Measuring market efficiency: The Shannon entropy of high-frequency financial time series

Andrey Shternshis *, Piero Mazzarisi, Stefano Marmi

Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa, Italy

ABSTRACT

When prices reflect all available information, the price dynamics is a martingale and the market is said to be efficient. However, much empirical evidence supports the conclusion about the inefficiency of financial markets, especially at high-frequency timescales. We investigate the sources and dynamics of the inefficiency of the ETF market at a 1 min timescale by proposing a computational methodology for a genuine estimation of the Shannon entropy. Since several sources of regularity lead to the detection of apparent inefficiencies, we build a multi-step filtering method, which allows (i) to remove the seasonality of volatility and heteroskedasticity, (ii) to detect and remove spurious effects due to price staleness, and (iii) to filter out microstructure noise. We corroborate our findings with an extensive analysis of the ETF market. We conclude that, after removing all known patterns of regularity, the market is not efficient at a one-minute time scale and on a weekly basis; however, the signal is weak.

Keywords:
Shannon entropy
Market efficiency
Zero returns
Price staleness
High frequency

1. Introduction

The fundamental price of a stock is a quantitative way to assess the intrinsic value of a company. In principle, complete information about a company permits us to know its fair price. When a company is quoted on a stock exchange, the market price of the stock is instead the result of a highly complex process of matching between the supply and demand of traders. In a market with complete information at each time, the matching of supply and demand *should* incorporate all information in the market price. Thus, the *best* forecast is the current observation and the price dynamics is a martingale. Very short, this is what is known as the *Efficient Market Hypothesis* (EMH) [1]. When this hypothesis is verified, a market is called *efficient*. However, the definition of an information set which is complete, i.e., including any variable having an impact on price, is usually unfeasible, especially for a quantitative approach. For this reason, it is preferred to work with a weak form of the EMH, that is, the information set is assumed to include only the past observations of the price dynamics [2].

As stated in [3], the hypothesis rejection of a martingale model suggests the existence of trading rules increasing the expected return of some actively managed portfolio with respect to a simple buy-and-hold strategy. In other words, forecasting patterns of price dynamics with a given level of certainty allows devising trading strategies with a positive profit on average.[1] If so, the Efficient Market Hypothesis is not verified and the market is said *inefficient*.

Different approaches have been proposed to test *market inefficiency*, all of them are with the common rationale of measuring how much an empirical price dynamics is far from the assumption of complete randomness (martingale hypothesis). A time varying autocorrelation of stock returns has been proposed as a measure of the degree of market inefficiency for the U.S. stock market [4]. The R/S statistics and the Hurst exponent have been used to rank the efficiency of emerging equity markets [5,6]. The algorithmic complexity of return time series has been applied as a measure of the relative efficiency of financial markets [7]. The algorithmic complexity has been used to check the Efficient Market Hypothesis [8]. Finally, the approximate entropy (ApEn) [9] has been proposed as a measure of the degree of market efficiency over time and for different markets [10–13].

A well-known measure of randomness for symbolic dynamics is the Shannon entropy. It represents the average amount of uncertainty removed with the transmission of each symbol. In the case of financial time series, price dynamics can be opportunely discretized and the Shannon entropy can be computed over the resulting sequence of symbols. This approach has been considered, for example, in [14] to evaluate predictability of financial time series. Many measures and methods based on the definition of the Shannon entropy have been proposed in

* Corresponding author.
E-mail addresses: andrey.shternshis@sns.it (A. Shternshis), piero.mazzarisi@sns.it (P. Mazzarisi), stefano.marmi@sns.it (S. Marmi).

[1] Within this context, we neglect the role played by trading costs.

https://doi.org/10.1016/j.chaos.2022.112403 .

recent years with the common goal of studying market efficiency. Risso has studied the Shannon entropy as a measure of efficiency for twenty markets, comparing emerging markets with the developed ones [15]. A time-varying entropy of crude oil market efficiency has been studied in [16]. Oh et al. [17,18] have connected the Shannon entropy with the probability of having market crashes and financial crises. The entropy calculated for energy markets has been associated with historical events and climatic factors in [19]. Ahn et al. [20] have used the entropy to state that the degree of market inefficiency in the Chinese stock market has a strong effect on the economic fundamentals. The relationship between the entropy and risk measures has been investigated in [21]. The Shannon entropy as a measure of the risk of a portfolio has been considered also in [22]. The conditional entropy has been used in [23] to measure the randomness in stock and exchange markets at different time scales. A generalization of the Shannon entropy, the Tsallis entropy [24], has been proposed as a risk measure during financial crises and crashes [25,26]. The permutation transition entropy has been introduced in [27] to measure the complexity of financial time series. Marschinski and Kantz [28] and Kwon and Yang [29] have studied the transfer entropy introduced in [30] to investigate the strength and the direction of the information transfer in the U.S. stock market. Finally, an entropy measure has been used to identify different types of trading behaviors based on historical prices and news in [31].

A naive computation of the Shannon entropy for opportunely discretized price dynamics is not, however, the end of the story. There are well-known regularity patterns in financial time series, for instance, daily seasonality or volatility clustering. When not filtered out, such patterns tend to decrease any measure of randomness; nevertheless, no profitable strategies can be built upon them. Thus, there is a need for devising a computational method for the evaluation of the Shannon entropy that takes into account such regularity patterns. The first study in this direction has been presented in [32], where volatility clustering, intraday seasonality, and microstructure noise are filtered out before the computation of the Shannon entropy as a measure of efficiency for the Exchange Traded Funds (ETF).

Here, we propose a computational methodology for entropy estimation, by accounting for many patterns of regularities in high-frequency financial time series, in particular including price staleness [33,34], as well as considering the evolution of entropy in time and at different time scales. Finally, the genuine estimation of the Shannon entropy is used to determine the degree of randomness of the time series of price returns.

More specifically, we start from the method introduced in [35] for the detection of outliers, then removing splits and merges. After that, we remove step by step both daily seasonality and volatility patterns. Then, we study the effect of price staleness on entropy estimation. In particular, the presence of persistent 0-returns in a row because of the lack of price adjustments or very small trading volumes tends to decrease any estimate of entropy: persistent 0-returns are converted into a persistent sequence of the same symbol with the effect of larger predictability associated with such a persistence pattern. Nevertheless, no trading strategy is able to exploit it. In fact, the presence of price staleness is possible because of very low liquidity or any trading order to be executed should go deep in the limit-order book, thus destroying the persistence pattern of 0-returns. We first show empirically that price staleness tends to decrease the estimate of entropy. Then, we build a method for filtering out 0-returns associated with price staleness and apply it to the Exchange Traded Funds market. We show that the number of times we detect that the market is inefficient decreases significantly after filtering out 0-returns due to price staleness. Finally, in the last step of the methodology, we study the effect of microstructure noise. After filtering out all mentioned sources of the apparent inefficiencies, it is possible to conclude that the ETF market is *not* efficient at a high frequency (1-min) on weekly, monthly, and quarterly time intervals; however, the signal of market inefficiency is weak.

The paper is organized as follows. We review the data handling process and the computation of the Shannon entropy in Section 2. In Section 3, we present the method of filtering 0-returns and apply it to simulated and real data. The analysis of the efficiency of the ETF market after filtering out microstructure noise is in Section 4. We present additional results of testing the Efficient Market Hypothesis in Section 5. Section 6 concludes the paper.

2. Data handling and the estimation of entropy

2.1. Financial datasets and data handling

We consider two high frequency datasets of return time series. The first one contains the time series of the prices of the 100 most liquid stocks belonging to the Russell 3000 Index from 02.01.1998 to 23.06.2017. The second dataset contains the time series of the prices of ETFs from 02.01.2003 to 01.12.2009. We consider 1-min closing price data during a regular US trading session, from 9:30 to 16:00. The choice of high liquidity is thus motivated by the need to consider stocks which are traded very frequently, in such a way that the price dynamics are observed at a 1 min time scale. If there is no trading at some specific minute, the missing value for the price is reconstructed as the last price available. By using this method, each trading day contains 390 data points. The tickers of ETFs and stocks are listed in the Appendix A.

Before applying our methodology for the entropy estimation, we perform a data handling process. We remove outliers, interpreted as values in the dataset with no economic sense, and splits. Then, well-known sources of the regularities in prices are filtered out, e.g., seasonality and volatility patterns [36–38], in order to focus on the hidden sources of market inefficiency.

The data handling process is in four steps.

Step 1. Removing outlier values from the dataset;

Step 2. Detecting possible splits where the return is >0.2. We delete such returns from the dataset;

Step 3. Filtering out the daily seasonalities;

Step 4. Filtering out the heteroskedasticity.

Steps 1 and 2 represent a data cleaning process. Steps 3 and 4 consist in filtering out the *apparent inefficiencies* presented in the data. The whitening procedure starts with removing the intraday volatility pattern getting deseasonalized returns and continues with removing the long memory contribution to returns due to volatility getting standardized returns. The details on each step are in the Appendix B. The approach of volatility estimation is described in Section 2.1.1.

We consider sub-intervals in order to detect the presence of market inefficiency in a given time interval. We concentrate on weekly non-overlapping intervals consisting of 5 working days. Weekly time interval consists of $5 \cdot 390 = 1950$ data points. If we detect the presence of inefficiency in a particular week, as described in Section 2.2.5, we will refer to such a week as an *inefficient week* or a *week with inefficiency*. Dividing a time series into short intervals also helps to measure the degree of market inefficiency. We calculate it as the percentage of inefficient weeks for the considered set of assets in the market. If the percentage is less or equal to 1 %, the level of significance for testing the EMH, we will interpret it as a perfect randomness of prices in the ETF market.

2.1.1. Volatility estimation

Volatility clustering refers to the fact that large returns tend to be followed by other large returns of either sign, and vice versa for small returns. The volatility clustering needs to be filtered out by re-scaling each observation by the estimated value of the volatility at that time.

