UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

QUANTITATIVE MEASURES OF EFFICIENCY AND COLLECTIVE ORDINAL DYNAMICS IN HIGH-FREQUENCY DIGITAL MARKETS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR (A) EN CIENCIAS

PRESENTA:
MARIO ALEJANDRO LÓPEZ PÉREZ

DIRECTOR
RICARDO MANSILLA CORONA (CEIICH, UNAM)

MIEMBROS DEL COMITÉ TUTOR
PEDRO MIRAMONTES VIDAL (FACULTAD DE CIENCIAS)
PABLO PADILLA LONGORIA (IIMAS)

CIUDAD DE MÉXICO, SEPTIEMBRE DE 2023

The purpose of studying economics is not to acquire a set of ready-made answers to economic questions, but to learn how to avoid being deceived by economists.

Joan Robinson, *Marx, Marshall and Keynes*

Hay dos panes. Usted se come dos. Yo ninguno. Consumo promedio: un pan por persona.

Nicanor Parra

Contents

1 Introduction **5**

2 Theoretical Framework and Data **13**

2.1 Hurst exponent 14

2.2 Ordinal Patterns 16

2.3 Data 20

3 Analysis of Efficiency in High-Frequency Digital Markets Using the Hurst Exponent **23**

3.1 Introduction 23

3.2 Methodology 25

3.3 Results and discussion 26

3.4 Conclusions 28

4 Ordinal Synchronization and Typical States in High-Frequency Digital Markets **37**

4.1 Introduction 38

4.2 Individual Analysis of Stocks through Ordinal Patterns 39

4.3 Collective Analysis of Stocks through Transcript Synchronicity Dynamical Networks . . 41

4.4 Clustering Analysis 47

4.5 A Markov Model for State Transitions 53

4.6 Conclusions 54

Appendices **59**

4.A Correlation Matrixes 59

4.B Multiscale Analysis 60

5 Conclusions **63**

Chapter 1

Introduction

> Capital by its nature drives beyond every spatial barrier. Thus the creation of the physical conditions of exchange – of the means of communication and transport – the annihilation of space by time – becomes an extraordinary necessity for it.
>
> Karl Marx, *Grundrisse*

> Speed is now taking primacy over quantity, as a factor in wealth. The hidden face of the maintenance of accumulation is the acceleration of circulation. The function of the control devices is thus to maximize the volume of commodity flows by minimizing the events, obstacles, and accidents that would slow them down. Cybernetic capitalism tends to abolish time itself, to maximize fluid circulation to the maximum: the speed of light. Such is already the case for certain financial transactions. The categories of "real time" of "just in time", show clearly this hatred of duration. For this very reason, time is our ally.
>
> Tiqqun, *The Cybernetic Hypothesis*

Durante mucho tiempo los mercados financieros modernos han fascinado tanto a investigadores como al público en general, desde los esfuerzos in situ para comprender las primeras burbujas financieras modernas [1] hasta los estudios sociológicos, económicos y técnicos de su estructura y función social

[2][3]. Han sido blanco de los elogios más entusiastas, como heraldos del desarrollo económico y la innovación [4], así como de la censura más severa, como un juego psicológico, especulativo y, dejado a sus anchas, socialmente dañino [5]; esto no debería de sorprendernos, toda vez que los últimos dos siglos han atestiguado su surgimiento como una de las características más destacadas de nuestra sociedad global contemporánea [6], tal como los aún sensibles recuerdos y consecuencias que dejó tras de sí la crisis financiera mundial de 2008 se encargan de recordarnos [7]. Más aún, los mercados financieros y su fauna sin escrúpulos han penetrado tan profundamente en nuestra imaginación como para convertirse en protagonistas de un género cinematográfico en crecimiento que abarca películas y documentales como *Wall Street Wolf*, *The Big Short* e *Inside Job*.

Con todo, es necesario recuperar una y otra vez una idea que, por obvia que pueda parecer, ha sido sorprendentemente negligida por muchos famosos teóricos del mercado [8]: los mercados financieros son, como cualquier otra institución social, de carácter histórico [9]. Cambian con el tiempo, tanto social como técnicamente. Es por eso que es inútil proponer teorías globales y ahistóricas de los mercados financieros. Se hace en cambio necesario continuar renovando nuestra comprensión de ellos a medida que evolucionan.

En este trabajo nos centraremos en una encarnación histórica muy específica de los mercados financieros: los contemporáneos y ya ubicuos Mercados Digitales de Alta Frecuencia. Desde los años setenta y principalmente desde los noventa, los mercados financieros han experimentado cambios importante en su estructura técnica y social: cada vez más, son algoritmos los que toman decisiones sobre qué, cuándo, cómo y cuánto comprar y vender en el mercado, todo esto en una escala temporal de milisegundos. En palabras de un historiador de tales transformaciones: "En los mercados financieros actuales, la lógica no es la de coordinar las interacciones interpersonales sino la de administrar las señales electrónicas puntuadas que codifican las órdenes de masas de inversionistas anónimos. El arte de las finanzas ya no se trata de miradas y señales manuales, sino de jugar con algoritmos ágiles, procesadores informáticos sofisticados, enrutadores pirateados y sistemas de telecomunicaciones especializados que son los cimientos materiales de la bolsa de valores contemporánea. A través de la tecnología, los pisos comerciales se convirtieron en una amalgama de cables y software; y a través de la automatización, las bulliciosas multitudes humanas se transformaron en colas electrónicas silenciosas y veloces" [10]. Nacidos en la encrucijada del desarrollo tecnológico, la búsqueda de ganancias, la opacidad institucionalizada, la fragmentación espacial y los cambios políticos, los Mercados Digitales de Alta Frecuencia y la veloz carrera armamentista que estos provocan son una característica fundamental, aunque relativamente reciente y no bien entendida, de la vida social contemporánea, cuyo desacoplamiento de los tiempos humanos y de comercio-a-la-velocidad-de-la-luz [11] conlleva consecuencias impredecibles y potencialmente catastróficas,

como quedó demostrado con el llamado Flash Crash del 6 de mayo de 2010 cuando, sin motivo aparente, el mercado cayó seiscientos puntos en pocos minutos, recuperando su estado anterior en solo unos minutos más. El dominio de los algoritmos ha reforzado la opacidad estructural de los mercados [12][13], al igual que en muchas otras áreas de la vida social contemporánea en las que la inteligencia artificial juega un papel importante [14]. Como se narra en [13], tan recientemente como en 2010 esta opacidad había impedido que muchos actores del mercado tuvieran una idea adecuada de lo que ocurría frente a ellos.

Por supuesto, una institución tecnosocial tan compleja merece ser y ha sido estudiada desde muchos puntos de vista diferentes por economistas, antropólogos, historiadores, estadísticos y, quizás sorprendentemente, físicos. Aunque la historia de la relación entre Física y Finanzas se remonta al menos al trabajo pionero de L. Bachelier [15] publicado hace más de un siglo, e incluso antes [16], no fue sino en las últimas décadas que se estableció un creciente campo del conocimiento que aplica herramientas de la Física, y particularmente de la Mecánica Estadística, para estudiar fenómenos económicos a través del estudio de series de tiempo [17][18][19]. Para citar solo algunos ejemplos recientes, en [20] y [21] los autores aplican la teoría de matrices aleatorias para estudiar correlaciones en mercados financieros con el fin de describir varios estados del mercado y transiciones entre estados, mientras que en [22] se presenta un modelo microscópico para transacciones de alta frecuencia realizadas automáticamente, utilizando herramientas de la teoría cinética de los gases. Nuestro trabajo se enmarca en esa tradición, conocida con el nombre de Econofísica.

Así, abordaremos los Mercados Digitales de Alta Frecuencia desde un punto de vista matemático, aplicando un conjunto de herramientas estadísticas y analíticas provenientes de diferentes campos, desde la Teoría de Sistemas Complejos hasta el Aprendizaje Estadístico No Supervisado, al estudio de nuestro conjunto de datos particular, que cubre un período de un año de transacciones totalmente automatizadas en los mercados de Estados Unidos y México, del 7 de marzo de 2018 al 7 de marzo de 2018.

Nos ocuparemos de dos cuestiones específicas. Primero abordaremos el clásico tema de la eficiencia del mercado. La Hipótesis del Mercado Eficiente (HME), un dogma de la economía ortodoxa, postula la incorporación inmediata de toda la información relevante para la formación de los precios a través de la interacción en los mercados de agentes económicos altamente racionales, lo que resulta en un "mercado justo" basado en precios en principio impredecibles. Para los mercados clásicos, existe amplia evidencia en contra de la eficiencia, comenzando con el trabajo pionero de Mandelbrot [23]. Sin embargo, en los Mercados Digitales de Alta Frecuencia contemporáneos, para cuyos agentes económicos "verdaderamente racionales" (esto es, los algoritmos) se podría esperar que se cumpliera la HME, la evidencia a favor o

en contra es menos abundante. En este trabajo ofreceremos evidencia estadística clara contra la HME en Mercados Digitales de Alta Frecuencia a través del Análisis del Exponente de Hurst de acciones individuales, discutiendo también algunas consideraciones metodológicas sobre su validación estadística.

Una vez establecida la ineficiencia para nuestros datos, profundizaremos en nuestro entendimiento de los Mercados Digitales de Alta Frecuencia al estudiarlos de forma colectiva, como una red dinámica, utilizando para ello herramientas como patrones ordinales y análisis de agrupamiento (clustering), avanzando así en la dirección señalada por los recientes estudios sobre detección de estados típicos del mercado, bien ejemplificados en [24], así como por el trabajo clásico de Bandt y Pompe [25]. Nuestra red dinámica de acciones, el peso de cuyos aristas se define por medio de una medida de sincronziación ordinal inspirada en Teoría de la Información, nos permite detectar días atípicos, así como dos temporadas separadas del año comercial, caracterizadas por su grado de sincronicidad centralizada o descentralizada. Todo esto se confirma cuantitativamente mediante la aplicación de un par de algoritmos de agrupamiento para encontrar estados significativos y cuantitativamente distinguibles del mercado, correspondientes a diferentes niveles dinámica colectiva centralizada/descentralizada, finalmente modelados mediante un proceso de Markov simple.

Nuestra investigación es metodológica tanto como aplicada: queremos comprender los Mercados Digitales de Alta Frecuencia, así como proponer nuevas rutas metodológicas para lograrlo, estadística y fenomenológicamente.

Este trabajo está organizado de la siguiente manera: el capítulo 2 contiene el marco teórico necesario para comprender nuestros resultados originales: presenta el análisis R/S y la teoría del exponente de Hurst, así como los patrones ordinales y algunas medidas de complejidad que aplican conceptos de Teoría de la Información, cerrando con los detalles de nuestro conjunto de datos. El cuerpo principal de este trabajo está contenido en los capítulos 3, que contiene nuestro análisis de Hurst de la eficiencia en los Mercados Digitales de Alta Frecuencia, y 4, que aborda el tema de la dinámica ordinal colectiva y los estados típicos en tales mercados. El capítulo 5 cierra el trabajo con las conclusiones generales.

Modern financial markets have fascinated researchers and the general public for a long time, from the in-situ efforts to understand the first modern financial bubbles [1] to the sociological, economical and technical accounts of their structure and social role [2][3]. They have been targets for the most enthusiast eulogies, as heralds of economic development and innovation [4], as well as for the more severe reprimands, as a psicological, speculative and, if let to its own, socially harmful game [5]; this is no surprise, as the last two centuries have witnessed their raising as one of the most salient features of our contemporary global society [6], as the memories and consequences left behind by the 2008 global financial crisis, still felt today, take care of reminding us [7]. Indeed, financial markets and their reckless fauna have penetrated so deeply in our imagination as to become main characters in a growing filming genre encompasing movies and documentaries such as *Wall Street Wolf*, *The Big Short* and *Inside Job*.

Even so, we have to bring again and again an idea which, obvious as it seems, has been astonishingly overlooked by many famous market theorists [8]: Financial markets are, as any other social institutions, historical in character [9]. They change through time, socially as well as technically. That is why it is futile to advance global, a-historical theories of financial markets. Thus, it is necessary to keep renewing our understanding of them as they evolve.

Here, we will focus in a very specific historical incarnation of financial markets: contemporary and already ubiquitous High-Frequency Digital Markets. Since the seventies, but mainly since the nineties, financial markets have gone through a major change in its technical and social structure: more and more, they are algorithms who make decisions about what, when, how and how much to trade in the market, all of this in the scale of miliseconds. In words of an historian of such transformations: "In present-day financial markets, the logic is not one of coordinating interpersonal interactions but of managing the punctuated electronic signals that encode the orders from masses of anonymous investors. The art of finance is no longer about gazes and hand signals, but about toying with the nimble algorithms, sophisticated computer processors, hacked routers, and specialized telecommunication systems that are the material foundations of the contemporary stock exchange. Through technology, trading floors became an amalgam of cables and software; and through automation, rowdy human crowds were refashioned into silent and speedy electronic queues" [10]. Born in the crossroads of technology development, profit seeking, institutionalized opaqueness, spatial fragmentation and political changes [26], High-Frequency Digital Markets and the speed arms race they provoque are a fundamental, yet relatively recent and not well understood, feature of contemporary social life, its decouplement of human and light-speed-trading times [11] bearing unpredictable, potentially catastrophic consecuences, as demonstrated by the so-called Flash Crash on May 6, 2010 when, for no obvious reason, the market fell six hundred points in a few minutes, recovering to its previous state in just a few minutes more. The rising of algorithms have

rendered markets structurally even more opaque [12][13], just as in many other areas of contemporary social life in which artifical intelligence has an important role [14]. As narrated in [13] and depicted in movies ans series such as *Margin Call* and *The Fear Index*, as recently as 2010 this opaqueness had prevented too many market players from knowing what was going on.

Of course, such a complex techno-social institution deserves being and has been studied from many different points of view by economists, anthropologists, historians, statisticians and, perhaps surprisingly, physicists. Although the history of the relation between Physics and Finance dates back at least to the pioneer work of L. Bachelier [15] published more than a century ago and even earlier [16], it was not until recent decades that a growing and well established body of research using tools from Physics, and particularly statistical mechanics, to understand economic phenomena through time series analysis had emerged [17][18] [19]. To cite just a few recent examples, in [20] and [21] the authors apply random matrix theory to study cross-correlations in financial markets, in order to describe various market states and state transitions, while in [22] a microscopic model for automatically done high-frequency transactions is presented, using tools from the kinetic theory of gases. Our work is framed in that tradition, branded as Econophysics.

So, we will approach High-Frequency Digital Markets from a mathematical point of view, applying a set of statistical and analytical tools coming from different fields from Complex Systems Theory to Unsupervised Statistical Learning to the study of our data set, covering a period of one year of fully automated transactions in the US and Mexican markets, from March 7, 2018 to March 7, 2018.

Two specific questions will concern us in this work. First, we will adress the classical issue of market efficiency. The Efficient Market Hypothesis (EMH), a dogma in mainstream economics, postulates the immediate incorporation of all relevant information for the formation of prices through the interaction in the markets of highly rational economic agents, which results in a "fair market" based on in-principle-unpredictable prices. For classical markets, there is ample evidence against efficiency, starting with the path-breaking work by Mandelbrot [23]. However, for contemporary High-Frequency Digital Markets, for whose " truly rational" economic agents (that is, algorithms) it could be expected that EMH should be fulfilled, the evidence is less abundant in either direction. We will offer clear statistical evidence against EMH in High-Frequency Digital Markets through Hurst Exponent Analysis of individual stocks, also offering some methodological considerations about its appropiate statistical validation.

After inefficiency is established for our data set, we will go deeper in our study of High-Frequency Digital Markets by studying our data set, this time collectivelly, as a dynamical network through Ordinal Patterns and Clustering Analysis, thus building on recent work on typical market states detection, well

exemplified in [24], as well as on the classical work of Bandt and Pompe [25]. Our dynamical network of stocks, the weight of whose edges is defined by means of an information-theoretic measure of ordinal synchronization, allows us to detect outlier trading days as well as two separate seasons of the treading year, characterized by their degree of centralized or decentralized synchronicity. All this is quantitatively confirmed by applying a couple of clustering algorithms to find meaningful and quantitatively distinguishable market states corresponding to different levels of centralized collective dynamics, to be modeled by a simple Markov process.

Our inquiry is both methodological and applied: we want to understand High-Frequency Digital Markets as well as to propose new methodological insights on how to accomplish that, both statistically and phenomenologically.

This work is organized as follows: chapter 2 contains the theoretical framework needed to understand our original results: it introduces R/S analysis and Hurst exponent theory, as well as ordinal patterns and information-theoretic measures of complexity, closing with the details on our data set. The main body of this work es contained in chapters 3, containing our Hurst analysis of efficiency in High-Frequency Digital Markets, and 4, tackling the issue of collective ordinal dynamics and typical states in such markets. Chapter 5 closes the work with general conclusions.

Chapter 2

Theoretical Framework and Data

My young son asks me: Must I learn mathematics?
What's the use, I'd like to say.
That two pieces
Of bread are more than one–
You can see that already.
My young son asks me: Must I learn French?
What is the use, I feel like saying. This State's collapsing.
And if you just rub your belly with your hand and
Groan, you'll be understood with little trouble.
My young son asks me: Must I learn history?
What is the use, I feel like saying. Learn to stick
Your head in the earth, and maybe you'll still survive.
Yes, learn mathematics, I tell him.
Learn your French, learn your history!

Bertolt Brecht, *1940*

In this chapter we introduce all the theory needed to understand the results of this investigation, which are exposed in the next two chapters. Accordingly, this chapter is divided in two sections, the first of which discusses the Hurst exponent, to be extensively applied in chapter 3 to test efficiency, while the second one discusses ordinal patterns and information-theoretic measures, which are the base for the study on market collective dynamics in chapter 4.

2.1 Hurst exponent

The method that we will use to test efficiency has its origins in the work of Hurst [27], framed in the context of his studies in hydrology and later refined by Mandelbrot and Wallis [28]. Given a time series x_n, $n = 1, \ldots, m$, with mean $\mu = (1/m) \sum_{i=1}^{m} x_i$ and variance $\sigma^2 = (1/m) \sum_{i=1}^{m} (x_i - \mu)^2$, we define its partial centered sums as $y_n = \sum_{i=1}^{n} (x_i - \mu)$ and its R/S statistics or rescaled range as the ratio between the range of the partial centered sums series and the standard deviation of the original series:

$$R/S = \frac{\max_{n \leqslant m} y_n - \min_{n \leqslant m} y_n}{\sigma},$$

Hurst noted that the rescaled range of the time series of annual flows of the Nilo river as a function of the length n of the series was asymptotically a power law when n tends to infinity: $\mathrm{E}(R/S(n)) \sim n^H$ for n sufficiently large, where $\mathrm{E}(R/S(n))$ is the mean of the R/S statistics calculated on the subseries of length n of the original series. The exponent H of the power law is known as the *Hurst exponent*. It is known that under the hypothesis of the original series being a random walk, the exponent is $H = 0.5$ [28]; instead, Hurst found $H > 0.5$.

The Hurst exponent of a time series is obtained in this work as follows: Given n less than the length of the series, we calculate R/S of all subseries of length n of the original series and define $E(R/S(n))$ as the average of these calculations. Finally, the Hurst exponent of the series is calculated as the exponent of the function $c \cdot n^H$ that best fits (in the least squares sense) the function $n \mapsto E(R/S(n))$ for n large enough. Taking into account the asymptotic nature of the Hurst exponent, as well as the sensitivity of the R/S statistic to the length of the time series [29], uniformly spaced values of n are taken on a logarithmic scale, with a minimum n of the order of 2^9.

In general, a Hurst exponent H greater than 0.5 is associated with the long-term persistence of the series: the range grows faster than expected from a random walk, that is, movements in one direction follow, with greater probability, movements in the same direction; while $H < 0.5$ is associated with long-term antipersistence: the range grows more slowly than that of a random walk, that is, movements in one direction are more likely to follow movements in the other direction, this on average and for large enough lengths. In both cases, the deviation of H from its hypothetical value 0.5 can be taken as a long-term memory measure: the movements of the series are not independent of the remote past.

The study of long-term correlations measured by Hurst exponent has been applied in Physics, for example in [30] to the ion saturation current fluctuations and in [31] to gamma ray data. In the context of financial time series that concerns us here, this long-term memory translates into deviations from market

efficiency (see chapter 3).

The remaining part of this section will be devoted to briefly discuss the approaches in the literature that will be used in the next section to define our methodology, as well as some results related to those of this work.

In [32] it is argued that, given a sequence of independent and identically distributed random variables, the shape of the probability distributions of the random variables affects the Hurst exponent of the series. The authors calculate the Hurst exponent H_{stock} for daily series of the S&P500 index. Then, they shuffle the series in order to remove its memory, after which they calculate the Hurst index of the shuffled series, denoted by H_{perm} (actually, this process of shuffling and calculating is repeated a certain number of times, H_{perm} is defined as the mean and the standard deviation is reported). The difference between H_{stock} and H_{perm} is an indicator of the memory of the original series, while if H_{perm} is different of 0.5 this is attributed to the distributions of the variables, since the shuffled series are, by construction, memoryless. It is proposed then to use H_{perm} as an indicator of lack of memory, instead of the canonical value 0.5, that is, $H_{\text{stock}} > (<) H_{\text{perm}}$ would be and indicator of (anti)persistency. In other words, the significant null hypothesis will be not $H_{\text{stock}} = 0.5$, but $H_{\text{stock}} = H_{\text{perm}}$.

