Using self-organizing maps to adjust intra-day seasonality

Walid Ben Omrane[1] Eric de Bodt[2,3]

December 6, 2005

Abstract

The existence of an intra-day seasonality component in financial market variables (volatility, volume, activity, etc.), has been highlighted in many previous studies. To adjust raw data for their cyclical component, many researchers start by using the intra-day average observations model (IAOM) and/or some smoothing techniques (e.g. the kernel method) in order to remove the day-of-the-week effect. When the seasonality involves only a deterministic component, the IAOM method succeeded in estimating the periodicity almost perfectly. However, when the seasonality contains both deterministic and stochastic components (e.g. closed days), both IAOM and the kernel method fail to capture it. We introduce self-organizing maps (SOM) as a solution. SOM are based on neural network learning and nonlinear projections. Their flexibility allows seasonality to be captured even in the presence of stochastic cycles.

Keywords: self-organizing maps, currency market, intra-day seasonality, high frequency data

JEL Classification : F31, C22

[1] Department of Business Administration (IAG), Finance Unit, Université catholique de Louvain, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium, e-mail: benomrane@fin.ucl.ac.be. Tel: +32 10 47 84 49, Fax: +32 10 47 83 24.

[2] Corresponding author. Department of Business Administration (IAG), Finance Unit, Université catholique de Louvain, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium, e-mail: debodt@fin.ucl.ac.be. Tel: +32 10 47 84 47, Fax: +32 10 47 83 24.

[3] ESA, University of Lille 2, Place Déliot, BP 381, Lille F-59020, France.

The authors would like to thank Luc Bauwens, Marie Cottrell, Christian Gouriroux, Michel Verleysen and participants at the 11^{th} ACSEG International Conference in Lille, and the 5^{th} workshop on the Self-Organizing Maps in Paris (France), for helpful discussions and suggestions. The work was supported in part by the European Community Human Potential Program under contract HPRN-CT-2002-00232, Microstructure of Financial Markets in Europe. The usual disclaimers apply.

1 Introduction

Evidence of intra-day seasonality in financial market behaviors has been highlighted in many previous studies and concerning many microstructure variables (e.g. volatility and quoting activity). Two categories of methods are most often used to remove this seasonality. Some studies (such as Degennaro and Shrieves, 1997, Andersen and Bollerslev, 1998, Melvin and Yin, 2000, Cai, Cheung, Lee, and Melvin, 2001, Bauwens, Ben Omrane, and Giot, 2005, and Ben Omrane and Heinen, 2004) adopt a linear projection technique. They regress variables (affected by the seasonal component) on a set of dummy variables (or flexible Fourier forms) in order to capture intra-day cycles. Other authors adjust the raw data for seasonality using a direct correction factor obtained from intra-day averages (Dacorogna, Müller, Nagler, Olsen, and Pictet, 1993, Eddelbuttel and McCurdy, 1998, Melvin and Yin, 2000, and Bauwens, Ben Omrane, and Giot, 2005) or a smoothing kernel (Engle and Russell, 1998, Bauwens and Giot, 2000, and Veredas, Rodriguez-Poo, and Espasa, 2002).

Our research builds on previous literature to explore the limits of the classical approaches and to introduce a solution for stochastic cycles. We show that the more the raw data involves a deterministic seasonality, the more the classical methods, particularly the intra-day average observations model, succeed in estimating the cycles. However, in the presence of stochastic cycles (or the combination of deterministic and stochastic cycles), such as those generated by closed days (among others), classical methods reveal their limits. We introduce a method based on the self-organizing maps algorithm (Kohonen, 1995). Self-organizing maps (SOM) allow both deterministic and stochastic cyclical components to be captured.

Our evidence is based both on Monte Carlo simulations and on application to a real data set (taken from the foreign exchange (FX) market). Our Monte Carlo simulations adopted a five-step framework. We began by generating an auto-regressive process. We then simulated either a deterministic seasonality, or both deterministic and stochastic cycles which we added to the auto-regressive variable. After that we deseasonalized the endogenous variable, using the three methods cited above. We finally re-estimated the process coefficients on the deseasonalized data series. The better the deseasonalization, the closer the estimated coefficient should be to the simulated one, and the lower the root mean square error (RMSE) should be. This allows us to compare the performance of our methods in a controlled setup. For our empirical study we used a high frequency data set of 5-minute regularly time-spaced Euro/Dollar quotes. The time period stretched from May 15, 2001 through May 15, 2002. Our results confirm that in the presence of both deterministic and stochastic cycles, the SOM method is more powerful at neutralizing the seasonality than either the intra-day average observations model (IAOM) or the Nadaraya-Watson kernel smoothing method. We based our comparison on an analysis of the autoregressive correlation function (ACF) of the deseasonalized variables. In this way, the quality of the adjustments is inferred from the persistence of the cycles in the ACF. Our results are consistent with the Monte Carlo simulation results.

This paper is divided into six sections. In Section 2 we present a brief review of literature related to intra-day seasonality (focusing on the FX market because we use it as the empirical setting in Section 5). We detail our deseasonalization methods in Section 3. The Monte Carlo simulation is presented in Section 4, and Section 5 describes the empirical application. Section 6 summarizes our conclusions.

2 Foreign Exchange Seasonality

A large segment of the foreign exchange microstructure literature documents that market opening and closing, news announcements and days of the week introduce significant cyclical factors into many microstructure variables such as volatility and quoting activity (Bollerslev and Domowitz, 1993, Andersen and Bollerslev, 1996, Degennaro and Shrieves, 1997, Melvin and Yin, 2000, and Ben Omrane and Heinen, 2004). A typical case is highlighted by Andersen and Bollerslev (1998) and Bauwens, Ben Omrane, and Giot (2005) who show that scheduled news announcements have a seasonal impact on volatility. These news events exhibit both a cyclical and a stochastic component, the latter being the news content not fully anticipated by the market participants. However, the cyclical news component could itself be either deterministic or stochastic, since the timing of some announcements changes from one week to the next.