For a reason that will be clear below, we choose an algorithm for volatility estimation in the case of missing observations [39]. It is based on the Expectation-Maximization algorithm (EM) [40], but the values of missing squared returns, r^2, are updated after each step of the numerical maximization of a likelihood function: The volatility is assumed to follow a GARCH(1,1) model.

$$\sigma^2(t) = \mu_0 + \alpha\sigma^2(t-1) + \beta r^2(t-1)$$

and the estimation of parameters $\theta = \{\mu_0, \alpha, \beta\}$ is obtained by using the following 4-steps algorithm.[2]

1. Choose initial values of θ and calculate $\sigma^2(\theta)$;
2. Estimate the missing values as $E[r_t^2] = \sigma_t^2$;
3. Using the maximum likelihood estimation, find new values of θ and, hence, new estimation of volatility;
4. Continue steps 2 and 3 until stopping criteria are satisfied.

2.2. The computation of the Shannon entropy

The unpredictability of asset returns in an efficient market implies maximum uncertainty, which can be captured by an entropy measure. The entropy attains its own maximum under the EMH hypothesis. A measure significantly smaller needs to be intended as a signal of market inefficiency.

2.2.1. The Shannon entropy

We consider the Shannon entropy computed over sequences of random variables. The Shannon entropy is defined as the average amount of information that the source transmits with each symbol [41]. The uncertainty of transmission is proportional to the expected value of the logarithm of the probability of receiving a sequence of symbols.

Definition 1. Let $X = \{X_1, X_2, \ldots\}$ be a stationary random process with a finite alphabet $A$and a measure μ. An n-th order entropy of X is

$$H_n(\mu) = -\sum_{x_1^n \in A^n} \mu(x_1^n) \log \mu(x_1^n)$$

with the convention $0 \log 0 = 0$. A process entropy (entropy rate) of X is

$$h(\mu) = \lim_{n\to\infty} \frac{H_n(\mu)}{n}.$$

We set the base of the logarithm to be equal to the size of the alphabet A.

2.2.2. Discretization

The Shannon entropy is defined over a finite alphabet. Prices move on a discrete grid and the minimum price variation is bounded by a tick size. However, the huge amount of possible discrete variations combined with the absence of an upper bound for them makes the computation of entropy infeasible in practical applications. Hence, we build a coarse-grained grid in such a way that the patterns of price variations have a more direct interpretation: "the price goes up", "the price is stationary", or "the price goes down". More specifically, we consider 2-symbols and 3-symbols alphabets.

We define price returns as $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$, where P_t is the price at time t and ln() is the natural logarithm. For the binary alphabet, we distinguish positive returns from negative returns.

$$\mathrm{s_t} = \begin{cases} 0, \mathrm{r_t} < 0, \\ 1, \mathrm{r_t} > 0 \end{cases}$$

The case $r_t = 0$ is not considered and is removed from the sequence of symbols. The sub-samples of the sequence splitted by the presence of 0-returns are then concatenated. This type of discretization is invariant to any multiplicative factor. In particular, the volatility of returns in a given period is not important for entropy computation.

The ternary alphabet is obtained by labelling returns according to two tertiles of the empirical distribution of returns, namely

$$s_t = \begin{cases} 1, r_t \le \theta_1, \\ 0, \theta_1 < r_t \le \theta_2, \\ 2, \theta_2 < r_t, \end{cases}$$

where θ_1 and θ_2 denote the two tertiles of the empirical distribution of the time series $r_t = \{r_1, \ldots, r_N\}$. In other words, θ_1 and θ_2 divide the sorted data r_t into three parts, each containing a third of the total number of the returns. This gives almost the same amount of unique symbols $\{0,1,2\}$ in the sequence s_t. We assume that $\theta_1 < 0$ and $\theta_2 > 0$, thus symbol 0 represents the interval of small price variations. This type of discretization is invariant under the addition of a constant term to each value of returns. Thus, the mean of return dynamics does not affect the entropy computation.

2.2.3. Empirical frequencies method

The Empirical Frequencies (EF) method is used to estimate an entropy from a given finite sequence of symbols [42]. The method includes the calculation of empirical probabilities of blocks of symbols and substituting them in the formula for the entropy in Eq. 1. We define a shift-invariant Borel probability measure μ on the space A^∞ of sequences $x = \{x_n\}$ drawn from a finite alphabet A. The process is ergodic[3] with a positive entropy h. Let μ_k be the true distribution of k-blocks (blocks with length of k).

Let $k \le n$ and $x_1^n \in A^n$, $x_i^{i+k-1} = x_i \ldots x_{i+k-1}$. For each $a_1^k \in A^k$, empirical frequencies are defined as

$$f\left(a_1^k | x_1^n\right) = \#\left\{i \in [1, n-k+1] : x_i^{i+k-1} = a_1^k\right\}.$$

By considering an empirical k-block distribution as

$$\hat{\mu}_k\left(a_1^k | x_1^n\right) = \frac{f(a_1^k | x_1^n)}{n-k+1},$$

an empirical k-entropy is defined by

$$\hat{H}_k(x_1^n) = -\sum_{a_1^k} \hat{\mu}_k\left(a_1^k | x_1^n\right) \log\left(\hat{\mu}_k\left(a_1^k | x_1^n\right)\right).$$

We exploit the following Theorem introduced in [43] (Theorem II.3.5–6) to obtain a consistent estimate of the Shannon entropy.

Theorem 1. If μ is an ergodic measure of entropy $h > 0$, if $k(n) \to \infty$ as $n \to \infty$, *and if* $k(n) \le \frac{\log(n)}{h}$, then

$$\lim_{n\to\infty} \frac{1}{k(n)} \hat{H}_{k(n)}(x_1^n) = h \text{ a.s.}$$

In practical applications, we set $k = \lfloor \log(n) \rfloor$. The estimation of the process entropy is

$$\hat{h}_k = \frac{\hat{H}_k}{k}.$$

2.2.4. Unbiased estimation

The estimation is biased when the sample of data is finite. The estimator introduced by Grassberger [44], $\hat{h}_k^G$, is defined in order to correct for the bias, so that $E\left(\hat{h}_{k(n)}^G\right) = h$ for samples of length n. More precisely,

[2] We make several changes in the method. First, we consider not only 0-returns as missing values but also returns after each 0-return. Second, we calculate the likelihood function using all available data including reconstructed returns. The comparison of performances of different approaches and the computations of errors for different parameters of optimization can be found in the Supplemental Material S-7.

[3] Statistical features of an ergodic process can be deduced from a single typical realization.

let f_i, $i = 1, \ldots, M$, be the empirical frequencies of all possible k-blocks, where $M = |A|^k$ and $n_b = n - k + 1$ represents the number of blocks in consideration. The entropy estimate

$$\widehat{H}_k = -\sum_{i=1}^{M} \frac{f_i}{n_b} \log \frac{f_i}{n_b} = \log(n_b) - \frac{1}{n_b} \sum_{i=1}^{M} f_i \log f_i$$

is then replaced by

$$\widehat{H}_k^G = \log(n_b) - \frac{1}{n_b} \sum_{i=1}^{M} f_i \log(\exp G(f_i)),$$

where the sequence $G(i)$ is defined recursively as

$$G(1) = -\gamma - \ln(2)$$

$$G(2) = 2 - \gamma - \ln(2)$$

$$G(2n+1) = G(2n)$$

$$G(2n+2) = G(2n) + \frac{2}{2n+1}, n \geq 1.$$

with the Euler's constant $\gamma = 0.577215\ldots$. We estimate the process entropy as

$$\widehat{h}_k^G = \frac{\widehat{H}_k^G}{k}.$$

2.2.5. Monte Carlo simulations

Given an estimate of the Shannon entropy, its statistical significance is studied by using Monte Carlo simulations.[4] A random walk after the discretization in 2-symbols or 3-symbols, as described in the Section 2.2.2, is a Bernoulli sequence with equal probabilities for the occurrence of each symbol. We define a time series of returns as unpredictable if the entropy estimate is consistent with the entropy of the corresponding Bernoulli process. Any violation is interpreted as a signal of inefficiency for that particular time series of returns.

The bound of significance was computed as follows. We consider lengths of sequences that are multiples of 10. For each considered length, we simulated 10^5 Bernoulli sequences with $p = [\frac{1}{2}, \frac{1}{2}]$ for the binary alphabet and $p = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$ for the ternary alphabet.

Then, we computed the Shannon entropy for all of them. For each length, we found the 99 % confidence interval (CI) associated with the distribution of entropy estimates. Then, to determine CI for lengths that are not multiples of 10, we used a piecewise linear interpolation. We define a time series as *inefficient* in a given time interval if the estimated entropy of the time series in this interval is less than the bound of 99 % one-sided CI of the Bernoulli process with the same length.

3. Entropy analysis results

Time series of price returns are characterized by some regularities related to market patterns, which may apparently suggest the possibility of building up trading strategies to make risk-free profits. For example, microstructure effects result in a non-zero autocorrelation of returns at a high frequency. However, any trading strategy that tries to exploit such an effect has a non-trivial impact on the price dynamics with the result of zero profit on average, see [45,46]. Similar considerations can be drawn also for intraday patterns and volatility clustering. The impact of such effects on the estimation of entropy for return time series has been already considered in the existing literature [32].

Interestingly, there is another source of regularity characterizing time series of returns. It leads to apparent inefficiency which cannot be however exploited to build profitable strategies at high frequency trading. This source of regularity is the presence of 0-returns in data, which lowers the estimate of entropy. It can be interpreted as a spurious effect and must be removed before consideration on market efficiency.

