ACSEG 2005, Connectionist Approaches in Economics and Management Sciences,
Aix-en-Provence (France), 17-18 November 2005, pp. 78-89.

Using the Self-Organizing Maps to prove empirically the market inefficiency: Evidence from Paris Stock Exchange

Didier Catteau
Université catholique de Louvain

Geoffroy Simon [*]
Machine Learning Group
Université catholique de Louvain

Walid Ben Omrane
Unité Finance d'Entreprise
Université catholique de Louvain

Michel Verleysen
Machine Learning Group
Université catholique de Louvain
and SAMOS
Université Paris I Panthéon-Sorbonne

Abstract

Market efficiency was the subject of long and continuing debate. In some recent work by Chordia, Roll and Subrahmanyam it has been proved empirically that the New-York Stock Exchange is inefficient over a short period of thirty minutes. The proof is based on linear models over short time intervals. In this paper, a linear regression is first applied to data from the Paris Stock Exchange that is found to be inefficient over an interval of five minutes. Then an original nonlinear method using Self-Organizing Maps is proposed. This new methodology allows detecting market inefficiency over longer time intervals, while making an intuitive graphical representation possible. The method is used to prove that the Paris Stock Exchange is inefficient on time intervals up to 30 minutes.

J.E.L. classification: C32, G14

Keywords: Market efficiency, order imbalances, Self-Organizing Maps.

[*] Corresponding author, simon@dice.ucl.ac.be

1 Introduction

Market efficiency was the subject of long and continuing debate since the 1970s ([2], [3] and [4] among others). The evidences of all the researches converge to the same conclusion: stock prices involve all the information shared by the market participants, they always incorporate the best information about the fundamental values, returns are random walk, and there is no arbitrage opportunity to make money by beating the market. However these researches were built on the general behaviour of the stock markets, and their empirical evidences involve low frequency returns (monthly, weekly or daily returns). In addition, [5] and [6] have mentioned too many anomalies triggered by the market microstructure and dealers behaviour.

Recently [1] showed that financial markets converge to efficiency according to a certain speed. Stock markets are inefficient over a very short period of time and are then pushed toward efficiency by the collective behaviour of traders. The latter undertake countervailing trades sufficient to remove all serial dependence. Their evidence is based on intraday returns and order imbalances (that are defined as the difference between buyer initiated orders and seller ones) for 150 NYSE stocks. They compute the serial dependence of returns and they regress it on lagged returns and order imbalance according to five different frequencies (5,10, 15, 30 and 60 minutes). They find that returns are negatively correlated over intervals of up to ten minutes, and order imbalance are highly positively dependent and they predict future returns over intervals of up to thirty minutes. These results show strong evidence that the market is not efficient over very short time intervals.

In this paper, we try to check for the speed of convergence to market efficiency on Paris Stock Exchange. We start by using almost the same methodology as in [1], and we add a nonlinear technique built on the Self-Organizing Map (SOM) algorithm introduced by Kohonen [7], in order to check for the robustness of our results carried out by the first methodology. SOM is based on neural network learning and nonlinear projections.

The estimation results corresponding to the regression framework carry out strong evidence that Paris Stock Exchange is inefficient over a short period of time of five minutes just after the trade. Then, the market reverts to efficiency: the returns lose their serial dependence and there is no longer predictive power triggered by order imbalances.

Self-Organizing Maps are used on vectors containing past values of returns and order imbalances. The obtained empirical results strengthen the conclusion obtained by linear models: market is inefficient for short period of 5 minutes. Furthermore, the SOM allows concluding that market is also inefficient over longer intervals up to 30 minutes. This resuld could not be deduced from linear models.

The remaining of this paper is divided in four sections. In Section 2 the models used to prove market inefficiency are presented. In addition to the linear and SOM models, a detailed explanation of the specific way SOM are used in this context is proposed. The data are then described in Section 3. Section 4 shows the experimental results on Paris Stock exchange and leads to the conclusions of this paper in section 5.

2 Models for market inefficiency

2.1 Linear technique

In a first step stock returns (R) are computed as the difference between the prices in time $t+1$ and t. Order imbalances (OIB) are also computed as the difference between the market purchase and sell orders. Then a linear model is build, based on regression technique and ordinary least square (OLS) estimation. Regressions for returns are obtained using past return (R) and lagged order imbalance (OIB) values:

$$R_t = \alpha + \sum_{i=1}^{T} \beta_i R_{t-i} + \sum_{j=1}^{T} \gamma_j OIB_{t-j} + \varepsilon_t, \tag{1}$$

where R_t is the stock return at time t, T is the total number of lags and ε_t is a white noise following a Gaussian distribution.

