=1}^{\infty}. \tag{2.2}$$

Hence, k denotes the frequency of aggregation so that $\varphi(t) - \varphi(t-1) = k_t$, that is, the number of operational time observations per sampling interval $(t-1, t]$. In traditional

fixed-interval aggregation, such as aggregation of monthly data into quarters, k is a fixed number (specifically, $k = 3$ for this example). However, FX data arrive at random intervals so that k_t is best thought of as a stochastic point process.

For simplicity, we do not distinguish here between "asks" and "bids" in which case, x_t is typically taken to be the average of the log ask and log bid quotes. The *change of price* or *return* is defined as:

$$r_t \equiv [\mathsf{x}_t - \mathsf{x}_{t-1}]. \tag{2.3}$$

The volatility associated with this process is defined as

$$v_t \equiv \frac{1}{k_t} \sum_{j=1}^{k_t} |r_{\tau-j}|, \tag{2.4}$$

where k_t corresponds to expression (2.2). The absolute value of the returns is preferred to the more traditional squared value because it captures better the autocorrelation and seasonality of the data (see Taylor, 1988; Müller *et al.*, 1990; Granger and Ding, 1996). Although there are other quantities of interest (such as the relative spread, the tick frequency, and the volatility ratio), these are more fundamental variables of interest. These variables display the following stylized characteristics:

1. The data is non-normally distributed with "fat tails." However, temporal aggregation tends to diminish these effects. At a weekly frequency, the data appears normal.

2. The data is leptokurtic although temporal aggregation reduces the excess kurtosis.

3. Seasonal patterns corresponding to the hour of the day, the day of the week and the presence of traders in the three major geographical trading zones (East Asia, Europe and America) can be observed for returns and particularly for volatility (see Dacorogna *et al.*, 1993, and 1996).

4. Let the *scaling law* reported in Müller *et al.* (1990) be defined as:

$$|\overline{x_\tau - x_{\tau-1}}| = \left(\frac{\Delta\tau}{m}\right)^D, \tag{2.5}$$

where m is a constant that depends on the FX rate and $D = 1/E$ is the drift exponent. For a Gaussian random walk, the theoretical value of $D = 0.5$. However, it is observed that $D \simeq 0.58$ for the major FX rates. The scaling law holds with a similar value of D for volatility.

5. Volatility is decreasingly conditionally heteroskedastic with the frequency of aggregation.

6. Seasonally filtered absolute returns data exhibits long-memory effects, that is, autocorrelations that decay at a slower than exponential rate (usually hypergeometric or even quasi-linear decay rates).

In order to investigate what mechanisms may give rise to these stylized facts, we experiment with a rather simple example. Specifically, under common forms of market efficiency, it is natural to assume that the price process of a financial asset follows a martingale. Therefore, assume that the driving process for FX prices is a random walk – a more stringent assumption than a martingale in that it does not allow dependence in higher moments. Accordingly, let

$$x_\tau = \rho x_{\tau-1} + \varepsilon_\tau \qquad \varepsilon_\tau \sim WN(0, \sigma_\varepsilon^2), \tag{2.6}$$

where the random walk condition would imply $\rho = 1$.

Consider now a simple scenario in which the frequency of aggregation is deterministic and cyclical, i.e., $k = k_1, k_2, ..., k_j, k_1, k_2, ..., k_j, ...$ This is a convenient way of capturing the seasonal levels of activity during different hours of the day, or days of the week and serves to illustrate some basic points. The (point-in-time) aggregated process resulting

from (2.6) and the frequency of aggregation described above result in a time-varying seasonal AR(1):

$$\begin{aligned}
\mathsf{x}_t &= \rho^{k_1}\mathsf{x}_{t-1} + u_t \qquad u_t \sim (0, \sigma^2_{u,t}), \\
\mathsf{x}_{t+1} &= \rho^{k_2}\mathsf{x}_t + u_{t+1} \qquad u_{t+1} \sim (0, \sigma^2_{u,t+1}), \\
&\ldots \\
\mathsf{x}_{t+j-1} &= \rho^{k_j}\mathsf{x}_{t+j-2} + u_{t+j-1} \qquad u_{t+j-1} \sim (0, \sigma^2_{u,t+j-1}), \\
\mathsf{x}_{t+j} &= \rho^{k_1}\mathsf{x}_{t+j-1} + u_{t+j} \qquad u_{t+j} \sim (0, \sigma^2_{u,t}), \\
&\ldots,
\end{aligned} \tag{2.7}$$

where the errors are uncorrelated and have variances, $\sigma^2_{u,t+(i-1)} = (1 + \rho^2 + ... + \rho^{2(k_i-1)})\sigma^2_{\varepsilon}$, $i = 1, ..., j$, and t is measured in small intervals of calendar time (such as one hour, say). Further calendar-time aggregation by point-in-time sampling (as is sometimes done to avoid intra-day seasonal patterns) with $\widetilde{k} = \sum_{i=1}^{j} k_i, \sum_{i=1}^{j} k_i, ...,$ yields the constant parameter AR(1) process

$$\mathsf{x}_T = \rho^{\widetilde{k}}\mathsf{x}_{T-1} + e_T \qquad e_T \sim WN(0, \sigma^2_e), \tag{2.8}$$

with $\sigma^2_e = \sum_{i=0}^{j-1} \rho^2 \sum_{l=0}^{i} {}^{k_l}\sigma^2_{u,t-i}$, $k_0 = 0$. The time scale T now refers to larger intervals of calendar-time (e.g. days or weeks) relative to the calendar-time intervals given by t.