A number of methods have recently been put forward in the literature to capture this cyclical behavior in high frequency data. Melvin and Yin (2000) and Bauwens, Ben Omrane, and Giot (2005), inter alia, use the intra-day average observations model (IAOM)[1] to adjust volatility and quoting activity variables for seasonality. They divide returns (quoting activity) by the square root of the cross sectional average volatility (cross sectional average quoting activity) to clean the series from its cyclical components (see Section 3.2). The more the data involve a deterministic seasonality,[2] the more the IAOM is successful in removing the seasonality. Degennaro and Shrieves (1997) use dummy variables to identify 'hour of the day' cyclical effects, but they do not discriminate between different days of the week. On the other hand Andersen and Bollerslev (1998) and Bauwens, Ben Omrane, and Giot (2005) show that workdays are characterized by specific cyclical behaviors. To deseasonalize volatility, they therefore allow for a specific seasonality for each day of the week (although they assume that the day of the week effect is constant from week to week).

On the same topic, Andersen and Bollerslev (1998), Cai, Cheung, Lee, and Melvin (2001) and Ben Omrane and Heinen (2004) use the flexible Fourier form (a sum of sinusoids) to capture intraday cycles. Dacorogna, Müller, Nagler, Olsen, and Pictet (1993) and Eddelbuttel and McCurdy (1998) deseasonalize volatility using an adjustment factor. This factor is proportional to the (mean) absolute value of the returns over a time interval divided by the size of the time interval. Engle and Russell (1998) and Bauwens and Giot (2000) adjust duration variables for seasonality using the cubic splines technique, but Veredas, Rodriguez-Poo, and Espasa (2002) adopt the kernel estimator to adjust duration and show that their method is more effective than the cubic splines.

To sum up, there are two broad categories of seasonality adjustment methods used in the literature. The first is a one-step procedure and consists of removing seasonality through a regression. Seasonality is captured through some added variables (such as dummies or the flexible Fourier form). The second category is a two-step procedure: before implementing the regression, the raw data is adjusted for seasonality; then the adjusted variable is regressed onto the set of exogenous

[1] This method is equivalent to one based on hourly dummy variables.

[2] Deterministic seasonality corresponds to continual cycles without gaps or discontinuities generated by, for instance, closed days

variables. In this study we focus on the seasonality-removal part of the procedure. We compare both the IAOM and the kernel methods to a new type of algorithm: self-organizing maps (SOM) introduced by Kohonen (1995). This algorithm has lead to many applications in physics and engineering as well as in finance (see, inter alia, de Bodt, Cottrell, Henrion, and Van Wijmeersch, 1998, de Bodt, Lendasse, Cardon, and Verleysen, 2004, and DeBoeck and Kohonen, 1998). The SOM model is based on neural network learning and nonlinear discrete projection (see Section 3.1). As show below, the SOM algorithm allows us to deal with the stochastic component of the seasonality.

3 Deseasonalization Methods

In this section we present the three seasonality identification methods. We start with a presentation of the SOM algorithm as its usage is the key contribution of this study. We then review the well known IAOM and kernel smoothing approaches. For a more convenient presentation of the details of the different method we use some examples picked from either the simulated or the real data series.

3.1 The Self-Organizing Maps Model (***SOM(p,q)***)

The self-organizing maps (SOM) method introduced by Kohonen (1995) is a method of data analysis which allows, through a (discrete) projection, the dimensions of the data space to be reduced (as principle component analysis methods do). Simultaneously it allows, through vector quantization, the data being summarized to be projected in specific mean profiles. The projection step is carried out on a discrete data space.

Before turning to a more formal presentation of the SOM algorithm, we will introduce it with an example. Imagine that we are faced with a two dimensional data matrix such as the one presented in Table 1. There are 21 observations and, for each observation, two measures have been taken (e.g., the size and the weight). Figure 1 Panel A shows the two-dimensional input data space (each observation being represented by a dot). Three natural clusters clearly emerge. The two situated on the left are closer to each other than to the one situated on the right of the chart. We use a self-organizing map (SOM) both to capture the proximity relations among clusters and to summarize (quantify) the information contained in each cluster. Our map is shown in Figure 1 Panel B. It is a SOM(2,2) - that is to say it is composed of two rows and two columns (four nodes). Each node is identified by its location in the map (the row and column indices). A vector of coordinates is associated to each node. This vector defines the location of the node in the input space. Figure 1 Panel C displays the locations of the four nodes after random initialization of their coordinates' vectors. In technical terms, the map is said to be folded: the proximity relations among the nodes in the map do not reflect the proximity relations in the input space. In the input space, the node (1,1) is closer than the node (1,2) to the node (2,2), while this is not the case in the map. Moreover, the locations of the nodes bear no relation to the clusters of data. The SOM algorithm is the numerical procedure by which the map is unfolded and displaced towards the data clusters (Figure 1 Panel D). After this, if everything goes right, the neighborhood relations in the map correspond to those observed in the input space. The node coordinates' vectors represent homogeneous clusters of data (as here for nodes (1,1), (2,2) and (2,1), although not (1,2)). Note that in the present case, in order to allow a visual representation of the process, we have used a projection of the two-dimensional input space onto a two-dimensional map. In real applications, the dimensions of the input space are usually far higher and the two dimensional map provides a convenient way of observing the neighborhoods relation among clusters of individuals. Nothing forbids the use of higher dimensional maps except that visual representation then becomes difficult (if not impossible).

In more formal terms, SOM defines a mapping from the input data space Ω, onto a K-dimensional array of output nodes. In order to visualize the outputs, K must not to be above two (a grid). If x represents one observation, then let $x \in \Omega$ be a stochastic data vector. A vector quantization φ is an application from the continuous space Ω, endowed with some probability density function $f(x)$, to a finite subset F composed of the n nodes $m_1, \ldots, m_n$. These nodes, which are located at specific points on the map, are associated to a coordinate vector, which will, after learning, represent the average profile of observations associated with a specific node. After unfolding the position of a node is a result of the neighborhood structure of the data in the input space. The SOM algorithm is defined as follows:

- the structure of the map is first defined (number of rows (p) and columns (q));
- the coordinate vectors of the nodes $m_1, \ldots, m_n$ are randomly initialized (in the input space);
- each node occupies a specific location in the map;
- at each iteration t of the algorithm:
  - an observation x is randomly drawn according to the density $f(x)$,
  - the 'winning' node $m_{k^*,t}$ is identified by minimizing the classical Euclidean norm:

    $$\|x_i - m_{k^*,t}\| = \min_k \|x_i - m_{k,t}\| \tag{1}$$

  - the class m_{k^*} and its neighbors in the map are updated by

    $$m_{k^*,t+1} = m_{k^*,t} + \varepsilon_t (x_i - m_{k^*,t}), \tag{2}$$

    where ε_t is an adaptation parameter which satisfies the Robbins and Monro (1951) conditions ($\sum \varepsilon_t = \infty$ and $\sum \varepsilon_t^2 < \infty$). Note that the set of the neighbors adapted at each iteration can either be kept fixed or progressively decreased throughout the procedure.