The speed of convergence to market efficiency is deduced from the statistical significance of the estimated coefficients β_i and γ_j corresponding to the lagged parameters. The market is efficient if and only if the estimated coefficients are all not significant according to a t-test. Otherwise future returns can be predicted through their past or by using the information involved in order imbalance. In such a case returns do not follow a random walk and the market is inefficient.

Note that lags are introduced in order to know the horizon over which the market reverts to efficiency. However, if all the estimated coefficients corresponding to the different lags are statistically not significant, then the market is efficient. Otherwise the market is inefficient over the interval corresponding to the observed lag.

2.2 Nonlinear Technique: Self-Organizing Maps

The Self-Organizing Map (SOM) can be defined as an unsupervised classification algorithm. Since its introduction in the 80's by Kohonen [7], SOM have been used in many different applications [7]. Their theoretical properties are now well established [8], [9].

Briefly speaking, a Self-Organizing Map places a fixed number of *centroids* or *prototype vectors* or *prototypes* in short $\bar{x}$ in a given data space. More formally, SOM performs a rough approximation of the local data density by means of the prototype density.

The key concept of SOM is the *neighbourhood* between prototypes. The prototypes are linked by neighbourhood relations using a predefined grid or *map*, usually a 1- or 2-dimensional one. This grid constrains the learning stage of the SOM algorithm.

The neighbour prototypes can be characterized by their distance on the grid with respect to the considered prototype. This grid distance is only related to prototype positions on the grid. Figure 1 shows the neighbours of a given prototype for a 2-dimensional square map. Neighbours at a grid distance of 1 and 2 from the considered prototype in position (4,4), using a squared neighbourhood, are highlighted. For the SOM learning stage, a maximum size is given to the

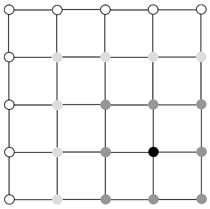

Figure 1: A 2-dimensional SOM square grid. The neighbourhood relations are illustrated by line segments between prototypes. The prototype (4,4) is shown in black. The 1-neighbours are in dark grey, the 2-neighbours are in light grey. If a 2-neighbourhood is considered, the last prototypes shown in white are too far away from the considered prototype in (4,4).

neighbourhood.

During the learning, the prototypes are moved iteratively within the data space according to the location of the considered x_t data and to neighbourhood constraints. After the learning stage, the set of prototypes has established a vector quantization of the data. Each prototype $\bar{x}$ is associated to a region of the space, namely a *cluster*, where data x_t share some similar features. Each data is usually associated to a cluster so that the corresponding prototype is the nearest one to the data. A clustering can thus be obtained by considering data subsets containing a single prototype and the associated x_t data.

The SOM prototypes have also the property of data topology preservation: two similar data belong either to the same cluster or to two neighbouring ones (on the grid). Furthermore, the SOM obtained after learning allows graphical representations that can be interpreted intuitively.

The SOM is usually used as a classification tool and have already been applied in the financial context. See [10] for example. In this paper, an original representation of the transition between prototypes will be provided as an empirical proof of market inefficiency. When build on time series, SOM associates a prototype to each data vector formed by successive values of the series. Sliding the window over time then generates successive vectors. Transitions between successive vectors can thus be analysed, as well as transitions between prototypes. This analysis is further detailed in the rest of the paper.

2.3 Prototype shape and prototype shape plot

A first graphical representation deals with the *shape* of the prototypes after learning. The goal of this representation is to provide a qualitative criterion for observing the SOM convergence.

The prototype shape is obtained by plotting the various components according to their ordering in the vector: the first component is the first dot of the plot; the second component is the second dot, etc. The final shape is obtained by linking the dots, as shown in the left part of Figure 2.

Shapes can be constructed for each prototype of the SOM grid and gathered together in a *prototype shape plot*. Such a prototype shape plot is shown in Figure 2 right for a SOM with 36 prototypes on a 6x6 grid obtained from the well known Santa Fe A time series benchmark [11], using 6-dimensional data vectors $x_t = (x(t), x(t-1), x(t-2), x(t-3), x(t-5), x(t-6))$.