In [33] it is argued that the R/S statistic is sensitive to short-term correlations, so that if for a time series is obtained $H \neq 0.5$, this is not enough to conclude the presence of long-term memory. In [34] the authors propose that, to ensure that the results of the R/S analysis are due to long-term correlations, the following experiment is carried out: the series is divided into blocks of, for example, 50 elements each one, and the elements within each block are permuted to destroy the short-range correlations. With this new series the previous analyzes are repeated and the long-term memory is corroborated if the change in the Hurst exponent is insignificant.

In [35] and [36] the need to observe the evolution of the efficiency, measured by the Hurst exponent, over time is stated. The first article uses one-minute resolution data from 1983 to 2009 from the SP500. The Hurst exponent of the daily subseries is calculated and a decreasing evolution is observed from 0.8 towards 0.5, with a statistically insignificant difference for the period 2005-2009. Similar results are obtained for the monthly exponents. For the purposes of this project, it is important to underline their explanation of the phenomenon: they attribute it to the growth of algorithmic trading. In the second article, daily exponent series are studied of eleven emerging markets between 1992 and 2002, with similar results.

Other perspectives on market efficiency by studying Hurst exponent had been proposed in the past. Very interesting is the discussion in [37], in which the authors study a measure of quantitative correlation

between theoretical inefficiency and empirical predictability for 60 financial indices from different countries. The Hurst exponent is taken as a measure of market inefficiency, while to measure predictability they use one of the most basic techniques of supervised machine learning: Nearest Neighbor (NN), and its proportion of correct answers. A considerable positive correlation (around 60%) between inefficiency and predictability is reported.

In [38] it is shown that before the great economic collapses of 1929, 1987 and 1998 a clear decrease in the Hurst exponent from the persistence regime ($H > 0.5$) to the anti-persistence regime ($H < 0.5$) is observed.

In [39] Peters carries out an extensive analysis of the R/S statistic, the relevance of which he argues through the Fractal Market Hypothesis (FMH) as an alternative to the EMH. The fractal properties of the financial time series would be due to the differences in the time horizons of the financial agents, who, according to their interests, incorporate certain pieces of relevant information into the price of an asset that are not relevant for other time horizons. Market stability is attributed to the dynamic interaction of these different scales.

2.2 Ordinal Patterns

For our study on collective dynamics of the US Market we will use a few tools coming from ordinal patterns series analysis and information theory.

Permutation entropy was introduced in [25] as a (non-parametric) complexity measure, robust to dynamical noise, invariant with respect to nonlinear monotonous transformation and computationally efficient. It is defined as follows: First, given a time series x_n for $n = 1, \ldots, N$ and two parameters $m < N$ and l, the pattern length and the time lag respectively, we consider, S_m, the group of permutations of length m and, for $0 < t \leqslant N - (m-1) \cdot l$, we say that the sliding window $x_m^l(t) = (x_t, x_{t+l}, \ldots, x_{t+(m-1)\cdot l})$ is of type $\pi_t \in S_m$ if $\pi_t = (i_1, \ldots, i_m)$ is the only m-permutation satisfying the following conditions:

(1) $x_{t+i_s \cdot l} \leqslant x_{t+i_{s+1} \cdot l}$ for $s = 1, \ldots, m-1$, and

(2) $i_s < i_{s+1}$ if $x_{t+i_s \cdot l} = x_{t+i_{s+1} \cdot l}$,

and we denote this with $\Phi(x_m^l(t)) = \pi_t$.

For example: if $x = (1.2, 3.2, 2.3, 1.4, 1.1, 4.3, 3.1)$ is a time series then, since $x_5 < x_1 < x_4 < x_3 <$

$x_7 < x_2 < x_6$, we have the 5-length ordinal patterns:

$$\pi_1 = \Phi(x_5^1(1)) = (4, 0, 3, 2, 1)$$
$$\pi_2 = \Phi(x_5^1(2)) = (3, 2, 1, 0, 4)$$
$$\pi_3 = \Phi(x_5^1(3)) = (2, 1, 0, 4, 3),$$

thus obtaining the 5-length ordinal pattern series

$$\pi_x = ((4, 0, 3, 2, 1), (3, 2, 1, 0, 4), (2, 1, 0, 4, 3)),$$

and similarly from the time series $y = (3.2, 4.2, 5.1, 0.4, 0.9, 2.3, 3.4)$ we obtain

$$\pi_y = ((3, 4, 0, 1, 2), (2, 3, 4, 0, 1), (1, 2, 3, 4, 0)).$$

The study of this new ordinal patterns series $\{\pi_t\}_{t<N-(m-1)\cdot l}$, has been applied to biomedicine [40] [41] [42] [43] [44] [45] [46] [47], paleoclimatology [48], economics [49] [50] [51] [52] [53] [54], geology [55] and engineering [56] for classification and prediction of deterministic and stochastic non-linear dynamics.

The choice for m is generally unproblematic, as it is understood that m must be as large as posible without compromising statistical reliability when measuring information-theoretic quantities on m-length ordinal patterns distributions (to be definied below), for which is enough to set $m! << N$, being $m!$ the cardinality of the permutation group S_m [57] [48]. Our data (see next section) dictates the choice $m = 5$. As for the lag time l, it has been shown to be critically important, since a naïve choice could lead to spurious results, making thus necessary to carry out a multiscale analysis (that is, varying l) before drawing any conclusions [58]. Our results, however, have shown to be highly independent of this parameter, and our conclusions, to be exposed in chapter 4, are the same no matter which value of l we pick, in a range going from $l = 1$ to $l = 100$ (see section 4.B for some figures supporting this claim). This is in itself an interesting result: we are observing here a phenomenon which is present in a wide range of time scales, which allows us to drop the l parameter in the subsequent discussion. The figures in chapter 4 correspond to $l = 1$.

So, for any $\pi \in S_m$ its relative frequency is defined as:

$$p_m(\pi) = \frac{\#\{t \mid t \leqslant N - m + 1, \Phi(x_m(t)) = \pi\}}{N - m + 1}. \tag{2.2.1}$$

Although the original proposal by Bandt and Pompe consisted basically in the study of time series through Shannon entropy of the ordinal patterns distribution, called *permutation entropy* (PE) and given by

$$\mathrm{PE}(m) = -\sum_{\pi \in S_m} p_m(\pi) \log p_m(\pi), \tag{2.2.2}$$

a vast stream of theoretical and methodological approaches has developed since then, giving place to a constellation of complexity measures, important examples of which are weighted permutation entropy [59] [60], symbolic transfer entropy [61] [62], statistical complexity [52] [63] [64], various measures of coupling and synchronicity [46] [65][62] [66] and network-based measures [55] (we will define some of them below). Other studies had combined these analytical tools with Machine Learning techniques [42] [67], and still others had exploited the algrebraic structure of S_m [66] [68] [69] [62].

We must be very careful in applying permutation entropy, the authors of [70] warn us. The presence of equal consecutive values in the original time series, which is frequently dealt with by preserving the temporal order in the corresponding permutation, could lead to draw false conclusions by detecting spurious patters. To adress this issue one can sum low amplitude noise to the original series in order to (randomly) break equalities, as proposed in [71], and that is exactly what we do in this work, with an artificially generated uniform distribution series of amplitude 10^{-7}.

We will need some (more or less) classic definitions for the next sections. Given two probability distributions p, q defined on a finite set, its Jensen-Shannon divergence is defined as

$$D_{\mathrm{JS}}(p, q) = \frac{D_{\mathrm{KL}}(p \,\|\, M) + D_{\mathrm{KL}}(q \,\|\, M)}{2},$$

where $M = (p + q)/2$ and D_{KL} is the Kullback-Leiber divergence:

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_i p_i \log \frac{p_i}{q_i},$$

or the relative entropy from q to p, of which D_{JS} is thus a smoothed, symmetric version.

So, if p_m is the probability distribution of ordinal patterns of the time series x_i, its *statistical complexity* (Com) is defined as [52] [63] [64]

$$\mathrm{Com}(p_m) = D_{\mathrm{JS}}(p_m, u_m)\ \mathrm{H}(p_m), \tag{2.2.3}$$

where u_m is the uniform distribution on S_m and $\mathrm{H}(\cdot)$ is the Shannon entropy. Thus, $\mathrm{Com}(p)$ aims to measure complexity through a trade-off between randomness and determinism: while Shannon entropy increases as p_m aparts away from determinism towards randomness, reaching a maximum for u_m, its divergence from u_m grows as p_m aparts away from randomness. $\mathrm{PE}(m)$ and $\mathrm{Com}(p)$ have been jointly used to classify dynamical regimes [52].

Missing Patterns Frequency (MPF) [72] is defined as

$$\mathrm{MPF}(p_m) = \frac{\#\{\pi \in S_m \,|\, p_m(\pi) = 0\}}{m!}. \tag{2.2.4}$$

As deterministic dynamics are expected to display a relatively small set of ordinal patterns in contrast to random dynamics [72] [50] [51], MPF is understood to quantify determinism degrees.

Yet another methodology for studying ordinal patterns, this time as nodes of a network, has been proposed in [73][74] [75] [55], where the directed edges are weighted according to the transition probability of passing from one pattern to another inmediatly in time, thus taking patterns as states of the process. So, for the ordinal sequence $\{\pi_t\}_{t<N-m+1}$ nodes are defined as the ordinal patterns and the weight of their edges are given by the transition probabilities $p_m(\pi'|\pi)$ of observing for some t that $\pi_{t+m} = \pi'$ given that $\pi_t = \pi$. We can then compute the node entropy as

$$\mathrm{NE}(\pi) = -\sum_{\pi' \in S_m} p_m(\pi'|\pi) \log(p_m(\pi'|\pi)). \tag{2.2.5}$$

These measures, unlike the previously defined, aim to quantify determinism not in terms of pattern frequency, but of pattern transitions in time. Minimum Node Entropy (MNE) and Global Node Entropy (GNE) are defined as [76]

$$\mathrm{MNE}(p_m) = \min\{\mathrm{NE}(\pi) \,|\, \pi \in S_m\} \tag{2.2.6}$$

and

$$\mathrm{GNE} = \sum_{\pi \in S_m} p_m(\pi)\, \mathrm{NE}(\pi), \tag{2.2.7}$$

respectively, so MNE measures how deterministic can a pattern be in a given network, while GNE gives us a global, weighted score of pattern transitions determinism. Finally, Missing Transitions Frequency (MTF) is defined as

$$\mathrm{MTF}(p_m) = \frac{\#\{(\pi, \pi') \in S_m \times S_m \,|\, p_m(\pi'|\pi) = 0\}}{(m!)^2}, \tag{2.2.8}$$

and is of course the analogous measure of MPF for pattern transitions.

A classification plane is proposed in [76], whose axes are given by GNE and the MNE, both of them measured using non-overlapping ordinal patterns to avoid transition constrains. This methodology seems to be very useful, not to mention intuitive, to distinguish between (linear) stochastic and (non-linear) deterministic dynamics. When applied to stock market time series, this methodology allows the authors to discriminate financial dynamics from fractional Gaussian noise, since financial scores lie below the diagonal around which the noises cluster. We will find this plane useful in the following discussion.

2.3 Data

The data used in this investigation are the time series of prices of the automated (algorithmic) operations that occurred from March 7, 2018 to March 7, 2019 in the Mexican and US markets (251 trading days). For the US market there were 539,834,024 records and for the Mexican market 78,863,574 records.

The importance for this work that our data comes from fully automated transactions cannot be overstated: it is for this alone that we are able to test efficiency and test our methodological insights specifically for Algorithmic High-Frequency markets, which is one of the main incentives for our inquiry, as stated in the Introduction.

The data belongs to 59 assets, of these, 35 correspond to companies listed on the Mexican stock exchange: AC, ALSEA, ALPEK, ALPHA, AMX, ASUR, BIMBO, BSMX, CEMEX, CUERVO, ELEKTRA, FEMSA, GAP, GCARSO, GENTERA, GFINBUR, GFNORTE, GMEXICO, GMXT, GRUMA, IENOVA, KIMBER, KOF, LALA, LIVEPOL, MEGA, MEXCHEM, NEMAK, OMA, PENOLES, PINFRA, RA, TLEVISA, VOLAR, WALMEX; while the other 24 are from the US market: ABT, BAC, BMY, C, CSCO, F, FB, FOXA, GE, GM, HPQ, INTC, KO, MDLZ, MO, MS, MSFT, ORCL, PFE, TWTR, T, USB, WFC, VZ. The details can be seen in tables 2.1 and 2.2.

The following analysis are carried out for the series of logarithmic returns, calculated in this way: If the original time series of prices of a given asset is $\{x_i\}_{i\in I}$, we first build an ordered partition $\{A_t^\tau\}$ of the original time series in which each A_t^τ is the set containing all data points which were registered between seconds $t\cdot\tau$ and $(t+1)\cdot\tau$, and then define

$$x(t,\tau) := \frac{1}{|A_t^\tau|}\sum_{x\in A_t^\tau}\log(x).$$

Finally, the logarithmic returns are given by

	Market code	Name
1	ABT	Abbott Laboratories
2	BAC	Bank of America Corporation
3	BMY	Bristol-Myers Squibb Company
4	C	Citigroup Inc.
5	CSCO	Cisco Systems, Inc.
6	F	Ford Motor Company
7	FB	Facebook, Inc.
8	FOXA	Tweenty-First Century Fox, Inc.
9	GE	General Electric Company
10	GM	General Motors Company
11	HPQ	HP Inc.
12	INTC	Intel Corporation
13	KO	The Coca-Cola Company
14	MDLZ	Mondelez International, Inc.
15	MO	Altria Group, Inc.
16	MS	Morgan Stanley
17	MSFT	Microsoft Corporation
18	ORCL	Oracle Corporation
19	PFE	Pfizer Inc.
20	T	AT&T Inc.
21	TWTR	Twitter, Inc.
22	USB	U.S. Bancorp
23	VZ	Verizon Communication, Inc.
24	WFC	Wells Fargo & Company

Table 2.1: Assets of the US market

$$r(t,\tau) = x(t+1,\tau) - x(t,\tau).$$

	Market code	Name
1	AC	Arca Continental, S.A.B de C. V.
2	ALFA	Alfa, S.A.B de C. V.
3	ALPEC	Alpek, S.A.B de C. V.
4	ALSEA	Alsea, S.A.B de C. V.
5	AMX	América Móvil, S.A.B de C. V.
6	ASUR	Grupo Aeroportuario del Sureste, S.A.B de C. V.
7	BIMBO	BIMBO & Grupo Bimbo, S.A.B de C. V.
8	BSMX	Grupo Financiero Santander, S.A.
9	CEMEX	Cemex, S.A.B de C. V.
10	CUERVO	Becle, S.A.B de C. V.
11	ELEKTRA	ELEKTRA & Grupo Elektra, S.A.B de C. V.
12	FEMSA	Fomento Económico Mexicano, S.A.B de C. V.
13	GAP	Grupo Aeroportuario del Pacífico, S.A.B de C. V.
14	GCARSO	Grupo Carso, S.A.B de C. V.
15	GENTERA	Gentera, S.A.B de C. V.
16	GFINBUR	Grupo Financiero Inbursa, S.A.B de C. V.
17	GFNORTE	Grupo Financiero Banorte, S.A.B de C. V.
18	GMEXICO	Grupo México, S.A.B de C. V.
19	GMXT	GMéxico Trasportes, S.A.B de C. V.
20	GRUMA	Gruma, S.A.B de C. V.
21	IENOVA	Infraestructura Energética Nova, S.A.B de C. V.
22	KIMBER	Kimberly Clark de México, S.A.B de C. V.
23	KOF	US Coca-Cola Femsa, S.A.B de C. V.
24	LALA	Grupo Lala, S.A.B de C. V.
25	LIVEPOL	El Puerto de Liverpool, S.A.B de C. V.
26	MEGA	Megacable Holdings, S.A.B de C. V.
27	MEXCHEM	MexicHem, S.A.B de C. V.
28	NEMAK	Nemak, S.A.B de C. V.
29	OMA	Grupo Aeroportuario del Centro Norte, S.A.B de C. V.
30	PENOLES	Industrias Peñoles, S.A.B de C. V.
31	PINFRA	Promotora y Operadora de Infraestructura, S.A.B de C. V.
32	RA	Regional, S.A.B de C. V.
33	TLEVISA	Grupo Televisa
34	VOLAR	Controladora Vuela Compañía de Aviación, S.A.B de C. V.
35	WALMEX	Walmart de México

Table 2.2: Assets of the Mexican market

Chapter 3

Analysis of Efficiency in High-Frequency Digital Markets Using the Hurst Exponent

> ["Neo-liberal" discourse] considered as a virtue the optimal market allocation of information -and no longer that of wealth- in society. In this sense, the market is but the instrument of a perfect coordination of players thanks to which the social totality can find a durable equilibrium. Capitalism thus becomes unquestionable, insofar as it is presented as a simple means -the best possible means- of producing social self-regulation.
>
> Tiqqun, *The Cybernetic Hypothesis*

3.1 Introduction

It is important to establish that when speaking of the efficiency of the markets, two great perspectives of analysis must be distinguished: the so-called distributive efficiency and the informational efficiency. It is to the second approach to which we dedicate this chapter. The informational efficiency of prices is defined as the immediate incorporation of all relevant information for the formation of prices through the interaction in the markets of highly sophisticated economic agents.

Introduced by Fama in [77], this Efficient Market Hypothesis has been widely questioned, for example in [78] and [79]. The informational efficiency of a market implies that the consecutive price differences

must be independent. Indeed, if there were any correlation between consecutive prices, it could be used to perform arbitrage, an action that would be in contradiction with the assumed efficiency. Thus, one can examine the short-term movement patterns that describe the returns of the assets in the market in question and attempt to identify the process underlying those returns. If the market is efficient, the model will not be able to identify a pattern and we will conclude that the returns follow a random walk process. If a model is able to establish a pattern, past market data can be used to predict future market movements and the market is therefore inefficient, since efficiency implies unpedictability.

Note that the observation of a random walk is a necessary condition for efficiency. There are studies that show that this condition is not sufficient [80], [81]. Consequently, the deviations of a random walk allow rejecting the informational efficiency of the assets under study.

One of the most common explanations for the inefficiency of real markets has been the "animal spirit" of Keynes [5], that is, the psychological and emotional factors that lead investors to make their decisions in capital markets when there is uncertainty, the ways in which human emotions can drive making financial decisions in uncertain and volatile environments.

Another explanation comes from the work of H. Simon [82], through the concept of limited rationality, which postulates that most people are only partially rational and act on emotional impulses without rational foundations in many of his actions.

Various authors, such as Lo in [83] and McCauley in [79] for example, recover the elementary fact that financial markets are, like all social phenomena, of a historical and dynamic nature, that is, agents respond to their specific social, political, psychological and technical conditions, which is why it is inappropriate to postulate general and anti-historical hypotheses about their behavior.

As indicated in chapter 1, in recent years most of the operations in the large financial markets have been automated and are now computers and not human beings who make decisions by executing certain algorithms. Although in the last decades evidence has accumulated against the efficiency of the traditional markets, one could imagine that, with the execution of orders controlled by computer algorithms, devoid of feelings and emotional decisions, efficiency could be achieved in financial markets. Indeed, High Frequency Digital Markets were brought to life precisely to exploit (human) market inefficiencies [13], and although this could be read as an historical refutation of classical markets inefficiencies, the effect of High Frequency Trading (HFT) dominance over financial markets remains controversial. For supporters of markets efficiency might argue that the very existence of HFT points to the asymptotic efficiency of markets, because it is meant to remove such human marginal inefficiencies [84] .Following such a logic, it must be expected that the hegemony of HFT would have the effect of rendering markets more efficient

than before. On the other hand, the Flash Crash of 2010, refered in chapter 1, as well as the raising frequency of big prices deviations known as “black swans” [85], have made clear that HFT generates its own new inefficiencies by predating and running out long-term financial strategies in fragmented markets [86][87]. This is a fundamental debate, since the confirmation or refutation of the Efficient Market Hypothesis bear important practical consequences: under-estimation of financial risks has been shown to play a fundamental role in financial crises such as the Black Monday [88].

The objective of this chapter is to offer evidence, through the analysis of time series of automated transactions in the US and Mexican markets, that the use of computational algorithms in automated high-frequency markets has not led these markets to the efficiency prescribed by the neoclassical theory. The organization of the chapter is as follows: Section 2 describes the characteristics of the data and the methodology used for our analysis. In Section 5 the results obtained are discussed. Section 6 contains the conclusions. The theoretical background has been discussed in section 2.1.

The results in this chapter are contained in a published paper by the author of this thesis and his main advisor, Dr. Ricardo Mansilla [89].

3.2 Methodology

In this chapter, we use $\tau = 1$ and $\tau = 5$ and study daily and weekly subseries to analyze daily and weekly dynamics throughout the trading year (recall that τ is the number of seconds to partition and average the original time series, see section 2.3).

