Using Self-Organizing Maps to Adjust Intra-Day Seasonality

Walid Ben Omrane, Eric de Bodt
Department of Business Administration (IAG), Finance Unit, Catholic University of Louvain, Place des Doyens 1, 1348 Louvain-la-Neuve, Belgium
benomrane@fin.ucl.ac.be.
debodt@fin.ucl.ac.be

Abstract - *The existence of an intra-day seasonality component within financial market variables (volatility, volume, activity,...), has been highlighted in many previous works. To adjust raw data from their cyclical component, many studies start by implementing the intra-daily 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 seasonality involves only a deterministic component, IAOM method succeed in estimating periodicity almost without estimation error. However, when seasonality contains both deterministic and stochastic components (e.g. closed days), we show that either the IAOM or the kernel method fail to capture it. We introduce the use of the self-organizing maps (SOM) as a solution. SOM are based on neural network learning and nonlinear projections. Their flexibility allows capturing seasonality even in the presence of stochastic cycles.*

Key words - self-organizing maps, currency market, intra-day seasonality, high frequency data

1 Introduction

Evidences of intra-daily seasonality in financial market behaviors has been highlighted in many prior studies. Some recent references include [7], [1], [11], [5], [2] and [4]. These works illustrate the existence of seasonality in many microstructure variables (e.g. FOREX volatility and quoting activity). Two categories of methods are most often used in order to remove this seasonality. Some studies like [7], [1], [5] and [4] adopt a linear projection technique. They regress variables (affected by the seasonal component) on a set of dummies variables (or flexible Fourier form) in order to capture intra-day cycles. Other authors adjust raw data from seasonality using a direct correction factor, obtained by intra-daily average ([6], [8], [11], and [2]) or a smoothing kernel ([9], [3], and [13]).

This work builds on the previous literature to explore the limits of the classical approaches and to introduce a solution in the case of stochastic cycles. We show that the more the raw data involves a deterministic seasonality, the more the classical methods, particularly the intra-daily 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 the ones generated by closed days (among others), classical methods reveal their limits. We introduce in this a method based on the self-organizing maps algorithm ([10]). The self-organizing maps (SOM) allows capturing both deterministic and stochastic cyclical

components and purging endogenous variables from the seasonal component.
Our evidences are based on a Monte Carlo simulations. Our Monte Carlo simulations adopt a five-step framework. We begin by generating an auto-regressive process. We then simulate either only a deterministic seasonality, or both deterministic and stochastic cycles which we add to the auto-regressive variable. Next, we partition the generated variable into block of observations, each one representing a day of the week. After that we deseasonalize the endogenous variable, using the three methods cited above. We finally re-estimate the process coefficients on the deseasonalized data series. Better is deseasonalization, closer should be the estimated coefficient to the simulated one, and lower should be the root mean square error (RMSE). This allows use comparing the performance of the above cited methods in a controlled setup.
This paper is divided in five sections. In Section 2 we detail our deseasonalization methods. We present the Monte Carlo simulation in Section 3 and we show the results in Section 4. We finally conclude.

2 Deseasonalization Methods

2.1 The Self-Organizing Maps Model ($\boldsymbol{SOM}$)

The self-organizing maps (SOM) introduced by [10] can be considered as a method of data analysis which allows, through a (discrete) projection, to reduce the dimension of the data space (as principle component analysis methods do). Simultaneously it allows, through vector quantization, to summarize the data projected in specific mean profiles. The projection step is carried out on a discrete data space.[1]

2.2 The Intra-daily Average Observations Model ($\boldsymbol{IAOM}$)

To estimate seasonality, we compute the intra-daily 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 have one observation for the random variable, Y. We thus compute in principle Q values mv_{n_k} for each day of the week, that makes a total of W $(5 \times Q)$ values over a week. Formally,

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

where

$$f(s,k,n_k) = W\,(s-1) + \sum_{j=1}^{k-1} N_j + n_k, \tag{2}$$

$s = 1, \ldots, S$. $k = 1, \ldots, 5$. $N_1 = N_2 = N_3 = N_4 = N_5 = Q$. $n_1 = 1, \ldots, Q$. n_2, n_3, n_4 and n_5 likewise.
To adjust the different variables for seasonality, we implement the same methodology used for the SOM adjustment. We just divide/withdraw them at the endpoint of each five minute interval by/from the corresponding value of the intra-daily average observation. That

[1] We refer the reader to [10] for a detailed presentation of SOM.

means, for example, that all quoting activity at 12h on Thursday in the sample are withdrawed/divided by the same value (the average quoting activity at 12h on Thursday).