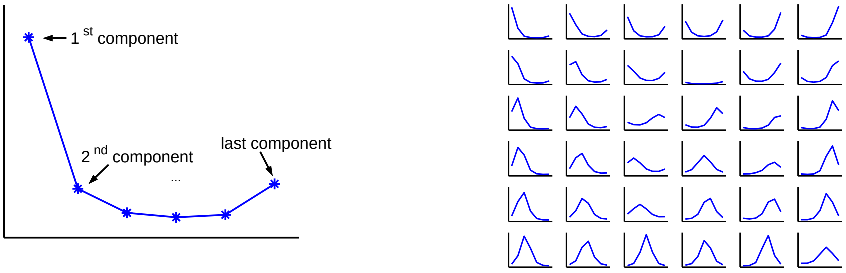

Figure 2: Left: Prototype shape construction: the prototype components are plotted according to their order in the vector. The final shape is obtained by linking the dots. Right: A prototype shape plot example obtained from a 6x6 square SOM on the Santa Fe A time series.

The prototype shape plot helps qualifying the correctness of the SOM convergence. Intuitively, a good convergence is obtained when similar data are grouped in the same cluster, or in neighbouring ones. The qualitative criterion is thus: if similar prototype shapes are grouped on the prototype shape plot, the convergence is satisfactory; if dissimilar prototype shapes are neighbours, the SOM has converged poorly. In the best case, the difference between neighbour prototype shapes should be smooth; a continuum should be observed in the prototype shapes while moving along a line, a column, or even a diagonal of the prototype shape plot. This continuum can be observed on Figure 2 right that shows a correct convergence.

2.4 Theoretical and empirical transition matrices

An original graphical representation based on SOM is now introduced. The main idea of this new representation is to observe the transition probabilities between clusters. These transition probabilities are computed from the data associated to each cluster. Two approaches will be introduced leading to two transition matrices. Comparisons between these two matrices will provide the second empirical proof of market inefficiency.

The first matrix is the theoretical transition matrix that represents the expectation of having a transition from a prototype to itself or to one of its 1-neighbours. This matrix is constructed as follows: Once the SOM shows a good convergence, a counter c_i is associated to each prototype $\bar{x}_i$. For each data x_t its nearest prototype counter is incremented. Once all data have been considered, the number of data associated to each cluster equals c_i. Then the total number of data n_i in the 1-neighouring of a given prototype $\bar{x}_i$ is the sum of the considered prototype counter c_i and its 1-neighbours counter c_j. The probability of transition from prototype $\bar{x}_i$ to a 1-neighbour $\bar{x}_j$ (or to itself) is equals to c_j (resp. c_i) divided by n_i. For each prototype a (3*3) matrix is thus created containing the theoretical probabilities of transition from this prototype to either itself or anyone of its 1-neighbours. These operations are repeated for each prototype, leading to the theoretical transition matrix that contains I (3*3) matrices arranged as the prototypes are arranged on the SOM grid where I is the total number of prototypes. Obviously, transition probabilities of prototypes on the grid edges are set to zero for directions leading outside of the grid. The left hand side of Figure 3 shows an example of such a theoretical

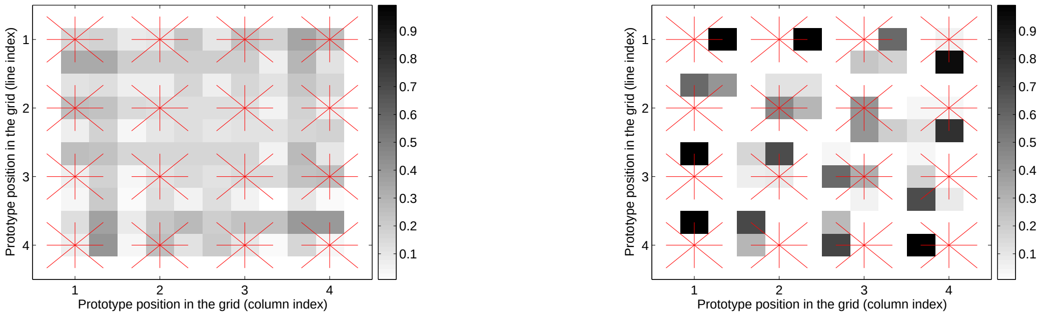

Figure 3: Illustration of the theoretical (left) and empirical (right) transition matrices computed on the Santa Fe A time series using a (4*4) square SOM. The relative significance of the transition probabilities are given by the greyscale legend.

transition matrix, obtained from the Santa Fe A time series [11]. Each prototype is symbolised by a big asterisk centred on it. The transitions to the eight 1-neighbours are represented by a numerical value presented in greyscale at the end of each line segment of the asterisk.