To analyze the evolution of the Hurst exponent throughout the period under study, which is one year, we calculate the Hurst exponent H_{stock} of subseries of a certain number N of days, slided one day at a time [36]. For example, if $N = 5$, the Hurst exponent of the first five days of the series is calculated, then that of the series that goes from the second to the sixth day, etc., and the last calculation is for the series of the last five days, with which the evolution of the weekly Hurst exponent H_{stock} throughout the year is obtained.

There is no satisfactory analytical theory for the R/S statistic; most of the results on the subject are derived from computer simulations, which implies that they depend on particular models. Thus, although R/S is non-parametric, it is usually used to test the null hypothesis of Gaussian random walk [39], so its rejection may be due to non-Gaussianity or short-term memory. That is why the methodology that will be used below to establish the statistical significance of our calculations, inspired by the the proposals

discussed in Section II, is based on global and local permutations of the series in question [34], [36], [32].

Continuing with the previous example with $N = 5$, for each subseries of five days of the original series, we shuffle its terms to destroy its memory, and the Hurst exponent is calculated for this new radomized subseries. The process of shuffling and calculating the Hurst exponent is repeated one hundred times, thus obtaining a statistical sample, which we will call H_{perm}, of the Hurst exponent of the subseries under the null hypothesis of lack of long-term memory, so we can use its quantiles to test the statistical significance of the difference between H_{stock} and H_{perm}.

To rule out that the results thus obtained are due to short-term memory, we obtain in a similar way a statistical sample of locally randomized Hurst exponents. Given a weekly subseries ($N = 5$) and a fixed length l, the subseries is divided into blocks of l elements and the elements within each block are shuffled to destroy short-term correlations, without altering the long-range memory structure. This process is repeated a hundred times to get a statistical sample which we will call $H_{locperm}$. Thus, if not only H_{stock}, but also $H_{locperm}$ is statistically different of H_{perm}, then it is ruled out that the rejection of the null hypothesis is due to short-range correlations.

We make all these calculations for each subseries of N days, so we can observe the evolution of H_{stock}, H_{perm} and $H_{locperm}$ throughout the year.

3.3 Results and discussion

In figures 3.1 and 3.2 we plot for $\tau = N = 1$ and $\tau = N = 5$, respectively, the evolution of H_{stock} (blue curve) and the area between the 0.1 and 0.9 quantiles of H_{perm} (purple zone). Thus, when the blue curve passes outside this area, it is concluded that the corresponding original (daily or weekly) subseries has long-term memory: the difference between H_{stock} and H_{perm} is statistically significant; while when the H_{stock} curve passes inside, the randomness of the subseries cannot be ruled out: the difference between H_{stock} and H_{perm} is not statistically conclusive.

Figure 3.1 shows that for the US market it is not possible in general to reject the null hypothesis $H_{stock} = H_{perm}$ when $\tau = N = 1$ (daily series of one-second averages), while the inspection of figure 3.2 allows to conclude the existence of a clear tendency to anti-persistence ($H_{stock} < H_{perm}$) for $\tau = N = 5$.

In what follows we will focus on the latter case. Figure 3.3 shows for $l = 300$ the effect of locally shuffling the series to destroy their short-term memory. As before, we plot H_{stock} and the 0.1 and 0.9 quantiles of $H_{locperm}$. Although the Hurst exponent tends to increase slightly after the local shuffling, the

global $H_{locperm}$ shape is considerably similar to H_{stock}, reinforcing the idea that its behavior reflects well the long-term memory from the original series. Note that this tendency to increase $H_{locperm}$ is not valid for all assets, for example, F.

Once we have visually detected the general trend towards anti-persistence and the effect of local shuffling, we can define two annual inefficiency indices, one of them given by the percentage of windows whose Hurst exponent H_{stock} is below the 0.1 quantil of H_{perm}, the second by the percentage of windows such that the mean of $H_{locperm}$ does the same. We will call each of them weak and strong anti-persistence indices, and we will denote them by I_d, I_f respectively.

To be more specific, if H^q_{perm} is the q-quantile of H_{perm} and

$$N_d = \#\{\text{sliding } N\text{-days windows such that } H_{stock} < H^{0.1}_{perm}\},$$

then, since our data consist of 251 trading days and therefore we have $251 - N + 1$ sliding N-days windows, we set $I_d := N_d/(251 - N + 1)$. Analogously we define $I_f := N_f/(251 - N + 1)$, where

$$N_f = \#\{\text{sliding } N\text{-days windows such that } \mathrm{E}(H_{locperm}) < H^{0.1}_{perm}\}$$

and $\mathrm{E}(H_{locperm})$ is the mean of the statistical sample $H_{locperm}$. Figure 3.4 shows the table of I_d and I_f results by asset and its average per market (US). It is concluded that most of the assets spend a significant part of the year under the anti-persistence regime.

It is important to note that even in the case of assets with a low level of inefficiency measured with these indices (TWTR for example), the antipersistence trend is clear given that the H_{stock} and $H_{locperm}$ curves remain in the lower part of the H_{perm} zone throughout the year (recall that the antipersistence threshold that we define: the 0.1 quantile of H_{perm}, is as arbitrary as the more traditional 0.05), a result that is hardly due to to statistical sensitivity of the methods used, since it is observed systematically in all assets and throughout the year. Thus, although they are useful as summary indicators, they should not be considered as the ultimate criterion of efficiency. These observations on the qualitative nature of the process are possible thanks to the use of a dynamic approach to observe the evolution of the Hurst exponent [36], as opposed to the more traditional method of calculating a single exponent for each series, a method that reduces the problem to a purely quantitative and static criterion.

To formalize this idea and obtain, also here, a quantitative indicator: considering the subset of the $M = \lfloor 251/N \rfloor$ consecutive weekly series without overlap (where $\lfloor n \rfloor$ is the largest integer less than or

equal to n) and given $q \leqslant 0.5$ and n_d the number of these weekly series such that H_{stock} is less than H^q_{perm}, the q-quantil of H_{perm}, and assuming the M consecutive weekly series as independent experiments, that is, assuming that these series are statistically independent of each other (efficiency, lack of memory at that scale), what is the probability of observing, as we do, at least n_d windows (realizations) in which H_{stock} is below H^q_{perm}? This problem is equivalent to determining the probability of obtaining at least n_d heads in a sequence of M tosses with an (unfair) coin with probability q of observing a head in each realization. This probability is modeled with the binomial distribution:

$$P^q(n_d) = \sum_{i=n_d}^{M} \binom{M}{i} q^i (1-q)^{M-i}.$$

Thus, this p-value indicates how likely it is to observe the behavior described above in an efficiency scenario given by the independence between non-overlapping weekly series. Once again, we define a strong version of this index given by $P^q(n_f)$, where n_f is the number of weekly series such that the mean of $H_{locperm}$ is less than the quantile q of H_{perm}. The results for $q = 0.1$ and $q = 0.5$ are shown in Figure 3.5. The evidence against the null efficiency hypothesis thus formulated is compelling. Except for the maximum value in the table, which is obtained for TWTR with $P^{0.1}(n_f) = 0.215$, all stocks clearly reject the null hypothesis with a 95% level of confidence, almost always by a considerable margin, and even TWTR does it for the other three parameter combinations.

Similar results were obtained for the Mexican market (figuras 3.6 y 3.7), although due to their lower resolution $\tau = 30$ and $N = 30$ are used. It is observed that the result of the local permutations is more ambiguous in this case, which can be interpreted as short-term memory lack.

In the next chapter we will inquiry into the structure of stock correlations to better understand their collective dynamics beyond the individual question on efficiency.

3.4 Conclusions

This chapter discussed the efficiency in high-frequency digital markets, quantified by the Hurst exponent measured by the R/S statistic. Results indicate that, in the period from March 7, 2018 to March 7, 2019 and for the 24 assets in the United States market and the 35 in the Mexican market studied here, the Efficient Market Hypothesis is clearly rejected: the presence of long-term memory, particularly of anti-persistence, is clear.

As noted before, the relevance of these results to the question of efficiency in automated digital markets lies in the nature of our data, coming from fully automated (algorithmic) transactions. It is because of this that we can draw the main conclusion of this chapter: automated digital markets do not meet the efficiency postulated by neoclassical theory. Thus, classical explanations of the inefficiencies of human markets, based on the psychological or emotional factors of human beings [5] or on their limited rationality [82], must be discarded, since the algorithms that have ordered the transactions here studied do not suffer from these human limitations. Market inefficiency therefore seems to be due to more fundamental factors of economic dynamics. This opens a new line of investigation in the search for the real sources of the lack of efficiency.

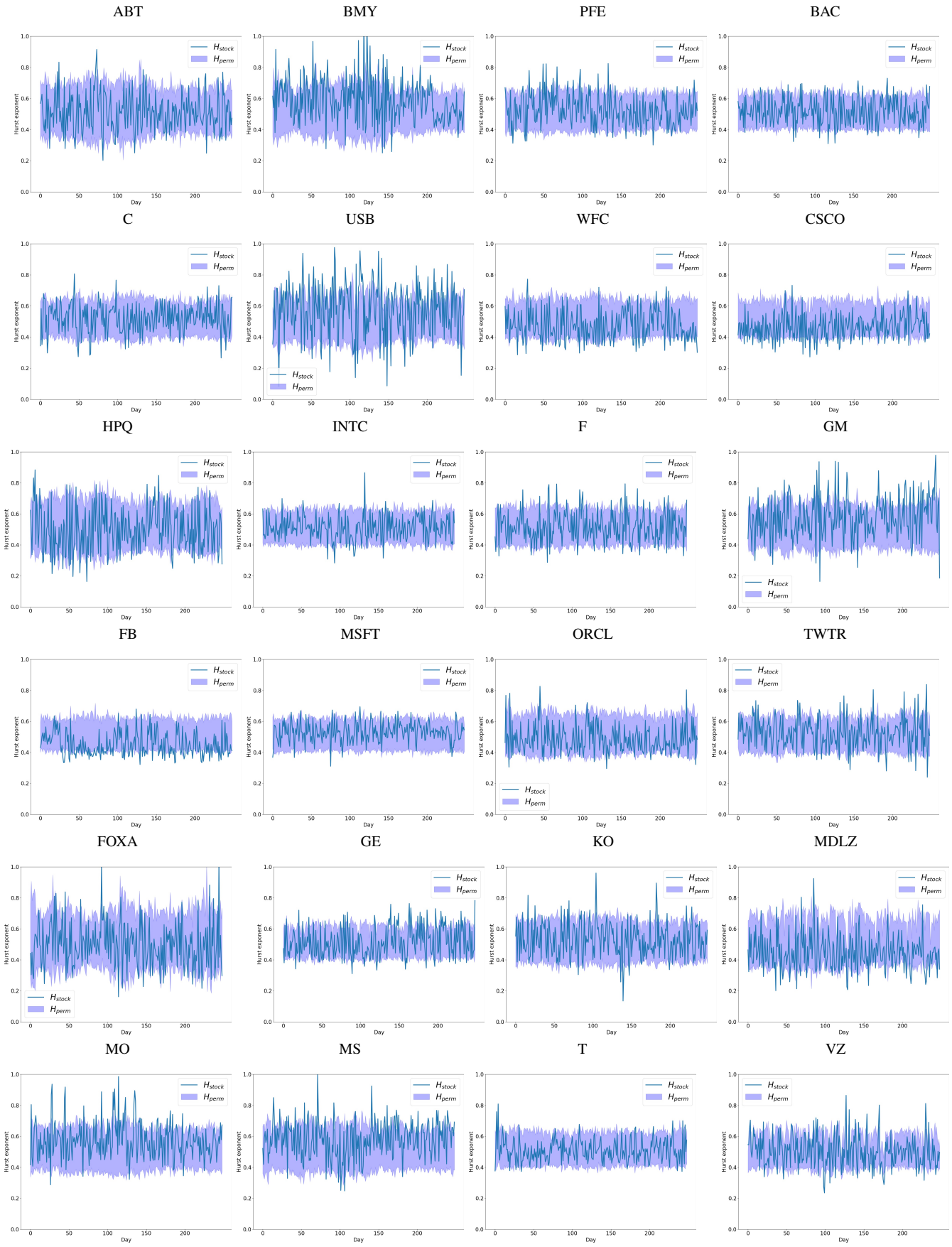

Figure 3.1: Evolution of H_{stock} and 0.1 and 0.9 quantiles of H_{perm} for the series of US market for $\tau = 1$ y $N = 1$.

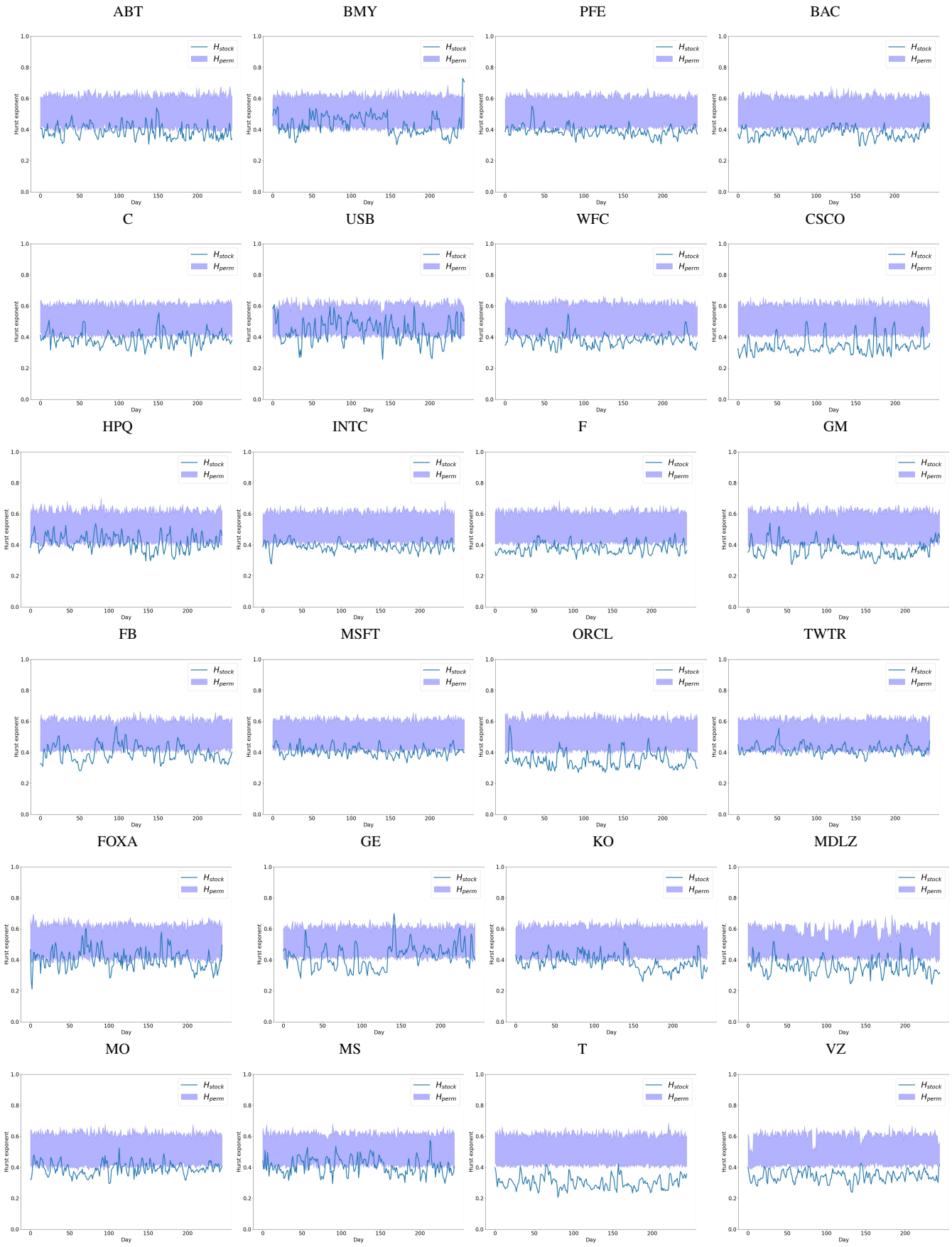

Figure 3.2: Evolution of H_{stock} and 0.1 and 0.9 quantiles of H_{perm} for the series of US market for $\tau = 5$ y $N = 5$.

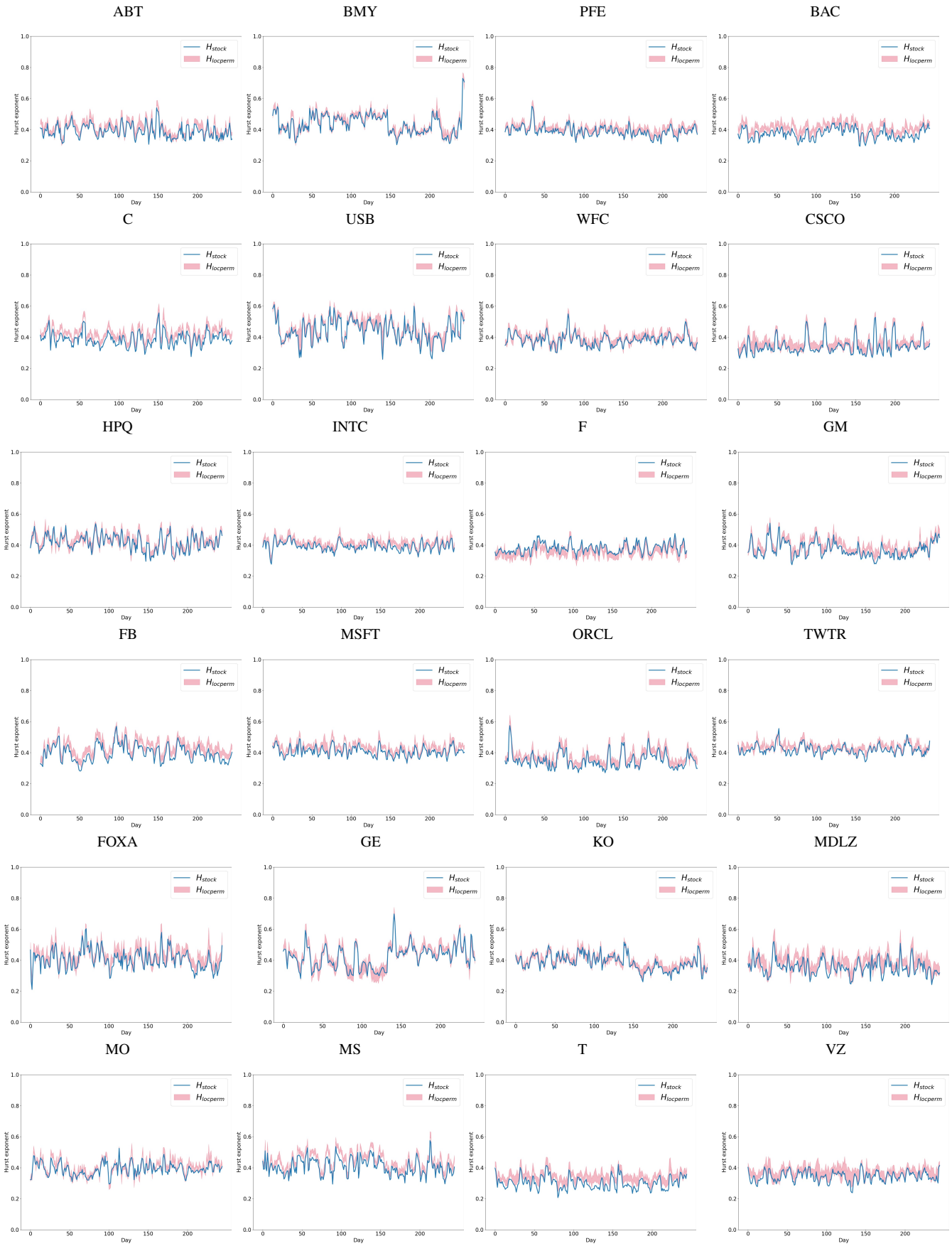

Figure 3.3: Evolution of H_{stock} and 0.1 and 0.9 quantiles of $H_{locperm}$ for the series of the US market for $\tau = 5$, $N = 5$ y $l = 300$.

	I_d	I_f
F	0.82	0.92
ABT	0.67	0.45
KO	0.67	0.61
FB	0.65	0.36
HPQ	0.42	0.33
MSFT	0.57	0.26
GE	0.44	0.4
MO	0.64	0.52
MS	0.5	0.26
VZ	0.98	0.91
INTC	0.74	0.45
BAC	0.83	0.48
MDLZ	0.85	0.69
GM	0.76	0.61
C	0.71	0.34
FOXA	0.51	0.36
USB	0.31	0.21
ORCL	0.88	0.77
WFC	0.78	0.66
CSCO	0.88	0.86
PFE	0.75	0.57
T	0.99	0.95
BMY	0.35	0.32
TWTR	0.4	0.12
Average US	0.67	0.52

Figure 3.4: Values of I_d and I_f for the US market for $\tau = 5$ y $N = 5$.