In addition, note that most of the stylized facts described at the top of this section refer to the first differences of the variables, and therefore, we also derive their generating mechanism. From (2.7) and after some rearrangements, we get:

$$\begin{aligned}
\Delta\mathsf{x}_{t+1} &= \frac{\rho^{k_2}-1}{\rho^{k_1}-1}\rho^{k_1}\Delta\mathsf{x}_t + u_{t+1} - \left(\frac{\rho^{k_2}-1}{\rho^{k_1}-1}\rho^{k_1} - \rho^{k_2} + 1\right) u_t, \\
\Delta\mathsf{x}_{t+2} &= \frac{\rho^{k_3}-1}{\rho^{k_2}-1}\rho^{k_2}\Delta\mathsf{x}_{t+1} + u_{t+2} - \left(\frac{\rho^{k_3}-1}{\rho^{k_2}-1}\rho^{k_2} - \rho^{k_3} + 1\right) u_{t+1}, \\
&\ldots,
\end{aligned} \tag{2.9}$$

that is, a time-varying seasonal ARMA(1,1) process, except for $\rho = 1$ (the model then collapses to a random walk with time-varying variance). Instead, further aggregation up to the time-scale T results in:

$$\Delta \mathsf{x}_T = \widetilde{\rho^k} \Delta \mathsf{x}_{T-1} + \Delta e_T. \tag{2.10}$$

Let us revisit then the six stylized facts at the top of the section in light of this simple manipulation:

1. Non normality of $\Delta \mathsf{x}_t$ and normality of $\Delta \mathsf{x}_T$ is coherent with the fact that u_t is a weighted sum of a smaller number of original errors (ε_τ) than e_T. The time-varying nature of (2.9) can also contribute to the generation of outliers, that in turn can determine the leptokurtosis in the distribution of $\Delta \mathsf{x}_t$.

2. (2.9) can also explain why the value of D in (2.5) is not 0.5: x_t is not a pure Gaussian random walk. It is more difficult to determine theoretically whether (2.9) can generate a value of D close to the empirical value 0.59. We will provide more evidence on this in the simulation experiment of the next subsection.

3. The long memory of $\Delta \mathsf{x}_t$ can be a spurious finding due to the assumption of a constant generating mechanism, even if particular patterns of aggregation can generate considerable persistence in the series.

4. The presence of seasonality in the behavior of $\Delta \mathsf{x}_t$ is evident from (2.9). (2.10) illustrates that this feature can disappear when further aggregating the data.

5. Conditional heteroskedasticity can also easily emerge when a constant parameter model is used instead of (2.9). That it disappears with temporal aggregation is a well known result, see e.g. Diebold (1988), but (2.10) provides a further reason for this to be the case, i.e., the aggregated model is no longer time-varying.

6. The time-scale seasonal transformations by Dacorogna *et al.* (1993, 1996) can be interpreted in our framework as a clever attempt to homogenize the aggregation frequencies, i.e., from $k = k_1, k_2, ..., k_j, k_1, k_2, ..., k_j, ...$ to $k = \widetilde{k}, \widetilde{k}, ...,$ and consist in redistributing observations from more active to less active periods. This changes the t time scale, which can still be measured in standard units of time, and makes the parameters of the $\Delta \mathsf{x}_t$ process stable over time. This transformation attenuates several of the mentioned peculiar characteristics of intra daily or intra weekly exchange rates.

In order to further investigate whether temporal aggregation alone can explain these features, we provide some simple simulations in the next subsection.

2.1. A Monte Carlo Study of FX Properties

This subsection analyzes the claims presented above and illustrates some of the theoretical results just derived via Monte-Carlo simulations. The D.G.P. we consider for the *price* series is the following operational time AR(1) model:

$$x_\tau = \mu + \rho x_{\tau-1} + \varepsilon_\tau,$$

where $\varepsilon_\tau \sim N(0,1)$. Under a strong version of market efficiency, it is natural to experiment with $\mu = 0$ and $\rho = 1$. However, we also consider $\mu = 0.000005$ and $\rho = 0.99$ to study the consequences of slight deviations from the random walk ideal. We simulated series of 50,000 observations in length. The first 100 observations of each series are disregarded to avoid initialization problems.

The operational time D.G.P. is aggregated three different ways:

1. **Deterministic, fixed interval aggregation:** This consists on a simple sampling scheme with $k_t = 100 \; \forall t$ or, if we define the auxiliary variable $s_\tau = 1$ if observation τ is recorded, 0 otherwise, then $s_\tau = 1$ if $\tau \in \{100, 200, ...\}, 0$ otherwise.

2. **Deterministic, seasonal, irregularly spaced aggregation:** Consider the following deterministic sequence that determines the point-in-time aggregation scheme:

$$\begin{cases} s_r = 1 \quad if \quad r \in \{1,2,3;26,27,28;36,37;41,42;56,57,58;76,77\} \\ s_r = 0 \qquad otherwise \end{cases},$$

and $s_{r+100n} = s_r$ for $r \in \{1,2,...,100\}$ and $n \in \{1,2,...\}$. In other words, the aggregation scheme repeats itself in cycles of 100 observations. Within the cycle there are periods of high frequency of aggregation and low frequency of aggregation that mimic the intensity in trading typical of the FX market. Note that from the sequence $\{s_\tau\}_{\tau=1}^{50,000}$ it is straight forward to obtain the sequence $\{k_t\}_{t=1}^{T}$. For example, the first few terms are: 1, 1, 23, 1, 1, 8, ...

3. **Random, seasonal, irregularly spaced aggregation:** Let $h_\tau \equiv P(s'_\tau = 1)$ which can be interpreted as a discrete time hazard.[1] Accordingly, the expected duration between recorded observations is $\psi_\tau = h_\tau^{-1}$. Think of the underlying "innovations" for the process that generates s'_τ as being an i.i.d. sequence of continuous-valued logistic variables denoted $\{v_\tau\}$. Further, suppose there exists a latent process $\{\lambda_\tau\}$ such that:

$$P(s'_\tau = 1) = P(v_\tau > \lambda_\tau) = (1 + e^{\lambda_\tau})^{-1}.$$

Notice, $\lambda_\tau = log(\psi_\tau - 1)$. Hamilton and Jordà (2002) show that one can view this mechanism as a discrete-time approximation that generates a frequency of aggregation that is Poisson distributed. For the sake of comparability, we choose λ_τ to reproduce the same seasonal pattern as in bullet point 2 but in random time.