The second matrix is the empirical transition matrix that represents the expectation of having a transition from one prototype to itself or to one of its 1-neighbours with respect to the temporal dependences between data. This second matrix is constructed accordingly: After the SOM convergence, the transitions from x_t to x_{t+1} are considered for each data x_t. A counter c_{ij} associated to the prototype couple $(\bar{x}_i, \bar{x}_j)$ is incremented if the transition from x_t associated to prototype $\bar{x}_i$ leads to x_{t+1} which has as nearest prototype $\bar{x}_j$, a 1-neighbour of $\bar{x}_i$. The counters are thus now associated to the temporal transitions between prototypes instead of being associated to the prototypes themselves, as in the theoretical transition matrix. Having considered all the transitions between two successive data, each counter associated to a couple $(\bar{x}_i, \bar{x}_j)$ is divided by the sum of all the 9 counters c_{ij} associated to $\bar{x}_i$. These operations lead to the empirical transition probabilities of each prototype, arranged in a (3*3) matrix. The empirical transition matrix is thus obtained again by grouping the I (3*3) matrices. Here again, transition probabilities of prototypes on the grid edges are set to zero for directions leading outside of the grid. The right-hand side of Figure 3 shows an example of empirical transition matrix, obtained from the Santa Fe A time series. The graphical conventions are the same as in the theoretical transition matrix case.

The theoretical and empirical transition matrices and their graphical representations give the second empirical proof of market inefficiency. By construction, the theoretical transition matrix is expected to contain only mid-range probabilities. Indeed, if the SOM converges well, it is expected to contain about the same number of data in each cluster. All transitions from one cluster to another are thus expected to be the similar. Suppose now that the market is efficient. Starting with data x_t, the following temporal data x_{t+1} can be any of the data reachable from x_t, as the transitions are all equiprobable under the efficiency hypothesis. As SOM preserves the data topology, the transitions should thus be equiprobable from one cluster to any other cluster in the neighbourhood even when temporal dependences are taken into account. As a consequence, under the hypothesis of efficiency, both the theoretical and empirical transition matrices obtained from the SOM are expected to be approximately uniformly distributed. If it is the case, this provide an empirical proof of market efficiency, otherwise the market is inefficient.

The efficiency or inefficiency conclusion can be deduced easily with the greyscale representations of the two transition matrices. See for example the left and right parts of Figure 3, obtained from the Santa Fe A time series, which is known to be predictable [11]. In the left figure, the expected transitions are approximately uniformly distributed. Moreover the empirical transition matrix on the right figure shows unevenly distributed probabilities: For each prototype, only a limited number of 1-neighbours are likely to be selected. Furthermore, the transitions to the very few reachable 1-neighbours happen with a high probability. Comparing figure 3 left and right thus lead to the conclusion that the Santa Fe A time series is predictable, which is indeed the case [11].

3 Data description: Paris stock exchange

The evidence for market inefficiency is based on five representative stocks from the Paris Stock Exchange. The latter is an electronic market that operates through a market and limit orders. Buy and sell orders are collected into the order book before their execution. The exact definition of the database content is detailed in [12].

Five out of the most traded stocks are selected, corresponding to relatively high capitalized firms: Carrefour, France Telecom, L'Oreal, Suez and Total. The frequency of the database is five minutes and the period of the study corresponds to 2 months per year, Mars and April, from 1998 until 2001. Focusing on a discontinued period instead of a continued one avoids having to face the problems triggered by market seasonality.

Returns are built on mid-quote prices, where the mid-quote is the average of the last bid and the last ask prices corresponding to each time interval. Order imbalances are estimated using the Lee and Ready algorithm [13], [14]. This algorithm classifies a trade as buyer (seller) initiated if it is closer to the ask (bid) of the prevailing quote. If the trade is exactly at the midpoint of the quote, a "tick test" is adopted whereby the trade is classified as buyer (seller) initiated if the last price change prior to the trade is positive (negative). Then, the order imbalance (OIB) corresponding to different stocks and related to a fixed time interval, is computed as the number of buyer less the number of seller-initiated trades. OIB measures, in turn, the size of the trades. Both returns and OIB are computed within a grid corresponding to a frequency of five minutes.

