Analysis of short term price trends in daily stock-market index data

H.R Olivares[1] Sánchez[1], H.F. Coronel-Brizio[1], E Scalas[2], A R Hernández Montoya[1], C.M. Rodríguez-Martínez[1]

[1]Facultad de Física. Universidad Veracruzana, Apdo. Postal 475. Xalapa, Veracruz. México.
[2]Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale "Amedeo Avogadro", Viale T. Michel 11, 15121 Alessandria, Italy. BCAM-Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Basque Contry, Spain.

E-mail: hectorolivares100@gmail.com, alhernandez@uv.mx

Abstract. In financial time series there are periods in which the value increases or decreases monotonically. We call those periods *elemental trends* and study the probability distribution of their duration for the indices DJIA, NASDAQ and IPC. It is found that the trend duration distribution often differs from the one expected under no memory. The expected and observed distributions are compared by means of the Anderson-Darling test.

1. Introduction

One of the goals of financial-market analysis is to predict the future movements of prices and financial indices. In order to achieve this goal, a huge variety of methods to forecast markets behavior were developed, ranging from complex mathematical models even to astrological pseudo-scientific techniques. An approach that has been recently growing in popularity is the statistical analysis of large sets of data, which has become now possible due to the increasing availability of computer power and high quality data sets. This approach has benefited from the contributions not only from economists, but also from many physicists and mathematicians who have applied methods and ideas of probability theory and statistical physics to finance. A set of nontrivial statistical properties of historical data was observed and classified as "stylized facts" [3], which are expected to provide a better insight on markets structure and behavior.
When observing the time series of the prices of an asset on a chart, it is common to see "trends" in which most of the values are greater (or smaller) than the previous ones. These trends are very popular within the so called *technical analysis*. Trends as those studied by technical analysis can be seen as composed by smaller elemental trends, periods in which the value increases or decreases monotonically. These kind of trends are the ones that will be studied in the present work. Among other things, technical analysts seek patterns in the charts of financial data, that are believed to be indicators of changes in the trend direction. The effectiveness of technical analysis is disputed and put at a stake by what is known as the Efficient Market Hypothesis. Before going further, it is necessary to give some definitions. In Subsections 1.1, 1.2 and 1.3 of this introduction these definitions and other useful information will be presented. In Section 2, a model for the distribution of trends durations will be developed from the efficient

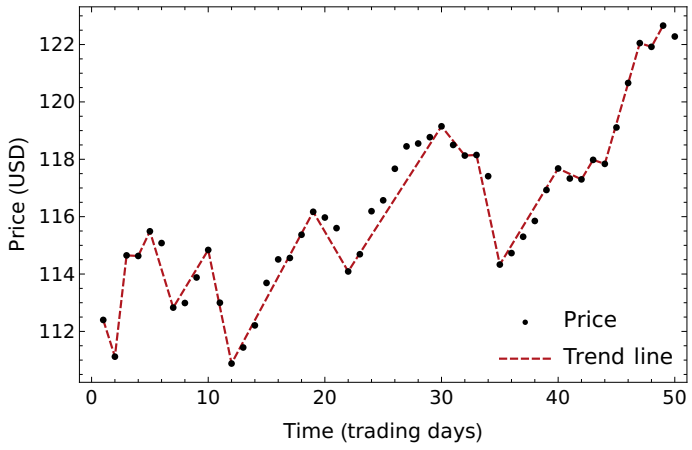

Figure 1: The line segments join the starting and ending points of each elemental trend.

market hypothesis. Section 3 will explain how the data were analyzed and section 4 provides an interpretation of the analysis.

1.1. Definitions

Let $S(t)$ be the price of an asset or an index value at time t and $X(t) = \log S(t)$ its logarithm. The log-return at time t is defined as:

$$r(t, \Delta t) = X(t + \Delta t) - X(t) \tag{1}$$

for a given time sampling scale Δt. If the price variation is small, the log-return is a good approximation of the return

$$R(t, \Delta t) = \frac{S(t + \Delta t) - S(t)}{S(t)}. \tag{2}$$

In this paper, we consider Δt equal to 1 day and we use the values of the indices corresponding to the close value further in the investigated markets. More details on the data set will be given in section 3.