[1] We use the notation s'_τ to distinguish it from its deterministic counterpart introduced in bullet point 2.

Accordingly:

$$\lambda_\tau = \lambda - 1.5\lambda s_\tau,$$

where $\lambda = log(15 - 1)$, since 15 is the average duration between non-consecutive records described by the deterministic, irregular aggregation scheme introduced above. In other words, the probability of an observation being recorded is usually 0.07 except when $s_\tau = 1$ in which case this probability jumps to 0.8.

Table 1 compares the following information for the original and aggregated data: (1) the coefficient of kurtosis of the simulated *price* series; (2) the p-value of the null hypothesis of normality from the Jarque-Bera statistic; (3) the estimated coefficient D of the scaling law; (4) the presence of ARCH in absolute returns ($|r_t|$ in (2.3)) ; and (5) the presence of ARCH in volatility for averages over 5 periods (v_t in (2.4)).

Several patterns are worth noting from this table. Both forms of irregularly spaced data generate fat tailed distributions away from gaussianity with excess kurtosis and ARCH in absolute returns. The coefficient for D is very close to the analytical level of 0.5 for the original and the regularly spaced data but it takes on values of approximately 0.55 for irregularly spaced data for both cases of $\rho = 1$. This is close to the 0.58 reported for most FX series. In addition, the seasonal patterns induced through the deterministic, and irregularly spaced aggregation, are readily apparent in the shape of the autocorrelation function of absolute returns but not for the returns series per se, in a manner that is also characteristic of FX markets. Consequently, this simple experiment along with the derivations in the previous section demonstrate that time aggregation may be behind many of the stylized facts common to high frequency FX data and that these statistical properties may not be reflective of the properties of the native D.G.P.

3. The Information Content of Order Flow

The previous sections demonstrate that the irregular nature of data arrival characteristic of FX data (as well as other financial data) instills rather unusual statistical properties to the data, even if these properties are not native to the operational time processes themselves. This section investigates a different modelling aspect – that of incorporating the information about the stochastic intensity of data arrival into classical fixed interval econometrics. We illustrate the proposed methods by examining an important issue in FX microstructure: the information content of order flow. We begin by briefly describing the data and the microstructure background to motivate the techniques that are proposed thereafter. The section concludes with the empirical results.

3.1. The Information Content of Quote Spreads and Intensity of Quote Arrivals: The HFDF-93 Data

Rational expectations and arbitrage-free theories of exchange rate determination suggest that all relevant information in exchange rates is common to all market participants, perhaps with the exception of central banks. However, as an empirical matter, these macroeconomic models tend to fail rather miserably (see Isard, 1995). By contrast, microstructure models focus on the role of asymmetric information, suggesting that order flow is an important factor in explicating exchange rate variation.

Without devoting too much time to developing microstructure theoretical models, we discuss the two main views on the relation between order flow and information. On one hand, Lyons (2001) suggests that innovations in nondealer order flow spark repeated interdealer trading of idiosyncratic inventory imbalances. Hence, a substantial amount of liquidity trading is generated with virtually no new information and as a consequence, periods of low intensity trades are viewed as more informative. On the other hand, Easly and O'Hara (1992) suggest the inverse relation to be true in the context of a signal-

extraction model of asymmetric information and competitive behavior. Thus, periods of high intensity in trading would correspond with periods in which the information content is high.

Before devoting more time to explaining how we plan to explore these issues empirically, it is important to describe the data in our analysis and its limitations. The data correspond to the HFDF-93 data-set available from Olsen & Associates. These data contain indicative quotes (rather than trades) that provide non-dealer customers with real-time information about current prices on the USD-DM FX rate[2]. These quotes lag the interdealer market slightly and spreads are roughly twice the size of interdealer spreads (see Lyons, 2001). Although most research done on these data has focused on forecasting, here we will explore the dynamics of the bid-ask spread as a function of quote-arrival intensity so as to get a measure of how information content varies with this intensity.

The FX market is a 24 hours global market although the activity pattern throughout the day is dominated by three major trading centers: East Asia, with Tokyo as the major trading center; Europe, with London as the major trading center; and America, with New York as the major trading center. Figure 1 displays the activity level in a regular business day as the number of quotes per half hour interval. The seasonal pattern presented is calculated non-parametrically with a set of 48 time-of-day indicator variables. Figure 2 illustrates the weekly seasonal pattern in activity by depicting a sample week of raw data.

The original data-set spans one year beginning October 1, 1992 and ending September 30, 1993, approximately 1.5 million observations on the USD-DM FX rate. The data has a two second granularity and it is pre-filtered for possible coding errors and outliers at the source (approximately 0.36% of the data is therefore lost). The subsample that

[2]The HFDF-93 contains other very interesting tick-by-tick data on other exchange rates and interest rates which are not explored here.

we consider contains 3,500 observations of half-hour intervals (approximately 300,000 ticks) constructed by counting the number of quotes in half hour intervals throughout the day. For each individual half-hour observation we then record the corresponding bid-ask spread. A comprehensive survey of the stylized statistical properties of the data can be found in Guillaume *et al.* (1997). Here, we only report some of the salient features.

The average intensity is approximately 120 quotes/half-hour during regular business days although during busy periods this intensity can reach 250 quotes/half-hour. The activity level significantly drops over the weekend although not completely. The bid-ask spread displays a similar seasonal pattern, with weekends exhibiting larger spreads (0.00110) relative to regular business days (0.00083).