3.1. Zeros as a source of inefficiency

0-returns in financial time series arise because of many effects including rounding, no trading, and price staleness. The 0-returns occurring because of no trading implies a spurious autocorrelation of time series; see [47]. Moreover, except for the non-trading, there is also the effect of price staleness in the data shown in [33]. The authors of the article define price staleness as a lack of price adjustments yielding 0-returns. The effect of staleness is one of the features that distinguish real data from prices following a random walk. To explain the phenomenon of price staleness in the data, we refer to the work [48]:

"Classical models of price formation postulate that informed traders react to new information not yet reflected in the transaction price of a security and transact if the trade guarantees a profit net of execution costs (e.g., [49,50]). Thus, due to lack of trading, a security with higher transaction costs should experience less frequent price updates and a larger number of "small" returns than a security with a lower cost of transacting. Similarly, uninformed traders may not just buy and sell randomly. They may also react to the size of transaction costs and choose not to trade should these costs be considered too large."

The presence of *spurious zeros*, zeros that appear due to no trading or no price adjustments, affects any estimate of the Shannon entropy. The value of entropy as a measure of randomness is affected by the 0-returns in the data since the large amount of 0-returns makes a time series predictable. When 0-returns are persistent in time because of no trading or no price adjustments, the price dynamics look predictable because the price is constant in time. However, such an effect cannot be seen as market inefficiency since no profitable strategy can be implemented in this case.

In the next sections, we show empirically that the 0-returns are one of the sources of apparent inefficiency. We construct a method of filtering out 0-returns due to price staleness in Section 3.3. We test the method of filtering out spurious 0-returns first on simulated data and then on the real dataset. The results for the simulated data are in Section 3.4. We show that spurious 0-returns generated non-uniformly change the measure of entropy of the return time series. However, the entropy as well as the amount of 0-returns due to rounding goes back to its genuine value by implementing the method. The main empirical result obtained on the real dataset in Section 4.2 is that the amount of inefficient weeks decreases significantly after filtering out 0-returns.

3.2. Influence of 0-returns on the entropy value

We first investigate the impact of 0-returns on the estimation of entropy for the 2-symbols discretization. The presence of clustering and intraday patterns in volatility does not influence the 2-symbols discretization. In this case, the data whitening process described in Section 2.1 has no impact on the estimation of the entropy, and any signal of the price inefficiency is not linked to the aforementioned patterns.

We focus the analysis on the set of 100 most liquid stocks belonging to the Russell 3000 index. For each stock, we consider the time series of returns for each week of the period from 02.01.1998 to 23.06.2017. For each stock, we average the fraction of 0-returns separately for the *inefficient* weeks and for the *efficient* weeks. We show the averaged fraction of 0-returns for weeks with inefficiency and for weeks without inefficiency in Fig. 1. The result points out an evident correlation between the fraction of 0-returns and low entropy, a signal of the inefficiency of market dynamics. This supports empirically that the inefficient

[4] We use simulations here because they take less computation than iterating over all cases. From combinatorics we know that the number of outcomes to distribute N blocks over M values is equal to $\binom{N+M-1}{M-1}$. This value increases rapidly with the increase of N. All programs including simulations are written using the MATLAB software.

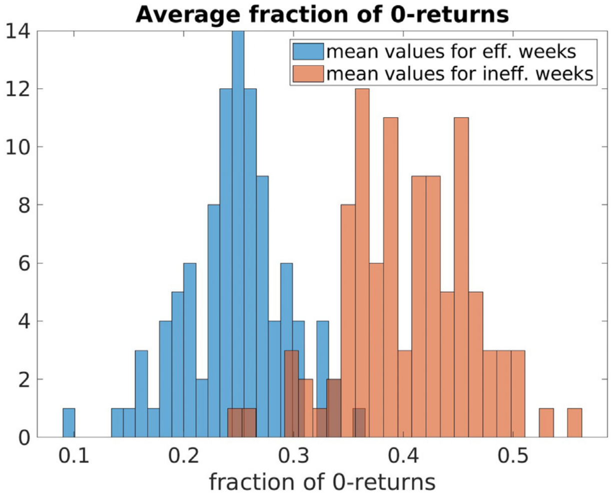


Fig. 1. The fraction of 0-returns for weeks with and without inefficiency presented in two histograms with 25 bins.

weeks are characterized by the presence of a large number of 0-returns.[5]

A larger fraction of 0-returns implies a smaller sample size for the time series of symbols. However, this is not the reason for a biased estimation. First, the estimator proposed by Grassberger has been applied to correct for the bias associated with finite sample sizes. Second, we show in the Appendix C that if one artificially shortens the sample length by aggregating returns on bootstrapping, the entropy estimate is not affected and has no downward bias.

3.3. Censoring spurious 0-returns

Since we found a correlation between the fraction of 0-returns and the entropy value, we developed an algorithm to filter out spurious 0-returns. There exist two effects resulting in 0-return for the price dynamics as pointed out in [33]: The first is the result of a price discretization due to the tick size, and the second is an economic phenomenon linked to traded volumes. The second channel consists of a high transaction cost and the absorption of bounded volumes causing no price changes and leading to price staleness. In the first case, since the price does not change due to rounding, the return is equal to zero at the minimum resolution available. On the other hand, the 0-returns due to the price staleness effect hide information about the underlying asset. Such 0-returns should be considered as *spurious*. Disentangling these two effects is thus crucial to the end of entropy estimation.

Following [33,34], we model price staleness by assuming that observed transaction prices are the result of the coupling of a random walk with a Bernoulli process for the occurrence of the spurious 0-returns.

$$\tilde{P}_i = P_i B_i + \underset{i-1}{\tilde{P}} (1 - B_i) \tag{1}$$

where B_i is Bernoulli variables with values 0 or 1, P_i is an efficient price discretized following the Geometric Brownian Motion, and $\tilde{P}_i$ is an observed price. According to this model, the efficient price is diffusive, even when we do not observe it because of price staleness.

By assuming the process (1) for the price dynamics of an asset, we aim to infer from real data the probability for each 0-return to be generated because of rounding or price staleness, then filtering out 0-returns which are likely associated with the second effect. To this end, the probability of obtaining a 0-return because of rounding within some given tick size is obtained under some approximations[6] as

$$p_t = erf(R_t) + \frac{1}{R_t\sqrt{\pi}}\left(\exp\left(- R_t^2\right) - 1\right) \tag{2}$$

where $R_t = \frac{d}{\overline{P}_t\overline{\sigma}_t\sqrt{2\Delta}}$, $erf(x)$ is the Gaussian error function, d is the tick size, Δ is a time step, $\overline{P}$ is the rounded price, and $\overline{\sigma}_t$ is an estimation of volatility at time t.[7] The result is obtained by considering the probability that the price moves so slightly so that it is rounded to the same value as one time step ago. The obtained probability is approximated using the observed price and the volatility estimation. See the Appendix D.1 for details.

Given p_t at each time step t, the expected number of 0-returns due to rounding is the sum of all p_t within the considered time period, i.e., $N_{save} = \sum_1^N p_t = \widehat{p}N$, where $\widehat{p}$ is an average probability. The variance of the amount of 0-returns is equal to $Var = \widehat{p}(1 - \widehat{p})N$. If the observed number of 0-returns N_{real} is not significantly larger than the expected one according to the model (2), i.e., $N_{real} \leq N_{save} + 1.96\sqrt{Var}$, we do not filter out any 0-return. Otherwise, in order to filter out 0-returns due to price staleness, we replace by missing values the 0-returns which appear not due to the rounding according to the approach below.

3.3.1. Approaches of identifying zeros

A 0-return is considered as spurious according to the following method called *probability-based*. An expected time when a 0-return appears due to rounding is determined when the expected number of 0-returns due to rounding, $Z(t) = \sum_{i=1}^t P_i$, jumps to a new integer value, $\lfloor Z(t) \rfloor - \lfloor Z(t-1) \rfloor = 1$. Then, moving from $t = 0$ to the final time, the expected time when a 0-return appears due to rounding is matched with the closest time with a real 0-return in the time series.[8] We save these 0-returns, but set other 0-returns at the amount $N_c = N_{real} - N_{save}$ as missing values. We assume that N_c 0-returns appear due to price staleness. We set not only 0-returns due to price staleness but also the values at the consecutive time step as missing for the reasons we will discuss below in Remark 1.

The main feature of this approach is that *we use only information on prices for its implementation*: we calculate the probability of getting 0-returns due to rounding using prices, the estimation of volatility, and the tick size. We consider this *probability-based* approach as basic and use it for the analysis. Modifications of this approach including the usage of the information about traded volumes and the estimation of a bid-ask spread are in the Supplementary material S-2.

3.4. Filtering 0-returns on simulated data

To include 0-returns into the analysis, we consider here the three-symbols alphabet, where one of the symbols corresponds to returns between the two tertiles of the empirical distribution. We aim to study here the effect of 0-returns on the estimation of entropy for simulated time series.

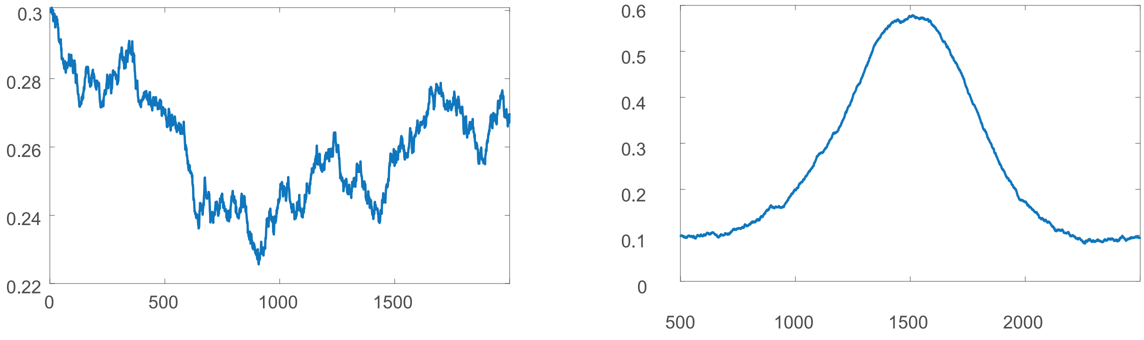
We model a price as the Geometric Brownian motion (GBM), $P_t = P_0 + \int_0^t P_s \sigma dW_s$, setting an initial price equal to 50 and a constant volatility equal to 10^{-3}. We take a time step equal to 1 min and simulate 2000 data points. Then, we generate additional 0-returns with a probability $pr_t = pr_0 + \int_0^t \mu_s ds + \int_0^t \nu dW_s$ with constant $\nu = 10^{-3}$ and three different values of μ_t and pr_0. The first two values present cases where the additional 0-returns are distributed uniformly

[5] Despite the fact that 0-returns are removed from the discretized sequence, they may affect the 2-symbols sequence by changing the ratio between positive and negative returns, the sample size of the resulting blocks of symbols after concatenation, and the values after 0-returns.