To check for the speed of convergence to efficiency lagged variables are introduced in order to guess the time horizon in which there is a serial dependence between the returns, their past values and the OIB ones. This horizon is fixed to thirty minutes.

Variable	Coefficient	t-Statistic
α	3.77E-06	0.149
β_1 (coef. for R_{t-1})	-2.13E-02	-1.782
β_2 (coef. for R_{t-2})	-1.83E-02	-1.569
β_3 (coef. for R_{t-3})	-5.65E-03	-0.428
β_4 (coef. for R_{t-4})	3.59E-03	0.363
β_5 (coef. for R_{t-5})	5.20E-03	0.469
β_6 (coef. for R_{t-6})	5.02E-03	0.449
γ_1 (coef. for OIB_{t-1})	1.26E-05	3.591
γ_2 (coef. for OIB_{t-2})	-4.32E-06	-1.257
γ_3 (coef. for OIB_{t-3})	-1.61E-06	-0.595
γ_4 (coef. for OIB_{t-4})	-1.57E-06	-0.486
γ_5 (coef. for OIB_{t-5})	7.50E-07	0.175
γ_6 (coef. for OIB_{t-6})	5.46E-07	0.287

Table 1: t-test on the coefficient of the linear regression for the aggregated market.

4 Experimental results

4.1 Linear model

The results presented here are obtained on the aggregated market. This aggregate is built upon an average computation, as in Chordia et al [1]. Table 1 contains the estimation results corresponding to the aggregate market corresponding to five representative stocks of Paris Stock Exchange (Carrefour, France Telecom, L'Oreal, Suez and Total).

Considering the rejection level of 5%, only the first lags corresponding to both returns R and order imbalance OIB are statistically significant. The first lagged estimated coefficient of R is negative and all the remaining lagged coefficients are not significant. By contrast, the first lagged estimated coefficient of OIB is positive. This implies that a day characterized by a high imbalance on the buy-side will likely be followed by several additional days of aggregate buy-side imbalances, and similarly for the sell-side imbalance. This means, as mentioned by [1], that investors continue buying or selling for quite long time, either because they are herding, or because they are splitting large orders across days, or both. Furthermore, the negative sign of the first order correlation for returns means that some astute investors must be correctly forecasting continued price pressure from order imbalances and conducting countervailing trades within the very first minutes, which could be sufficient to remove all serial dependence in returns.

To summarize, returns present a first order serial correlation up to five minutes, and order imbalance have a positive significant impact on returns which does not go beyond five minutes. In turn, there is evidence, according to our estimation results, that the Paris Stock Exchange is not efficient over a short interval of five minutes just after the trade. In no more than five minutes order imbalances lose their predictive power and returns are no longer serially dependent; in other words the market reverts to efficiency.

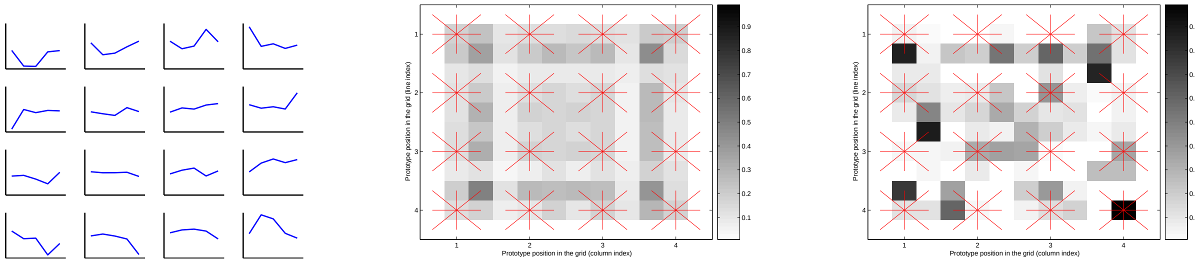

Figure 4: From left to right: Prototype shape plot for the 5 firms aggregated to represent the global market, using a (4*4) SOM; Theoretical and empirical transition matrices for 5 minutes intervals.