Although we do not observe the levels of trading activity directly, these are naturally associated with the intensity of quote arrivals. Hence, to obtain a measure of information content, we will use the bid-ask spread. The explanations for the width of the spread vary widely (see O'Hara, 1995), and while undoubtedly inventory and transaction costs are important factors, the notion that information costs affect prices is perhaps the most significant development in market structure research. In fact, evidence in Lyons (1995), Yao (1998), and Naranjo and Nimalendran (2000) all suggest that dealers increase their spreads to protect themselves against informative, incoming order flow. As we mentioned above, Lyons (2001) reports that the quote-spread to non-dealers (which corresponds to our data) is twice the spread quoted to dealers. This is consistent with the notion that dealer risk aversion against informed trading generates wider spreads and thus cements our confidence in the interpretation of the width of the bid-ask spread as a measure of information flow.

3.2. Modeling the Intensity of Quote Arrival: The Autoregressive Conditional Intensity Model

A common approach in the empirical finance literature is to model the data as being generated by a time deformation model. 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. The nonlinearities introduced into the discrete-time representations of these time deformation processes can be summarized in the following expression:

$$\mathsf{x}_t = \mu(k_t) + \Phi(k_t; L)\mathsf{x}_{t-1} + \theta(k_t; L)\varepsilon_t, \tag{3.1}$$

where $\mu(k_t)$ is the intercept, $\Phi(k_t; L)$ and $\theta(k_t; L)$ are lagged polynomials in which k_t is the aggregation frequency described in (2.2) that describes the correspondence between economic time (or as we have denoted above, operational time) and calendar time. Note that when $k_t = k$, as is typical in conventional fixed-interval aggregation, the model in (3.1) delivers a typical constant-parameter representation. However, for a generic point process k_t the dependency on k can be quite complex (see Stock, 1988).

A question that naturally arises in this context is whether the parameters of the generating mechanism can be uncovered from the aggregated data. Although some papers address this issue in a discrete-time, time-domain framework (e.g. Wei and Stram, 1990 and Marcellino, 1998), it is usually analyzed as a discretization of a continuous-time process and in the frequency domain as is done in Bergstrom (1990) and Hinich (1999).

A common consequence of aggregation of high-frequency components is a phenomenon known as aliasing. Standard methods exist to smooth point processes to produce unaliased, equally-spaced aggregates. Hinich (1999) in particular, determines the mini-

mum sufficient sampling rate that allows the discrete-time representation of the system to be identified, while Hinich and Patterson (1989) show the relevance of adopting a proper sampling scheme when analyzing high-frequency stock returns. The idea implicit in these filtering methods is that the underlying D.G.P. is a constant-parameter, continuous-time model. Approximations with continuous-time models in finance are common but conceptually, they are generally ill-suited to describe high-frequency irregularly spaced data since the data already appear in their native frequency. Furthermore, because our analysis focuses on semi-structural issues related to the effects of quote intensity and information flow, we prefer to follow the tradition in the microstructure literature and avoid these filtering methods since they distort the very microstructure relationships that we wish to investigate.

In this sense and with regard to the issues discussed above, we share Engle's (2000) view that the joint analysis of quote arrival intensity and the size of the bid-ask spread generalizes 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 other observables in the model. But while Engle (2000) investigates the interarrival times themselves (such as is done in Engle and Russell, 1998), we advocate in favor of analyzing the point process directly and of modeling this process dynamically. Hence, instead of looking at the duration in time between observations, we investigate the dual of this problem, that is, its associated count process.

Therefore, the measurements of the number of quotes per unit time (in our investigation, 30-minute intervals) is an example of a count variable such as when one measures the number of customers that arrive at a service facility, the arrival of phone calls at a switchboard, and other analogous variables that describe infrequent events that occur at random times within the interval of observation. Denoting the number of quotes per 30-minute interval as k_t, the benchmark for count data is the Poisson distribution (see Cameron and Trivedi, 1998 for an excellent survey on count data models) with density,

$$f(k_t|\mathbf{x}_t) = \frac{e^{-\lambda_t}\lambda_t^{k_t}}{k_t!} \qquad k_t = 0, 1, 2, ..., \tag{3.2}$$

and conditional expectation

$$E(k_t|\mathbf{x}_t) = \lambda_t = \exp(\mathbf{x}_t'\boldsymbol{\gamma}), \tag{3.3}$$

so that $\log(\lambda_t)$ depends linearly on $\mathbf{x}_t$, a vector of explanatory variables that may include the constant term and lags of the dependent variable k. Expression (3.3) is called the exponential mean function and together with expression (3.2) they form the Poisson regression model, the workhorse of count data models. The model can be easily estimated by maximum likelihood techniques since the likelihood is globally concave.

However, unlike most applications of the Poisson regression model, the variable k is a time series that exhibits remarkable persistence (the Ljung-Box statistic takes on the values $Q_5 = 9607$,and $Q_{10} = 13,694$ and the autocorrelation function only dips below 0.15 after 16 periods or equivalently, eight hours). One solution to this problem is to endow expression (3.3) of a more conventional time series representation, similar in concept to the specification common in ACD, ACH[3], and ARCH models. Thus, we propose replacing expression (3.3) with

$$\log(\lambda_t) = \alpha \log(\lambda_{t-1}) + \beta k_{t-1} + \mathbf{x}_t'\boldsymbol{\gamma}. \tag{3.4}$$

Thus, we refer to the model in expressions (3.2) and (3.4) as the autoregressive conditional intensity model of order (1,1) or ACI(1,1). Extensions of expression (3.4) to more general ACI(p,q) lag structures is straight-forward as we will see in the empirical application. Expression (3.4) ensures that the intensity parameter λ_t remains strictly

[3]ACH stands for Hamilton and Jorda's (2002) autoregressive conditional hazard model, which is a dynamic, discrete-time duration model.

positive for any values of α, β, and $\boldsymbol{\gamma}$ while allowing the dependence of $\log(\lambda_t)$ to be linear in past values.