The iteration or 'learning' process combines a projection and a quantization. Nodes begin by being distant from each other and then converge gradually to the barycenter of the clusters of observations. At the end of the learning process, their coordinate vectors represent the 'average individual' of a cluster of observations. The adaptation parameter, ε_t (also called the learning coefficient), drops progressively. At the starting of the learning, nodes are moved by large steps in order to bring them closer to their convergence zone, and then their positions are progressively computed with more precision. Once learning has been achieved, each observation i is attached to its winning node $m_{k^*,t}$, identified by its map coordinates. This correspondence is a kind of projection on a discrete subspace.

In our empirical study (Section 5), Ω is the data matrix containing the number of trading days during the year (258) in its rows, and the number of five-minute observations (288) in its columns. The seasonality is captured by the value of the node coordinate vector after learning. The deseasonalization is achieved through a two-step process. Each observation is attached to a winning node, and then, to adjust the observation for seasonality, we divide it by (or subtract it from) the mean profile for the corresponding winning class. We implement the division or subtraction according to how the cyclical component is involved in the raw data. For instance, volatility and quoting activity have to be adjusted for seasonality by division. However, both the simulated AR(P) processes (y_t and z_t), computed in Section 4 have to be subtracted from the cyclical component.

Finally, note that selecting a given map structure (number of rows (p) and number of columns (q)) is a trial and error process. The autoregressive correlation function (ACF) is a convenient guidance tool for selecting the right number of lags within an ARMA(p,q) model.

3.2 The Intra-Day Average Observations Model ($IAOM$)

To estimate seasonality, we computed the intra-day average observations at time n_k of day k (called mv_{n_k}). We divide each day into Q intervals of time. We assume for simplicity that we have exactly S weeks of data. For each interval endpoint per day of the week over the S week period, we had one observation for the random variable, Y. We thus compute in principle Q values mv_{n_k} for each day of the week, making a total of W ($5 \times Q$) values over a week. Formally,

$$mv_{n_k} = \frac{1}{S} \sum_{s=1}^{S} Y_{f(s,k,n_k)}, \tag{3}$$

where

$$f(s,k,n_k) = W\,(s-1) + \sum_{j=1}^{k-1} N_j + n_k, \tag{4}$$

$s = 1, \ldots, S$. $k = 1, \ldots, 5$. $N_1 = N_2 = N_3 = N_4 = N_5 = Q$. $n_1 = 1, \ldots, Q$ and similarly for n_2, n_3, n_4 and n_5.

To adjust the different variables for seasonality, we used the same methodology as for the SOM adjustment. We just divided/subtracted the endpoint of each five-minute interval by/from the corresponding value of the intra-day average observation. This means, for example, that all quoting activities at 12.00 on Thursday in the sample were divided by the same value (the average quoting activity at 12.00 on Thursdays).

3.3 The Smoothing Method

The smoothing method consists of smoothing the raw data using the Nadaraya-Watson kernel estimator and then adjusting each raw observation by the corresponding value on the smoothed curve. The adjustment is made in the same way as for the SOM and IAOM methods. The Nadaraya-Watson kernel estimator $\hat{Y}_t$ of $Y(t)$ is:

$$\hat{Y}_t = \frac{\sum_{j=1}^{T} K_h(t-t_j)Y_t}{\sum_{j=1}^{T} K_h(t-t_j)} \,. \tag{5}$$

where t is the vector of time, T corresponds to the number of observations, and h is the bandwidth parameter. Choosing the appropriate bandwidth is an important aspect of any local-averaging technique. In our case we selected a Gaussian kernel with a bandwidth, h, as computed by Silverman (1986):

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}} \tag{6}$$

$$h = \left(\frac{4}{3}\right)^{1/5} \sigma_k \, l^{-1/5}, \tag{7}$$

where σ_k is the standard deviation for the observations.

4 Monte Carlo Simulation

4.1 Simulation Procedure

In order to compare the three seasonality identification methods (IAOM, SOM, NW-kernel) we implemented a five-step simulation procedure.

1) We start by generating a P-lag autoregressive process, y_t^*, (AR(P), $P = 1, 5$):

$$y_t^* = \sum_{p=1}^{P} \beta_p y_{t-p}^* + \epsilon_t, \tag{8}$$

where β equals 0.95 if $P = 1$, and, if $P = 5$, $\beta_1 = 0.5$, $\beta_2 = 0.09$, $\beta_3 = 0.08$, $\beta_4 = 0.07$, and $\beta_5 = 0.06$. ϵ_t has a standard normal distribution.

2) We partitioned y_t^* by blocks of Q observations, each representing one day of the week.

3) We simulated a deterministic seasonality $S_{t,i}^{det}$ and we added it to the AR(P) process described above, so that

$$y_{t,i} = y_{t,i}^* + S_{t,i}^{det}. \tag{9}$$

If $y_{t,i}^*$ represents one such block, where i is an index corresponding to the trading days of the week ($i = 1, \ldots, 5$), then $S_{t,i}^{det}$ is generated by the following procedure. The block of $y_{t,i}^*$ observations, corresponding to each day of the week, is divided into three periods (say, the morning, noon, and the afternoon). Then, a defined constant is added to the AR(P) process depending on the specific time period in which the observation is located. One set of constants is chosen for each day of the week, since we generate a deterministic weekly cycle (seasonality). In this way, y_t becomes an autoregressive variable which involves a deterministic seasonality.