An *elemental trend* of duration k will be defined here as a subseries of $k + 1$ values within the series $S(t)$ in which every value is greater (for an uptrend) or smaller or equal (for a downtrend) than the preceding one (Figure 1). The aim of this work is to study with a statistical approach the kind of short term trends defined above.

1.2. The Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) claims that the market quickly finds the rational price for a traded asset [14]. The most important consequence of this hypothesis was shown by P. Samuelson [16] and it is the fact that the best forecast for the future price of an asset is its present price.

$$\mathbb{E}(S(t + \Delta t)|\mathcal{F}_t) = S(t), \tag{3}$$

where $\mathbb{E}(\cdot|\mathcal{F}_t)$ is the conditional expectation with respect to the filtration $\mathcal{F}_t$, namely with respect to the known history up to time t. Indeed, it is easy to derive the EMH from a simple no-arbitrage argument. Suppose we have two assets, a risky one, with price $S(t)$ and a risk-free one giving a constant interest rate r_F. To avoid arbitrage, one has to require that the expected return of the risky asset is equal to the risk-free interes rate, that is

$$\mathbb{E}(R(t, \Delta t)|\mathcal{F}_t) = r_F; \tag{4}$$

the latter equation immediately yields, for non vanishing $S(t)$,

$$\mathbb{E}(S(t+\Delta t)|\mathcal{F}_t) = (1+r_F)S(t), \tag{5}$$

which reduces to (3) for $r_F = 0$. Equations (3) and (5), jointly with the integrability of the process $S(t)$, are known as martingale and sub-martingale (remember that $r_F \geq 0$) conditions, respectively.
The EMH would invalidate the attempts of technical analysis to predict future prices or trends; in fact, in Samuelson's words, "there is no way of making an expected profit by extrapolating past changes in the futures price, by chart or any esoteric devices of magic or mathematics" [16] as the best forecast of the future price would be the current price.

1.3. Stylized facts
As mentioned before, financial time series share some nontrivial statistical properties called stylized facts. Although those properties are often formulated qualitatively, they are so constraining that it is difficult to reproduce all of them by means of a stochastic process[3]. As a matter of fact, none of the market models, including analytical models, Monte Carlo simulations and multi-agent based models, created before 1990, when awareness of such regularities gradually started to appear, could reproduce all of these stylized facts [15]. As an interesting issue, some studies suggest that stylized facts appear not only in financial time series, but also in other complex systems such as Conway's Game of Life [5]. To fix the ideas, some of the stylized facts, taken from reference [3], are listed below:

Absence of linear autocorrelations: Autocorrelations of returns are often negligible, except for very small time scales, depending on the market and on the time horizon.

Heavy tails: The return distribution is leptokurtic and some authors claim that the tails decay as a power-law.

Gain-loss asymmetry: Large downward jumps in stock prices and stock index values are observed, but not equally large upward movements. (In exchange rates there is a higher symmetry in up/down movements).

Volatility clustering: High volatility events do cluster in time.

2. An 'Efficient Market' model for the duration distribution

Among all the possible martingale or sub-martingale models that can describe price fluctuations, the geometric random walk is the simplest one. A geometric random walk is just a product of independent and identically distributed positive random variables. If the expected value of these variables is 1, then the geometric random walk is a martingale; otherwise, if the expected value is larger than 1, the geometric random walk is a submartingale. However, the geometric random walk hypothesis is neither necessary nor sufficient for an efficient market, as shown by many authors among whom Leroy [6], Lucas [7] and Lo and Mckinlay [8]. To understand this point, it is enough to consider Equation (5) allowing for any martingale model.
At each step of a series of index values, there are two possible outcomes: the index either increases or does not increase. In an efficient market, the expected future price depends only on information about the current price, not on its previous history. Therefore, it should be impossible to predict the expected direction of a future price change given the history of the price process. In formula, from Equation (3) (after discounting for the risk-free rate), we have