[6] The price follows the Geometric Brownian Motion, so that the returns are distributed normally. The bid-ask spread is set to be 0.

[7] Returns used for the volatility estimation have been deseasonalized in a preliminary step.

[8] If we need to keep a 0-return from the sequence of 0-returns, we place it at the beginning of this sequence.



(a) uniform probability

(b) bell-shaped probability

Fig. 2. Examples of the probabilities of getting spurious 0-returns.

but with different extents. The third case simulates a scenario where 0-returns cluster together. The first choice of the probability function is $pr_0 = 0.3$, $\mu(t) = 0$. The second choice of the probability function is $pr_0 = 0.2$, $\mu(t) = 0$. The third choice of the probability function is

$$pr_0 = 0.1\ \ \mu(t) = -\frac{1}{400^2}(t-1000)\ \ \exp\left\{\left(-(t-1000)^2/400^2\right)\right.$$

The examples of the first and third functions are in Fig. 2a and Fig. 2b, respectively.

Remark 1. There are two reasons to consider returns after 0-returns as missing values too.

I. If, according to Eq. 1, the real price is hidden for one or more minutes, the first non-zero return is the sum of all returns that were hidden. Thus, we do not have the value for the return at this minute, but only know the aggregate information.
II. Denoting as missing values returns after 0-returns increases the accuracy of the estimation of volatility. See Fig. 3 as an example. The value of volatility is on the x-axis in the range from $5 \cdot 10^{-4}$ to $2 \cdot 10^{-3}$. Two mean values of the estimations of volatility in the case of the second choice of the probability function are on the y-axis. The closer scatter plot to the diagonal, the more precise the estimation.

We aim to test if the method we have developed identifies spurious 0-returns well. After detecting the spurious 0-returns and setting them as missing values, the entropy value should increase so that the time series should be indistinguishable from a realization of a random walk with some missing values. We calculate the entropies only for the last 1950 data points. First, we calculate the entropy of initial return time series, then with the additional 0-returns, and then after setting the spurious 0-returns as missing values. We simulate 1000 time series with ≤ 1/3 of 0-returns.

How can we estimate the entropy of a sequence with missing values? We keep missing values in the sequence but consider the partitions of the time series in blocks that do not contain missing values of symbols.[9] The description of the chosen method of the entropy estimation and the proof of the consistency of the entropy estimator are in the Appendix D.2. The results are shown in Tables 1- 3 below. The columns of the tables represent the mean entropy for all samples, its standard deviation, the number of samples that are not defined as inefficient, the number of 0-returns averaged, and its standard deviation.

The large amount of spurious 0-returns added uniformly does not sufficiently decrease the entropy value. We take two probability functions with different mean values 0.2 and 0.3. In both cases, on average, sequences with additional 0-returns have the entropy value close to the value of the initial time series. The return series with missing values have the entropy lower than the initial sequence. Since the lower bound of CI also decreases, all sequences after filtering out 0-returns appear to have no inefficiencies. (We compare each estimate with the lower bound of 99 % CI for the entropy of Bernoulli sequences, which is about 0.9935 for the length of 1950). Moreover, when we make the distribution sharper and bell-shaped, as in the third case, the estimate of entropy decreases significantly after adding spurious 0-returns. However, when we use *probability-based* method described in the Section 3.3.1, the entropy becomes closer to its initial value. The other important aspect is that the *probability-based* method keeps the amount of 0-returns in the sequence quite close to the number of 0-returns appearing by rounding the efficient price in all three cases.

We expect similar results for real data. If spurious 0-returns are distributed uniformly, setting the spurious 0-returns as missing values does not affect the value of entropy. However, if the 0-returns cluster together, for example, in the presence of high transaction costs, then a predictable pattern due to spurious 0-returns needs to be removed from the time series for a genuine estimation of entropy.

3.5. Filtering 0-returns on real data

To test the *probability-based* approach on real data, we take the SPY ETF, which aims to track the Standard & Poor's 500 Index, and SPDR

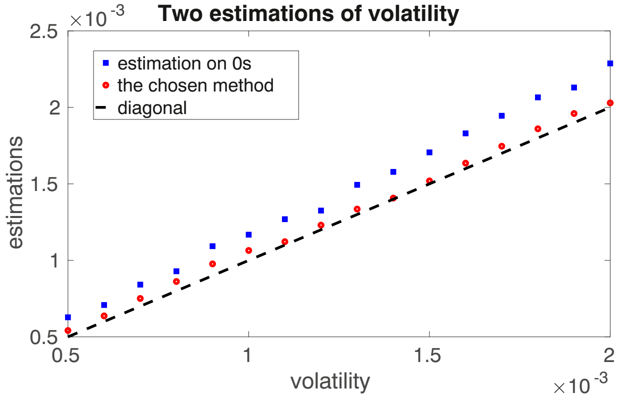


Fig. 3. The volatility and two estimations. The blue squares are the estimations of volatility using only 0-returns as missing values. The red circles are the estimations of volatility using 0-returns and the values after 0-returns as missing values.

[9] Other common approaches to deal with missing values are concatenating observed sequences and using an interpolation to replace a missing observation with its reconstructed value. The former may create new patterns containing parts of concatenated blocks and the latter incorporates predictable patterns instead of the missing values.

Table 1
Results of filtering 0-returns with the 1st choice of probability function.

Time series	Mean entropy	Std. of entropy	N of eff. Series	N. of 0-returns	Std. for 0-returns
GBM	1.0002	0.0025	993	153.38	13.09
After adding 0-returns	1.0002	0.0026	994	619.77	23.77
After setting missing values	0.9978	0.0083	1000	147.96	4.99

Table 2
Results of filtering 0-returns with the 2nd choice of probability function.

Time series	Mean entropy	Std. of entropy	N of eff. Series	N. of 0-returns	Std. for 0-returns
GBM	1.0001	0.0027	987	155.49	12.31
After adding 0-returns	1.0002	0.0026	994	508.49	51.77
After setting missing values	0.9992	0.0061	1000	147.65	4.69

Dow Jones Industrial Average ETF Trust (DIA). The tick size is $d = 0.01$, and the time step is $\Delta = 1$ minute. We discretize returns after filtering out the daily seasonalities and the heteroskedasticity. We use the *probability-based* approach described in Section 3.3.1.[10] Fig. 4a and Fig. 4b show the entropy of these returns including all 0-returns and the entropy calculated after filtering out 0-returns.

For the ETF SPY, we observe 330 weeks. 36 of them are associated with the estimate of entropy lower than the 99 % confidence bound after the step of filtering out heteroskedasticity: we refer to these weeks as inefficient weeks. After applying the method of 0-filtering there are 10 inefficient weeks, but only 4 of them are from the group of previously inefficient weeks. For the ETF DIA, we observe 333 weeks with 42 inefficient weeks after the step of filtering heteroskedasticity. After applying the method of 0-filtering there are 12, but only 8 of them are from previously inefficient weeks. The main conclusions we can make from this section are the following.

1. On average, the algorithm of filtering out 0-returns increases the entropy value.
2. There are weeks where a low entropy value still cannot be explained by intraday volatility pattern, heteroscedasticity, and 0-returns.
3. For some reason, our method can determine inefficiency for a week which we considered as “efficient” in the sense described in Section 2.2.5. Two possible explanations are random fluctuations of the entropy measure as a random variable and detecting significantly low values of the entropy that were not detected before due to a high level of confidence. We discuss this issue in the Appendix E.

We investigate another reason for appearing new inefficiencies and the possible source of remaining inefficiencies in the next section.

4. Periodic patterns, microstructure noise, and the ETF market

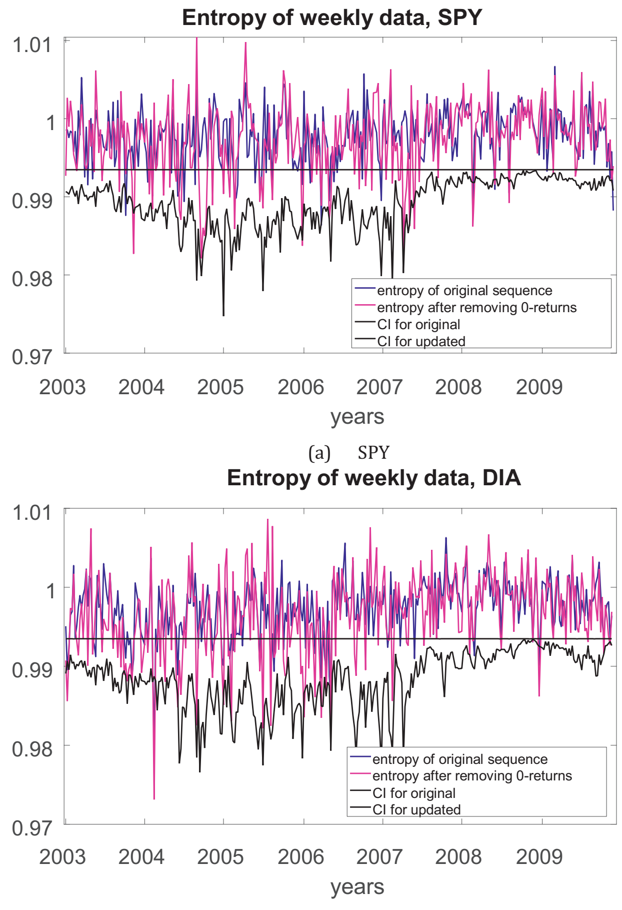
4.1. Periodic patterns

The goal of this section is to investigate a reason for low entropy values by considering the process of calculating the empirical frequencies. When estimating entropy, the frequencies of all possible k-blocks are calculated. We are interested in finding some repeating patterns in blocks that appear to be the most frequent, which may cause a decrease in the entropy estimation. For each inefficient week of the ETF DIA after filtering out 0-returns, we write down the most frequent block(s) of symbols of length k which are met while calculating the entropy. The complete table with results is in the Supplementary material S-8. The most frequent blocks in weeks with the new inefficiencies are **000121**, 011210, **020121**, **021002**, 201211, 202111, **210120**. We highlight in bold blocks where the same characters 1 or 2 do not appear in a row. This may mean that the price fluctuates around its average value. We assume that such an effect can be observed, for example, with a bid-ask bounce. That is, such patterns occur when there is no movement in the efficient price, but transactions occur at both the bid and ask prices. We show when the existence of this pattern is statistically significant.