	$q = 0.1, n = n_d$	$q = 0.5, n = n_d$	$q = 0.1, n = n_f$	$q = 0.5, n = n_f$
F	9.38e-27	8.88e-16	4.86e-36	8.88e-16
ABT	2.7e-27	8.88e-14	2.38e-09	2.18e-12
KO	3.41e-23	1.78e-15	4.64e-16	8.88e-14
FB	1.85e-19	2.18e-12	9.56e-08	2.86e-08
HPQ	5.36e-07	3.49e-11	0.008	4.11e-10
MSFT	3.71e-17	8.88e-14	0.00269	3.49e-11
GE	1.57e-08	3.8e-09	1.57e-08	9.82e-07
MO	1.15e-20	1.78e-15	4.64e-16	8.88e-14
MS	2.38e-09	1.78e-15	0.0215	9.82e-07
VZ	1e-49	1.78e-15	1.4e-40	1.78e-15
INTC	6.94e-26	1.78e-15	4.28e-11	1.78e-15
BAC	3.55e-28	8.88e-16	8.87e-12	8.88e-16
MDLZ	7.55e-37	1.78e-15	1.15e-20	1.78e-15
GM	2.7e-27	8.88e-14	1.85e-19	2.18e-12
C	2.7e-27	8.88e-14	0.000226	3.49e-11
FOXA	5.61e-13	3.49e-11	5.69e-05	2.86e-08
USB	1.31e-05	7.1e-05	0.0215	0.0047
ORCL	1.4e-40	8.88e-14	2.95e-30	8.88e-14
WFC	1.98e-33	8.88e-14	6.54e-22	8.88e-14
CSCO	7.55e-37	8.88e-14	7.55e-37	3.49e-11
PFE	2.7e-27	8.88e-14	3.71e-17	8.88e-14
T	1e-49	1.78e-15	9.57e-45	1.78e-15
BMY	5.36e-07	2.86e-08	5.36e-07	9.82e-07
TWTR	9.56e-08	2.18e-12	0.215	2.18e-12
Average US	5.96e-07	2.96e-06	0.0112	0.000196

Figure 3.5: Values of $P^q(n)$ for the US market $\tau = 5$, $N = 5$ y $q = 0.1, 0.5$.

	I_d	I_f
GMEXICO	1.0	1.0
ELEKTRA	0.25	0.35
PENOLES	0.94	0.76
BSMX	0.54	0.92
MEXCHEM	0.23	0.34
GFNORTE	0.44	0.44
AMX	0.93	0.95
ALFA	0.84	0.97
CEMEX	0.88	0.86
PINFRA	0.49	0.5
WALMEX	0.54	0.53
KIMBER	0.67	0.74
GAP	0.47	0.62
RA	0.23	0.33
ALSEA	0.68	0.9
CUERVO	0.77	0.8
IENOVA	0.29	0.41
AC	0.53	0.94
MEGA	0.36	0.52
VOLAR	0.6	0.51
GFINBUR	0.65	0.78
LIVEPOL	0.54	0.61
GMXT	0.65	0.67
NEMAK	0.57	0.64
LALA	0.58	0.54
BIMBO	0.73	0.53
ASUR	0.46	0.7
GRUMA	0.69	0.89
TLEVISA	0.73	0.77
KOF	0.67	0.6
OMA	0.62	0.58
GENTERA	0.15	0.1
ALPEK	0.79	0.78
FEMSA	0.65	0.66
GCARSO	0.37	0.36
Average MX	0.59	0.65

Figure 3.6: Values of I_d and I_f for the Mexican market for $\tau = 30$ y $N = 30$.

	$q = 0.1,\ n = n_d$	$q = 0.5,\ n = n_d$	$q = 0.1,\ n = n_f$	$q = 0.5,\ n = n_f$
GMEXICO	1e-45	2.84e-14	1e-45	2.84e-14
ELEKTRA	8.2e-05	0.00805	3.74e-06	2.71e-07
PENOLES	8.06e-41	2.84e-14	8.45e-24	1.31e-12
BSMX	4.51e-12	4.33e-10	1.04e-38	2.84e-14
MEXCHEM	0.00404	0.0676	6.95e-07	0.0178
GFNORTE	2.62e-09	0.116	2.62e-09	0.0676
AMX	9.88e-37	2.84e-14	1.04e-38	2.84e-14
ALFA	9.5e-30	1.31e-12	8.06e-41	2.84e-14
CEMEX	4.4e-33	4.67e-09	4.4e-33	4.67e-09
PINFRA	2.62e-09	2.94e-11	1.84e-08	2.94e-11
WALMEX	4.55e-13	3.29e-05	4.55e-13	3.29e-05
KIMBER	4.01e-21	3.29e-05	3.3e-25	3.94e-08
GAP	2.62e-09	0.000124	1.97e-17	1.31e-12
RA	0.032	0.116	6.95e-07	0.00805
ALSEA	8.45e-24	1.31e-12	9.88e-37	2.84e-14
CUERVO	1.15e-26	1.31e-12	3.53e-28	2.84e-14
IENOVA	0.00122	0.000412	3.74e-06	3.94e-08
AC	1.97e-17	1.31e-12	4.06e-43	2.84e-14
MEGA	8.2e-05	2.71e-07	4.55e-13	4.33e-10
VOLAR	2.77e-16	3.94e-08	4.1e-11	1.56e-06
GFINBUR	1.97e-17	2.71e-07	3.3e-25	4.33e-10
LIVEPOL	4.21e-14	2.94e-11	1.27e-18	2.94e-11
GMXT	1.27e-18	2.71e-07	1.27e-18	3.94e-08
NEMAK	4.21e-14	4.67e-09	2.77e-16	2.84e-14
LALA	1.97e-17	4.33e-10	4.21e-14	4.33e-10
BIMBO	1.94e-22	2.71e-07	4.51e-12	0.000412
ASUR	4.1e-11	1.56e-06	4.01e-21	4.33e-10
GRUMA	1.94e-22	2.71e-07	4.4e-33	2.84e-14
TLEVISA	3.3e-25	4.67e-09	3.3e-25	2.94e-11
KOF	1.94e-22	1.31e-12	3.57e-15	2.94e-11
OMA	1.97e-17	4.67e-09	4.21e-14	4.33e-10
GENTERA	0.159	0.186	0.671	0.186
ALPEK	3.53e-28	4.67e-09	3.53e-28	4.67e-09
FEMSA	7.5e-20	4.67e-09	1.27e-18	4.33e-10
GCARSO	1.84e-05	2.71e-07	3.74e-06	2.71e-07
Average MX	0.0056	0.0141	0.0192	0.00799

Figure 3.7: Values of $P^q(n)$ for the Mexican market for $\tau = 30$, $N = 30$ y $q = 0.1, 0.5$.

Chapter 4

Ordinal Synchronization and Typical States in High-Frequency Digital Markets

> No one but man himself–with his own hands– produces these commodities and determines their prices, except that, here again, something flows from his actions which he does not intend or desire; here again, need, object, and the result of the economic activity of man have come into jarring contradiction.
> In the entity which embraces oceans and continents, there is no planning, no consciousness, no regulation, only the blind clash of unknown, unrestrained forces playing a capricious game with the economic destiny of man. Of course, even today, an all-powerful ruler dominates all working men and women: capital. But the form which this sovereignty of capital takes is not despotism but anarchy. And it is precisely this anarchy which is responsible for the fact that the economy of human society produces results which are mysterious and unpredictable to the people involved. Its anarchy is what makes the economic life of mankind something unknown, alien, uncontrollable.
>
> Rosa Luxemburg, *What is Economics?*

4.1 Introduction

In the previous chapter we have analyzed the issue of informational efficiency in high-frequency digital markets t and individual level. Now we want to go further in our understanding of high-frequency digital market dynamics

Financial markets are highly complex evolving systems, which means that their statistical dynamics is constantly redefined by the interaction of economic agents. Thus, it is not surprising that, in order to understand these dynamics, much effort have been recently dedicated to study them as dynamical networks which can be analyzed and classified, either topologically to quantify crisis periods [90] [91] or by clustering them to detect market states [92][93] [94] [24] [95] and posibly early precursors for the catastrophic ones [21] [20].

This approach looks very promising for studying Algorithmic High-Frequency Trading, whose rise during the last decades was recently shown to display an even higher degree of networked structure and complexity than those of traditional markets [96]. However, those network-based studies have been carried out mainly by defining edge weights through correlation matrixes, which amounts to ignore non-linear interactions (see [97] for an exception, which does not use cluster analysis, but Machine Learning techniques). This is not adequate for high-frequency data, normally expected to be very noisy and highly non-linear and non-stationary. Thus, in order to successfully apply the same network-clustering pipeline of those previous works and, at the same time, be able to detect locally complex non-linear interactions, we define dynamical networks of stocks by means of their transcript synchronicity, a measure of pairwise coupling of time series defined through ordinal patterns, a tool which has been successfully applied to discern non-linear deterministic and stochastic dynamics in real-world data [98].

Once this has been done, we propose to study the obtained dynamical network through the distribution of its Eigenvector Centrality and Degree, two well known measures of connectivity for network nodes which we use to define suitable phase representation spaces in order to detect meaningful market states through a couple of clustering algorithms. This allows us to detect two whole coherent and quantitatively distinguishable seasons, characterized by their degree of centralized/decentralized synchronicity.

So, the goal of this chapter is twofold: to adapt an increasingly popular methodology for studying financial markets as dynamical networks and clusters as market states to the needs of Algorithmic High-Frequency Trading Data, and to show with its application to a particular data set of fully automated transactions its potential for detection of collective dynamical regimes.

In order to underline the necessity of collective analysis, we include a section in which ordinal pattern

analysis for individual stocks is carried out as a start point, and whose findings are latter shown to be reproduced, extended and further explained at the collective level. For this, we use some of the most common information-theoretic measures related to ordinal patterns.

The chapter is organized as follows: Section 3 describes the characteristics of the data. In Section 4 we apply the previosly defined measures to individual stocks, in order to detect anomalous behaviors. Section 5 contains the definition of our transcript synchronicity coefficient, which we use to define our dynamical network and discuss the more convenient phase representation spaces for the sake of our analysis. In Section 6 we carry out the clustering analysis to detect typical market states, which in Section 7 are modeled as first order Markov processes. Section 8 contains the conclusions. The theoretical background for this chapter is to be found in section 2.2

The results in this chapter are contained in paper published by the author of this thesis and his main advisor, Dr. Ricardo Mansilla [99].

4.2 Individual Analysis of Stocks through Ordinal Patterns

In this chapter, we use $\tau = 5$ and study daily subseries to analyze daily dynamics throughout the trading year; we also restrict ourselves to the US market, because of its higher frequency data. Again, recall that τ is the number of seconds to partition and average the original time series, see section 2.3.

Before addressing the analysis of collective behaviors in our data set, which is the main concern of this chapter, lets take a look at the stocks individually. To do this we calculate, for the daily series of five seconds average logarithmic returns of each stock, Permutation Entropy (PE, equation 2.2.2), Statistical Complexity (Com, equation 2.2.3), Missing Patterns Frequency (MPF, equation 2.2.4), Minimum Node Entropy (MNE, 2.2.6), Global Node Entropy (GNE, equation 2.2.7) and Missing Transitions Frequency (MTF, equation 2.2.8), the last three without overlapping patterns, just as in [76]. We use $m = 5$, that is, we study 5-length ordinal patterns of daily series of returns of 5-seconds averages of logarithms of prices, the choice for m being dictated by considerations about statistical reliability and the problem of undersampling for PE, Com and MPF [57] [48] (although the same problem remains for MNE, GNE and MTF [76], we keep $m = 5$ for the sake of a better visualization). Next, we plot the evolution of such quantities along the year, as well as the MNE vs GNE plane.

For the vast majority of stocks, MNE, GNE and MTF clearly present an unusual (outlier) behavior for the days 82, 181 and 201, which we will call outlier days, while we will refer to the other days as

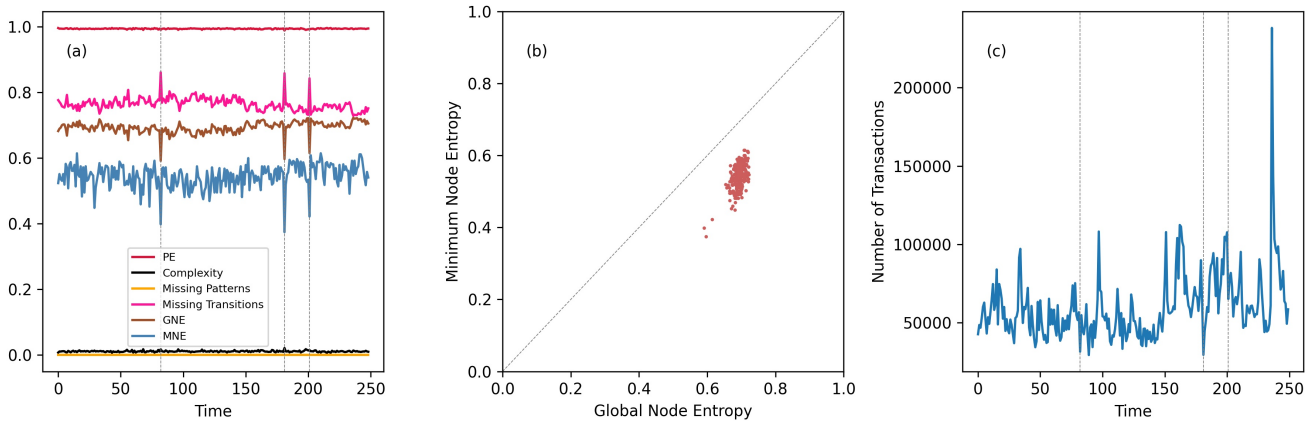

Figure 4.1: Ordinal Entropy Measures for KO. (a) Permutation Entropy, Complexity, Missing Pattern Frequency, Missing Transition Frequency, Global Node Entropy, Minimum Node Entropy; (b) Minimum Node Entropy vs Global Node Entropy plane; (c) Daily number of transactions.

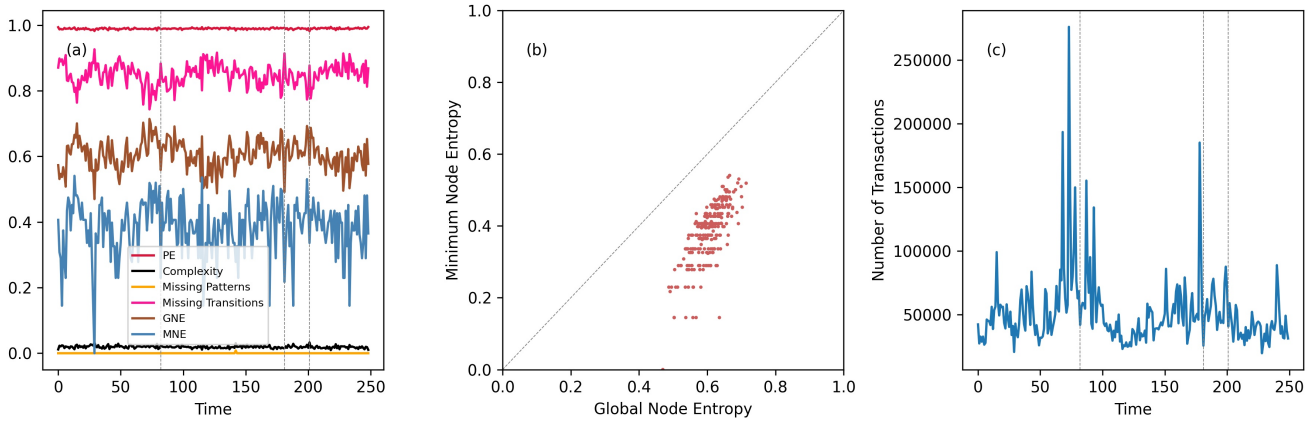

Figure 4.2: Ordinal Entropy Measures for FOXA. (a) Permutation Entropy, Complexity, Missing Pattern Frequency, Missing Transition Frequency, Global Node Entropy, Minimum Node Entropy; (b) Minimum Node Entropy vs Global Node Entropy plane; c) Daily number of transactions.

typical (figure 4.1 shows this for KO (The Coca-Cola Company); outlier days are indicated by vertical gray lines in the left and right panels). MNE and GNE abruptly decrease for the outlier days, while MTF does the opposite, thus supporting the idea that these days present complex, semi-deterministic behavior: given a certain pattern, the transition to the next one has candidates much more probable than others. In some of the stocks we can further visualize what appears to be a small number of discrete levels for MNE (see second panel in figure 4.2 for an example). This suggests the existence of a finite number of typical market states, each corresponding to a specific level of MNE. The problem with these measures is that they need length series $n >> (m!)^2$, and, as already mentioned, we could be undersampling for $m = 5$,

when we begin to classify outlier days. PE, Com and MPF are not capable of detecting (visually at least) the outlier days.

By inspecting in the same figure the graphs of the daily number of transactions of each stock (right panel), we can notice that the first two outlier days are low-liquidity days for many of them, although not necessarily global minima or even specially severe drops in some others, while the last outlier does not present this behavior. Thus, low liquidity is not enough to explain this phenomenon.

4.3 Collective Analysis of Stocks through Transcript Synchronicity Dynamical Networks

To get a deeper insight on this behavior and figure out if it is, as it seems to be, a collective behavior, we study the synchronization index given by the transcript entropy of pairs of stocks, that is: for each pair (i, j) of stocks and each trading day T, we obtain their daily ordinal pattern sequences $(\pi_i(t))_{t=1}^n$ and $(\pi_j(t))_{t=1}^n$ to obtain the transcript series [66] as $\tau_{i,j}(t) = \pi_j(t) \circ \pi_i(t)^{-1}$, where the product and the inverse are those of the permutation group S_m: for $\pi, \rho \in S_m$, $\pi = (\pi_1, \ldots, \pi_m)$ and $\rho = (\rho_1, \ldots, \rho_m)$,

$$\pi \circ \rho = (\pi_{\rho_1}, \ldots, \pi_{\rho_m}), \text{ and}$$

$$\pi^{-1} : \pi_k \to k \text{ for } k < m$$

is the sorting operation. In this way, the transcript $\tau_{i,j}(t)$ is the result of ordering $\pi_j(t)$ according to the ordinal type of $\pi_i(t)$.

Lets illustrate this with an example: if, as in Section 2, $x = (1.2, 3.2, 2.3, 1.4, 1.1, 4.3, 3.1)$ and $y = (3.2, 4.2, 5.1, 0.4, 0.9, 2.3, 3.4)$ are two time series then, as we have seen, their respective 5-length ordinal pattern sequences are

$$\pi_x = ((4, 0, 3, 2, 1), (3, 2, 1, 0, 4), (2, 1, 0, 4, 3))$$

and

$$\pi_y = ((3, 4, 0, 1, 2), (2, 3, 4, 0, 1), (1, 2, 3, 4, 0)).$$

The group inverse of $\pi_x(1) = (4, 0, 3, 2, 1)$, which denotes the function

$$0 \to 4$$
$$1 \to 0$$
$$2 \to 3$$
$$3 \to 2$$
$$4 \to 1,$$

is just the inverse function

$$4 \to 0$$
$$0 \to 1$$
$$3 \to 2$$
$$2 \to 3$$
$$1 \to 4,$$

or, conveniently, expresed, $\pi_x(1)^{-1} = (1, 4, 3, 2, 0)$. After similar calculations for the remaining ordinal patterns we obtain the series of group inverses

$$\pi_{x^{-1}} = ((1, 4, 3, 2, 0), (3, 2, 1, 0, 4), (2, 1, 0, 4, 3)).$$

Now, to obtain our first transcript we compose the two functions $\pi_x(1)^{-1} = (1, 4, 3, 2, 0)$ and $\pi_y(1) = (3, 4, 0, 1, 2)$, thus obtaining the function

	$\pi_x(1)^{-1}$		$\pi_y(1)$	
0	$\to$	1	$\to$	4
1	$\to$	4	$\to$	2
2	$\to$	3	$\to$	1
3	$\to$	2	$\to$	0
4	$\to$	0	$\to$	3

better represented as $\pi_y(1) \circ \pi_x(1)^{-1} = (4, 2, 1, 0, 3)$. After repeating the operations with the other pairs of ordinal patterns, we finally get the transcript series:

$$\tau_{x,y} = ((4, 2, 1, 0, 3), (0, 4, 3, 2, 1), (3, 2, 1, 0, 4))$$

We take the normalized entropy of this transcript series (that is, the usual permutation entropy of this transcript series) as a (daily) coefficient of desynchronization, and one minus that quantity as our measure of synchronization, which we will call transcript synchronization and denote by $H_T^{transcript}(i, j)$, so

$$H_T^{transcript}(i,j) = 1 + \sum_{\pi \in S_m} p_m(\pi) \log p_m(\pi),$$

where p_m is as defined in 2.2.1, but for the transcript series:

$$p_m(\pi) = \frac{\#\{t \mid t \leqslant N - m + 1, \tau_{i,j}(t) = \pi\}}{N - m + 1}.$$

Transcript synchronization measures the diversity of transcripts: low transcript synchronization means high variety of transcripts, that is, a lot of different transcripts and then a lot of information is needed to deduce the (ordinal) dynamics of one series given complete knowledge of the other, and analogously for high transcript synchronization. Transcripts have been applied to study synchronization in time series in [68], [69], [100], [62].

Thus we obtain for each trading day a (symmetric) transcript synchronization matrix $H_T^{transcript}$ of dimension $n_{stock} \times n_{stock}$ whose ij value is the transcript synchronization of stocks i and j during day T. By considering each of these daily matrixes as adjacency matrixes, we obtain a dynamical weighted network through the year, the nodes of which are the stocks and the weight of whose edges are given by our transcript synchronization coefficient. We can thus analyze our time series with classical network-based measures, following [92] [93] [94] [20] [24] [90] [21] [95], which have done that for correlation matrixes, and [97], where mutual information networks are studied. We consider two well known such network measures here: degree and eigenvector centrality.