The ACI(1,1) endows the original expression (3.3) with rather rich dynamics in a very parsimonious manner: the process $\log(\lambda_t)$ depends on infinite lags of k_{t-1} and $\mathbf{x}_t$ at a geometrically decaying rate β. Note that stationarity will require the condition $\alpha+\beta < 1$. Estimation of the ACI(1,1) can be done by conditional maximum likelihood techniques by setting λ_0 to the unconditional mean of k (alternatively, λ_0 can be estimated as an additional parameter) and is disarmingly simple. For example, one can simply specify the following three lines of code in the LogL object in EViews version 4.0 (see EViews manual, chapter 18):

```
@log ll
log(lambda) = c(1) + c(2)*log(lambda(-1)) + c(3)*k(-1) + c(4)*x
ll = log(@dpoisson(k,lambda))
```

3.3. Empirical Results

According to the discussion in previous subsections, let k_t denote the number of quotes per half hour interval and let y_t denote the bid-ask spread that corresponds to the half hour interval t. Thus, the problem of measuring the information content of order flow can be translated into that of modeling the joint density of k_t and y_t. This joint density can be decomposed without loss of generality as,

$$h(k_t|\mathbf{k}_{t-1}, \mathbf{y}_{t-1}, \theta_1) = \frac{e^{-\lambda_t}\lambda_t^{k_t}}{k_t!}, \tag{3.5}$$

and

$$g(y_t|\mathbf{y}_{t-1}, \mathbf{k}_t, \theta_2), \tag{3.6}$$

where $\mathbf{y}_{t-1} = \{y_{t-1}, y_{t-2}, ...\}$ and $\mathbf{k}_{t-1} = \{k_{t-1}, k_{t-2}, ...\}$ and the conditional density (3.6) corresponds to the process described in expression (3.1). In particular, the speci-

fication of the conditional mean for the point process k_t corresponds to a version of the ACI model described in the previous subsection in expression (3.4), namely :

$$\log(\lambda_t) = seasonals + \theta(L)\log(\lambda_{t-1}) + \Psi(L)k_{t-1} + \Pi(L)y_{t-1}, \tag{3.7}$$

where the *seasonals* are a collection of indicator variables for time-of-day effects, day-of-week effects, and holiday effects. The corresponding lag polynomials are

$$\begin{aligned} \theta(L) &= \left(\theta_1 + ... + \theta_7 L^7\right)\left(1 - \theta_d L^{48}\right)(1 - \theta_w L^{336}), \\ \Psi(L) &= \left(\psi_1 + ... + \psi_7 L^7\right)\left(1 - \psi_d L^{48}\right)\left(1 - \psi_w L^{336}\right), \\ \Pi(L) &= \left(\pi_1 + ... + \pi_7 L^7\right), \end{aligned} \tag{3.8}$$

that is, the dynamic formulation of the intensity allows for deterministic as well as multiplicative, stochastic, time-of-day and day-of-week seasonal effects. We include up to 7 lags to capture some of the periodicity in the "lunch" and other breaks that recur across the trading areas. The model for y_t is the following:

$$y_t = seasonals(1 + F_0(k_t)) + \Phi(L, k_t)y_{t-1} + \varepsilon_t,$$

with

$$\Phi(L, k_t) = \phi_1(1 + F_1(k_t)) + \phi_2(1 + F_2(k_t))L + \phi_3(1 + F_3(k_t))L^2, \tag{3.9}$$

where $F_i(k_t)$ for $i = 0, 1, 2, 3$ is a non-parametric estimate based on a sixth order polynomial designed to capture the effects of the intensity level on the short-run dynamics of the spread variable. There are at least two ways in which the formulation of the model in (3.9) may appear incomplete. One is that we do not consider multiplicative, stochastic seasonal effects. However, these are implicit in the manner the coefficients are time-varying in k. The second is that we do not specify the variance in a dynamic way. However, the residuals of the fitted model did not show any evidence of ARCH effects which indicates that modelling the variance may not be central to learning about

the short-run dynamics of the bid-ask spread. This result is also consistent with our discussion on time deformation in section 2.

Table 2 compares the estimates of a simple Poisson count regression model exclusively based on the seasonal dummies against the ACI model in equations (3.6) and (3.7). These results are rather encouraging. The improvement on the overall fit of the data is quite remarkable by any measure. The Ljung-Box statistics reveal that the ACI model dramatically reduces the amount of left-over serial correlation in the residuals although there seems to be room left for improvement.

The second part of the exercise examines the dynamics of the spread as a function of the level of activity. Figure 3 depicts the estimated autoregressive parameters as a function of the intensity. In the limit, as $k_t \to 0$ then $\phi_1 \to 1, \phi_2 \to 0$, and $\phi_3 \to 0$ as we should expect when the sampling frequency is so high as to record observations were no activity has elapsed. However, as the aggregation frequency becomes higher, the parameter estimates display a fair amount of non-monotonic variation, ranging from high persistence to negative correlation and back into higher levels of persistence. Figure 4 reports the fluctuation in the average, seasonally-adjusted residual spread as a function of the intensity. After accounting for the intra-day trading patterns, the spread exhibits two well defined peaks: One at low levels of activity and another when the intensity reaches 140 quotes per half-hour (recall that the average trading intensity in a regular business day is around 120 quotes per half hour). The first peak is thus consistent with the view that inventory imbalance adjustment generates uninformative activity as is suggested in Lyons (2001). However, the second peak is consistent with Easly and O'Hara's (1992) signal-extraction model. Although our enthusiasm for this result has to be guarded due to the limitations that these data impose, we believe that it is of considerable importance because previous studies (see references in Lyons, 2001) have not considered non-monotonicities in the manner in which we report here.

4. Conclusions

In this paper we have shown how temporal aggregation of high-frequency, irregularly-spaced data can generate non-normality, conditional heteroskedasticity, and leptokurtosis even when these features are absent in the original D.G.P. In addition, we have introduced a new approach to modeling high-frequency irregularly spaced data based on the Poisson regression model. The new model is called the autoregressive conditional intensity (ACI) model and it has the advantage of being simple and of maintaining the calendar time scale. When applied to high frequency FX data, the model works well and highlights the variation in information content associated with changes in the intensity of activity levels.