Table 3
Results of filtering 0-returns with the 3rd choice of probability function.

Time series	Mean entropy	Std. of entropy	N of eff. Series	N. of 0-returns	Std. for 0-returns
GBM	1.0002	0.0027	995	153.78	13.04
After adding 0-returns	0.9905	0.0051	287	613.02	27.64
After setting missing values	0.9984	0.0058	1000	149.27	5.05

If we consider only positive and negative returns, we move back for a moment to the 2-symbols discretization. Let choose $k = \lfloor \log_2(L) \rfloor$, where L is the length of a 2-symbols sequence. Let's consider 2 sequences ‘1010…’ and ‘0101…’. We know that the expected amount of blocks with these sequences for the process with the entropy $h = 1$ is



(a) SPY

(b) DIA

Fig. 4. The entropies calculated for the 3-symbols discretization before and after filtering out 0-returns for the ETFs SPY and DIA with the corresponding 99 % Confidence Intervals.

[10] We apply the approach only to weeks with <1/3 of 0-returns and <10 consecutive minutes with 0-volume.

$n_b/2^{(k-1)}$, where $n_b = L - k + 1$. Also, we construct a 99 % CI using the formula for the standard deviation from the binomial distribution.

$$\begin{aligned} p &= \frac{1}{2^{(k\,-\,1)}} \\ q &= 1 - p \\ \widehat{\sigma}^2 &= n_b p q \end{aligned} \tag{3}$$

Definition 2. If the actual amount of blocks '1010…' and '0101…' in a sequence, n_{pp}, is greater than the upper bound of CI, that is,

$$n_{pp} > n_b/2^{(k\,-\,1)} + q_\alpha \widehat{\sigma}$$

where q_α is a quantile of the normal distribution, we determine the sequence as one with periodic patterns (PP).

These periodic patterns may appear due to a low price value with respect to the tick size and a low volatility (which are also reasons for 0-returns generated by rounding). If a price fluctuates randomly around the mean value, crossing the same tick size twice in the opposite directions is more likely than crossing two tick sizes in the same direction.

We have markers for inefficiency and periodic patterns for each week. Using the hypergeometric distribution,[11] we may conclude if there is a dependence between a detected inefficiency and PP. Thus, we test the following hypotheses.

H_0. : the appearance of a week with inefficiency and the appearance of a week with periodic patterns are *independent* against an alternative hypothesis.

H_a. : the appearance of a week with the inefficiency and the appearance of a week with periodic patterns are *dependent*.

We say that H_0 is rejected if

$$n_{ineff+pp} > m + q_\alpha \widehat{\sigma} \tag{4}$$

where $n_{ineff+pp}$ is the number of weeks with inefficiency and periodic patterns simultaneously; $m = \frac{nK}{N}$; $\hat{\sigma}^2 = \frac{nK(N\,-\,n)(N\,-\,K)}{N^2(N\,-\,1)}$; N is the total number of weeks; K is the number of weeks with inefficiency; n is the number of weeks with PP.

The results for SPY are that we have **no** 95 % confidence that the random processes "occurrence of a week with inefficiency" and "occurrence of a week with periodic patterns" are dependent. On the contrary, for DIA we do have 95 % confidence that there is the dependence for the occurrence of weeks with inefficiency and weeks with PP.

4.2. Analysis of the ETF market

The goal of this section is to assess the impact of the 0-filtering process on the ETF market and to test the dependence between the remaining weeks with inefficiency and the presence of periodic patterns. We take the set of 10 ETFs and test them for detecting weeks with inefficiency before and after the 0-filtering process. The characteristics we are interested in are the total amount of weeks; the amount of weeks with inefficiency before the 0-filtering; the amount of weeks with inefficiency after the 0-filtering; the amount of new inefficient weeks that appear only after the 0-filtering process; those new inefficient weeks which contain periodic patterns according to Definition 2. Finally, we test the hypothesis H_0 presented in the previous section about the independence of the events of occurring an inefficient week and a week with periodic patterns according to Eq. 4. We construct Table 4 with the results for all 10 ETFs.[12]

Table 4
Results of filtering out 0-returns with the probability-based approach.

ETF	Total weeks	Ineff. before 0-filtering	Ineff. after 0-filtering	New ineff.	New ineff. explained by PP	Reject H0 with 95 %	Reject H0 with 99 %
SPY	330	36	10	6	0	0	0
DIA	333	42	12	4	1	1	0
IWM	294	34	5	3	0	0	0
EWJ	14	3	3	1	1	1	0
XLE	231	12	7	3	1	0	0
XLF	120	23	16	4	1	1	1
XLU	94	10	6	4	0	0	0
IVV	161	6	0	0	0	0	0
XLB	124	4	2	2	0	0	0
IWO	148	13	3	3	0	0	0

For the last two columns, 1 indicates the rejection of H_0; 0 indicates a failure to reject H_0.

We conclude that for all 10 ETFs the number of inefficient weeks decreases after applying the filtering of 0-returns. The number of detected weeks with inefficiency remains the same only in the case of the ETF EWJ. The algorithm filters out apparent inefficiencies from all 6 weeks considered for the ETF IVV. That is, we cannot detect any statistically significant decrease in entropy for the ETF IVV by using non-overlapping weekly time windows.

Finally, we note that the percentage of the total number of inefficient weeks is 3.46. However, we detect a predictable time series with 99 % of confidence. Since 3.46 % is significantly >1 %, we can conclude that there are other unaccounted sources of inefficiency in the price dynamics. For this reason, we discuss below the role of microstructure noise, another stylized fact of financial markets which is key for the entropy estimation of return time series at a high frequency.

4.3. Filtering microstructure noise

An observed price includes various microstructure effects caused by transaction costs and price rounding. The difference between the efficient price and the observed price with the microstructure effects is called *microstructure noise*. In general, each new observation in a return time series depends on the previous values. A model that can explain the presence of a positive autocorrelation of data is the following. Assume to observe a price $\tilde{P}_t$ that differs from the efficient price P_t by some error term u_t, namely

$$\ln\,\tilde{P}_t = \ln\,P_t + u_t$$

and the observed market return $\tilde{r}_t$ is

$$\tilde{r}_t = r_t + u_t - u_{t-1}$$

where r_t is the return of the efficient price. The observed return is affected by the error term associated with the log-price at the previous time step. The microstructure noise tends to be positively autocorrelated in time [51]. If the error term u_t follows an autoregressive $AR(1)$ process, then the observed returns are described by an $ARMA(1,1)$ process.

The effect of such a noise term on the estimation of entropy has been described in [32] by considering both $AR(1)$ and $MA(1)$ models. In particular, the authors have found that larger (in absolute value) autoregressive coefficients are associated with lower values of the Shannon entropy. This intuition is exploited by M. Ito and S. Sugiyama [4], which use a time-varying autocorrelation of stock returns as a measure of market inefficiency for the U.S. stock market.

Here, we consider a further step of filtering based on the estimation of an ARMA model on the time series of returns after the 0-filtering. After selecting the best (P,Q) of an ARMA(P,Q) model describing the data by using the BIC criterion [52], we study the residuals in order to remove any autocorrelation pattern from data. We consider only ARMA(P,Q) models with $P + Q \leq 5$. After filtering out 0-returns and replacing them

[11] The hypergeometric distribution describes the probability of k successes in n draws without replacement from the finite population of size N that contains exactly K objects.

[12] See the Supplementary material S-2.1 for the comparison of results for the different approaches of the 0-filtering.

Table 5
Resulting table after filtering out the microstructure noise.

ETF	Total weeks	Ineff. before ARMA	Ineff. after ARMA	New ineff.	Reject H0 with 95 %	Weeks with ineff.
SPY	330	10	4	0	0	03-May-2007 - 09-May-2007 24-May-2007 - 31-May-2007 23-Feb-2009 - 27-Feb-2009 13-Nov-2009 - 19-Nov-2009
DIA	333	12	6	2	0	14-Apr-2003 - 21-Apr-2003 14-Jan-2004 - 21-Jan-2004 12-Feb-2004 - 19-Feb-2004 04-Feb-2008 - 08-Feb-2008 17-Nov-2008 - 21-Nov-2008 **16-Dec-2008 - 22-Dec-2008**
IWM	294	5	2	0	0	02-Dec-2008 - 08-Dec-2008 **16-Dec-2008 - 22-Dec-2008**
EWJ	14	3	1	0	0	16-Oct-2009 - 22-Oct-2009
XLE	231	7	1	0	0	21-Sep-2005 - 27-Sep-2005
XLF	120	16	5	0	0	22-Sep-2008 - 26-Sep-2008 20-Oct-2008 - 24-Oct-2008 03-Jun-2009 - 09-Jun-2009 **11-Sep-2009 - 17-Sep-2009** 09-Oct-2009 - 15-Oct-2009
XLU	94	6	1	0	0	**11-Sep-2009 - 17-Sep-2009**
IVV	161	0	1	1	0	22-Feb-2006 - 28-Feb-2006
XLB	124	2	4	2	0	28-Feb-2007 - 06-Mar-2007 16-Jul-2007 - 20-Jul-2007 16-Apr-2008 - 22-Apr-2008 21-May-2008 - 28-May-2008
IWO	148	3	0	0	0	

The presence of periodic patterns in weeks with inefficiency is defined with 95 % of confidence.

by missing values, we use the methodology introduced in [53] to deal with the estimation of an ARMA(P,Q) model with missing observations, in particular by using the Kalman filter. Other approaches of determining the order of an ARMA model including the maximization of entropy and an out-of-sample testing are presented in the Supplementary material S-3.