To simulate an AR(P) process which contains stochastic seasonality in addition to the deterministic one, we used the following procedure:

- we generated an AR(P) process:

$$z_t^* = \sum_{p=1}^{P} \beta_p z_{t-p}^* + \epsilon_t, \tag{10}$$

- we added a deterministic and stochastic seasonality to this process:

$$z_{t,i} = z_{t,i}^* + S_{t,i}^{det} + S_{t,i}^{sto}, \tag{11}$$

where $S_{t,i}^{det}$ is generated as described above, and $S_{t,i}^{sto}$ is the stochastic seasonality. The difference between these two seasonalities depends on the manner in which constants are added to the time periods of weekdays. For stochastic seasonality, days were selected randomly to the subject of added variation. Moreover, the variation changes from week to week.

4) The fourth step consists of estimating and removing seasonality from two simulated processes (y_t and z_t) using the IAOM, NW-kernel and SOM methods. The deseasonalization methodology consists of using a linear subtraction of the estimated seasonalities, ϕ_t^{det} and ϕ_t^{sto} respectively, from the analyzed variables y_t and z_t, such that:

$$y_t^{'} = y_t - \phi_t^{det}, \tag{12}$$

$$z_t^{'} = z_t - \phi_t^{sto}. \tag{13}$$

5) Finally, we estimated both AR(P) processes, based on their respective deseasonalized variables, using ordinary least square estimation:

$$y_t^{'} = \sum_{p=1}^{P} \beta_p^{'} y_{t-p}^{'} + \epsilon_t^{'}, \tag{14}$$

$$z_t^{'} = \sum_{p=1}^{P} \gamma_p^{'} z_{t-p}^{'} + \nu_t^{'}. \tag{15}$$

The whole procedure was iterated 1000 times. To assess the performance in terms of seasonality adjustment of each of the three methods, we computed the root mean square error (RMSE) of the estimated coefficients $\beta_p^{'}$ and $\gamma_p^{'}$ relative to the initially simulated β_p. The closer the estimated coefficients were to β_p, the lower the RMSE and better the seasonality adjustment. It is worth pointing out that the SOM algorithm was initialized with the IAOM outputs (which in practice, seems to have been a judicious choice).

4.2 Results

Estimation results for the Monte Carlo simulation are presented in Tables 2 and 3. Table 2 displays the estimation results for AR(1) and Table 3 presents those of AR(5) process. In both tables Panel A displays the mean, standard deviation, and RMSE for 1000 iterations of the autoregressive coefficient for Equation (14) in the presence of deterministic seasonality. Panel B illustrates the same results for the stochastic seasonality (Equation (15)). The variables, in this case, are deseasonalized from both their deterministic and their stochastic seasonality. The RMSE in both panels characterizes the estimation error generated by the added seasonality.

Starting with the results for deterministic seasonality, the estimated coefficients for the non-deseasonalized autoregressive parameters corresponding to Equations (14) and (15) (see the second column of Tables 2 and 3) show a higher level of error in Panel B than in Panel A. The more seasonality there is in the process, the larger is the error in the estimated coefficients. This is why previous studies have tried to remove the cyclical component from their microstructure variables.

The IAOM deseasonalization displays interesting results in Panel A. The estimated coefficients are very close to the simulated ones, with an insignificant error equal to 0.09% for AR(1) and around 0.33% for the different coefficients of the AR(5). We conclude that the IAOM method succeeds in capturing almost the whole of the deterministic seasonality component. The IAOM method can therefore be recommended as an effective tool for seasonality adjustment when the cyclical component is strictly deterministic. This means that the time series should not include gaps due to missing values (due, for example, to closed days or data recording problems). Panel B presents very different results. In the presence of stochastic cycles, the IAOM method has a significant level of estimation error. The corresponding RMSE is much higher than that obtained by estimating the model with deterministic seasonality. When the seasonality involves both deterministic and stochastic elements, the IAOM does not even capture the whole cyclical component of the process. This suggest that, when there are good reasons to think that the seasonality could display some stochastic behavior, the IAOM approach should not be used.

The SOM method seems to be far more robust in the presence of stochastic cycles. Panel B of Tables 2 and 3 exhibits, in the third column, the estimation result for the seasonality adjusted AR(1) and AR(5) processes respectively. The corresponding RMSE is low compared to the IAOM case. Unlike the IAOM method, SOM(1,5) succeeds in capturing seasonality involving both deterministic and stochastic cycles. The results in Panel A show that SOM(1,5) is, however, less efficient than the IAOM method when the seasonality involves only deterministic cycles. In such a case, the estimation error generated by SOM(1,5) is much higher than that generated by the IAOM method. The choice between IAOM and SOM therefore depends on the presence of stochastic cycles.

The kernel results displayed in Table 2, Panel A, show that this method captures deterministic seasonality with a low error level. This finding is consistent with previous research which opted for the kernel method as a step in the deseasonalization process, particularly when the samples exhibited some deterministic cycles. However, the kernel adjustment is less accurate than IAOM. Nevertheless, Table 3, Panel A displays a higher RMSE for the kernel method, especially for the first three coefficients of the AR(5). Panel B results for Tables 2 and 3 show that the kernel method generates an estimation error level much higher than that generated by SOM and IAOM, but smaller than the non-adjusted data.

These findings are consistent with intuition. By construction, the IAOM method, built on the computation of cross-sectional means, can easily capture deterministic seasonality. The IAOM algorithm relies on the law of large numbers: the deterministic component estimation amounts to an estimation of the expected value of the hour-by-hour cycle by its sample average. The SOM algorithm and kernel methods can also capture deterministic cycles, but much less efficiently than the IAOM. However, when cycle irregularities are present (as often occurs in financial data), using the hour-by-hour sample average to capture the seasonality becomes problematic. The SOM model goes beyond the limits of the IAOM and kernel models. It estimates, efficiently, seasonality which contains both deterministic and stochastic cycles.

5 Empirical Evidence

5.1 Data Description

In this section we use the same data set as that used by Bauwens, Ben Omrane, and Giot (2005). The data chosen are two microstructure variables (volatility and quoting activity) from the currency market. The euro/dollar foreign exchange market is a market-maker based trading system, where three types of market participants interact around the clock (i.e. in successive time zones): the dealers, the brokers and the customers from whom the primary order flow originates. The most active trading centers are New York, London, Frankfurt, Sydney, Tokyo and Hong Kong. A complete description of the FX market is given by Lyons (2001).