Given a stock labeled as i and a trading day T, we define its degree as

$$\text{Degree}(i,T) = \frac{1}{C} \sum_{j=1,\, j \neq i}^{n_{stocks}} H_T^{transcript}(i,j),$$

so the degree of a node is just the sum of its transcript synchronization with all the other stocks and the normalizing constant C is such that $\sum_i \text{Degree}(i,T) = 1$, while its eigenvector centrality is defined as the i-th component of the normalized solution to the equation

$$\text{EVC}(i,T) = \frac{1}{\lambda} \sum_{j=1}^{n_{stocks}} H_T^{transcript}(i,j)\, \text{EVC}(j,T),$$

where λ is the largest eigenvalue of the adjacency matrix $H_T^{transcript}$. This is equivalent to find the normalized eigenvector with positive components corresponding to λ, which is known to exist by the Frobenius Theorem. So for each trading day T we have Degree and EVC, two n_{stocks}-dimensional vectors widely applied as measures of the importance, centrality or connectedness of a node in the network.

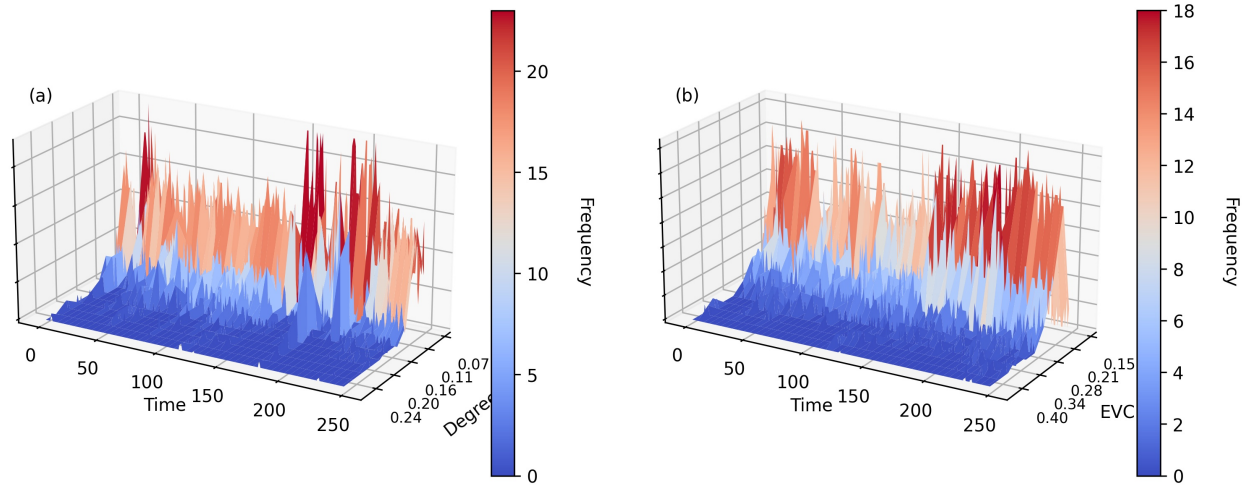

Figure 4.3: Evolution Through the Trading Year of Histograms of (a) Degree and (b) Eigenvector Centrality for Transcript Synchronization Dynamical Network.

In what follows we remove the first day of the year, for it turns out to be a very peculiar outlier. If we plot the evolution of the histograms of Degree and EVC (figure 4.3) we can observe a very important increment in the mode of Degree (this will be much more clear in figure 4.4) during the outlier days and further detect not just those three days, but what appears as two complete consecutive, although intermitent, regimes of highly centralized and decentralized connectivity, the latter weakly present at the beginning of the year and again with much more persistence roughly from day 150 to day 210, and the former just between those periods, approximately from day 50 to day 150. Thus, our outlier days seem to be an extreme manifestation of these collective dynamics. Recall that the detection of outlier days and different states at this collective level was by no means an obvious thing to expect, since the only measures capable of detecting them in the individual level were those measuring transitions between consecutive ordinal patterns, while our transcript synchronization coefficient focuses only in simultaneous pairs of patterns across the market.

To further illustrate this behavior, in figures 4.4 and 4.5 we plot the mean, standard deviation, minimum value and maximum frequency of the previously plotted daily histograms. First of all, for Degree one can easily confirm the presence of the outlier days, as its mean and minimum abruptly increase. Also, and this holds for both figures, at the beginning of the year (March) and during the highly decentralized season we can observe that the mean of the distributions of EVC and Degree increases, while standard deviation decrease, that is to say, the histograms shrink.

Recall that here highly decentralized synchronization means a more uniform distribution of degree-

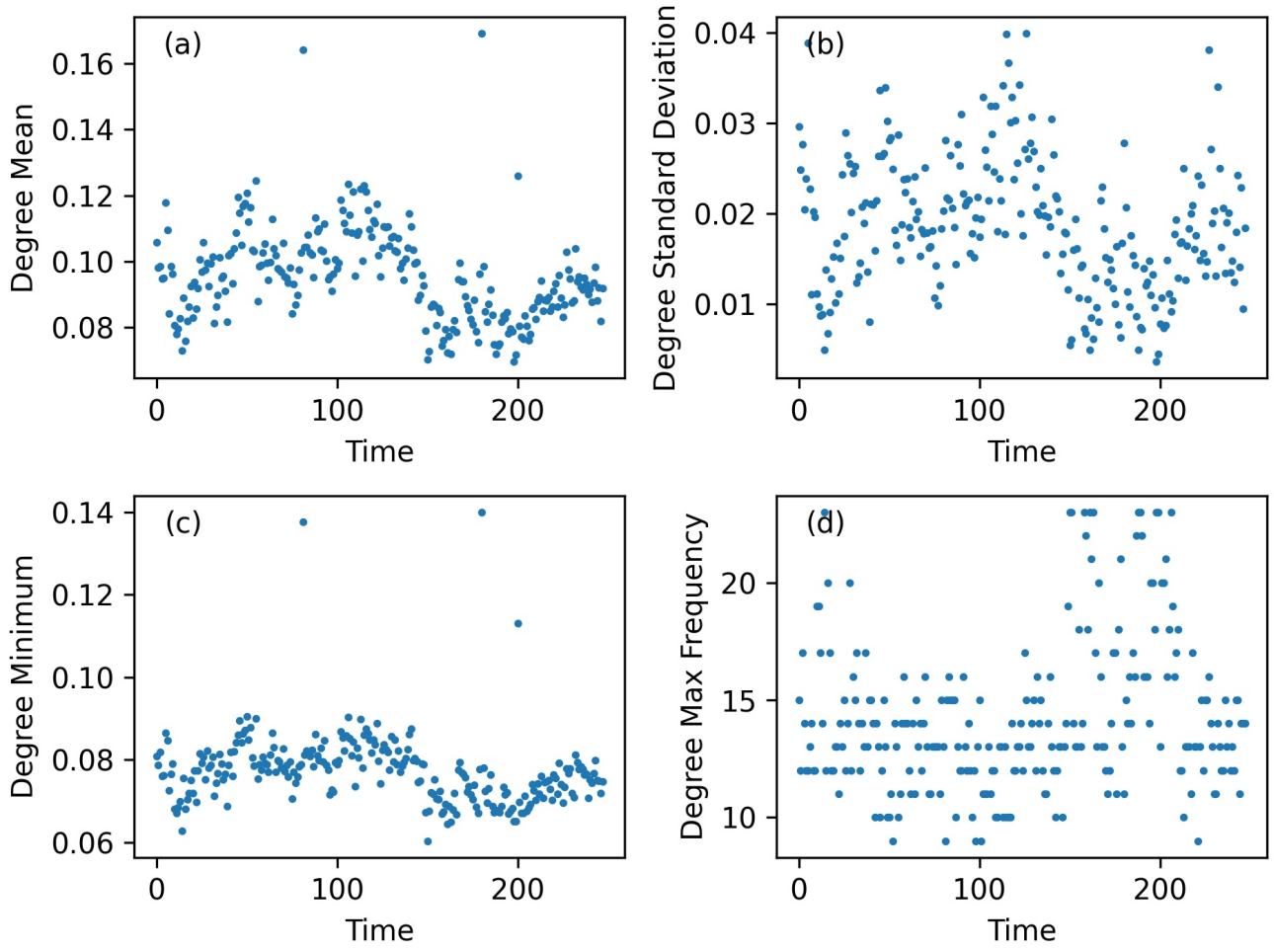

Figure 4.4: Evolution of Degree (a) Mean, (b) Standard Deviation, (c) Minimun and (d) Maximum.

eigenvector centrality among stocks, that is, most of the nodes are equally important in the network. That's why high connectedness can be associated with the shrinking of histograms, in opposition to centralized synchronicity, which happens when there are clearly dominating stocks, much more centrals to the network than the others.

Also in the evolution of Degree and EVC per stock (figure 4.6), we can see a different regime in about the same period, when their values seem to be distributed more uniformly between stocks than before: both high and low scored stocks tend to equalize each other towards a intermediate value (in the heat graph, blues and reds moves towards white). An exceptional stock, that seems to remain very central to the network through the entire year but with particular force during the highly centralized season is Citigroup (C). The next stock in importance seems to be Morgan Stanley (MS), also from the Financial sector. Interestingly, Facebook (FB) and U. S. Bancorp (USB) increase their influential score just the day

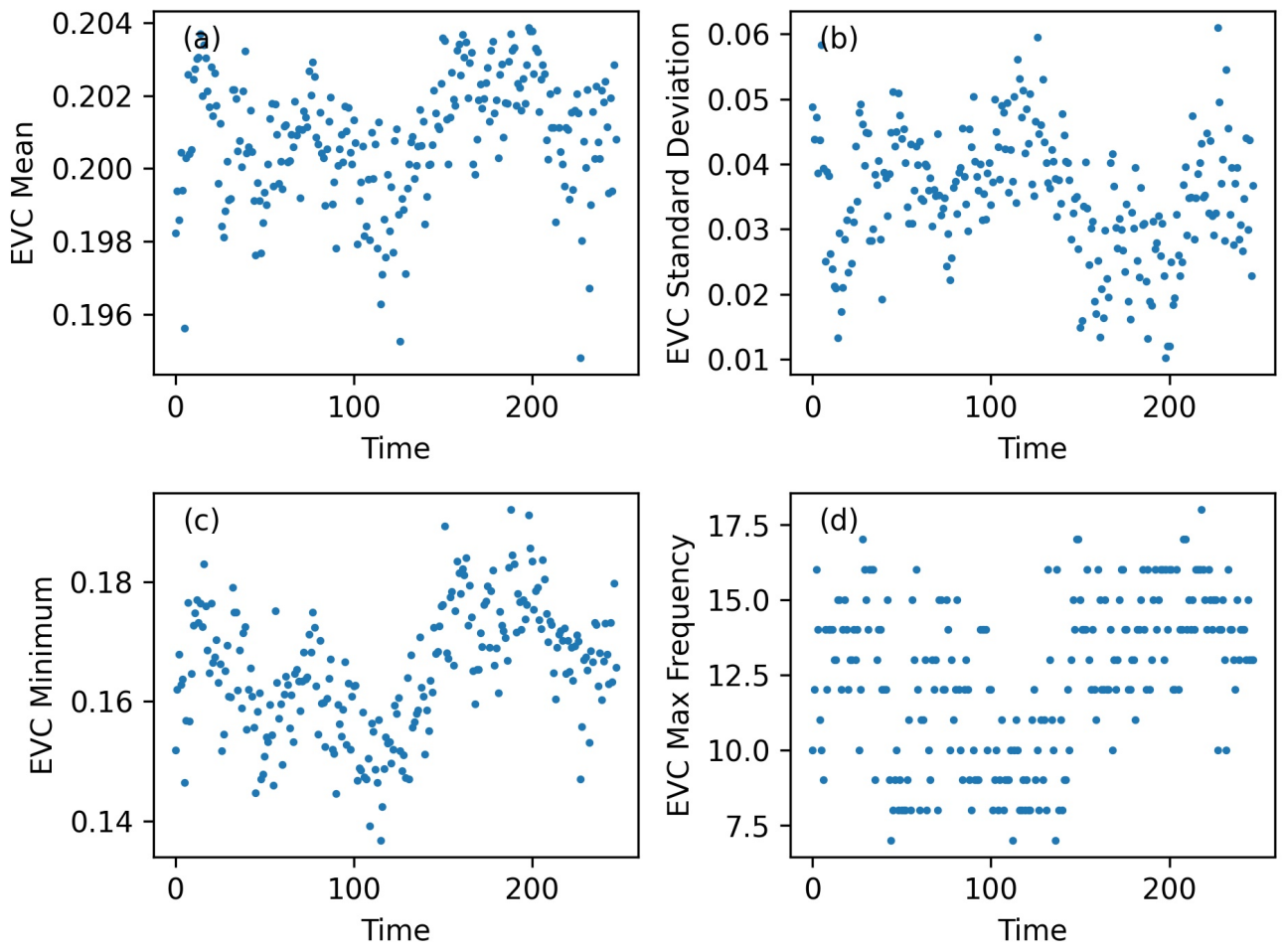

Figure 4.5: Evolution of Eigenvector Centrality (a) Mean, (b) Standard Deviation, (c) Minimun and (d) Maximum.

after the last two oulier days. Degree very clearly shows again the presence of outlier days, and we can spot the stocks driving their dynamics: the already noted Citigroup and Morgan Stanley, but also Cisco Systems (CSCO), Oracle (ORCL), Microsoft (MSFT) and Ford (F): with the exception of Ford, all of them belonging to the digital technology sector. The last outlier day seems to be a very decentralized one, its corresponding row displaying a remarkably uniform color.

Of course, the visual evidence here presented is not clear enough as to conclude in any formal way the existence of such dynamical regimes. It is to formalize and further investigate this structure that the next sections are dedicated.

Let's say that if the normalized eigenvector centrality or degree vectors are understood to be measures of market direction and strength, by measuring which specific stocks are driving it and how much, their

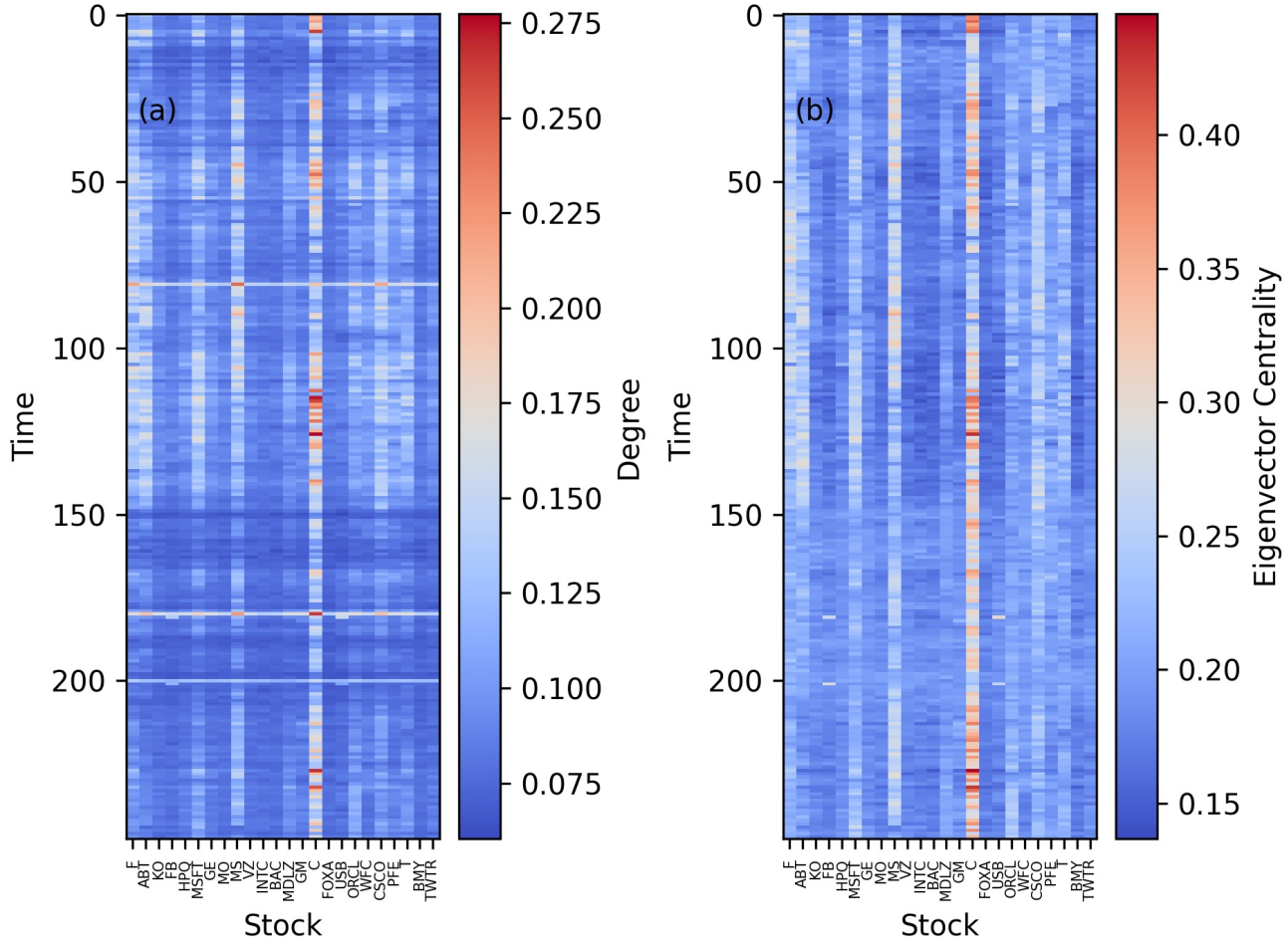

Figure 4.6: Evolution Through the Trading Year of (a) Degree and (b) Eigenvector Centrality per stock.

histograms, by disregarding the latter information, can be understood to measure just the intensity of this "market force", how much stocks are well connected and so on.

4.4 Clustering Analysis

Next, we want to apply a couple of clustering algorithms to our dynamical network, looking for typical market states, for which we need first to define a matrix distance $\zeta(A, B)$ to measure similarity between daily transcript synchronization networks. After inspection of similarity matrixes with different metrics (L^1, L^2, Jensen-Shannon Distance (the square root of D_{JS})) and different phase representation (distances are measured between the whole transcript synchronization matrixes, the EVC or degree vectors, and also between their histograms) one can conclude that whether we use L^2 or Jensen-Shannon Distance we

arrive to similar results, and the difference lies in the phase representations, of which the histograms of degree or/and EVC seem to be the most informative ones. Thus, in what follows we will be studying not the adjacency matrixes themselves, but the histograms of their EVC and Degree vectors. Since it is more meaningful than L^2 norm when measuring distances between probability distributions (it can be understood as a minimum redundancy measure [101]), we will use the Jensen-Shannon distance for the dendogram clustering. This is done also in order to reflect the structure found in the previous section.

If we then plot the Jensen-Shannon distance matrix $D_{ij} = \zeta_{JS}(H_i^{transcript}, H_j^{transcript})$, whose ij term equals the Jensen-Shannon distance between networks corresponding to days i and j, on Degree phase space (figure 4.7) we can clearly distinguish our outlier days as particularly distant of the typical days, which are very close to each other, thus confirming our previous idea: these days we observe a collective behavior in terms of determinism. In the same figure, as also in figure 4.8, which shows the same for EVC phase space, we observe at least two well separated seasons, previously identified as centralized and decentralized connectedness seasons. They exhibit oscillations but are clearly distinguisable by their internal coherence (low distances detected as blue blocks on diagonal) as well as the high distances between them (yellow strips in the figures). For comparison, see section 4.A for analogous figures when classical correlation coefficients are used instead of transcript synchronization.

While both phase spaces are good detecting structure in periods (highly decentralized season for instance), degree is far better highlighting outlier days.

Of course, to choose the phase space as that of the histograms is problematic in that it adds an extra, posibly very sensitive parameter, and more generally a whole new problem: the number of bins and the binning process. Here we choose ten uniformly sized bins covering the whole range of the corresponding quantity throughout the trading year.