For the analysis of results, we introduce *coinefficiency* for a group of assets. We say that two or more assets have coinefficiency if they have the same week with inefficiency. Moreover, if the hypothesis of the independence of appearing inefficiencies for a given time period is rejected with the significance of 0.05 according to the binomial distribution, we say that the coinefficiency is a statistically significant event (by analogy of defining periodic patterns with Eq. 3). New inefficient weeks are found in the comparison with the results obtained with the probability-based approach. Results for 10 ETFs are in Table 5.[13] Filtering microstructure noise using BIC gives a low percentage of new inefficient weeks equal to 0.27 %, which makes the results consistent with the previous step of filtering. Moreover, filtering microstructure noise removes the dependence between weeks with periodic patterns and weeks with inefficiency. The number of coinefficiencies is equal to 2 for this approach; both of them are statistically significant and are highlighted in bold in the table.

Filtering microstructure noise gives a low percentage of inefficient weeks equal to 1.35 %. This is a clear signal that the Efficient Market Hypothesis is not totally realistic for real-world market dynamics. Indeed, after filtering all known sources of apparent inefficiencies, the percentage of weeks detected as inefficient is >1 %, namely the confidence level used in our testing procedure. Thus, the market is not totally efficient at a 1 min time scale. Regarding the last step of filtration, namely, considering residuals of an ARMA model, we should also notice that we filter out all possible linear dependencies for returns together with the effect of the microstructure noise. As a consequence, the resulting measure of market inefficiency excludes as well any linear effect of predictability.

5. Analysis of longer intervals and cointegration

In the previous sections, we discussed the issue of (in)efficiency of the ETF market for weekly time intervals. Some questions may require more careful study. For instance, reasons for the occurrence of coinefficiencies are not yet investigated. Since the procedure of detecting an inefficient week for a specific asset is made with a high level of confidence, the occurrence in the data of 10 ETFs cases of the coinefficiency should be investigated better. Thus, there can be some exogenous cause that leads to more predictability for the several assets at the same time. A research interest is to determine if there is some correlation between assets at a time when both of them have a week with inefficiency. Since some ETFs aim to track indexes having some stocks in common, it is natural to expect that some inefficient weeks appear simultaneously due to exogenous events in the markets. We test assets for cointegration in the Supplementary material S-4.

Moreover, we concentrated on weekly time intervals. On a weekly basis, we conclude that we filter out the main sources of apparent inefficiency, so that the percentage of intervals with inefficiency is close to the level of significance. However, we restricted ourselves by the short time intervals. In the Supplementary materials S-5 we consider monthly and quarterly time intervals to analyze market inefficiency using rolling windows but with greater length. The results are that the measure of market inefficiency, by analogy with the case of weeks, is equal to about 11 and 10 % for months and quarters, respectively. A transition from weeks to months significantly increases the measure of inefficiency. Important differences between the two time intervals are the length of blocks taken into consideration and the amount of data in which different predictable patterns may be found. Finally, the dataset with the period of about 2.5 years is considered in the Supplementary material S-6.

6. Conclusions

We have studied the efficiency of the ETF market. To this aim we applied the Shannon entropy as a measure of randomness that has also

[13] See Supplementary material S-3.1 for the comparison of results for different approaches of filtering microstructure noise.

been used in the range of articles [15–18]. In contrast to the mentioned works, we filtered different sources of apparent inefficiency of market dynamics before calculating the degree of market inefficiency. The method of filtering out apparent inefficiencies was first introduced by Calcagnile et al. [32]. In our work, we have introduced price staleness [33,34] as a source of apparent inefficiency besides the apparent inefficiencies caused by daily seasonalities and heteroscedasticity.

Price staleness creates spurious 0-returns in the data. We constructed the method for detecting spurious 0-returns according to the probability of their occurrence. We set spurious 0-returns as missing values. Thus, we built and applied the modification of the Empirical Frequencies method [42] for calculating entropy in the case of the presence of missing values in the data. We have shown that 0-returns cause a false detection of inefficiency.

The last step of data whitening is filtering out microstructure noise. We have shown that for some ETFs there is a clear dependence between a low estimate of entropy for discretized returns and the presence of periodic patterns, i.e., the switching sign of non-zero returns for each trading minute. Both periodic patterns and microstructure noise may have their origin in a bid-ask bounce. We use a fitting ARMA model to get rid of these effects. We have measured market inefficiency on a weekly basis as the percentage of inefficient weeks among all assets taken into consideration.

After the implementation of the multi-step filtering method, which allows to remove daily patterns, heteroscedasticity, 0-returns, and microstructure noise, the fraction of weeks with inefficiency decreases to a value slightly >1 %. Taking into account the 99 % level of confidence for the test for inefficiency, we can conclude that there exists slight evidence of inefficiency in the ETF market at a high-frequency time scale. Molgedey and Ebeling [14] came to the similar conclusion and stated that the Dow Jones Index is not fully random, but nearly random.

We plan to use the methodology developed in this article to analyze the predictability of other markets using more recent data. Another direction for future research is related to the search for the optimal length of the time interval for detecting market inefficiency.

Finally, we plan to develop our methodology on ultra-high frequency time series, for example, by using tick-by-tick data.

CRediT authorship contribution statement

Andrey Shternshis: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Piero Mazzarisi:** Conceptualization, Validation, Writing – review & editing, Supervision. **Stefano Marmi:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Data availability

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Datasets

We use a proprietary intraday financial time series dataset provided by kibot.com.

Tables 6 and 7 show the list of tickers for ETFs and Stocks respectively.

Table 6
List of ETFs.

Ticker	Name ETF	Asset tracked
SPY	SPDR S&P 500	S&P 500 Index
DIA	DIAMONDS Trust Series 1	Dow Jones Industrial Average Index
IWM	iShares Russell 2000 Index	Russell 2000 Index
EWJ	iShares MSCI Japan Index	MSCI Japan Index
XLE	Energy Select Sector SPDR	Energy Select Sector Index
XLF	Financial Select Sector SPDR	Financial Select Sector Index
XLU	Utilities Select Sector SPDR	Utilities Select Sector Index
IVV	iShares S&P 500 Index	S&P 500 Index
XLB	Materials Select Sector SPDR	Materials Select Sector Index
IWO	iShares Russell 2000 Growth Index	Russell 2000 Growth Index

Table 7
List of stock tickers.

Tickers for 100 Stocks									
MSFT	MO	AAPL	LLY	AIG	CAT	ADBE	CL	FDX	EA
CSCO	HD	BAC	AMZN	SPLS	SCHW	GLW	PAYX	KR	NKE
INTC	HPQ	TXN	MCD	XLNX	LOW	CA	DUK	BBBY	MXIM
ORCL	DIS	PG	ABT	COST	IP	RIG	EMR	NEM	MAT
GE	MRK	JNJ	SLB	WFC	MMM	LUV	DOW	NTAP	COF
AMAT	WMT	DD	MDT	JPM	ALL	WMB	INTU	SO	SYMC
IBM	KO	BMY	MSI	WBA	GPS	CTXS	ADI	CAG	LMT
PFE	AMGN	MU	AXP	SBUX	BBY	KMB	CVS	LRCX	CCL
C	BA	T	HAL	XRX	BK	BHI	USB	BSX	TER
TWX	PEP	QCOM	AMD	KLAC	AA	UTX	HON	NE	JCP

Data for stocks are provided by Kibot. For the description of Stocks' symbols, we refer to http://www.kibot.com/Historical_Data/Russell_3000_Historical_Intraday_Data. aspx

Appendix B. Data cleaning and whitening

B.1. Outliers

We use the Brownlees and Gallo's algorithm of an outlier detection [35] with parameters $k = 20$, $\delta = 10\,\%$, $c = 5$, $\gamma = 0.05$. The algorithm identifies price values which are too distant from the mean value with respect to the standard deviation. The algorithm removes a price P_i if

$$\left|P_i - \overline{P_i}(k)\right| \geq cs_i(k) + \gamma$$

where $\overline{P_i}(k)$ and $s_i(k)$ are respectively the δ-trimmed sample mean and standard deviation of the k price records closest to time i. The δ lowest and the δ highest observations are discarded when the mean and the standard deviation are calculated from the sample.

B.2. Stock splits

A stock split is a change in the number of company's shares and in the price of the single share such that a market capitalization does not change. If the number of stocks increases, it is called a forward split. Otherwise, the split is a stock merge. We check the condition $|r| > 0.2$ in the return series to detect unadjusted splits.

B.3. Intraday volatility pattern

The volatility of intraday returns has periodic behavior. It is higher near the opening and the closing of the market, showing a typical U-shaped profile every day. For empirical evidence of the U-shaped intraday pattern of stock returns in NYSE, see [37]. We filter out the intraday volatility pattern from the return series by using the following model with intraday volatility factors. If $R_{d,\,t}$ is the raw return of day d and intraday time t, we define deseasonalized returns as

$$\tilde{R}_{d,t} = \frac{R_{d,t}}{\xi_t} \tag{5}$$

where

$$\xi_t = \frac{1}{N_{days}} \sum\nolimits_{d'} \frac{\left|R_{d',t}\right|}{s_{d'}}$$

N_{days} is the number of days in the sample and s_d is the standard deviation of the returns of day d. The procedure also normalizes the values of overnight returns that tend to have larger magnitudes than the other 389 returns. Fig. 5 shows ξ_t, for the ETF SPY where t passes from 9:31 to 15:59. All picks appear every half hour (from the largest at 10:00 to 15:30).