$$\mathbb{E}(S(t+\Delta t) - S(t)|\mathcal{F}_t) = 0; \tag{6}$$

if we consider the sign of the price change $Y(t, \Delta t) = \text{sign}(S(t + \Delta t) - S(t))$, which coincides with the sign of returns, we accordingly have

$$\mathbb{E}(Y(t, \Delta t)) = 0. \tag{7}$$

If the price follows a geometric random walk, then the series of price-change signs can be modeled as a Bernoulli process. This process could be biased to take the presence of a risk free interest rate into account. To be more specific, let us consider a log-normal geometric random walk and let us use the assumption $\Delta t = 1$. Let S_0 be the initial price. The price at time t will be given by

$$S(t) = S_0 \prod_{i=1}^{t} Q_i \tag{8}$$

where Q_i are independent and identically distributed random variables following a log-normal distribution with parameters μ and σ. These two parameters come from the corresponding normal distribution for log-returns. As a direct consequence of the EMH in the form (5), we have

$$\mathbb{E}(Q) = 1 + r_F, \tag{9}$$

and for a log-normal distributed random variable, we have also

$$\mathbb{E}(Q) = \mathrm{e}^{\mu} \mathrm{e}^{\sigma^2/2}. \tag{10}$$

This leads to a dependence between the two parameters

$$\mu = \log(1 + r_F) - \frac{\sigma^2}{2}. \tag{11}$$

Note that, when $r_F = 0$, it is impossible to get $\mu = 0$. This reflects a more general result, if the price process is a martingale, the log-price process cannot be a martingale and viceversa. Starting from the cumulative distribution function for a log-normal random variable

$$F_Q(u) = \mathbb{P}(Q \leq u) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\log(u) - \mu}{\sqrt{2\sigma^2}}\right), \tag{12}$$

the probability of a negative sign would be given by

$$q = F_Q(1) = \mathbb{P}(Q \leq 1) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\sigma}{2\sqrt{2}} - \frac{\log(1 + r_F)}{\sigma\sqrt{2}}\right), \tag{13}$$

which yields $q = 1/2$ for $r_F = \mathrm{e}^{\sigma^2/2} - 1$.
It becomes natural to use the biased Bernoulli process as the null hypothesis for the time series of signs [9]. It is well known that the distribution of the number k of failures needed to get one success for a Bernoulli process with success probability $p = 1 - q$ is the geometric distribution $\mathcal{G}(p)$; the number of failures N is given by

$$P(k) = \mathbb{P}(N = k) = p(1 - p)^k = pq^k. \tag{14}$$

The duration of a elemental downward trend in daily data is the number of days before the price increases, so the distribution of such trend durations should follow a geometric distribution. An identical argument applies to the duration of a upward trend. Such sequences of identical outcomes are also known as *runs* or *clumps* in the mathematical literature.

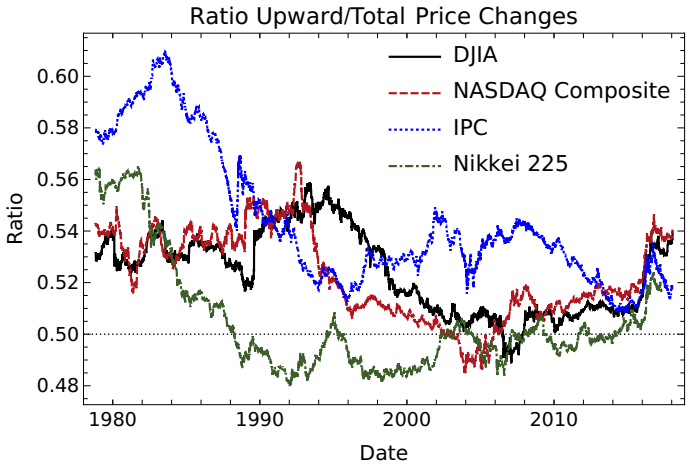

Figure 2: Ratio of upward to total price changes in daily data, plotted against time for the interval from 1978-10-30 to 2018-01-23, calculated over a time window of 1000 trading days.