We can now use ζ to cluster our daily matrixes (and thus our dynamical weighted network) looking for distinctions between collective states [92] [93] [20][24]. For this, we use two very different algorithms: first, an aglomerative dendogram with the Jensen-Shannon distance, whose merging process is set to stop by a cut-off threshold, chosen after careful inspection of the dendograms (figures 4.9 and 4.10, the threshold is shown as an horizontal gray line, and is set to 0.13 for Degree and 0.14 for EVC) and the L^2-based K-Means algorithm. As the dendogram gives us lots of clusters with just a couple of elements as their members, we, in order to keep the number of clusters reasonable, merge all of those with 3 elements or less into a unique set labeled as “Noise”, yet at the cost of losing some information about outliers, which as we will see will be recovered by the K-Means algorithm. The other clusters, which are our states and after the last merge into the noise set turn out to be 12 for Degree and 14 for

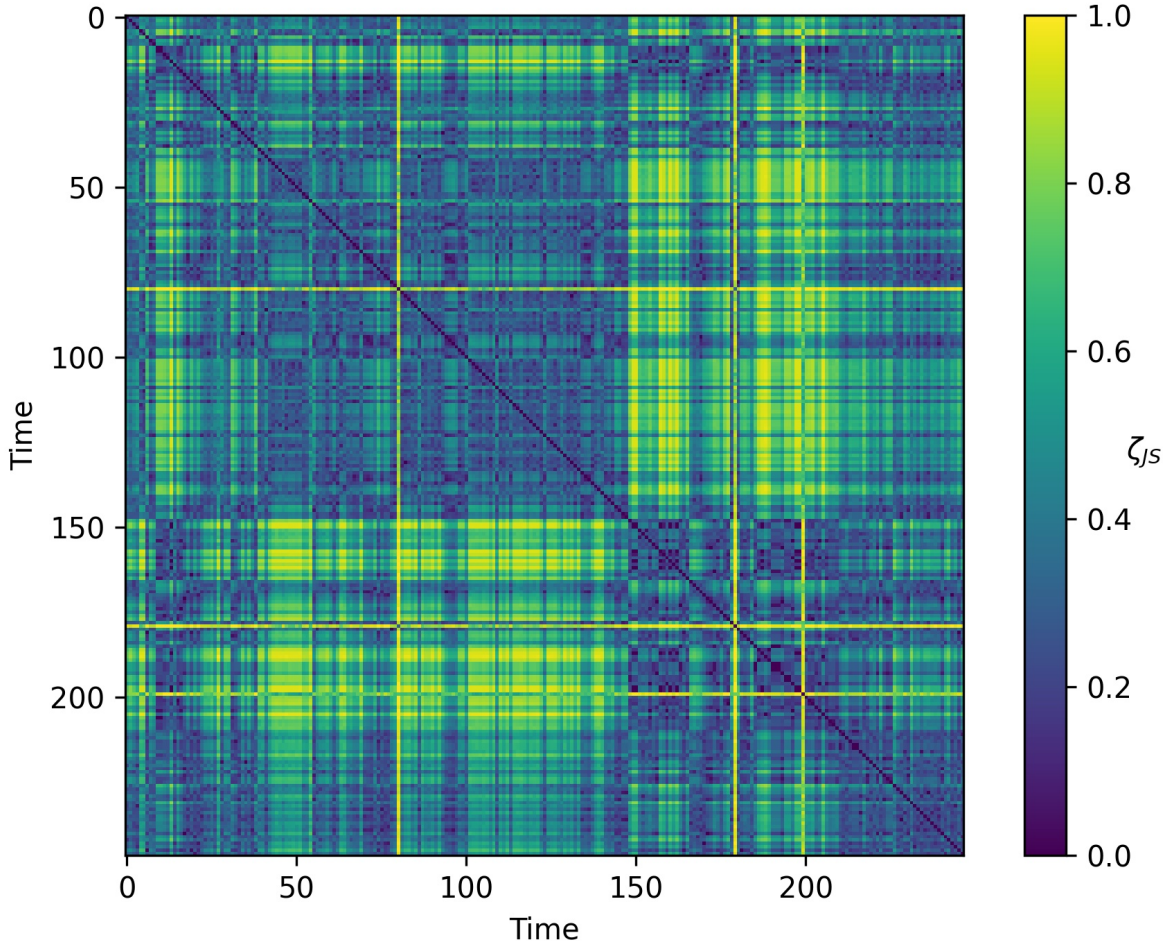

Figure 4.7: Jensen-Shannon Distance Matrix in Degree Phase Space. Its ij term equals the Jensen-Shannon distance between networks corresponding to days i and j

EVC, are ordered according to the Jensen-Shannon distance between the centroid (mean) of each of them and the corresponding uniform distributions, while the clusters obtained by the K-Means algorithm are ordered by their L^2-norm. Thus, low states are those whose centroids lie nearer to the uniform distribution, that is, they are centralized synchronicity states, and the high states are for the same reason decentralized states. The state of a given trading day reflects then, as was our intention, the level of centralization/decentralization. The cophenetic correlation of the JS-based dendograms are 0.41 and 0.43 for Degree and EVC phase spaces respectively. When applying K-Means algorithm, we choose the same number of clusters as that obtained by the dendogram algorithm.

In figures 4.11- 4.14), the upper panel displays the evolution of states throughout the year: black small dots at the bottom of the graph indicate noise days when dendogram is used, blue medium-sized

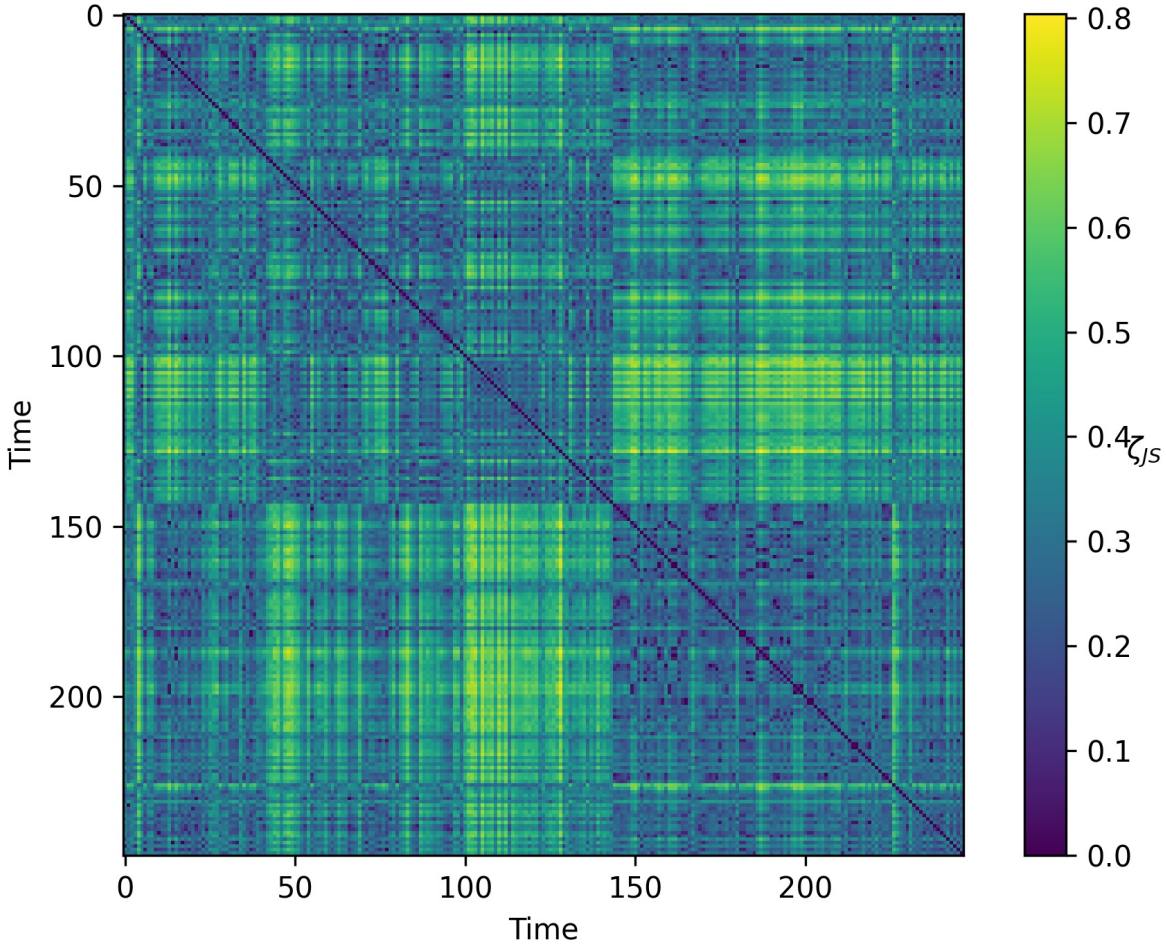

Figure 4.8: Jensen-Shannon Distance Matrix in EVC Phase Space. Its $i\,j$ term equals the Jensen-Shannon distance between networks corresponding to days i and j

dots indicate typical-and-not-noisy days and red big dots indicate outlier days; when a noisy day happen to be also an outlier day, it is shown as a red big dot. The lower panel of the same figures displays the ζ_{JS} or L^2-ordered centroids of the clusters (states). While both clustering methods agree in the detection of a low-states season and a high-states one, as well as in the shape of the centroids (of non-singleton clusters), they display different, complementary features.

First of all, it is noteworthy that both phase spaces and with both clustering methods, very different in nature and using two different distances, display very similar dynamics. This is a strong feature supporting our findings, as we can observe the highly decentralized season with particular persistence and clarity, as well as roughly monthly oscillations, less clear for dendogram algorithm due to the way we labeled days as noisy.

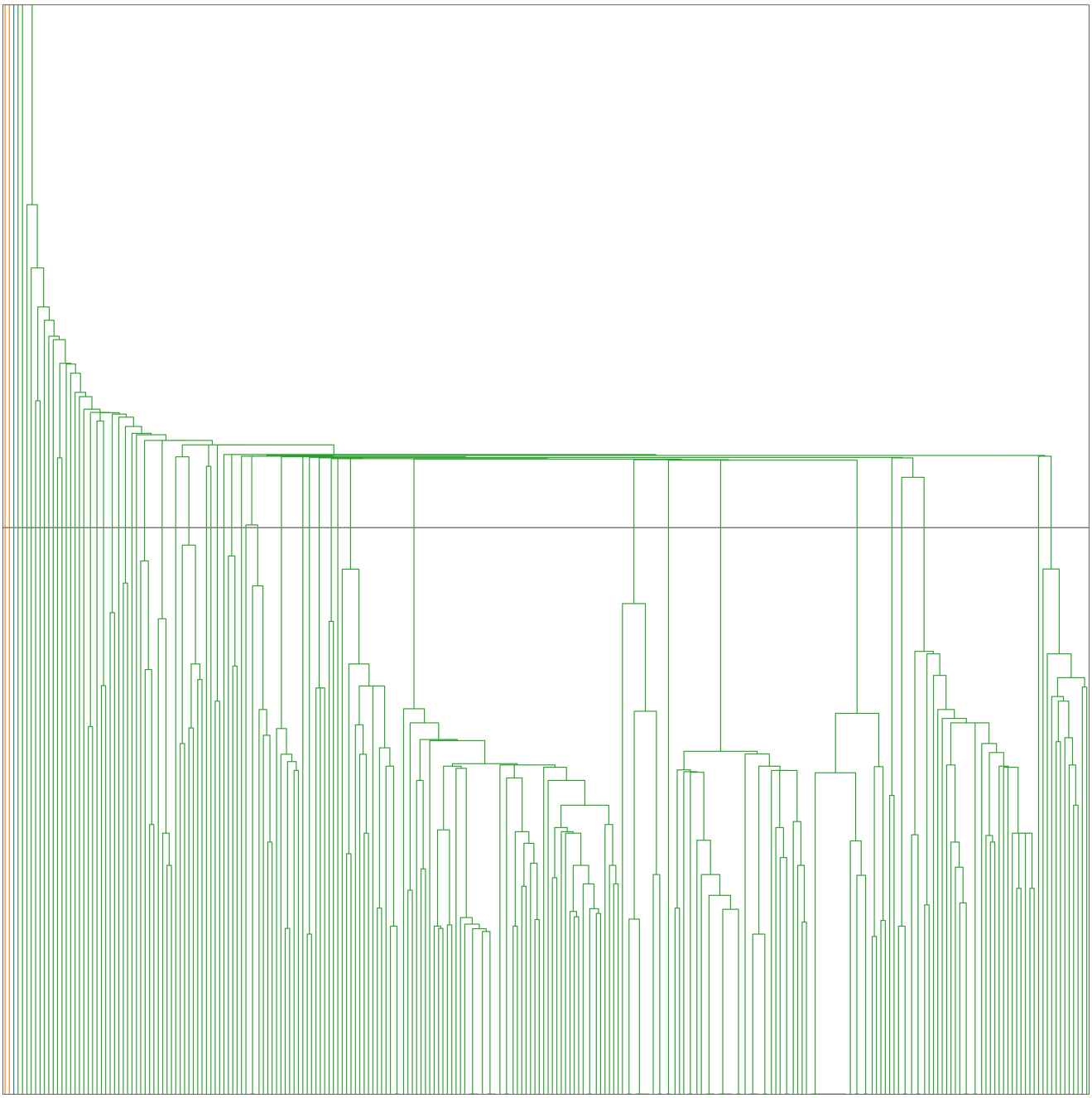

Figure 4.9: Clustering Dendogram on Degree Phase Space with Jensen-Shannon Distance. x axis represents the daily networks to be clustered, y axis represents distance. Labels are omitted for better visualization. The cut-off threshold is set to 0.13 and displayed as an horizontal gray line.

We have now more information to further classify outlier days and grasp a little more about their nature, but first some comments are required. First of all, both clustering algorithms on both phase spaces agree in this: the last outlier day belongs to a higher state than that of the other outlier days. This

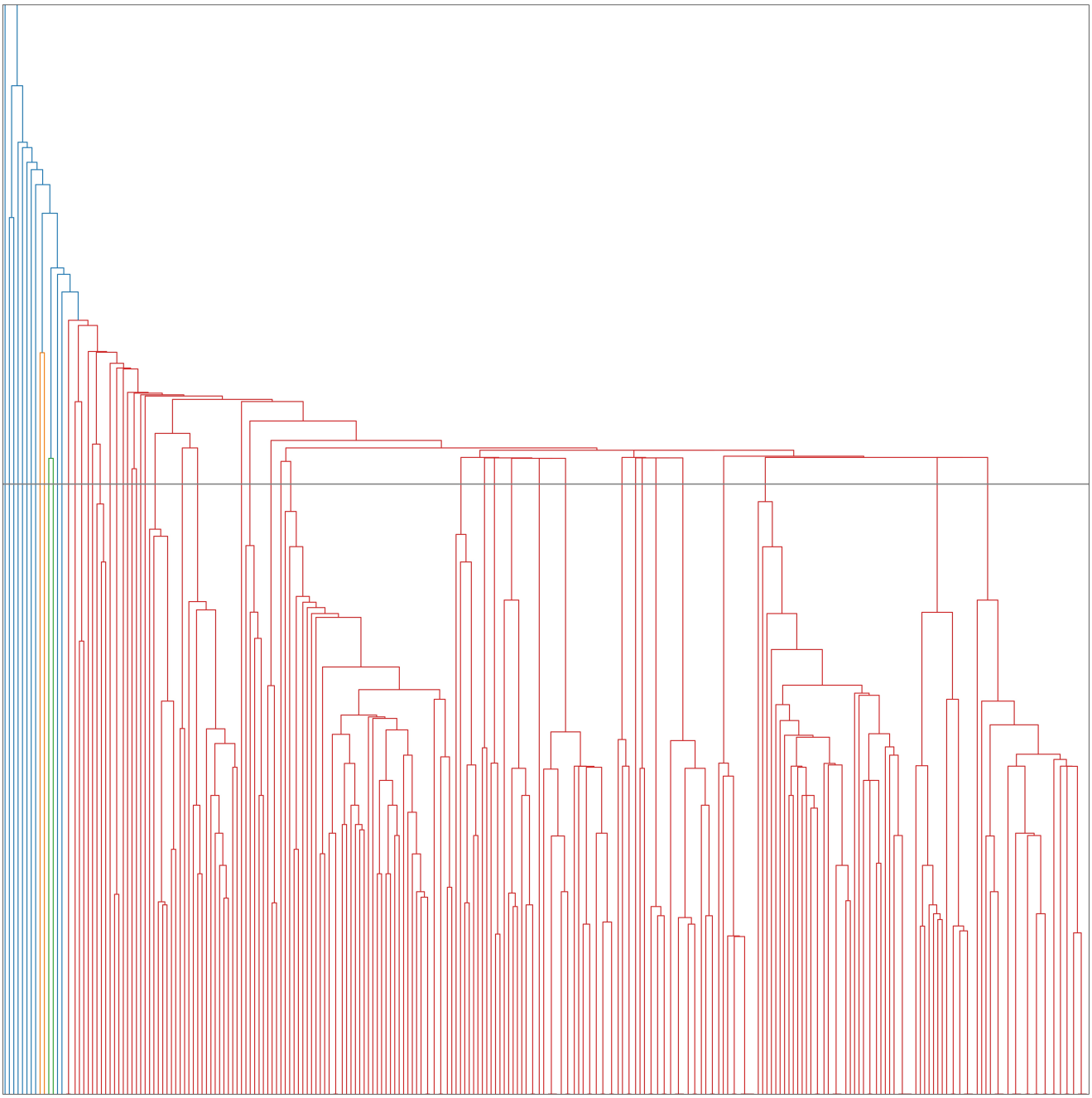

Figure 4.10: Clustering Dendogram on EVC Phase Space with Jensen-Shannon Distance. x axis represents the daily networks to be clustered, y axis represents distance. Labels are omitted for better visualization. The cut-off threshold is set to 0.14 and displayed as an horizontal gray line.

allows us to conclude that the last outlier day is of a particularly decentralized nature when compared with the others. That been said, the figures disagree in how different are those states: both algorithms finds high states for all three outlier days in EVC phase space; while this could be accepted without

further thoughts for K-Means algorithm, this is unlikely for dendogram algorithm, since we would have expected the outlier days being classified as noise because of the way in which we defined the Noise set. Moreover, they are not only not classified as noise by dendogram algorithm: K-Means algorithm itself does not recognize them as outliers in any significant way, since they are not singled out as members of clusters with no more elements than themselves; and it is at this point that we should remember that EVC phase space was the one with more difficulties when it came to detect outlier days. On the contrary, in Degree phase space, which since the beginning was the strongest for outlier detection, our dendogram correctly recognizes all three outlier days as noise, which encourages us to affirm that it is to this phase space that we must turn in order to better understand outlier days. This insight is confirmed by inspecting the findings of K-Means algorithm on Degree phase space: it finds the first two oulier days belonging to a state, the lowest one, of which they are the only members, just as the last outlier day conforms a (singleton) state of its own, and a very high one indeed. Thus, outlier days, at first singled out by the individual analysis of stocks, are again and independently detected as outliers during the clustering analysis.

The previous discussion should let clear that our collective analysis is able, not just to reproduce, but to further explain the nature of our individually detected outlier days as extreme manifestations of a collective behavior present throughout the trading year (centralized and decentralized synchronicity), as well as to discriminate between them in terms of that observed behavior: the first two outlier days are shown to be the most centrally synchronized of the whole year, while the last one is of a highly centralized nature, and a very peculiar one since it constitutes a market state on its own.

4.5 A Markov Model for State Transitions

Finally, in order to model these state dynamics in a simple way, we briefly propose a first order Markov model for prediction of the next day state given that we know today market state.

To check whether this Markovian approach is adequate or not we, just as in [20], compute the empirical transition probability matrix P from one state to the next, given that we know the current state, as well as the theoretical stationary probability distribution of states for such a Markov model given by the first order transition matrix P, that is, the probability distribution π giving the expected probabilities of finding the market in a given state over a very long period (provided it is Markovian), and satisfying the linear equation:

$$\pi_j = \sum_i P_{ij}\pi_i,$$

and compare it to the empirical frequencies of states through the year (figure 4.15), just to find that they are indeed very similar to each other for every combination of phase spaces and clustering algorithms. We then conclude that the states dynamics is consistent with a Markov process in which enough time has passed as to reach its steady state, and that to guess tomorrow state given knowledge about today state is in general a reasonable bet.

A considerable part of our work on ordinal patterns has been done with the Python package ordpy [102], while clustering analysis has been carried out through scikit-learn [103]. We also have made extensive, though elementary, use of NumPy [104].

4.6 Conclusions

In order to analyze collective states dynamics of stocks in high-frequency digital markets, and to overcome the limitations of correlation matrixes in detecting non-linear interactions in noisy time series, we proposed to study transcripts synchronization dynamical networks and their eigenvector centrality and degree vectors distributions.

After measuring different information theoretic quantities on the daily ordinal pattern series of individual stocks and detecting three outlier, semi-deterministic days in most of them and discrete levels in some, we have shown them to be extreme manifestations of a collective, emergent behavior not entirely reflected in individual dynamics. Applying two very different clustering algorithms, we were able to detect specific, persistent and quantitatively distinguishable market states throughout the one-year period of study, as well as two well defined and quantitatively distinguishable seasons of the trading year, characterized by their degree of centralized/decentralized synchronicity, with remarkable similar results for both algorithms. We also succesfully classified our previously found outlier days in terms of centralized and decentralized synchronicity. Finally, we showed that state transitions dynamics can be well described as a simple first order Markov process.

Of course, our work has several limitations that ought to be highlighted. As already mentioned, to choose the phase space as that of the histograms of EVC and Degree leaves open the question of the correct number of bins to be used. Also, it would be desirable to find a more objective criterion to choose the number of clusters; various purely quantitative ones are discussed in the litearature (see for a summary [105]), but none of those are convincing from our viewpoint; and ultimately, as explained in [106], clustering analysis is a problem dependent process and should not be subordinated to an abstract, global score. As our aim in this work was to propose and illustrate a methodology, as well as to adapt the

network-clustering pipeline often mentioned thorughout this work [24] to high-frequency digital markets, we did not deepen into this questions; instead, in both cases we contented ourselves with confirming the robustness of our results by varying the number of bins and dendogram threshold and founding similar qualitative results in a reasonable range of values and with two very different clustering algorithms for our particular data set.