To compute the returns used to estimate the volatility, we use the Olsen and Associates database made up of 'tick-by-tick' euro/dollar quotes for the period ranging from May 15, 2001 to May 15, 2002 (i.e. one year). It is worth pointing out that our sample involves five closed days.[3] This database includes 6,088,382 observations. As in most empirical studies on FX data, the euro/dollar quotes are market makers' quotes and not transaction quotes (which are not widely available).[4] More specifically, the database contains the date, the time-of-day time stamped to the second in Greenwich mean time (GMT), the dealer bid and ask quotes, the identification codes for the country, city and market-maker bank, and a return code indicating the filter status. According to Dacorogna, Müller, Nagler, Olsen, and Pictet (1993), when trading activity is intense, some quotes are not entered into the electronic system. If traders are too busy or the system is running at full capacity, quotations displayed in the electronic system may lag prices by between a few seconds and one or more minutes. We retained only the quotes that have a filter code value greater than 0.85.[5]

From the tick data, we computed mid-quote prices, where the mid-quote is the average of the bid and ask prices. As we used five-minute returns, we obtained a daily grid of 288 points. At the end of each interval, we used the closest previous and next mid-quotes to compute the relevant prices by interpolation. The mid-quotes were weighted by their inverse relative time distance to the interval endpoint. The return at time t was then computed as the difference between the logarithms of the interpolated prices at times $t-1$ and t, multiplied by 10,000 to avoid small values. Volatility was computed as the square of returns.

Because of scarce trading activity during the week-ends, we excluded all returns computed

[3] The dates of the closed days are 25 and 26 December 2001, 1 January, 18 April and 1 May 2002.

[4] Danielsson and Payne (2002) show that the statistical properties of 5-minute dollar/DM quotes are similar to those of transaction quotes.

[5] Olsen and Associates recently changed the structure of their high frequency (HF) database. While they used to provide a 0/1 filter indicator (for example in the 1993 database), they now provide a continuous indicator that lies between 0 (worst quote quality) and 1 (best quote quality). Despite a value larger than 0.5 is deemed acceptable by Olsen and Associates, we chose a 0.85 threshold to ensure high quality data. in practice we removed virtually no data records (Olsen and Associates supplied us with data which had already been filtered at 0.5), as most filter values are very close to 1.

between Friday 21.30 and Sunday 24.00. In addition, we excluded the first return on each Monday and of each day following a closed day (other than week-ends) to avoid possible biases due to the lack of activity during the week-ends and closed days. We took the daylight saving time adjustment into consideration to account for the time changes (to winter and summer time) that occurred on October 29, 2001 and March 25, 2002. This affected GMT hours between 6.00 and 21.00 (corresponding to market times in Europe and the USA). As well as return volatility, a second important variable is quoting activity. FX quoting activity, measured by the number of quotes in five-minute time intervals, is often considered as a proxy for volatility and/or as a proxy for private information. Adjustments for week-ends and holidays were computed in the same way as for returns. The total number of observations for volatility and quoting activity was 72,675.

Table 4 presents summary statistics for the euro/dollar returns and quoting activity. The mean of the returns is almost equal to zero, although their distribution has fatter tails than the normal distribution and features a positive skewness coefficient. The quoting activity mean and standard deviation are relatively high. Its distribution is less leptokurtic than that of the returns, but much more asymmetric.

5.2 Results

Our empirical results based on FX volatility and quoting activity rely on the ACF analysis of the adjusted data. The presence of closed days generates a discontinuity in cycles, being the source of both stochastic seasonality.

Figure 2 illustrates the ACF for both deseasonalized volatility and deseasonalized quoting activity. The seasonality adjustment was done by the IAOM method. It is clear that, despite the adjustment, cycles remain in the series, particularly in quoting activity. Figure 3 displays the ACF for the same variables, adjusted by the kernel method. The cycles for volatility persist but are less pronounced than in the previous figure. Figure 4 shows an ACF from which the cycles have been almost completely removed. This was achieved by adjusting the volatility through a SOM(2,5) and the quoting activity by SOM(6,6).

Table 5 shows the mean, standard deviation and autocorrelation coefficient (AC) computed with one-day, two-days and three-day lags. The idea is to quantify the peaks in the different ACF cycles in order to simplify the comparison. The ACs corresponding to the different adjustment method are consistent with the figures. These results are consistent with the known features of our sample in terms of the discontinuity of the cycles involved. The SOM model is more efficient than the IAOM or the kernel models in term of seasonality adjustment, particularly when the seasonality has both deterministic and stochastic components.

6 Conclusion

This paper has focused on three seasonality identification methods: the self-organizing maps algorithm (SOM), the intra-day average observation method (IAOM) and the Nadaraya-Watson kernel method. The IAOM and the kernel methods have appeared in the literature before. We introduced the SOM algorithm in order to overcome some of their shortcomings. We studied the ability of each method to capture cycles involving deterministic and stochastic components. We implemented a Monte Carlo simulation in which we generated AR(1) and AR(5) processes infected by a seasonality involving both deterministic and stochastic cycles. Then, we captured and removed the cycles using each of the three methods. This allowed us to estimate the deseasonalization process and to compute and compare the estimation errors each one generate. In addition we used the three seasonality identification methods to capture and remove the cyclical components of two microstructure FX variables: volatility and quoting activity for the 5-minute euro/dollar currency quotes in the period from May 15, 2001 to May 15, 2002. The simulation results show that:

1. the IAOM model is much more efficient than the kernel or SOM methods when the seasonality contains only deterministic cycles;

2. when the seasonality involves both deterministic and stochastic cycles, the SOM model outperforms the other methods in capturing and identifying seasonality.

The empirical results for the real financial data yielded results consistent with those obtained from the simulations. The real data sample contained five closed days which triggered discontinuities and stochastic cycles. This explains, inter alia, why the SOM method outperformed the IAOM and kernel methods at identifying seasonality in the real data.

References

Andersen, T., and T. Bollerslev (1996): "Heterogeneous information arrivals and returns volatility dynamics: uncovering the long-run in high frequency rendements," NBER Working paper 5752.

——— (1998): "Deutsche mark-dollar volatility: intraday volatility patterns, macroeconomic announcements and longer run dependencies," *The Journal of Finance*, 1, 219–265.

Bauwens, L., W. Ben Omrane, and P. Giot (2005): "News Announcements, Market Activity and Volatility in the Euro/Dollar Foreign Exchange Market," *Journal of International Money and Finance*, 24, 1108–1125.