3. Methodology

3.1. Data

Three indices were analyzed, namely Dow Jones Industrial Average (DJIA), NASDAQ Composite, the Mexican Índice de Precios y Cotizaciones (IPC) and Nikkei 225 during the period 30 October 1978 - 23 January 2018. The data was obtained from Yahoo Finance.

3.2. Building the sample

For each series of index values, several time windows of 252 trading days (which in average represent one trading year) were created, each one shifted forward by one trading day with respect to the previous one. This procedure resulted in 32102 time windows for the DJIA, 11569 for the NASDAQ and 5885 for the IPC. Two histograms were built for each time window, one for upward trends and the other one for downward trends. For every sample in the time windows, the number of days before every sign change is measured. For instance, given the sequence $--++++++-$, the recorded times will be $(2, 5, 1)$. Finally, all the histograms were normalized. Examples of both uptrend and downtrend histograms for different time windows are shown in Figure 4.

3.3. The Anderson-Darling goodness of fit test

In order to compare the observed and expected distributions of trend durations, the Anderson-Darling test described in reference [1] was used. The Anderson-Darling test was found to be the most suitable for this purpose because it places more weight on the tails of a distribution than other goodness of fit tests. The critical values of the Anderson-Darling statistic A_n^2 were dependent on the parameter of the geometric distribution which is chosen for every window as the ratio of upward to total price changes. This is necessary to take into account the possible fluctuations of the parameter p of the Bernoulli process. For a more complete discussion on Anderson-Darling Methodology see appendix 5.

4. Discussion

In Figure 2, the ratio of the upward to total price changes in daily data is plotted against time for the years 1978 - 2017. This ratio is calculated over a time window of 1000 trading days. It can be seen that variations are greater than those expected for the same time windows in

a Bernoulli process with parameter $p = 1/2$ (± 0.05), but it might be interesting to find out whether the hypothesis of a geometric distribution holds for smaller periods (such as each of the 1000 days time windows individually), because it would mean that in those periods the direction of price changes was not predictable using historical prices of the index. Plots in figure 4 show the distribution of trend durations corresponding to different indices and periods of 1000 days. Figure 3 display the p-values of the Anderson-Darling statistic for different periods. In order to avoid confusion between the parameter of the geometric distribution and the p-values for the distribution of A_n^2, the latter will be referred to as π-values. The meaning of π-values is the probability of obtaining a value of A_n^2 at least as big as the one that was really obtained, given that the probability distribution is actually geometric.