4.2 Self-Organizing Maps

As for the linear model, SOMs are learned on vectors composed of past values of the returns $R$and order imbalances OIB. In all experiments these vectors are created according to:

$$SOM - vector_t = \{R_t, R_{t-1}, OIB_{t-1}, OIB_{t-2}, OIB_{t-3}\}, \tag{2}$$

where t varies for different time lags, from 5, to 10, 20 and 30 minutes.

To observe the market efficiency, the evolution of the five firms considered is aggregated as for the linear case. The results obtained using the SOM for the aggregated firms are presented in Figure 4. On the left part of the figure, it can be seen that the SOM has converged correctly. The middle part of Figure 4 is the theoretical transition matrix that is more or less uniformly grey, as expected. Looking to the right hand side of Figure 4 it is clear that the empirical transition matrix is very different from its theoretical counterpart. It is thus observed that the market is inefficient over 5 minute intervals.

The same experiment is repeated for intervals of 10, 20 and 30 minutes. Figure 5 shows a summary of all the experiments for the market, obtained as aggregation of the 5 considered firms. In all cases, comparing the theoretical and empirical transitions matrices lead to the conclusion of market inefficiency: the theoretical transition matrices are always more or less uniformly grey while the empirical transition matrices are not uniform at all.

Experimental results for 60 minutes intervals for the market as aggregation of the 5 series are provided in Figure 6. In this case, the theoretical transition matrix is no longer uniformly grey. The upper half of the plot is far from being uniform. This behaviour is so far away from the one expected by construction that it is no more possible to make any conclusion for inefficiency from these figures. Furthermore, the empirical transition matrix is very poorly populated. This can be explained by the fact that transitions occur between 2-neighbours, instead of 1-neighbours as in all the previous cases. Thus for time intervals longer than 30 minutes, this empirical tool based on SOM is not able to detect market inefficiency.

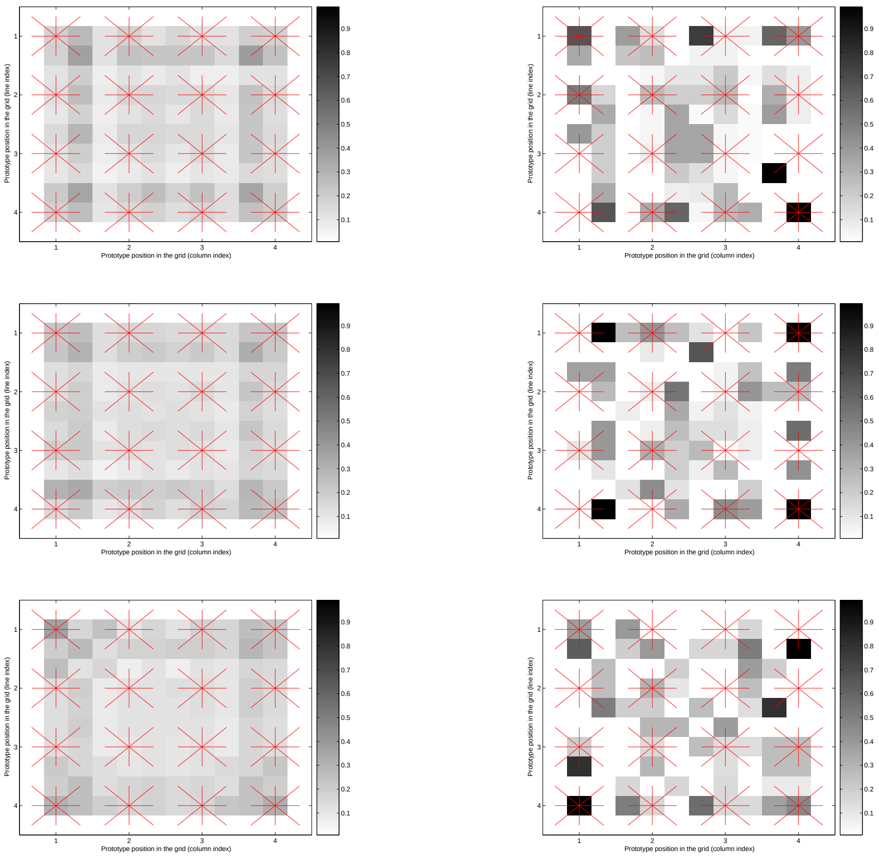

Figure 5: Top to bottom: Theoretical (left) and empirical (right) transition matrix for the market (obtained from 5 aggregated firms). Top: 10 minutes interval; middle: 20 minutes interval; bottom: 30 minutes interval. In all cases, (4*4) SOM were used.