Bauwens, L., and P. Giot (2000): "The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks," *Annales d'Economie et Statistique*, 60.

Ben Omrane, W., and A. Heinen (2004): "The Information Content of Individual FX Dealers' Quoting Activity," IAG Working paper 120/04.

Bollerslev, T., and I. Domowitz (1993): "Trading patterns and prices in the inter-bank foreign exchange market," *The Journal of Finance*, 4, 1421–1443.

Cai, J., Y. Cheung, R. Lee, and M. Melvin (2001): "Once in a generation yen volatility in 1998: fundamentals, intervention and order flow," *Journal of International Money and Finance*, 20, 327–347.

Dacorogna, M., U. Müller, R. Nagler, R. Olsen, and O. Pictet (1993): "A geographical model for the daily and weekly seasonal volatility in the foreign exchange market," *Journal of International Money and Finance*, 12, 413–438.

Danielsson, J., and R. Payne (2002): "Real trading patterns and prices in the spot foreign exchange markets," *Journal of International Money and Finance*, 21, 203–222.

de Bodt, E., M. Cottrell, E. Henrion, and C. Van Wijmeersch (1998): "Self-Organizing Maps for Data Analysis : an Application to the Belgium Leasing Market," *Journal of Computing Intelligence in Finance*, 6, 5–24.

de Bodt, E., A. Lendasse, P. Cardon, and M. Verleysen (2004): "Self-organizing feature maps for the classification of investment funds," *Journal of Economic and Social Systems*, 17, 183–195.

DeBoeck, G., and T. Kohonen (1998): *Visual Explorations in Finance with Self-Organizing Maps*. Springer.

Degennaro, R., and R. Shrieves (1997): "Public information releases, private information arrival and volatility in the foreign exchange market," *Journal of Empirical Finance*, 4, 295–315.

Eddelbuttel, D., and T. McCurdy (1998): "The impact of news on foreign exchange rates : evidence from high frequency data," Working paper, University of Toronto.

Engle, R., and J. Russell (1998): "Autoregressive conditional duration: a new approach for irregularly spaced transaction data," *Econometrica*, 66, 1127–1162.

Kohonen, T. (1995): *Self-Organizing Maps*. Springer.

Lyons, R. (2001): *The Microstructure Approach to Exchange Rates*. MIT Press.

Melvin, M., and X. Yin (2000): "Public information arrival, exchange rate volatility and quote frequency," *The Economic Journal*, 110, 644–661.

Robbins, H., and S. Monro (1951): "A stochastic approximation model," *Annals of Mathematical Statistics*, 22, 400–407.

Silverman, B. (1986): *Density estimation for statistics and data analysis.* Chapman and hall.

Veredas, D., J. Rodriguez-Poo, and A. Espasa (2002): "On the (Intradaily) Seasonality and Dynamics of a Financial Point Process: A Semiparametric Approach," CORE DP, 2002/23.

Table 1: Input data for self-organizing maps: example observations

observations	Xi	Yi
1	18	46
3	21	47
4	19.5	53
5	20	52
6	18	51
7	20	45
8	5	12
9	6	16
10	5.5	5
11	6.7	8
12	4.9	10
13	6.1	9
14	7	13
15	5	52
16	6	56
17	5.5	45
18	6.7	48
19	4.9	50
20	6.1	49
21	7	55

Table 2: Estimation results for the AR(1) process with seasonality

$$y_t^{'} = \beta^{'} y_{t-1}^{'} + \epsilon_t^{'},$$
$$z_t^{'} = \gamma^{'} z_{t-1}^{'} + \nu_t^{'}.$$

	Non-Deseas.	*Deseas. IAOM*	*Deseas. SOM(1,5)*	*Deseas. Kernel*
		Panel A	(deterministic seasonality)	
$\beta^{'}$	0.9596	0.9499	0.9433	0.9510
σ	0.10%	0.07%	0.12%	0.12%
$RMSE$	0.96%	0.09%	0.67%	0.14%
		Panel B	(stochastic seasonality)	
$\gamma^{'}$	0.9702	0.9672	0.9464	0.9640
σ	0.22%	0.30%	0.12%	0.20%
$RMSE$	2.02%	1.72%	0.36%	1.40%

$y_t^{'}$ and $z_t^{'}$ are two AR(1) processes generated, through the Monte Carlo simulation. $\beta^{'}$ and $\gamma^{'}$ are the means of the estimated AR(1) coefficients after 1000 simulations. The first column presents the estimated coefficients for the non-deseasonalized AR(1) process. Columns two to four show the estimated coefficients for the AR(1) process deseasonalized by the IAOM, SOM(1,5) and kernel methods respectively. RMSE is the root mean square difference between the estimated coefficient and the simulated one.