It would be desirible to have an ecomomic explanation for the behavior here observed; unfortunately, that is significantly difficult for high-frequency digital markets, because algorithms are particularly opaque in their trading decisions [13] [12]. At the moment, we can just stand for a phenomenological approach such as that of this chapter.

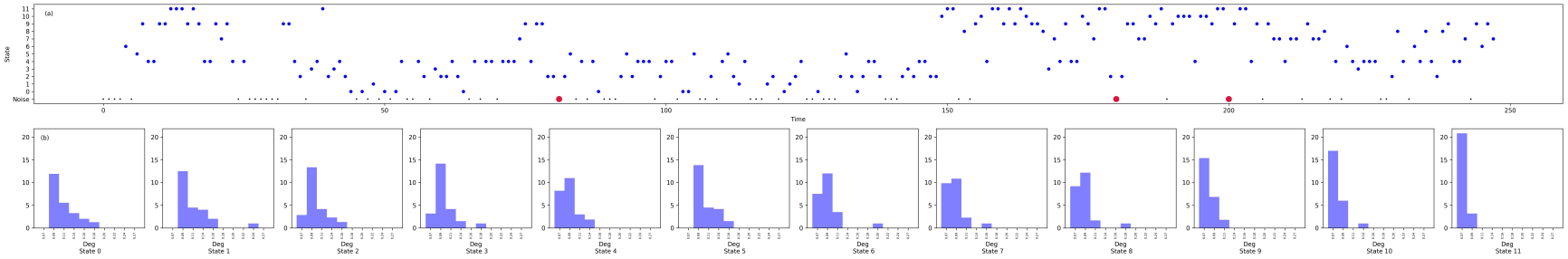

Figure 4.11: (a) Evolution of States in Degree Phase Space and (b)ζ_{JS}-ordered Centroids, Dendogram Algorithm.

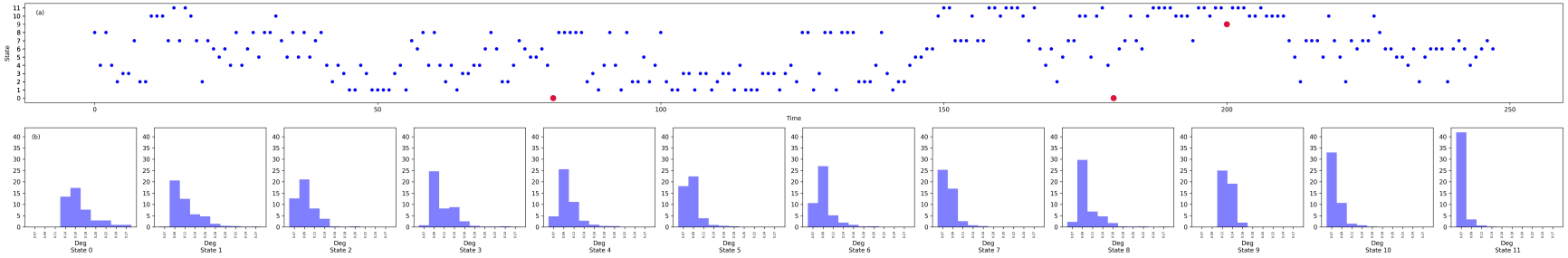

Figure 4.12: (a) Evolution of States in Degree Phase Space and (b) L^2-ordered Centroids, K-Means Algorithm.

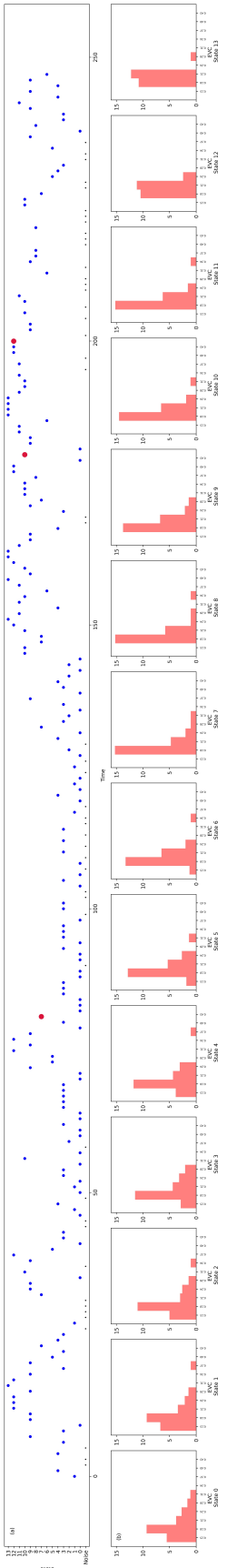

Figure 4.13: (a) Evolution of States in EVC Phase Space and (b) ζ_{JS}-ordered Centroids, Dendogram Algorithm.

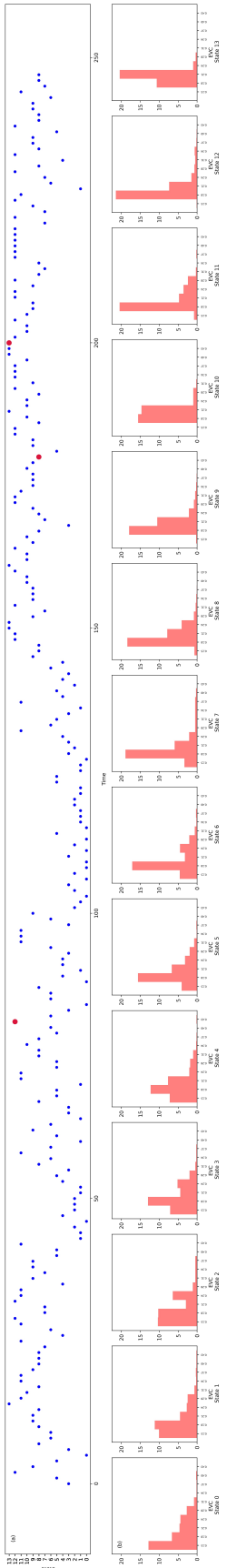

Figure 4.14: (a) Evolution of States in EVC Phase Space and (b) L^2-ordered Centroids, K-Means Algorithm.

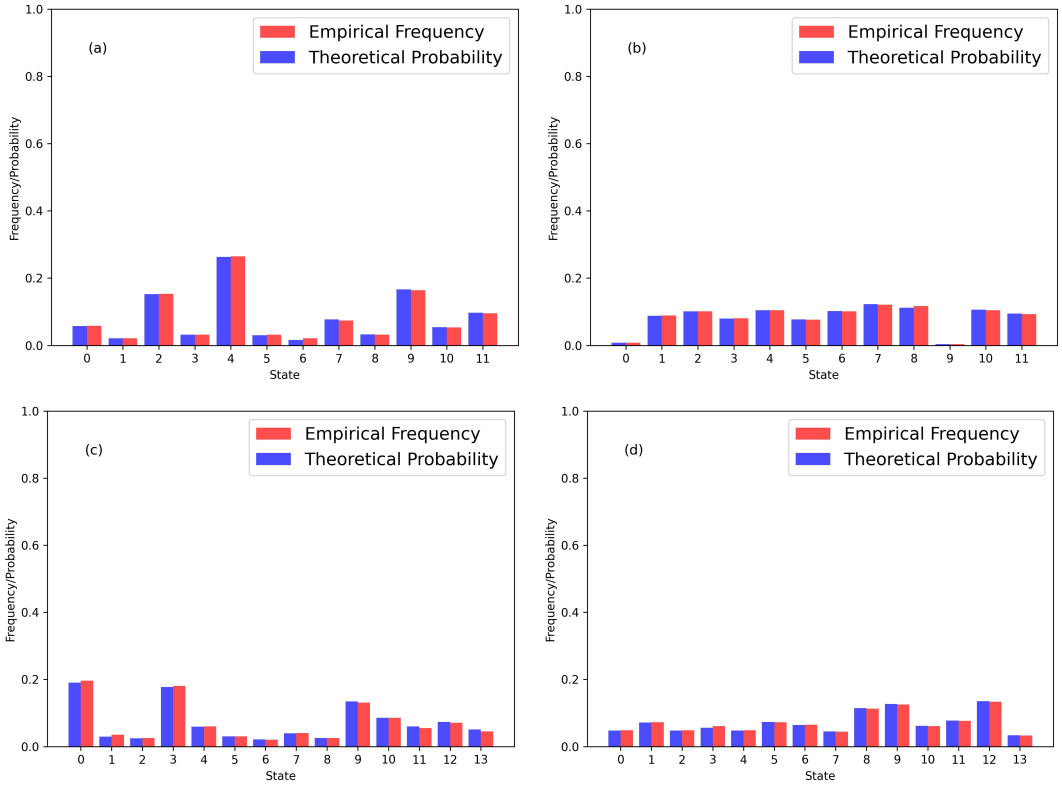

Figure 4.15: Empirical Frequency of States and Theoretical Stationary Distribution for the Markov Model. (a) Degree, Dendogram; (b) Degree, K-Means; (c) EVC, Dendogram; (d) EVC, K-Means.

Appendix

4.A Correlation Matrixes

For the sake of comparison, and to make clear why we use transcript synchronicity as our pairwise coupling measure instead of the more classic and straightforward correlation coefficient, we include here a few figures similar to those displayed above, but this time for correlation matrixes. However, since we are talking here about synchronization, we use the absolute value of such correlation coefficients, since a correlation coefficient equal to -1 should be understood as two perfectly (linearly) synchronized time series.

In figure 4.A.1 we can see that, althought some structure is still present in the evolution of the histogram of Degree for correlation matrixes, its presence is less clear than that observed in 4.3, while the histogram of EVC is not useful at all when correlation coefficients are used.

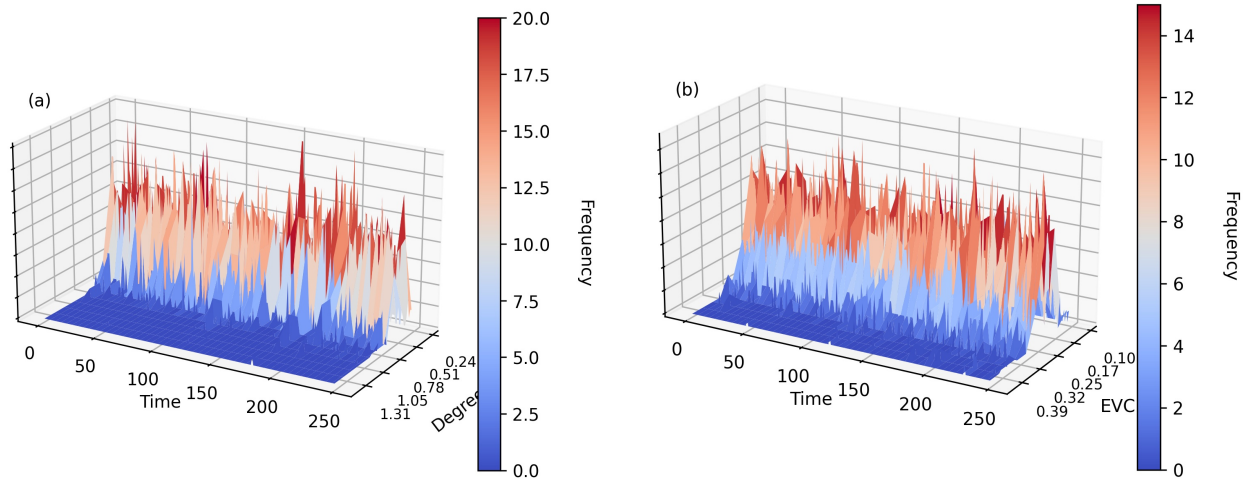

Figure 4.A.1: Evolution Through the Trading Year of Histograms of (a) Degree and (b) Eigenvector Centrality for Correlation Dynamical Network.

This should be clearer in figure 4.A.2, which shows the matrix of Jensen-Shannon distances between daily correlation matrixes, analogously to figures 4.7 and 4.8. The first panel of this figure still correctly detects oulier days clearly enough, but the knowledge of the centralized and decentralized seasons is almost totally lost, while the second panel is too noisy to conclude anything. Consequently, clustering analysis yields poorer results when compared to our previous transcript-based analysis.

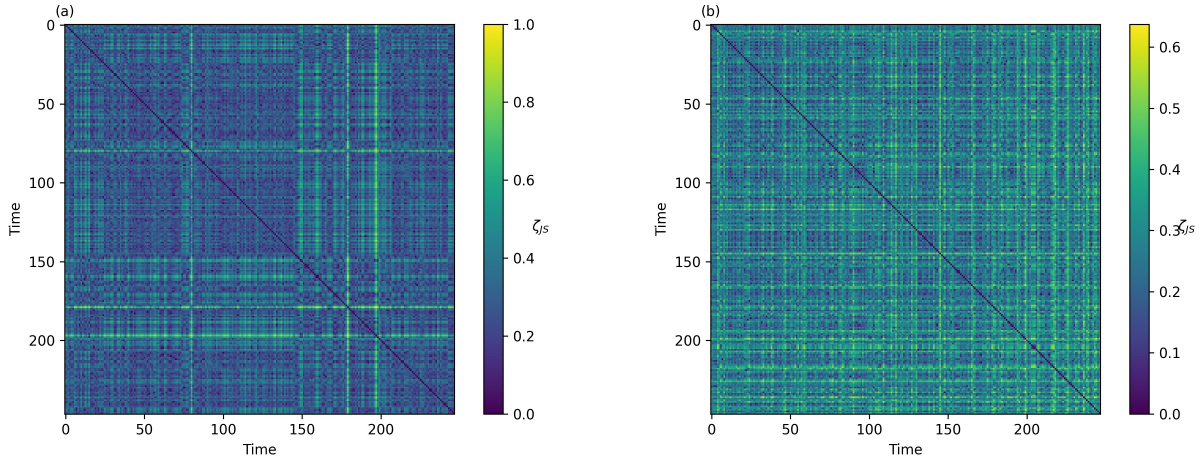

Figure 4.A.2: Jensen-Shannon Distance Matrix in (a) Degree and (b) EVC Phase Spaces. Its ij term equals the Jensen-Shannon distance between correlation networks corresponding to days i and j.

4.B Multiscale Analysis

In section 2.2 we made the statement that our results are robust to variations of the time lag parameter l. As already mentioned, in [58] the authors make it clear that a multiscale analysis is unavoidable if we want to guarantee the validity of our results. Thus, we plot here a couple of figures obtained when $l = 5, 25, 50, 75$ and 100 and everything else is kept as before. For brevity and space, we limit ourselves to plot the distance matrixes analogous to figures 4.7 and 4.8.

Figure 4.B.1, displays Jensen-Shannon distance matrixes for Degree (upper panel) and EVC (lower panel) phase spaces. It is clear that these matrixes are pretty similar to those just refered, detecting outlier days as well as highly centralized and decentralized seasons. Labels and colorbars have been removed to improve visualization. Clustering analysis yields similar results.

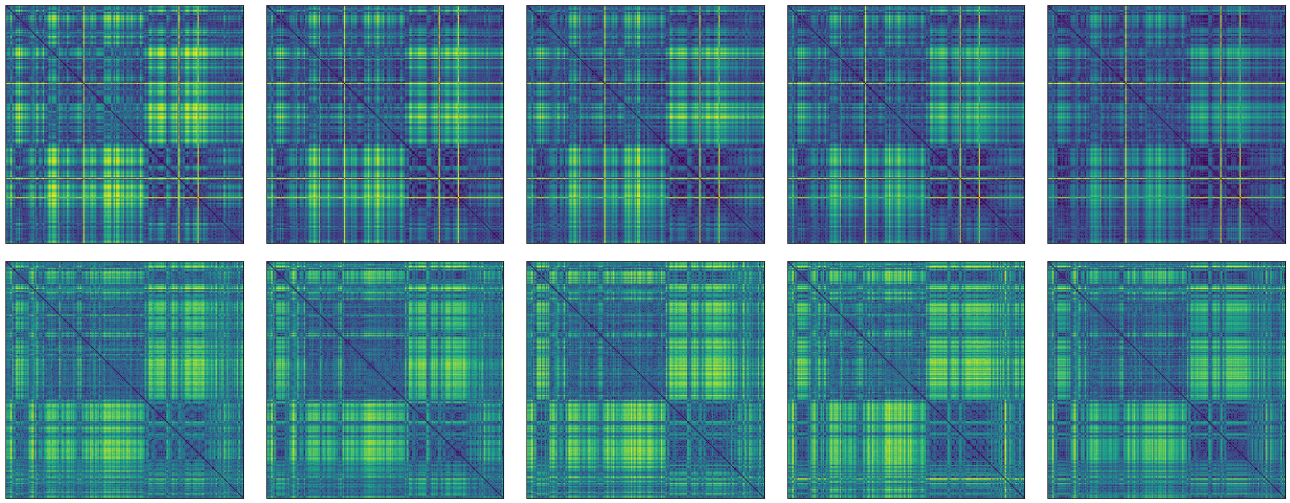

Figure 4.B.1: Jensen-Shannon Distance Matrixes in Degree (upper panel) and EVC (lower panel) Phase Spaces. Its ij term equals the Jensen-Shannon distance between transcript sychronization networks corresponding to days i and j for $l = 5, 25, 50, 75$ and 100, plotted in that order from left to right.

Chapter 5

Conclusions

Terminó la comedia:
Dentro de unos minutos parto para Chillán en bicicleta.
No me quedo ni un día más aquí
Sólo estoy esperando que se me sequen un poco las plumas.
Si preguntan por mí
Digan que ando en el sur, y que no vuelvo hasta el próximo mes.
Digan que estoy enfermo de viruela.
Atiendan el teléfono
¿Qué no oyen el ruido del teléfono?
¡Ese ruido maldito del teléfono va a terminar volviéndome loco!
Si preguntan por mí, pueden decir que me llevaron preso
Digan que fui a Chillán a visitar la tumba de mi padre.
Yo no trabajo ni un minuto más. Basta con lo que he hecho
¿Que no basta con todo lo que he hecho?
¡Hasta cuándo demonios quieren que siga haciendo el ridículo!
Juro no escribir nunca más un verso. Juro no resolver más ecuaciones
Se terminó la cosa para siempre
¡A Chillán los boletos!
¡A recorrer los lugares sagrados!

Nicanor Parra, *Hombre al Agua*

As stated in the Introduction, the spreading of Algorithmic Trading has arisen a lot of interesting

questions in the most different scientific areas, as well as a reasonable feeling of uncertainty about its potential consequences. Here we have tried, by mathematical means and from a phenomenological point of view, to tackle a couple of them: market efficiency and collective ordinal dynamics.

The first issue is studied through the Hurst exponent (chapter 3). Results indicate that, in the period from March 7, 2018 to March 7, 2019 and for the 24 assets in the United States market and the 35 in the Mexican market studied here, the Efficient Market Hypothesis is clearly rejected: the presence of long-term memory, particularly of anti-persistence, is clear. This is an important result for economic theory: wheter human or algorithmic, market efficiency remains elusive to say the least. Although that should be clear since the Flash Crash of 2010, the statistical evidence here offered bring us one step closer to understand market inefficiency in "normal" conditions.

Next, we took a deeper look into the structure of correlations between stocks. After an individual analysis of 24 stocks of the US market during a trading year of fully automated transactions by means of ordinal pattern series, we defined an information-theoretic measure of pairwise synchronization for time series which allows us to study this subset of the US market as a dynamical network. We applied to the resulting network a couple of clustering algorithms in order to detect collective market states, characterized by their degree of centralized or descentralized synchronicity. This collective analysis has shown to reproduce, classify and explain the anomalous behavior previously observed at the individual level. We also found two whole coherent seasons of highly centralized and descentralized synchronicity, respectively. Finally, we modeled these states dynamics through a simple Markov model.

So it is our hope to have contributed a little to the understanding of this recent historical stage of financial markets characterized by algorithms and speed-light trading, through our analysis of a particular data set, as well as to have proposed useful methodological insights potentially applicable to a broad variety of phenomena.

Of course, a lot of questions remain open, the three more obvious being: if human "imperfect rationality" is not the cause of market inefficiency, then which is? What is the economic explanation of outlier days and centralized/decentralized seasons of the trading year? And finally, are these phenomena specific to our data set, or are they a common feature of High-Frequency Digital Markets? To answer these questions will require a lot of collective efforts by scholars from different areas of science, efforts well beyond the intentions and capacities of this work. Let those questions be the inspiration for future research.

References

[1] Archibald Hutcheson. *A collection of calculations and remarks relating to the South Sea scheme*. London, 1720.

[2] Max Weber. Stock and commodity exchanges. *Theory and Society*, 29(3), 2000. URL: http://www.jstor.org/stable/3108485.

[3] T. Veblen. *The Theory of Business Enterprise*. Scribner's Sons, 1904.

[4] Joseph A. Schumpeter. *The Theory of Economic Development*. Harvard University Press, Cambridge, 1911.

[5] J. Keynes. *The General Theory of Employment, Interest and Money*. Palgrave Macmillan, 1936.

[6] R. Hilferding. *Finance Capital. A Study of the Latest Phase of Capitalist Development*. Routledge, 1981.