B.4. Heteroskedasticity

The deseasonalized returns $\tilde{R}$ defined by Eq. 5 are still heteroskedastic since different days can have different levels of volatility. In order to remove this heteroskedasticity, we estimate the volatility σ_t and define the standardized returns by

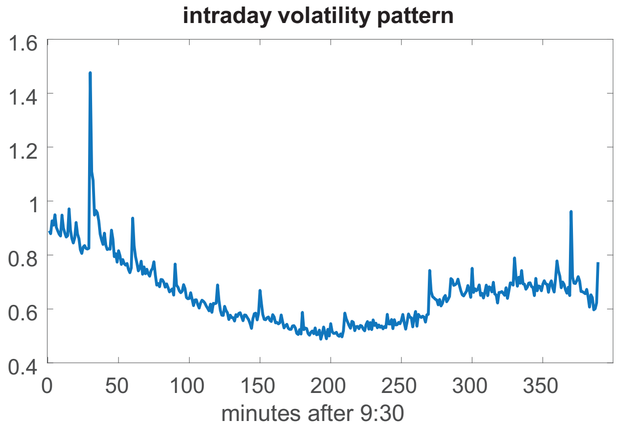


Fig. 5. Intraday volatility pattern for the ETF SPY.

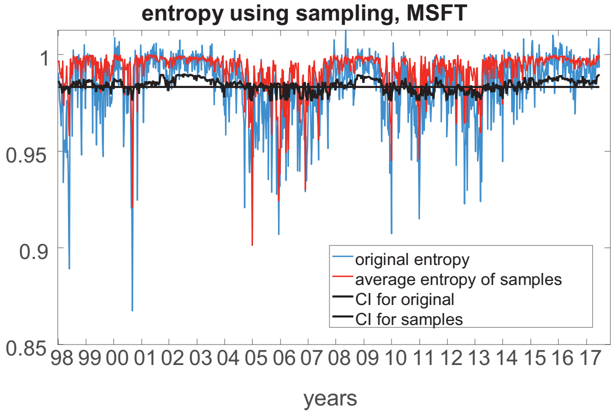


Fig. 6. The estimated entropy and the averaged entropy of samples for the stock MSFT with the corresponding 99 % confidence intervals.

$$r_t = \frac{\tilde{R}_t}{\sigma_t}$$

Appendix C. Testing dependence between the entropy estimation and the length of sequence

In this section we aim to test dependence between the entropy estimation and the length of sequence. For the study we take the price of the Microsoft Corporation stock (MSFT). Then, we discretize returns using the 2-symbols discretization. There are 980 weeks in the considered time interval from 02.01.1998 to 23.06.2017. The minimum length of a 2-symbols sequence is obtained in week 352 and is equal to 896. First, we calculate the entropy value for every week. Then, we use sampling of the binary sequence to decrease its length for each week. For every week, we consider 1000 sub-samples of binary symbols choosing only 896 symbols for the sign of returns (keeping their ascending order in time). Then, we calculate the average value of entropy of all samples and plot it for each week with the estimated value of entropy of the week. See Fig. 6.

The entropy value calculated using sampling does not decrease. On the contrary, the entropy of shortened sequences has a larger value. A probable explanation is the destruction of existing dependencies in the returns since we choose samplings with some gaps between symbols. The other way to show that entropy does not decrease with the amount of non-zero returns decreasing is by aggregating data to a less frequency. The amount of information does not increase with aggregation. Thus, the entropy value does not decrease with a smaller length. To show it empirically, we aggregate data to 5 min, so we reduce the lengths of sequences by about five times. See Fig. 7.

When we aggregate data for 5 min, the maximum possible length of sequence is 390. In spite of the small length of sequence, the entropy is close to 1. We may make the conclusion that a low entropy value may be characterized by a large fraction of 0-returns, but not by the length of the sequence.

Appendix D. Zeros

D.1. Expected 0-returns of rounded efficient price

Here, we calculate the approximate value of the amount of 0-returns generated by rounding of an efficient price. We consider the model for the efficient price following the Geometric Brownian Motion.

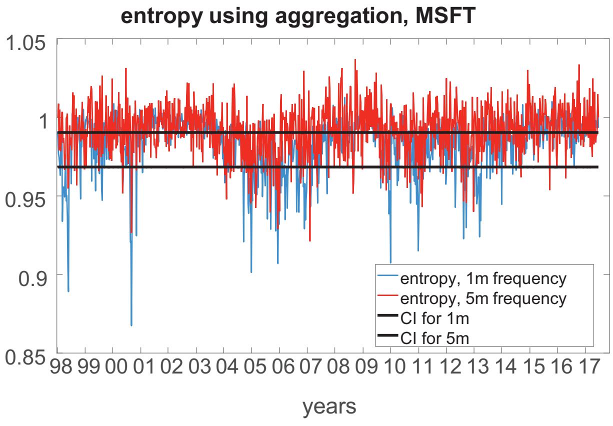


Fig. 7. The entropies calculated for 1 min and 5 min frequencies for the stock MSFT with the corresponding 99 % confidence intervals.

$$P_t = P_0 + \int_0^t \sigma_s P_s dW_s$$

Assuming that the price is rounded up to tick size d, we can find the probability that the efficient price will not change using the rounded (observed) price $\overline{P}$, the spot volatility σ, and the sampling frequency Δ. The probability that $\overline{P}_{i+1} = \overline{P_i}$ given $\overline{P_i}$ is

$$p_i = P\left[\bar{P}_i - \frac{d}{2} \ < \ P_{i+1} \ < \ \bar{P}_i + \frac{d}{2}\right]$$
$$\bar{P} = P + x$$
$$x \in U_{-\frac{d}{2},\frac{d}{2}}$$

that is, as shown in [33], equal to

$$\int_{-\frac{d}{2}}^{\frac{d}{2}} \int_{x-\frac{d}{2}}^{x+\frac{d}{2}} f_{N\left(0, P_i\sigma_i\sqrt{\Delta}\right)}(z) dz dx,$$

where f_N is the Normal density function.

$$I_i(x) = P\left[\overline{P_i} - \frac{d}{2} < P_{i+1} < \overline{P_i} + \frac{d}{2}\right] = P\left[P_i + x - \frac{d}{2} < P_{i+1} < P_i + x + \frac{d}{2}\right] = P\left[x - \frac{d}{2} < P_{i+1} - P_i < x + \frac{d}{2}\right] = \int_{x-\frac{d}{2}}^{x+\frac{d}{2}} f_{N\left(0, P_i\sigma_i\sqrt{\Delta}\right)}(z) dz$$

Then, we estimate $P_i\sigma_i$ by $\overline{P_i}\overline{\sigma_i}$, where the returns that are used for the estimation of volatility are deseasonalized.

$$I_i(x) \approx \frac{1}{2}\left[erf\left(\frac{x + \frac{d}{2}}{\sqrt{2}s_i}\right) - erf\left(\frac{x - \frac{d}{2}}{\sqrt{2}s_i}\right)\right],$$

where $s_i = \overline{P_i}\overline{\sigma}_i\sqrt{\Delta}$ and $erf(x) = \int_0^y \exp\left(-t^2\right)dt$ is the Gaussian error function. Using integration by parts $\int erf(y)dy = y \cdot erf(y) + \frac{1}{\sqrt{\pi}} \exp\left(-y^2\right)$, we obtain the result

$$\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} erf\left(\frac{x + \frac{d}{2}}{\sqrt{2}s}\right) dx = \frac{\sqrt{2}s}{d}\left(\frac{d}{\sqrt{2}s} \ erf\left(\frac{d}{\sqrt{2}s}\right) + \frac{\exp\left(-\frac{d^2}{2s^2}\right) - 1}{\sqrt{\pi}}\right)$$

$$\frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} erf\left(\frac{x - \frac{d}{2}}{\sqrt{2}s}\right) dx = \frac{\sqrt{2}s}{d}\left(\frac{d}{\sqrt{2}s} erf\left(-\frac{d}{\sqrt{2}s}\right) + \frac{-\exp\left(-\frac{d^2}{2s^2}\right) + 1}{\sqrt{\pi}}\right)$$

$$\mathrm{p_i}(R_i, 0) = \frac{1}{d}\int_{-\frac{d}{2}}^{\frac{d}{2}} I(x) dx = erf(R_i) + \frac{1}{\sqrt{\pi}R_i}\left(\exp\left(-R_i^2\right) - 1\right)$$

where $R_i = \frac{d}{s_i\sqrt{2}} = \frac{d}{\overline{P_i}\overline{\sigma}_i\sqrt{2\Delta}}$ and the second argument 0 stands for the amount of ticks that the price moves. The result is extended for the approximation of probability that the price moves by k ticks, $p_i(R_i, k)$ and by including a bid-ask spread in the Supplemental material S-1.

D.2. Entropy estimation in case with missing values

We adopt the method of Empirical Frequencies for the case of the presence of missing values in data. First, we need to choose a suitable value for the length of blocks, k. After correctly choosing the value of k, we consider partitions of the time series in blocks that do not contain missing values.
Definition 3. A non-decreasing sequence $k(n) \leq n$ is admissible if

$$\lim_{n\to\infty}\left|\widehat{\mu}_{k(n)}\left(\cdot|x_1^n\right) - \mu_{k(n)}\right| = 0 \text{ a.s.},$$

where the distance between two measures p and q on A^{k} is

$$|p - q| = \sum_{a_1^k}\left|p\left(a_1^k\right) - q\left(a_1^k\right)\right|.$$

We will rely on two theorems formulated for the case of complete data. See [42], Theorems III.2.1–2 for their proofs.