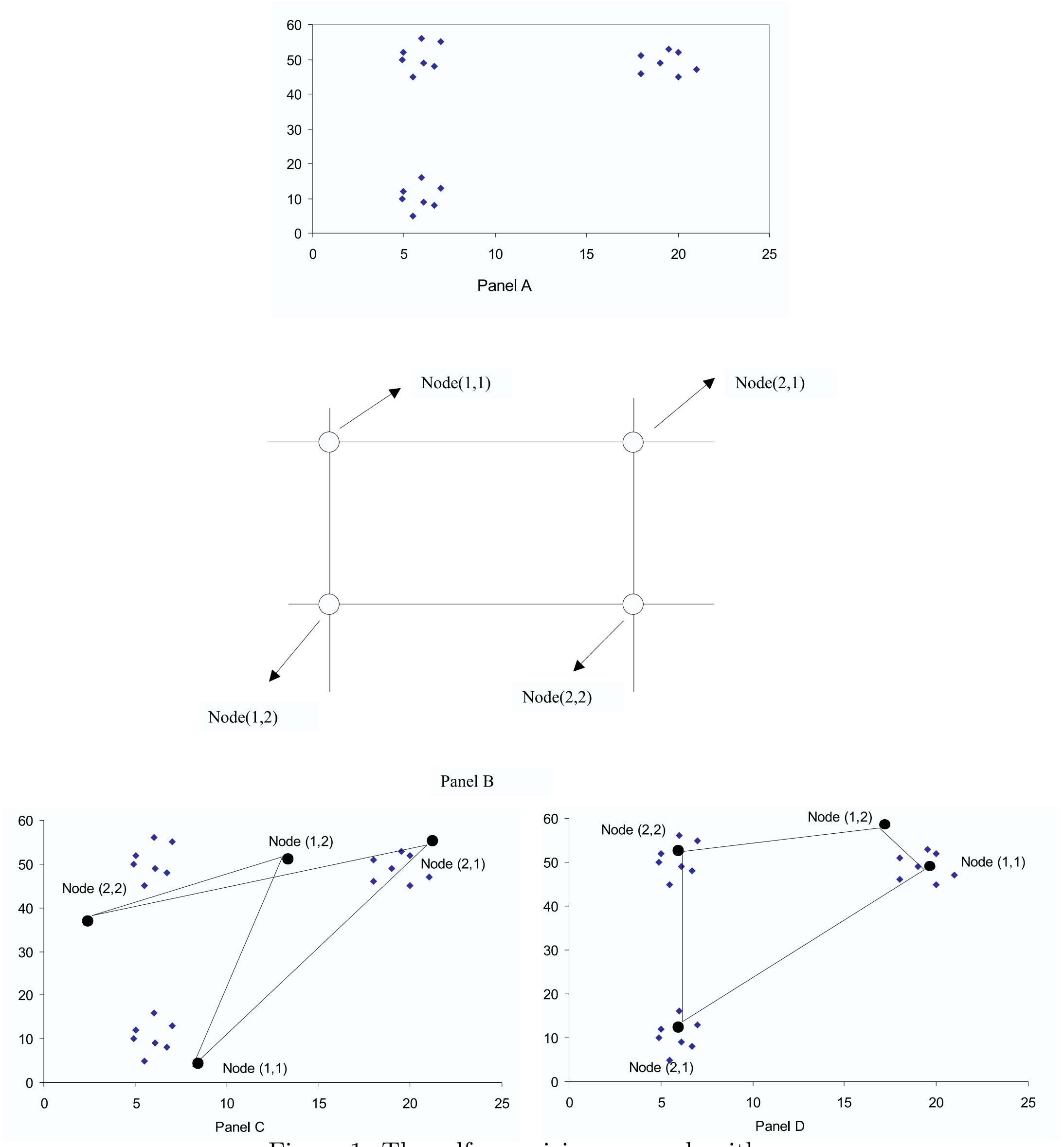

Figure 1: The self-organizing maps algorithm

Table 3: Estimation results for the AR(5) processes with seasonality

$$y_t^{'} = \sum_{p=1}^{5} \beta_p^{'} y_{t-p}^{'} + \epsilon_t^{'},$$
$$z_t^{'} = \sum_{p=1}^{5} \gamma_p^{'} z_{t-p}^{'} + \nu_t^{'}.$$

	Non-Deseas.	*Deseas. IAOM*	*Deseas. SOM(1,5)*	*Deseas. Kernel*
		Panel A	(deterministic seasonality)	
$\beta_1^{'}$	0.595	0.500	0.501	0.590
σ	0.35%	0.36%	0.52%	0.35%
$RMSE$	9.47%	0.29%	0.36%	9.00%
$\beta_2^{'}$	0.114	0.0901	0.0904	0.114
σ	0.42%	0.40%	0.44%	0.38%
$RMSE$	2.42%	0.33%	0.35%	2.37%
$\beta_3^{'}$	0.091	0.079	0.0802	0.088
σ	0.42%	0.40%	0.43%	0.36%
$RMSE$	1.09%	0.33%	0.34%	0.84%
$\beta_4^{'}$	0.073	0.070	0.0702	0.072
σ	0.44%	0.42%	0.44%	0.40%
$RMSE$	0.46%	0.34%	0.35%	0.38%
$\beta_5^{'}$	0.064	0.059	0.0601	0.059
σ	0.38%	0.39%	0.42%	0.38%
$RMSE$	0.43%	0.31%	0.33%	0.32%
		Panel B	(stochastic seasonality)	
$\gamma_1^{'}$	0.687	0.653	0.516	0.685
σ	2.13%	2.52%	0.61%	2.31%
$RMSE$	18.66%	15.26%	1.63%	18.49%
$\gamma_2^{'}$	0.114	0.116	0.0904	0.114
σ	0.47%	0.45%	0.47%	0.52%
$RMSE$	2.35%	2.55%	0.72%	2.37%
$\gamma_3^{'}$	0.077	0.084	0.085	0.077
σ	0.61%	0.61%	0.45%	0.59%
$RMSE$	0.52%	0.58%	0.59%	0.52%
$\gamma_4^{'}$	0.054	0.062	0.075	0.054
σ	0.67%	0.73%	0.45%	0.70%
$RMSE$	1.60%	0.89%	0.54%	1.57%
$\gamma_5^{'}$	0.036	0.047	0.067	0.037
σ	0.72%	0.87%	0.42%	0.80%
$RMSE$	2.41%	1.36%	0.71%	2.34%

See caption of Table 2.