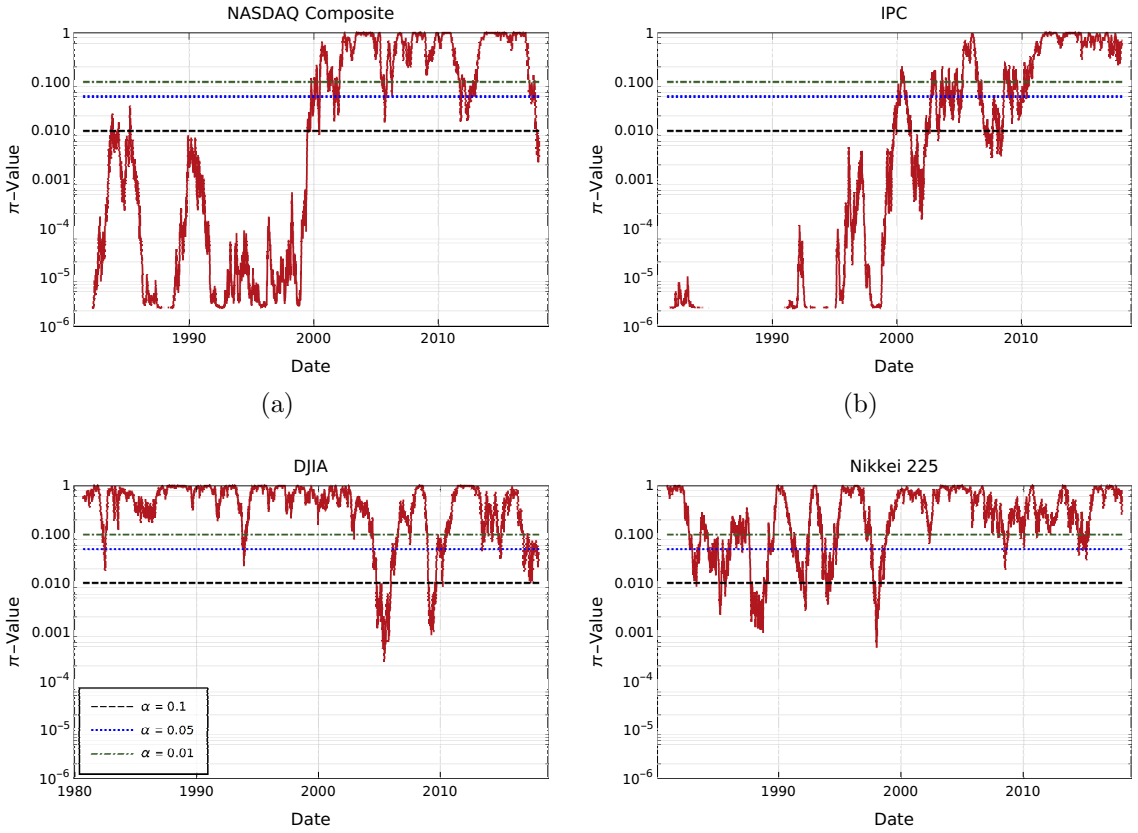

Figure 3: (Color online) 3a: π-values of the Anderson-Darling statistic for the NASDAQ plotted against time. As time passes, the data agree better with the geometric distribution; 3b: π-values of the Anderson-Darling statistic for the IPC plotted against time. As for NASDAQ, agreement between data and the geometric distribution increases with time; 3c: π-values of the Anderson-Darling statistic for the DJIA plotted against time. The greatest deviations from the geometric distribution occurred between the years 2000-2011.

It was observed that as time passes, the direction of price changes for the IPC and the Nasdaq is better described by a geometric distribution (Figures 3a and 3b). The distribution of trend durations for the Dow Jones is generally reasonably well fitted by the geometric distribution. This fact can be interpreted as a possible evidence that the Mexican stock market (that has become public and regulated since 1975) has been increasing its efficiency, as reported by previous research [4]. The same claim can be made about the NASDAQ, given that it is also a market of relatively recent creation. In contrast, the Dow Jones Industrial Average index represents a more

mature market. However, there is also evidence that the New York Stock Exchange, represented by the Dow Jones, has swiftly increased its efficiency between the beginning of the 1980s and the end of the 1990s [18]. Figure 3c shows that for the Dow Jones, the greatest deviations from the geometric distribution in the studied period (almost the whole XXth Century) occurred between the years 2000-2011.

5. Conclusions

The probability distributions for the duration of elemental trends were studied for the market indices Dow Jones Industrial Average (DJIA), NASDAQ Composite and for the Mexican Índice de Precios y Cotizaciones (IPC). These distributions are expected to be geometric and memoryless according to the discussion in section 2. The IPC and the NASDAQ present periods in which the memoryless hypothesis must be definitively rejected.

Appendices

The discrete version of the Anderson-Darling goodness of fit test

Determining whether or not a given probability model fits the observed data that it attempts to describe is one of the most important problems of applied statistics.

Even though extensive research has been done regarding this problem, most of it deals with continuous distributions. Unfortunately, the studies done for fitting discrete distributions are much less numerous, or at least harder to find.

Discrete distributions, however, are important in many fields as medicine, psychology and engineering. Scientific papers and textbooks often prescribe the chi-square test as the option to use when testing goodness of fit for these distributions.

Nevertheless, it is well known that chi-square tests suffer from low power, especially when applied to data in which there are bins with very small content (as a single event), and when the expected distribution predicts a very low probability for one of the categories. Data drawn from a process that can be described by a geometric distribution precisely exhibits both features.