[7] Yanis Varoufakis. *The Global Minotaur*. University of Chicago Press, 2011.

[8] K. J. Arrow and G. Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3), 1954. doi:https://doi.org/10.2307/1907353.

[9] Karl Marx. *Capital: A Critique of Political Economy, vol. III*. International Publishers, NY, 1967.

[10] Juan Pablo Pardo-Guerra. *Automating Finance. Infrastructures, Engineers and the Making of Electronic Markets*. Cambridge University Press, 2019.

[11] Thomas Skou Grindsted. Algorithms and the anthropocene: Finance, sustainability, and the promise and hazards of new financial technologies. apr 2019. American Association of Geographers. URL: http://www.aag.org/cs/events/event_detail?eventId=1259.

[12] Frank Pasquale. *The Black Box Society*. Harvard University Press, 2015.

[13] Michael Lewis. *Flash Boys: A Wall Street Revolt*. W. W. Norton & Company, 2014.

[14] Cathy O'Neil. *Weapons of Math Destruction*. Crown Books, 2016.

[15] L. Bachelier. *Théorie de la Spéculation*. Annales de l'Ecole Normale Supérieure, 1900.

[16] Philip Mirowski. *More Heat than Light: Economics as Social Physics, Physics as Nature's Economics*. Cambridge University Press, 1989.

[17] Emmanuel Farjoun and Moshé' Machover. *Laws of Chaos. A Probabilistic Approach to Political Economy*. Verso, 1983.

[18] R. N. Mantegna and H. E. Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, 2000.

[19] R. Mansilla. *Una breve introducción a la Econofísica*. Equipo Sirius, S. A., 2003.

[20] Hirdesh K Pharasi, Kiran Sharma, Rakesh Chatterjee, Anirban Chakraborti, Francois Leyvraz, and Thomas H Seligman. Identifying long-term precursors of financial market crashes using correlation patterns. *New Journal of Physics*, 20(10):103041, 2018. doi:10.1088/1367-2630/aae7e0.

[21] Anirban Chakraborti, Kiran Sharma, Hirdesh K Pharasi, K Shuvo Bakar, Sourish Das, and Thomas H Seligman. Emerging spectra characterization of catastrophic instabilities in complex systems. *New Journal of Physics*, 22(6):063043, jun 2020. doi:10.1088/1367-2630/ab90d4.

[22] Kiyoshi Kanazawa, Takumi Sueshige, Hideki Takayasu, and Misako Takayasu. Derivation of the boltzmann equation for financial brownian motion: Direct observation of the collective motion of high-frequency traders. *Phys. Rev. Lett.*, 120:138301, 2018. doi:10.1103/PhysRevLett.120.138301.

[23] Benoit Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419, 1963. URL: http://www.jstor.org/stable/2350970.

[24] H. Masuda and P. Holme. Detecting sequences of system states in temporal networks. *Scientific Reports*, 9:795, 2019. doi:https://doi.org/10.1038/s41598-018-37534-2.

[25] Christoph Bandt and Bernd Pompe. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.*, 88:174102, Apr 2002. doi:10.1103/PhysRevLett.88.174102.

[26] Justin Joque. *Revolutionary Mathematics: Artificial Intelligence, Statistics, and the Logic of Capitalism*. Verso, 2022.

[27] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1):770–799, 1951. doi:10.1061/TACEAT.0006518.

[28] Benoit B. Mandelbrot and James R. Wallis. Robustness of the rescaled range r/s in the measurement of noncyclic long run statistical dependence. *Water Resources Research*, 5(5):967–988, 1969. doi:https://doi.org/10.1029/WR005i005p00967.

[29] M.A. Sánchez Granero, J.E. Trinidad Segovia, and J. García Pérez. Some comments on hurst exponent and the long memory processes on capital markets. *Physica A: Statistical Mechanics and its Applications*, 387(22):5543–5551, 2008. doi:https://doi.org/10.1016/j.physa.2008.05.053.

[30] G. Wang, G. Antar, and P. Devynck. The hurst exponent and long-time correlation. *Physics of Plasmas*, 7, 2000. doi:https://doi:10.1063/1.873927.

[31] Chun-Feng Li. Rescaled-range and power spectrum analyses on well-logging data. *Geophysical Journal International*, 153(1):201–212, 2003. doi:10.1046/j.1365-246X.2003.01893.x.

[32] M. A. Sánchez, J. E. Trinidad, J. García, and M. Fernández. The effect of the underlying distribution in hurst exponent estimation. *PLoS ONE*, 10, 2015. doi:https://doi:10.1371/journal.pone.0127824.

[33] A. W. Lo. Long-term memory in stock market prices. *Econometrica*, 59, 1991. doi:https://doi.org/10.2307/2938368.

[34] W. Willinger, M. S. Taqqu, and V. Teverovsky. Stock market prices and long-range dependence. *Finance and Stochastics*, 3(1), 1999. doi:https://doi.org/10.1007/s007800050049.

[35] Jozef Barunik and Ladislav Kristoufek. On hurst exponent estimation under heavy-tailed distributions. *Physica A: Statistical Mechanics and its Applications*, 389(18):3844–3855, 2010. doi:https://doi.org/10.1016/j.physa.2010.05.025.

[36] Daniel O Cajueiro and Benjamin M Tabak. The hurst exponent over time: testing the assertion that emerging markets are becoming more efficient. *Physica A: Statistical Mechanics and its Applications*, 336(3):521–537, 2004. doi:https://doi.org/10.1016/j.physa.2003.12.031.

[37] Cheoljun Eom, Sunghoon Choi, Gabjin Oh, and Woo-Sung Jung. Hurst exponent and prediction based on weak-form efficient market hypothesis of stock markets. *Physica A: Statistical Mechanics and its Applications*, 387(18):4630–4636, 2008. doi:https://doi.org/10.1016/j.physa.2008.03.035.

[38] D Grech and Z Mazur. Can one make any crash prediction in finance using the local hurst exponent idea? *Physica A: Statistical Mechanics and its Applications*, 336(1):133–145, 2004. doi:https://doi.org/10.1016/j.physa.2004.01.018.

[39] E. Peters. *Fractal Market Analysis. Applying Chaos Theory to Investment and Economics*. John Wiley and Sons, INC., 1994.

[40] D. Cysarz, P. Van Leeuwen, F. Edelhäuser, N. Montano, and A. Porta. Binary symbolic dynamics classifies heart rate variability patterns linked to autonomic modulations. *Computers in Biology and Medicine*, 42(3):313–318, 2012. doi:https://doi.org/10.1016/j.compbiomed.2011.04.013.

[41] U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths, and N. Wessel. Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics. *Computers in Biology and Medicine*, 42(3):319–327, 2012. doi:https://doi.org/10.1016/j.compbiomed.2011.03.017.

[42] Nicoletta Nicolaou and Julius Georgiou. Detection of epileptic electroencephalogram based on permutation entropy and support vector machines. *Expert Systems with Applications*, 39(1):202–209, 2012. doi:https://doi.org/10.1016/j.eswa.2011.07.008.

[43] Gaoxiang Ouyang, Xiaoli Li, Chuangyin Dang, and Douglas A. Richards. Deterministic dynamics of neural activity during absence seizures in rats. *Phys. Rev. E*, 79:041146, Apr 2009. doi:10.1103/PhysRevE.79.041146.

[44] Luca Faes, Giandomenico Nollo, and Alberto Porta. Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series. *Computers in Biology and Medicine*, 42(3):290–297, 2012. doi:https://doi.org/10.1016/j.compbiomed.2011.02.007.

[45] Chunhua Bian, Chang Qin, Qianli D. Y. Ma, and Qinghong Shen. Modified permutation-entropy analysis of heartbeat dynamics. *Phys. Rev. E*, 85:021906, Feb 2012. doi:10.1103/PhysRevE.85.021906.

[46] Xiaoli Li and Gaoxiang Ouyang. Estimating coupling direction between neuronal populations with permutation conditional mutual information. *NeuroImage*, 52(2):497–507, 2010. doi:https://doi.org/10.1016/j.neuroimage.2010.05.003.

[47] E. Olofsen, J. W. Sleigh, and A. Dahan. Permutation entropy of the electroencephalogram: a measure of anaesthetic drug effect. *British Journal of Anaesthesia*, 101:810–821, 2008. doi:https://doi.org/10.1093/bja/aen290.

[48] Joshua Garland, Tyler R. Jones, Michael Neuder, Valerie Morris, James W. C. White, and Elizabeth Bradley. Anomaly detection in paleoclimate records using permutation entropy. *Entropy*, 20(12), 2018. doi:10.3390/e20120931.

[49] María del Carmen Ruiz, Antonio Guillamón, and Antonio Gabaldón. A new approach to measure volatility in energy markets. *Entropy*, 14(1):74–91, 2012. doi:10.3390/e14010074.

[50] Massimiliano Zanin. Forbidden patterns in financial time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(1):013119, 2008. doi:10.1063/1.2841197.

[51] Luciano Zunino, Massimiliano Zanin, Benjamin M. Tabak, Darío G. Pérez, and Osvaldo A. Rosso. Forbidden patterns, permutation entropy and stock market inefficiency. *Physica A: Statistical Mechanics and its Applications*, 388(14):2854–2864, 2009. doi:https://doi.org/10.1016/j.physa.2009.03.042.

[52] Luciano Zunino, Massimiliano Zanin, Benjamin M. Tabak, Darío G. Pérez, and Osvaldo A. Rosso. Complexity-entropy causality plane: A useful approach to quantify the stock market inefficiency. *Physica A: Statistical Mechanics and its Applications*, 389(9):1891–1901, 2010. doi:https://doi.org/10.1016/j.physa.2010.01.007.

[53] Luciano Zunino, Aurelio Fernández Bariviera, M. Belén Guercio, Lisana B. Martinez, and Osvaldo A. Rosso. On the efficiency of sovereign bond markets. *Physica A: Statistical Mechanics and its Applications*, 391(18):4342–4349, 2012. doi:https://doi.org/10.1016/j.physa.2012.04.009.

[54] Luciano Zunino, Benjamin M. Tabak, Francesco Serinaldi, Massimiliano Zanin, Darío G. Pérez, and Osvaldo A. Rosso. Commodity predictability analysis with a permutation information theory approach. *Physica A: Statistical Mechanics and its Applications*, 390(5):876–890, 2011. doi:https://doi.org/10.1016/j.physa.2010.11.020.

[55] Arthur A. B. Pessa and Haroldo V. Ribeiro. Characterizing stochastic time series with ordinal networks. *Phys. Rev. E*, 100:042304, Oct 2019. doi:10.1103/PhysRevE.100.042304.

[56] Ruqiang Yan, Yongbin Liu, and Robert X. Gao. Permutation entropy: A nonlinear statistical measure for status characterization of rotary machines. *Mechanical Systems and Signal Processing*, 29:474–484, 2012. doi:https://doi.org/10.1016/j.ymssp.2011.11.022.

[57] Mariano Matilla-García and Manuel Ruiz Marín. A non-parametric independence test using permutation entropy. *Journal of Econometrics*, 144(1):139–155, 2008. doi:https://doi.org/10.1016/j.jeconom.2007.12.005.

[58] Felipe Olivares and Luciano Zunino. Multiscale dynamics under the lens of permutation entropy. *Physica A: Statistical Mechanics and its Applications*, 559:125081, 2020. doi:https://doi.org/10.1016/j.physa.2020.125081.

[59] Bilal Fadlallah, Badong Chen, Andreas Keil, and José Príncipe. Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. *Phys. Rev. E*, 87:022911, Feb 2013. doi:10.1103/PhysRevE.87.022911.

[60] Liu Xiao-Feng and Wang Yue. Fine-grained permutation entropy as a measure of natural complexity for time series. *Chinese Physics B*, 18(7):2690–2695, jul. doi:10.1088/1674-1056/18/7/011.

[61] Matthäus Staniek and Klaus Lehnertz. Symbolic transfer entropy. *Phys. Rev. Lett.*, 100:158101, Apr 2008. doi:10.1103/PhysRevLett.100.158101.

[62] José M. Amigó, Roberto Monetti, Beata Graff, and Grzegorz Graff. Computing algebraic transfer entropy and coupling directions via transcripts. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(11):113115, 2016. doi:10.1063/1.4967803.

[63] R. Lṕez-Ruiz, H. L. Mancini, and X. Calbet. A statistical measure of complexity. *Physics Letters A*, 209(5):321–326, 1995. doi:https://doi.org/10.1016/0375-9601(95)00867-5.

[64] P.W Lamberti, M.T Martin, A Plastino, and O.A Rosso. Intensive entropic non-triviality measure. *Physica A: Statistical Mechanics and its Applications*, 334(1):119–131, 2004. doi:https://doi.org/10.1016/j.physa.2003.11.005.

[65] A. Bahraminasab, F. Ghasemi, A. Stefanovska, P. V. E. McClintock, and H. Kantz. Direction of coupling from phases of interacting oscillators: A permutation information approach. *Phys. Rev. Lett.*, 100:084101, Feb 2008. doi:10.1103/PhysRevLett.100.084101.

[66] Roberto Monetti, Wolfram Bunk, Thomas Aschenbrenner, and Ferdinand Jamitzky. Characterizing synchronization in time series using information measures extracted from symbolic representations. *Phys. Rev. E*, 79:046207, Apr 2009. doi:10.1103/PhysRevE.79.046207.

[67] Karsten Keller, Teresa Mangold, Inga Stolz, and Jenna Werner. Permutation entropy: New ideas and challenges. *Entropy*, 19(3), 2017. doi:10.3390/e19030134.

[68] José M. Amigó, Roberto Monetti, Thomas Aschenbrenner, and Wolfram Bunk. Transcripts: An algebraic approach to coupled time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(1):013105, 2012. doi:10.1063/1.3673238.

[69] W. Bunk, J. M. Amigó, T. Aschenbrenner, and R. Monetti. A new perspective on transcripts by means of their matrix representation. *The European Physical Journal Special Topics*, 222:363–381, 2013. doi:https://doi.org/10.1140/epjst/e2013-01847-6.

[70] Luciano Zunino, Felipe Olivares, Felix Scholkmann, and Osvaldo A. Rosso. Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions. *Physics Letters A*, 381(22):1883–1892, 2017. doi:https://doi.org/10.1016/j.physleta.2017.03.052.

[71] C. Quintero-Quiroz, S. Pigolotti, M. C. Torrent, and C. Masoller. Numerical and experimental study of the effects of noise on the permutation entropy. *New Journal of Physics*, 17(9):093002, sep 2015. doi:10.1088/1367-2630/17/9/093002.

[72] J. M Amigó, S Zambrano, and M. A. F Sanjuán. True and false forbidden patterns in deterministic and random dynamics. *Europhysics Letters*, 79(5):50001, jul 2007. doi:10.1209/0295-5075/79/50001.

[73] M. McCullough, M. Small, H. H. C. Iu, and T. Stemler. Multiscale ordinal network analysis of human cardiac dynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375:20160292, 2017. doi:https://doi.org/10.1098/rsta.2016.0292.

[74] Andriana S. L. O. Campanharo, M. Irmak Sirer, R. Dean Malmgren, Fernando M. Ramos, and Luís A. Nunes. Amaral. Duality between time series and networks. *PLOS ONE*, 6(8):1–13, 08 2011. doi:10.1371/journal.pone.0023378.

[75] J. Zhang, J. Zhou, M. Tang, H. Guo, M. Small, and Y. Zou. Constructing ordinal partition transition networks from multivariate time series. *Scientific Reports*, 7:7795, 2017. doi:https://doi.org/10.1038/s41598-017-08245-x.

[76] F. Olivares, M. Zanin, L. Zunino, and D. G. Pérez. Contrasting chaotic with stochastic dynamics via ordinal transition networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6):063101, 2020. doi:10.1063/1.5142500.

[77] E. Fama. The behavior of stock market prices. *The Journal of Business*, 38, 1965. doi:https://doi.org/10.1086/294743.

[78] A. C. MacKinlay A. W. Lo. *A Non-Random Walk Down Wall Street*. Princeton University Press, 1999.

[79] J. L. McCauley. *Dynamics of Markets, The New Financial Economics*. Cambridge University Press, 2009.

[80] S. Leroy. Risk aversion and the martingale property of stocks rerturns. *International Economic Review*, 14, 1973. doi:https://doi.org/10.2307/2525932.

[81] R. Lucas. Asset prices in an exchange economy. *Econometrika*, 46, 1978. doi:https://doi.org/10.2307/1913837.

[82] Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955. doi:10.2307/1884852.

[83] Andrew W. Lo. The adaptive markets hypothesis. *The Journal of Portfolio Management*, 30(5):15–29, 2004. doi:10.3905/jpm.2004.442611.

[84] Terrence Hendershott, Charles M. Jones, and Albert J. Menkveld. Does algorithmic trading improve liquidity? *The Journal of Finance*, 66(1), 2011. doi:https://doi.org/10.1111/j.1540-6261.2010.01624.x.

[85] Neil Johnson, Guannan Zhao, Eric Hunsader, Jing Meng, Amith Ravindar, Spencer Carran, and Brian Tivnan. Financial black swans driven by ultrafast machine ecology, 2012. doi:10.48550/ARXIV.1202.1448.

[86] J. Doyne Farmer and Spyros Skouras. An ecological perspective on the future of computer trading. *Quantitative Finance*, 13(3), 2013. doi:10.1080/14697688.2012.757636.

[87] A. Kirilenko, A.S. Kyle, M. Samadi, and T. Tuzun. The flash crash: The impact of high-frequency trading on an electronic market. *Journal of Finance, Forthcoming*, 2017. doi:10.2139/ssrn.1686004.

[88] Michael Lewis. How the eggheads cracked. *The New York Times Sunday Magazine*. URL: https://www.nytimes.com/1999/01/24/magazine/how-the-eggheads-cracked.html?pagewanted=all&src=pm.

[89] Mario López and Ricardo Mansilla. Analysis of effciency in high-frequency digital markets using the hurst exponent. *Revista Mexicana de Física*, 67(6):061402, 2021. doi:https://doi.org/10.31349/RevMexFis.67.061402.

[90] Rosario N. Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B*, 11:193–197, 1999. doi:https://doi.org/10.1007/s10051005092.

[91] J.-P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész. Dynamic asset trees and black monday. *Physica A: Statistical Mechanics and its Applications*, 324(1):247–252, 2003. Proceedings of the International Econophysics Conference. doi:https://doi.org/10.1016/S0378-4371(02)01882-4.

[92] Matteo Marsili. Dissecting financial markets: sectors and states. *Quantitative Finance*, 2(4):297–302, 2002. doi:10.1088/1469-7688/2/4/305.

[93] M. C. Münnix, T. Shimada, R. Schäfer, F. Leyvraz, T. H. Seligman, T. Guhr, and H. E. Stanley. Identifying states of a financial market. *Scientific Reports*, 2, 2012. doi:https://doi.org/10.1038/srep00644.

[94] Desislava Chetalova, Rudi Schäfer, and Thomas Guhr. Zooming into market states. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(1):P01029, jan 2015. doi:10.1088/1742-5468/2015/01/p01029.

[95] Anirban Chakraborti, Hrishidev, Kiran Sharma, and Hirdesh K Pharasi. Phase separation and scaling in correlation structures of financial markets. *Journal of Physics: Complexity*, 2(1):015002, nov 2020. doi:10.1088/2632-072x/abbed1.

[96] Federico Musciotto, Jyrki Piilo, and Rosario N. Mantegna. High-frequency trading and networked markets. *Proceedings of the National Academy of Sciences*, 118(26), 2021. doi:10.1073/pnas.2015573118.

[97] M. Kim and H. Sayama. Predicting stock market movements using network science: an information theoretic approach. *Applied Network Science*, 2, 2017. doi:https://doi.org/10.1007/s41109-017-0055-y.

[98] Massimiliano Zanin, Luciano Zunino, Osvaldo A. Rosso, and David Papo. Permutation entropy and its main biomedical and econophysics applications: A review. *Entropy*, 14(8):1553–1577, 2012. doi:10.3390/e14081553.

[99] Mario López Pérez and Ricardo Mansilla. Ordinal synchronization and typical states in high-frequency digital markets. *Physica A: Statistical Mechanics and its Applications*, 598:127331, 2022. preprint: arXiv:2110.07047 [nlin.AO]. doi:https://doi.org/10.1016/j.physa.2022.127331.

[100] Roberto Monetti, Wolfram Bunk, Thomas Aschenbrenner, Stephan Springer, and José M. Amigó. Information directionality in coupled time series using transcripts. *Phys. Rev. E*, 88:022911, Aug 2013. doi:10.1103/PhysRevE.88.022911.

[101] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, 2003. doi:10.1109/TIT.2003.813506.

[102] Arthur A. B. Pessa and Haroldo V. Ribeiro. ordpy: A python package for data analysis with permutation entropy and ordinal network methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(6):063110, 2021. doi:10.1063/5.0049901.

[103] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[104] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane,

Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature*, 585:357–362, 2020. doi:https://doi.org/10.1038/s41586-020-2649-2.

[105] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013. doi:https://doi.org/10.1016/j.patcog.2012.07.021.

[106] Ulrike von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or art? In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, editors, *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 65–79, Bellevue, Washington, USA, 02 Jul 2012. PMLR.