5 Conclusion

In this paper, the question of market efficiency of the Paris Stock Exchange has been observed. An original method based on Self-Organizing Maps has been proposed to prove empirically the market inefficiency.

Linear models are used in a way comparable to [1] in order to provide a first argument for the Paris Stock Exchange inefficiency at 5 minutes time intervals.

The SOM are then used to provide another empirical proof over longer time intervals. An original representation of the theoretical and empirical transition matrices, deduced from the clusters created by the SOM algorithm, is then introduced using intuitive plots. Comparisons of the two matrices are used as a criterion for proving market inefficiency. The proposed intuitive criterion leads to the conclusion of market inefficiency over 30 minutes. Experiments on 60 minutes intervals lead to the conclusion that market inefficiency can not be proved using the proposed SOM tool.

Further works include a more in depth research for the longest time interval such that the new methodology based on SOMs is still enable to detect inefficiency. Another field of research would be the definition of a mathematical criterion for matrix comparison in addition to the current graphical one.

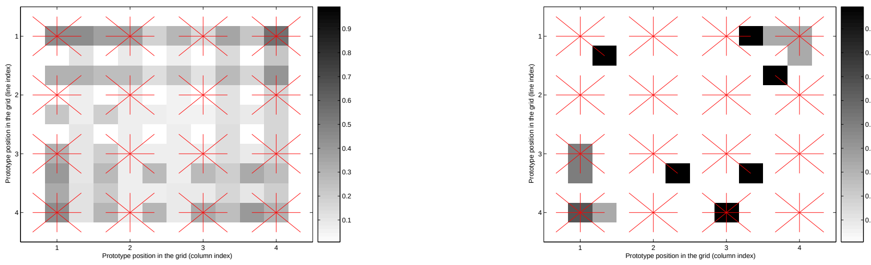

Figure 6: Theoretical and empirical transition matrix for 60 minutes time intervals for the market (obtained from as the 5 aggregated firms). (4*4) SOM were also used for these final experiments.

Acknowledgements

G. Simon is funded by the Belgian F.R.I.A.; M. Verleysen is Senior Research Associate of the Belgian F.N.R.S.

References

[1] Chordia, T., Roll, R., and Subrahmanyam, A., Evidence on the Speed of Convergence to Market Efficiency, *Journal of Financial Economics*, 76:271–292, 2005.

[2] Fama, E., On the Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance*, 25:383–417, 1970.

[3] Black, F., Capital Market Equilibrium with Restricted Borrowing, *Journal of Business*, 45:444–455, 1972.

[4] Lucas, R., Asset Prices in an Exchange Economy, *Econometrica*, 46:1429–1445, 1978.

[5] Black, F., Noise, *Journal of Finance*, 41:529–543, 1986.

[6] Shiller, R. J., From Efficient Markets to Behavioral Finance, *Journal of Economic Perspectives*, 17:59–82, 2003.

[7] Kohonen, T., *Self-organising Maps* (3rd ed.), Springer Series in Information Sciences, Springer, 2001.

[8] Cottrell, M., Fort, J.-C., and Pagès, G., Theoretical aspects of the SOM algorithm, *Neurocomputing*, 21:119–138, 1998.

[9] Cottrell, M., de Bodt, E., and Verleysen, M. (1997), Kohonen maps versus vector quantization for data analysis, in M. Verleysen (Ed.), *Proceedings of European Symposium on Artificial Neural Networks*, Bruges (Belgium), D-Facto pub., 187-193, 2002.

[10] Deboeck, G. and Kohonen, T., *Visual Explorations in Finance with self-organizing maps*, Springer-Verlag, 1998.

[11] Weigend, A. S., and Gershenfeld, N. A., *Times Series Prediction: Forecasting the future and Understanding the Past*, Addison-Wesley Publishing Company, 1994.

[12] ParisBourse SA, *Base de données : Premier marché - Second marché - Nouveau marché - MONEP*, 1999.

[13] Lee, C. M. C., and Ready, M. J., Inferring Trade direction from Intraday Data, *Journal of Finance*, 46:733–746, 1991.

[14] Boehmer, E., Broussard, J. and Kallunki, J. P. (2002), Analysis of transaction data, in Boehmer, E., Broussard, J. and Kallunki, J. P. *Using SAS in financial research*, SAS enterprises, 119-130, 2002.