Bracquemond *et. al.* (2002) [21] make a review of eight alternatives to the chi-square test proposed over the past years and perform a simulation-based comparative study specifically for the geometric distribution. It consists first on checking the empirical significance level against the nominal one for each test, and then performing a power study. They analyzed three tests based on the empirical distribution function, three based on the empirical generating function, the Neyman smooth test, and a test by Nikulin (1992) [23] based on the generalized Smirnov transformation.

The tests that had an overall better performance were the Baringhaus-Henze (BH) test, the Anderson-Darling (AD) test and Nikulin's test. Among these, Nikulin's test was considered to have a satisfying power, but they recommend not to use it for small data. The two other tests have the disadvantage of requiring a numerical procedure called parametric bootstrap, that is relatively expensive computationally speaking. However, since the BH test involves by far many more operations than the AD test, we preferred to use the latter for our analysis. The computer power required to carry out the AD tests in this work is reasonable with current technology. The code we made to perform the test gives the results within about one second on an ordinary laptop.

The AD test belongs to a family of goodness of fit tests called the Cramér-von Mises tests, which includes the Anderson-Darling test, Watson's test and the Cramér-von Mises test itself.

The family was originally developed to test continuous distributions, but a generalization for discrete distributions appeared for the first time in an article by Choulakian *et.al.* [22].

The principle behind this kind of tests is defining a statistic that serves to measure the distance between a theoretical distribution function $F_0(k)$ and the empirical (cumulative) distribution function for n events, $\mathbf{F}_n(k)$. Every value of the statistic is associated with a p-value, that can be interpreted as the probability of obtaining a value of the statistic at least as large as the one obtained, given that the null hypothesis

$$\mathcal{H}_0 : \mathbf{F}_n(k) = F_0 \tag{15}$$

is true. If the p-value is smaller than a previously defined threshold value α, the null hypothesis is rejected.
For the case of the discrete Anderson-Darling test, this statistic is the *Anderson-Darling statistic*:

$$A_n^2 = n \sum_{k=1}^{\infty} \frac{[\mathbf{F}_n(k) - F_0(k)]^2 p_0(k)}{F_0(k)(1 - F_0(k))}, \tag{16}$$

where $p_0 = F_0(k) - F_0(k-1)$.
If instead what is being tested is whether the observed data comes from a distribution belonging to a parametric family $F(\cdot;\theta)$, then the parameter θ must be estimated first.
For the case of the geometric distribution $\mathcal{G}(p)$, this is an additional complication, since the distribution of A_n^2, and therefore the correspondence between it and the p-values, depends both on n and on the parameter p.
In order to get around this problem, a numerical technique called parametric bootstrap is used. First, p is estimated using the maximum-likelihood estimator

$$\hat{p}_n = \frac{n}{\sum_{i=1}^{n} K_i}, \tag{17}$$

where K_i are the values of the random variable in the sample. Then, a large number of copies of the sample is generated and filled with random numbers taken from the actual geometric distribution $\mathcal{G}(\hat{p})$, and the AD statistic is calculated for every one of them. The distribution of the statistic found from the samples can be then integrated up to the value of A_n^2 calculated from the empirical data, in order to find the p-value for our case of interest. For the parametric bootstrap, we used 500 copies of the sample, which was the same number used by Bracquemond *et. al.* [21] for their tests. As our random number generator, we used the routine *ran2* from *Numerical Recipes in C* [24], which breaks serial correlations to a great extent and has a period of $\approx 2.3 \times 10^{18}$.

Acknowledgements
This work was supported by Conacyt-Mexico and MAE-Italy under grant 146498. We also thank Conacyt-Mexico for the support under project grant 155492, Universidad Veracruzana under project 41504 and PRIN 2009 Italian grant "Finitary and non-finitary probabilistic methods in Economics".