References

[1] Bergstrom, A. R. (1990) *Continuous Time Econometric Modelling*, Oxford: Oxford University Press.

[2] Cameron, A. C. and Pravin K. Trivedi (1998) *Regression Analysis of Count Data,* Econometric Society Monographs, no. 30, Cambridge: Cambridge University Press.

[3] Clark, Peter K. (1973) "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica,* 41, 135-156.

[4] Dacorogna, Michel M., Ulrich A. Müller, R. J. Nagler, Richard B. Olsen, and Olivier V. Pictet (1993) "A Geographical Model for the Daily and Weekly Seasonal Volatility in the FX Market," *Journal of International Money and Finance*, 12(4), 413-438.

[5] Dacorogna, Michel M., C. L. Gauvreau, Ulrich A. Müller, Richard B. Olsen, and Olivier V. Pictet (1996) "Changing Time Scale Short-Term Forecasting in Financial Markets," *Journal of Forecasting*, vol. 15, 203-227.

[6] Diebold, Francis X. (1988) *Empirical Modelling of Exchange Rate Dynamics,* Lecture Notes in Economics and Mathematical Systems, vol. 303, Berlin: Springer-Verlag.

[7] Easly, David and Maureen O'Hara (1992) "Time and the Process of Security Price Adjustment," *Journal of Finance,* 47, 577-606.

[8] Engle, Robert F. (2000) "The Econometrics of Ultra-High Frequency Data," *Econometrica,* 68, 1-22.

[9] Engle, Robert F. and Jeffrey R. Russell (1998) "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66(5), 1127-1162.

[10] Ghysels, Eric and Joana 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.

[11] Granger, Clive W. J. and Zhuanxin Ding (1996) "Some Properties of Absolute Return: An Alternative Measure of Risk," *Annales d'Economie et de Statistique*, 0(40), Oct.-Dec., 67-91.

[12] Guillaume, Dominique M., Michel M. Dacorogna, Rakhal R. Davé, Ulrich A. Müller, Richard B. Olsen and Olivier V. Pictet (1997) "From the Bird's Eye to the Microscope: A Survey of New Stylized Facts of the Intra-Daily Foreign Exchange Markets," *Finance and Stochastics*, 1, 95-129

[13] Hamilton, James D. and Òscar Jordà (2002) "A Model for the Federal Funds Rate Target," *Journal of Political Economy,* forthcoming.

[14] Hinich, Melvin J. (1999) "Sampling Dynamical Systems," *Macroeconomic Dynamics*, 3, 602-609.

[15] Hinich, Melvin J. and D. M. Patterson (1989) "Evidence of Nonlinearity in the Trade-by-Trade Stock Market Return Generating Process," in William A. Barnett, John Geweke, and Karl Shell (eds.) *Economic Complexity: Chaos, Sunspots, Bubbles and Nonlinearity*, Cambridge: Cambridge University Press.

[16] Harris, L. (1986) "A Transaction Data Study of Weekly and Intradaily Patterns in Stock Returns, "*Journal of Financial Economics,* 16, 99-117.

[17] Hasbrouck, J. (1988) "Trades, Quotes, Inventories and Information," *Journal of Financial Economics,* 22, 229-252.

[18] Hasbrouck, J. (1991) "Measuring the Information Content of Stock Trades," *Journal of Finance,* 46, 179-207.

[19] Hausman, Jerry A., Andrew W. Lo and A. Craig MacKinlay (1992) "An Ordered Probit Analysis of transaction Stock Prices," *Journal of Financial Economics,* 31, 319-330.

[20] Isard, P. (1995) *Exchange Rate Economics,* Cambridge: Cambridge University Press.

[21] Lyons, Richard K. (1995) "Tests of Microstructural Hypotheses in the Foreign Exchange Market," *Journal of Financial Economics,* 39, 321-351.

[22] Lyons, Richard K. (2001) *The Microstructure Approach to Exchange Rates,* Cambridge, MA: MIT Press.

[23] Marcellino, Massimiliano, (1998), "Temporal Disaggregation, Missing Observations, Outliers, and Forecasting A Unifying Non-Model Based Approach," *Advances in Econometrics*, vol. 13, 181-202.

[24] Müller, Ulrich A., Michel M. Dacorogna, Richard B. Olsen, Olivier V. Pictet, M. Schwarz, and C. Morgenegg (1990) "Statistical Study of Foreign Exchange Markets, Empirical Evidence of a Price Change Scaling Law, and Intraday Analysis," *Journal of Banking and Finance*, 14, 1189-1208.

[25] Naranjo A. and M. Nimalendran (2000) "Government Intervention and Adverse Selection Costs in Foreign Exchange Markets," *Review of Financial Studies,* 12, 453-477.

[26] O'Hara, Maureen, (1995), *Market Microstructure Theory*, Oxford: Blackwell Publishers.

[27] 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.

[28] Stock, James H. (1988) "Estimating Continuous-Time Processes Subject to Time Deformation," *Journal of the American Statistical Association, 77-85.*

[29] Tauchen, George E. and Michael Pitts (1983) "The Price Variability Volume Relationship on Speculative Markets," *Econometrica,* 51, 485-505.

[30] Taylor, S. J., (1988), *Modelling Financial Time Series*, Cichester: J. Wiley and Sons.

[31] Wei, W. W. S. and D. O. Stram (1990) "Disaggregation of Time Series Models." *Journal of the Royal Statistical Society*, Series B, 52, 453-467.

[32] Yao, J. (1998) "Spread Components and Dealer Profits in the Interbank Foreign Exchange Market," New York University Salomon Center Working Paper # S-98-4.