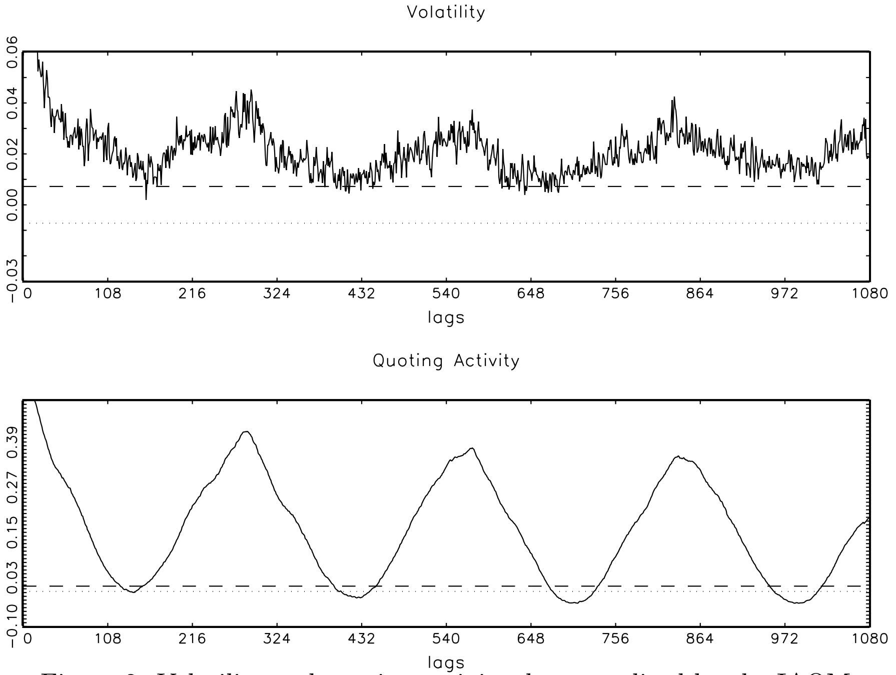

Figure 2: Volatility and quoting activity deseasonalized by the IAOM

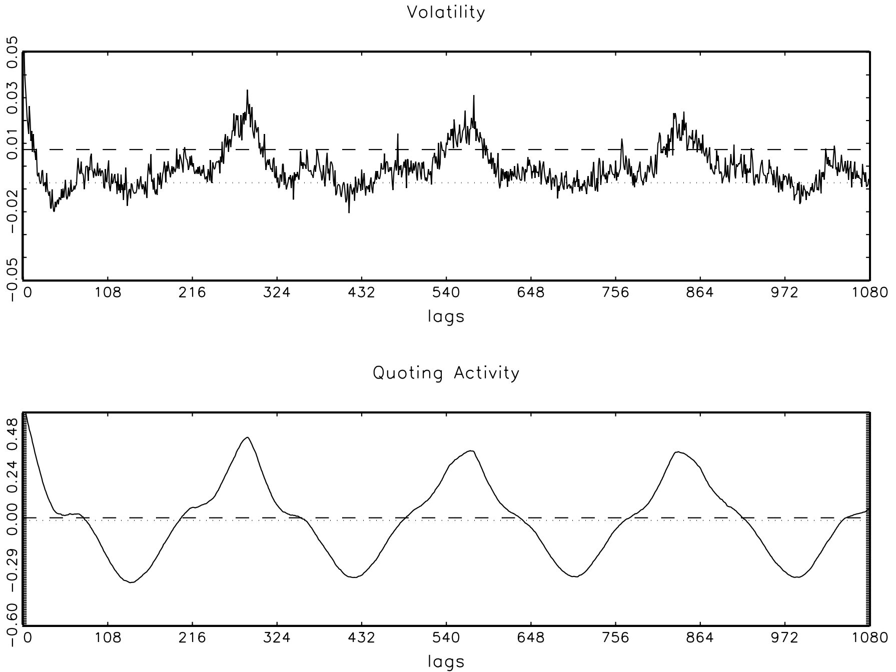

Figure 3: Volatility and quoting activity deseasonalized by the Nadaraya-Watson kernel smoothing method

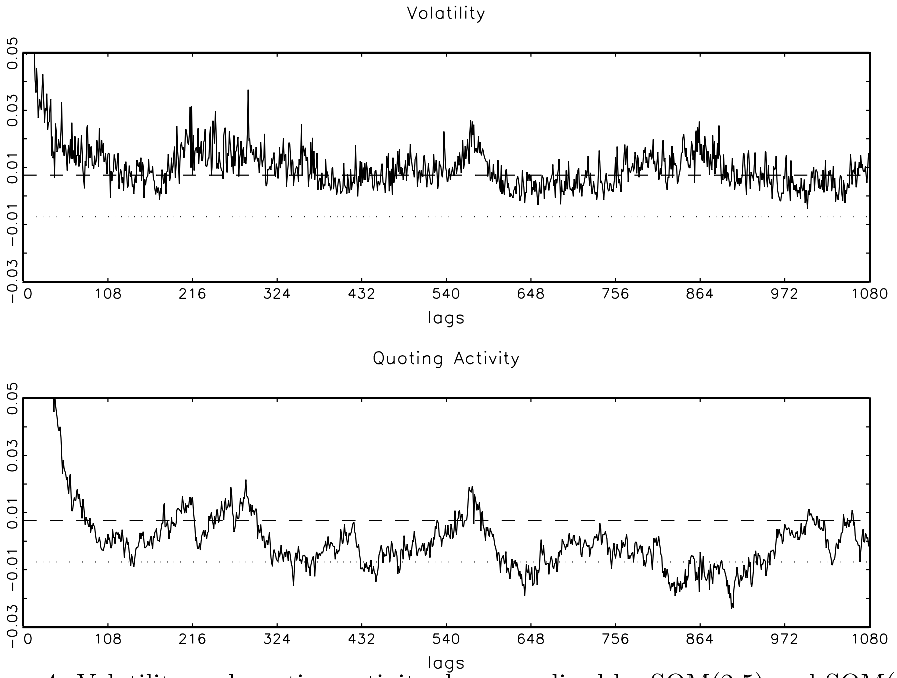

Figure 4: Volatility and quoting activity deseasonalized by SOM(2,5) and SOM(6,6)

Table 4: Moments of the euro/dollar returns and quoting activity

	Returns	Quoting activity
Mean	0.007	82.31
Standard deviation	3.91	60.11
Skewness coefficient	0.21	1.23
Kurtosis coefficient	15.0	5.43

The 5-minute returns have been pre-multiplied by 10,000 (to avoid small values). The number of observations is 72,675, corresponding to the period from May 15, 2001 to May 15, 2002.

Table 5: Moments and autocorrelation coefficient for the euro/dollar deseasonalized volatility and quoting activity

	IAOM	*IAOM*	*SOM(2,5)*	*SOM(6,6)*	*Kernel*	*Kernel*
	V	QA	V	QA	V	QA
μ	0.999	1.000	1.038	0.998	0.918	0.923
σ	2.180	0.554	2.629	0.387	2.045	0.440
ρ_{288}	0.037	0.417	0.037	0.014	0.027	0.461
ρ_{576}	0.032	0.372	0.025	0.006	0.031	0.382
ρ_{864}	0.030	0.320	0.026	-0.017	0.011	0.295

The number of observations is 72,675, corresponding to the period from May 15, 2001 to May 15, 2002. The seasonality adjustments were made by using the SOM model presented in Section 3.1, the intra-day average observations model (IAOM) presented in Section 3.2, and the kernel smoothing method detailed in Section 3.3.