References

[1] Bracquemond C, Crétois E and Gaudoin O 2002 'A comparative study of goodness-of-fit tests for the geometric distribution and application to discrete time reliability' *Preprint* http://www-ljk.imag.fr/SMS/preprints.html.
[2] Chen M K, Lakshminarayanan V and Santos L R 2006 'How Basic Are Behavioral Biases? Evidence from Capuchin Monkey Trading Behavior' *J. Political Econ.* **144** 517–537
[3] Cont R 2001 Empirical properties of asset returns: stylized facts and statistical issues' *Quantitative Finance* **1** 223–236

[4] Coronel-Brizio H F, Hernández-Montoya A R, Huerta-Quintanilla R, and Rodríguez-Achach M E 2007 'Evidence of increment of efficiency of the Mexican Stock Market through the analysis of its variations', *Physica A* (380) 391–398
[5] Hernández Montoya A R, Coronel-Brizio H F, Rodríguez-Achach M E, Stevens-Ramírez G A, Politi M, and Scalas E 2011 'Emerging properties of financial time series in the Game of Life' *Phys. Rev.* E **84** 066104
[6] Leroy S F 1973, 'Risk aversion and the martingale property of stock prices' *International Economic Review* **14** 436–446.
[7] Lucas R E 1978, 'Asset prices in an exchange economy' *Econometrica* **46** 1429–1445.
[8] Lo A W and Mackinlay A C 1999, *A Non-Random Walk Down Wall Street*, (Princeton: Princeton University Press).
[9] Scalas E 1998, 'Scaling in the market of futures', *Physica* A **253** 394–402.
[10] Hołyst J A and Sieczka P 2008 'Statistical properties of short term price trends in high frequency stock market data' *Physica* A **387** 1218–1224
[11] Instituto Nacional de Estadística, Geogrfía e Informática. Índice Nacional de Precios al Consumidor. (2011, 5 de noviembre). http://www.inegi.org.mx/est/contenidos/proyectos/inp/inpc.aspx.
[12] Jensen M H, Johansen A and Simonsen I 2002 'Inverse Statistics in Economic: The gain-loss asymmetry', *Physica* A **324** 6.
[13] Levy M, Levy H and Solomon S 2000 *Microsimulations of Financial Markets: From Investor Behavior to Market Phenomena* (London: Academic Press)
[14] Mantegna R N and Stanley H E 2000 *An Introduction to Econophysics.* (Cambridge: University Press)
[15] Samanidou E, Zschischang E, Stauffer D and Lux T 2006 'Microscopic Models of Financial Markets'. Economics Working Paper 2006-15. Department of Economics, University of Kiel.
[16] Samuelson P A 1965 'Proof that Properly Anticipated Prices Fluctuate Randomly', *Industrial Management Rev.* **6** 41–45.
[17] Samuelson P A and Nordhaus W D 2005 *Economics* (New York: McGraw-Hill)
[18] Totha B and Kerteszb J 'Increasing market efficiency: Evolution of cross-correlations of stock returns' *Physica* A **320** 505–515
[19] U. S. Bureau of Labor Statistics, Consumer Price Index. (2011, 5 de noviembre). http://data.bls.gov/timeseries/CUSR0000SA0?output_view=pct_1mth.
[20] Voit J 2003 *The Statistical Mechanics of Financial Market.* (New York: Springer-Verlag)
[21] Bracquemond, C., Crétois, E., Gaudoin O., (2002) "A comparative study of goodness-of-fit tests for the geometric distribution and application to discrete-time reliability" Technical Report, URL: www-ljk.imag.fr/SMS/ftp/BraCreGau02.pdf.
[22] Choulakian, V., Lockhart, R. A., Stephens, M. A., (1994) *The Canadian Journal of Statistics*, 22, 1, pp. 125-137.
[23] Nikulin, M. S. (1992), *C. R. Math. Rep. Acad. Sci. Canada*, 14, 4, pp. 151-156.
[24] Press, W. H., Teukolsky, S. A., Vetterling, W. T., *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 2002 pp. 281-282.