Table 1. Monte Carlo Simulations of Different Irregularly Spaced Aggregation Schemes

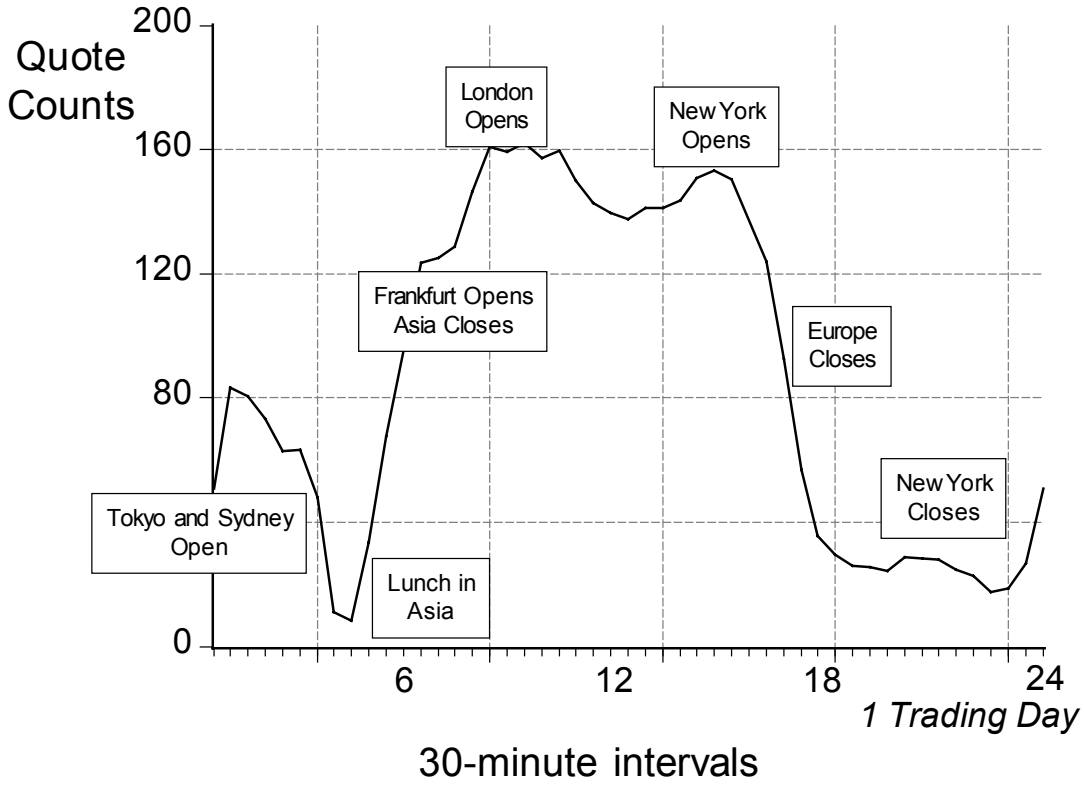
Kurtosis	**Aggregation Type**			
	Operational Time	*Deterministic Fixed Interval*	*Deterministic Irregular*	*Random Irregular*
$\rho = 1; \mu = 0$	3.0041	2.9406	7.7214	6.4773
$\rho = 1; \mu = 5 \times 10^{-6}$	3.0019	2.9371	7.7385	6.4035
$\rho = 0.99; \mu = 0$	3.0108	2.8761	7.3839	6.1091
$\rho = 0.99; \mu = 5 \times 10^{-6}$	2.9985	2.9657	7.4858	6.2055
Jarque-Bera (p-val.)	**Aggregation Type**			
	Operational Time	*Deterministic Fixed Interval*	*Deterministic Irregular*	*Random Irregular*
$\rho = 1; \mu = 0$	0.4283	0.5149	0.0000	0.0000
$\rho = 1; \mu = 5 \times 10^{-6}$	0.5797	0.5457	0.0000	0.0000
$\rho = 0.99; \mu = 0$	0.4213	0.5000	0.0000	0.0000
$\rho = 0.99; \mu = 5 \times 10^{-6}$	0.4214	0.5414	0.0000	0.0000
D	**Aggregation Type**			
	Operational Time	*Deterministic Fixed Interval*	*Deterministic Irregular*	*Random Irregular*
$\rho = 1; \mu = 0$	0.5002	0.5105	0.5506	0.5351
$\rho = 1; \mu = 5 \times 10^{-6}$	0.5044	0.7246	0.5716	0.5531
$\rho = 0.99; \mu = 0$	0.4842	0.0467	0.4464	0.4454
$\rho = 0.99; \mu = 5 \times 10^{-6}$	0.4832	0.0502	0.4483	0.4462
ARCH $\lvert r_t \rvert$	**Aggregation Type**			
	Operational Time	*Deterministic Fixed Interval*	*Deterministic Irregular*	*Random Irregular*
$\rho = 1; \mu = 0$	No	No	Yes	Yes
$\rho = 1; \mu = 5 \times 10^{-6}$	No	No	Yes	Yes
$\rho = 0.99; \mu = 0$	No	Yes*	Yes	Yes
$\rho = 0.99; \mu = 5 \times 10^{-6}$	No	Yes*	Yes	Yes
ARCH v_t	**Aggregation Type**			
	Operational Time	*Deterministic Fixed Interval*	*Deterministic Irregular*	*Random Irregular*
$\rho = 1; \mu = 0$	No	No	Yes	No
$\rho = 1; \mu = 5 \times 10^{-6}$	No	No	Yes	No
$\rho = 0.99; \mu = 0$	No	No	Yes	No
$\rho = 0.99; \mu = 5 \times 10^{-6}$	No	No	Yes	Yes*

Comments: The title "operational time" refers to the original data; "Deterministic, Fixed Interval" refers to aggregating the original data every 100 periods; "Deterministic, Irregular" refers to aggregating according to a seasonal cycle that repeats itself every 100 periods; and "Random, Irregular" refers to true random intervals of aggregation with a seasonal pattern. The * indicates the test was barely significant at the conventional 95% confidence level.