2.3 The Smoothing Method

It consists in smoothing the raw data using the Nadaraya-Watson kernel estimator and then adjusting each raw observation by the correspondent value on the smooth curve. The adjustment is done as for the SOM and IAOM methods.
The *Nadaraya-Watson* kernel estimator $\hat{Y}_t$ of $Y_(t)$ is:

$$\hat{Y}_t = \frac{\sum_{j=1}^{T} K_h(t - t_j) Y_t}{\sum_{j=1}^{T} K_h(t - t_j)} \,. \tag{3}$$

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 select a Gaussian kernel with a bandwidth, h, computed by [12]:

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}} \tag{4}$$

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

where σ_k is the standard deviations for the observations.

3 Monte Carlo Simulation

3.1 Simulation Procedure

In order to compare the three seasonality identification methods (IAOM, SOM, NW-kernel) we implement a five-step simulation procedure.
1) We start by generating a P-lag autoregressive process, y_t^*, (AR(P), $P = 1, 5$):

$$y_t^* = \sum_{p=1}^{P} \beta_p y_{t-p}^* + \epsilon_t, \tag{6}$$

where β is equal to 0.95 if $P = 1$, and if $P = 5$, then $\beta_1 = 0.5$, $\beta_2 = 0.09$, $\beta_3 = 0.08$, $\beta_4 = 0.07$, and $\beta_5 = 0.06$. ϵ_t is distributed as a standard Normal.
2) We partition y_t^* by block of Q observations each one representing a day of the week.
3) We simulate a deterministic seasonality $S_{t,i}^{det}$ and we add it to the above AR(P) process, such that:

$$y_{t,i} = y_{t,i}^* + S_{t,i}^{det} \,. \tag{7}$$

Let $y_{t,i}^*$ represent one such block, where i is an index corresponding respectively to the open days of the week ($i = 1, \ldots, 5$). $S_{t,i}^{det}$ is generated by the following procedure: we divide the block of observations, corresponding to each day of the week, into three time frame (let say the morning, the noon, and the afternoon). Then, we add a defined constant to the AR(P) process depending on the specific time frame in which the observation is located. One set of constants is chosen for each day of the week, since we generate a deterministic seasonality. In such a way, y_t becomes an autoregressive variable which involves a deterministic seasonality. To simulate an AR(P) process which contains stochastic seasonality added to the deterministic one, we go through the following procedure:

- We generate an AR(P) process:

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

- We add to this process a deterministic and stochastic seasonality:

$$z_{t,i} = z_{t,i}^* + S_{t,i}^{det} + S_{t,i}^{sto}, \tag{9}$$

$S_{t,i}^{det}$ is generated as described above, and $S_{t,i}^{sto}$ is the stochastic seasonality. The difference between the latter seasonality and the former one consists on the manner of which we add constants to the time frame in the weekdays. In the stochastic seasonality case, days are selected randomly to be subject for added seasonality. Moreover, seasonality changes from a week to another.

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

$$y_t^{'} = y_t - \phi_t^{det}, \tag{10}$$

$$z_t^{'} = z_t - \phi_t^{sto}. \tag{11}$$

5) Finally, we estimate both AR(P) processes, based respectively on the deseasonalized variables, using ordinary least square estimation:

$$y_t^{'} = \sum_{p=1}^{P} \beta_p^{'} y_{t-p}^{'} + \epsilon_t^{'}, \tag{12}$$

$$z_t^{'} = \sum_{p=1}^{P} \gamma_p^{'} z_{t-p}^{'} + \nu_t^{'}. \tag{13}$$

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

4 Results

Estimation results for the Monte Carlo simulation are presented in Tables 1 and 2. The latter table displays the estimation results for AR(5) process, and the former one presents those of AR(1). The panels A in both tables display the mean, the standard deviation, and the RMSE corresponding to 1000 estimation of the autoregressive coefficient for equation (12) in presence of deterministic seasonality. The panels B illustrates the same results for the stochastic seasonality (equation (13)). The variables, in this case, are deseasonalized from their deterministic and stochastic seasonality. The RMSE in both panels characterize the estimation error generated by the added seasonality. It is the root mean square difference between the simulated coefficients and the recovered ones after adding, estimating and removing seasonality, β' and γ'.

Table 1: Estimation results for the AR(1) processes 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)	
β'	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)	
γ'	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%

Starting with deterministic seasonality results, the estimated coefficients for the non-deseasonalized autoregressive parameters corresponding to equation (12) and (13) (see the second column of Tables 1 and 2) shows a higher error level in Panel B results than in panel A. The more there is seasonality into the process, the more important is the error in the estimated coefficients. This is the reason why previous studies try to get rid from the cyclical component involved in their microstructure variables.