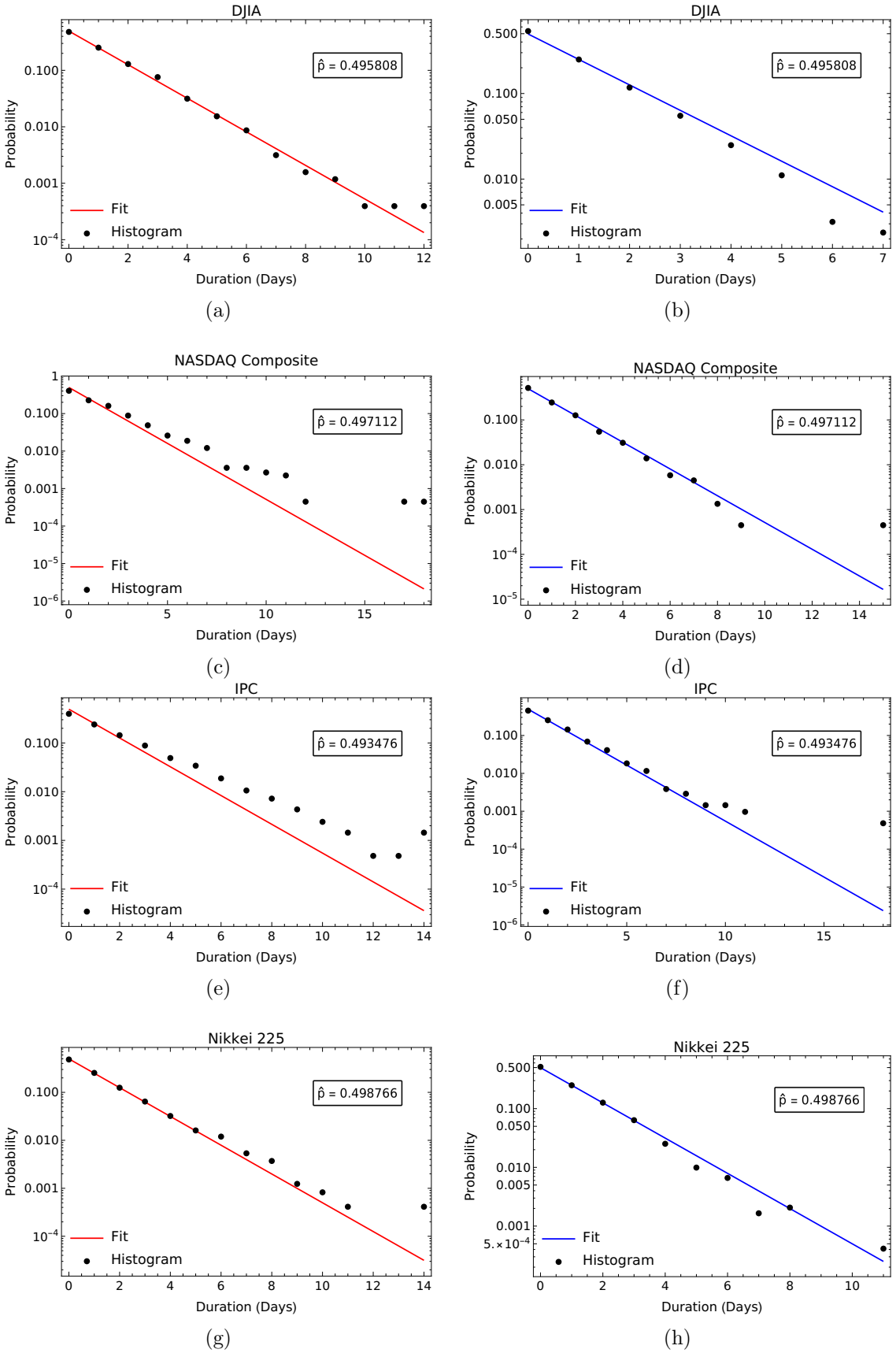

Figure 4: Distribution of trend durations for different indices and a time windows of 1000 days. (a), (c), (e) and (g) present the upward distribution and (b), (d), (f) and (h) the downward distribution. The solid lines are the expected geometric distributions. The parameters p and q are shown on each figure.