Table 2. Comparing the Poisson Regression with the ACI

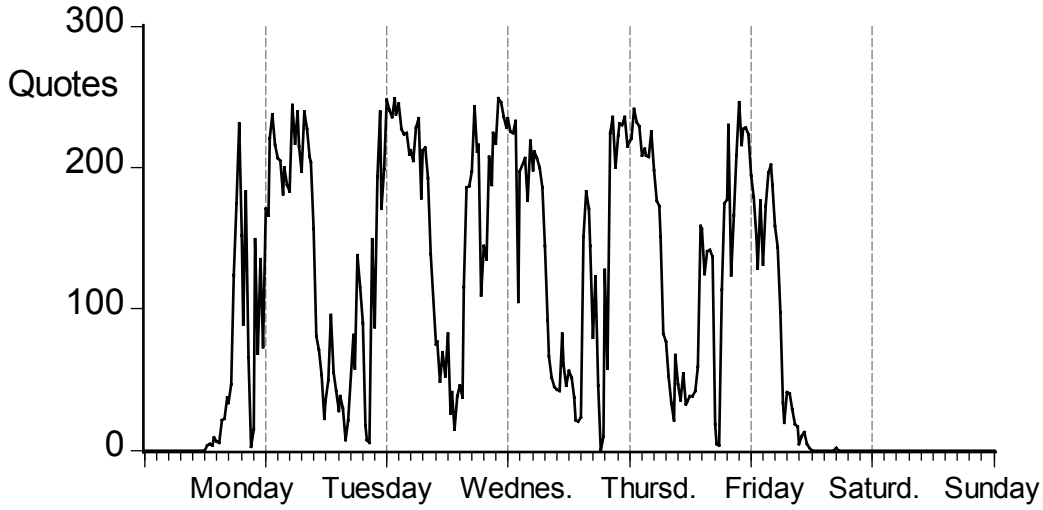
	Poisson Regression	***ACI***
Log-Likelihood	-28562.37	-18145.52
No. of parameters	55	80
AIC	19.729	12.565
SIC	19.842	12.730
Ljung-Box Q_5	1608	99
Ljung-Box Q_{10}	1903	128
Ljung-Box Q_{50}	2398	372
LR test ACI vs. Poisson (p-val)	0.000	

Figure 1. Quote Arrival Rate in the USD-DM FX Market: Seasonal Intraday Pattern



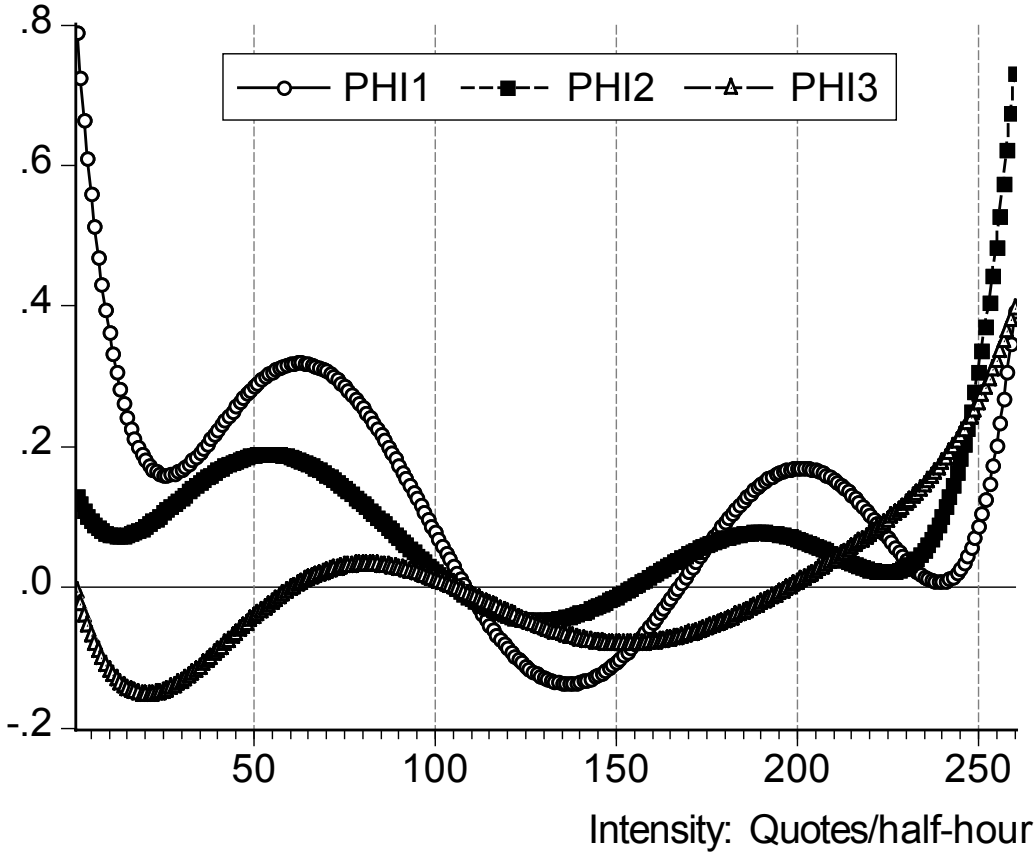
The graph displays non-parametric estimates of the seasonal pattern of a typical business day in intervals of 30-minutes. The opening and closing of major trading centers for the USD-DM FX is indicated.

Figure 2. Weekly Pattern of Quote Arrival in the USD-DM FX Market: Raw Data



Raw quote-arrival for a typical week in the USD-DM FX Market

Figure 3. Nonparametric Estimates of the Autoregressive Parameters of the USD-DM FX Spread Model as a Function of the Intensity of Quote Arrival per 30-minute Interval



Value of the autoregressive parameters ϕ_1, ϕ_2, and ϕ_3 of the model for the USD-DM FX spread variable as a nonparametric function of the intensity of quote arrival.

Figure 4. Nonparametric Estimate of the Mean USD-DM FX Spread as a Function of the Intensity of Quote Arrival per 30-minute Interval.