The IAOM deseasonalization displays interesting results in panel A. The estimated coefficients is very close to simulated ones with an insignificant error equal to 0.01% for AR(1) and around 0.35% for the different coefficients of the AR(5). We conclude that the IAOM method succeeds in capturing almost the whole 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, e.g., to closed days, data recording problems, ...). Panel B presents very different results. In the presence of stochastic cycles, the IAOM method leads to a significant estimation error level. The corresponding RMSE is much higher than the error obtained by estimating the model with deterministic seasonality. When the seasonality

Table 2: 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	0.901	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	0.0.059	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%

involves both deterministic and stochastic elements, the IAOM does in fact not capture the whole cyclical component of the process. When there are good reasons to think that the seasonality could display some stochastic behavior, the IAOM approach should no be used.
The SOM method seems to be far more robust to the presence of stochastic cycles. The panel B of Tables 1 and 2 exhibits, in the forth row, respectively the estimation result for the seasonality adjusted AR(1) and AR(5) processes. The corresponding RMSE is low compared to the IAOM case. Contrary to the IAOM method, SOM(1,5) succeeds in capturing seasonality involving both deterministic and stochastic cycles. Results displayed by panel A, for both tables, show that SOM(1,5) is however less efficient than the IAOM method when seasonality involves only deterministic cycles. In such a case, the estimation error generated by SOM(1,5) is much higher than the one generated by IAOM method. The choice between the IAOM and SOM depends therefore on the presence of stochastic cycles.
The kernel results displayed in Table 1, Panel A, show that the method captures deterministic seasonality with a low error level. This finding is consistent with previous research which opt for the kernel method as a step in their deseasonalization process in particular when their samples exhibit some deterministic cycles. However, the kernel adjustment is less accurate than IAOM. Nevertheless, Table 2, Panel A, displays a higher RMSE for the kernel method, especially for the first three coefficients of the AR(5).
Panel B results, corresponding to Tables 1 and 2, show that the kernel method generates an estimation error level much higher than the one generated by SOM and IAOM but quite smaller than non-adjusted data.
These findings are consistent with the intuition. By construction, in the case of deterministic seasonality, the IAOM method, built on the computation of the cross sectional means, can easily capture the seasonality. The IAOM algorithm relies indeed on the law of large numbers: the deterministic component estimation amounts to an estimation of the hour by hour cycle expected value by its sample average. The SOM algorithm and kernel methods can as well capture deterministic cycles but much less efficiently than the IAOM. Nonetheless, in case of cycle irregularities (as often observed in financial data), using hour by hour sample average to capture the seasonality becomes problematic. The SOM model goes beyond the limits of IAOM and the kernel models. It estimates, efficiently, the seasonality which contains both deterministic and stochastic cycles.

5 Conclusion

This paper focus 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 been used previously in the literature. We introduce the SOM algorithm in order to overcome some of their shortcomings.
We study the ability of each method to capture cycles involving deterministic and stochastic components. We implement a Monte Carlo simulation in which we generate an AR(1) and AR(5) processes infected by a seasonality involving deterministic and stochastic cycles. Then, we capture and remove the cycles by implementing the three methods. We estimate, afterward, the deseasonalized process and we compute and compare the estimation generated errors.
The simulation outputs carry out the following results: 1-the IAOM model is much more

efficient than the kernel and the SOM methods when seasonality contains only deterministic cycles. 2-When seasonality involves both deterministic and stochastic cycles, SOM model outperforms the other methods in capturing and identifying seasonality.

References

[1] T. Andersen, and T. Bollerslev (1998), Deutsche mark-dollar volatility: intraday volatility patterns, macroeconomic announcements and longer run dependencies, *Journal of finance*, **vol. 1** p. 219-265.

[2] L. Bauwens, W. Ben Omrane, and P. Giot (2003), News announcements, market activity and volatility in the euro/dollar foreign exchange market, *Journal of international money and finance*, **forthcoming**.

[3] L. Bauwens, and P. Giot (2000), The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks, *Annales d'économie et de statistiques*, **vol. 60**.

[4] W. Ben Omrane, and A. Heinen (2004), The information content of individual FX dealers' quoting activity, *IAG WP*, **120/04**.

[5] J. Cai, 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*, **vol. 20** p. 327-347.

[6] M. Dacorogna, U. Müller, R. Nagler, R. Olsen, and 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*, **vol. 12** p. 413-438.

[7] R. Degennaro, and R. Shrieves (1997), Public information releases, private information arrival and volatility in the foreign exchange market, *Journal of empirical finance*, **vol. 4** p. 295-315.

[8] D. Eddelbuttel, and T. McCurdy (1998), The impact of news on foreign exchange rates: evidence from high frequency data , *University of Toronto WP*.

[9] R. Engle, and J. Russell (1998), Autoregressive conditional duration: a new approach for irregularly spaced transaction data , *Econometrica* **vol. 66** p. 1127-1162.

[10] T. Kohonen (1995), *Self-organizing maps*, Berlin, Springer.

[11] M. Melvin, and X. Yin (2000), Public information arrival, exchange rate volatility and quote frequency , *The economic journal* **vol. 110** p. 644-661.

[12] B. Silverman (1986), *Density estimation for statistics and data analysis*, London, Chapman and hall.

[13] D. Veredas, J. Rodriguez-Poo, and A. Espasa (2002), On the (intradaily) seasonality and dynamics of a financial point process: a semiparapetric approach, *CORE DP*, **2002/23**.