Theorem I. If $k(n) \geq \log (n)/(h - \epsilon)$, where h is the entropy of the process μ, $\epsilon > 0$, then $k(n)$ is not admissible for μ.

Theorem II. If μ is i.i.d., Markov, or ϕ-mixing and $k(n) \leq \log (n)/(h + \epsilon)$, then $k(n)$ is admissible for μ.

Now, let's denote the amount of k-blocks as $n_{blocks}(k)$. The exact value of $n_{blocks}(k)$ is unknown without knowing the location of missing values. The proof of the Theorem I is based on the fact that, when k is large, the number of all blocks is bounded by the value $3^{k(h-\epsilon)}$ while the amount of all unique blocks is 3^k. Replacing $n(k)$ by $n_{blocks}(k)$ we can repeat the proof and update the lower bound of not admissible $k(n_{blocks})$ to be equal to log $(n_{blocks})/(h - \epsilon)$. We set 3 as a base, since we usually assume that the alphabet is ternary.

We aim to prove the following theorem for the case of having missing values in the output of the random process μ.

Theorem 2. Assume that the process of generating missing values and μ are independent. If μ is i.i.d., Markov, or ϕ-mixing and $k(n_{blocks}) \leq \log (n_{blocks})/(h + \epsilon)$, then $k(n_{blocks})$ is admissible for μ.

Proof of Theorem 2. When passing to the limit, it is assumed that $n \to \infty$. Since $k(n_{blocks}) \leq \log (n_{blocks})/(h + \epsilon) \leq \log (n)/(h + \epsilon)$, thus $k(n_{blocks})$ is *admissible* if the data is complete. Note that the number of blocks is greater than or equal to $3^{(k/(h+\epsilon))} > 0$.

Let's fix a_1^k. Without missing values $\hat{\mu}_k^n = \hat{\mu}_k(a_1^k|x_1^n) = \frac{f_k^n}{n-k+1}$ and $\hat{\mu}_k^n \to \mu_k$ a.s. If $N(n)$ values are missing, then

$$\bar{\mu}_k\left(a_1^k|x_1^n\right) = \frac{f_k^n - c(n)}{n-k+1-d(n)} = \frac{\hat{\mu}_k^n(n-k+1) - c(n)}{n-k+1-d(n)} = \frac{\hat{\mu}_k^n(n-k+1)}{n-k+1-d(n)} - \frac{c(n)}{n-k+1-d(n)}$$

where $d(n)$ is the total number of blocks eliminated, $N(n) \leq d(n) \leq k(n)N(n)$, and $c(n)$ is the number of blocks a_1^k eliminated, $0 \leq c(n) \leq k(n)N(n)$. $n - k + 1 - d(n) = n_{blocks} > 0$.

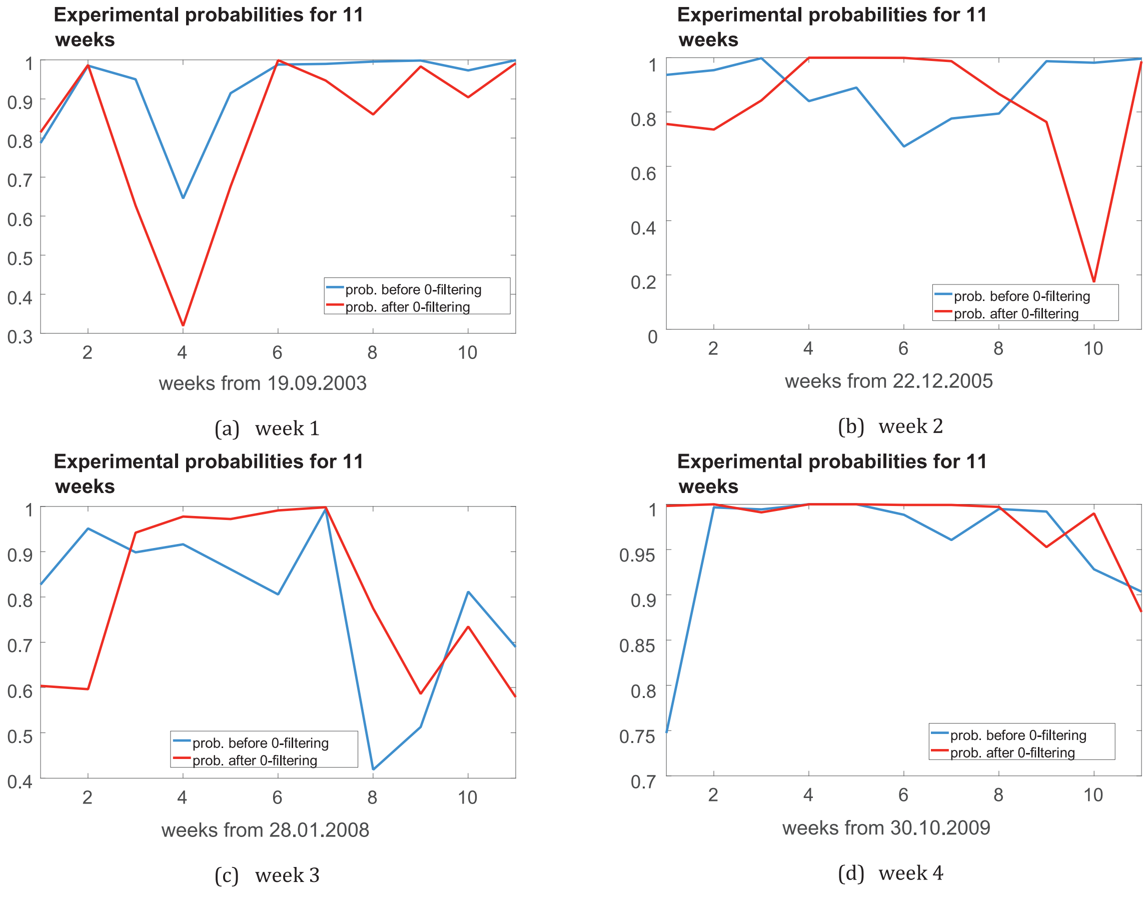
There are three possible cases: I. $d(n)/n \to 0$; II. $d(n)/n \to C$, $0 < C < 1$; III. $d(n)/n \to 1$.

I.

$$\bar{\mu}_k\left(a_1^k|x_1^n\right) = \frac{\hat{\mu}_k^n(n-k+1)}{n-k+1-d(n)} - \frac{c(n)}{n-k+1-d(n)}$$

Note that $0 \leq c(n) \leq d(n)$, thus $c(n)/n \to 0$. Dividing by n the numerator and the denominator of both fractions and noting that $k(n)/n \to 0$, we can conclude that $\bar{\mu}_k(a_1^k|x_1^n) \to \mu_k$.

II. Assume that if the block x_i is censored, it has a label (subindex) equal to 1, $I(x_i) = 1$, and otherwise $I(x_i) = 0$. Let introduce $B_k \in A^k$ such that $B_k = \{x_i^{i+k-1}, \exists j \in \{i,\ldots,i+k-1\} : I(x_j) = 1, i \in \{1\ldots n-k+1\}\}$. B_k are all blocks eliminated. Applying the Birkhoff's ergodic theorem (see [43], Section 1.3) with the characteristic function χ_B, we get that $P(B_k|x_1^n) \to \mu(B_k)$. Taking $D_{a_1^k} = B_k \cap [a_1^k]$ and proceeding in the similar way, we also get that $P(D_{a_1^k}|x_1^n) \to \mu(D_{a_1^k})$. Here $[a_1^k] = \{x_i^{i+k-1} : a_j = x_{j+i-1}, j \in \{1,\ldots,k\}, i \in \{1,\ldots,n-k+1\}\}$.



(a) week 1

(b) week 2

(c) week 3

(d) week 4

Fig. 8. Experimental probabilities associated with the entropy for 4 cases with the new inefficiencies of the ETF DIA.

Now, $\frac{d(n)}{n} \to \mu(B_k) = C$ and by independence $\frac{c(n)}{n} \to \mu\left(D_{a_1^k}\right) = C\mu_k\left(a_1^k\right)$. Therefore,

$$\frac{\widehat{\mu}_k^n(n-k+1)-c(n)}{n-k+1-d(n)} = \frac{\widehat{\mu}_k^n\frac{(n-k+1)}{n} - \frac{c(n)}{n}}{\frac{n-k+1}{n} - \frac{d(n)}{n}} \to \frac{\mu_k - C\mu_k}{1-C} = \mu_k$$

Case III is in contradiction with the fact $n-k+1-d(n) = n_{blocks} > 0$ since $\frac{n-k+1}{n} - \frac{d(n)}{n} \to 0$.

Therefore, the values of k such that $k(n_{blocks}) \leq \log(n_{blocks})/(h+\epsilon)$ are admissible in the case of missing data. Q.E.D.

In practice, we take $max(k : k < log(n_{blocks}(k)))$. The length of Bernoulli sequences, l needed to construct CI is also taken according to the number of blocks. More precisely, we take $l = n_{blocks} + k - 1$.

Appendix E. About new inefficiencies

In this section we analyze such particular weeks which are detected as inefficient after filtering out spurious 0-returns, but the standard methodology without filtering such 0-returns classifies them as efficient. We define an *experimental probability* as the fraction of entropy values calculated for 10^5 Bernoulli sequences that are greater than the entropy of the time series.

When we filtered out 0-returns, we determined 12 weeks with inefficiency for the ETF DIA, but 4 of them were not classified in such a way before the 0-filtering procedure. In order to investigate the dynamics of entropy, we take a new week with inefficiency, one week before, and one week after. We move the weekly time window day by day, so that the week with the new inefficiency is the 6th. For each week, we find the experimental probability associated with the entropy value. See Fig. 8 for the results for all four new weeks with inefficiency. As before, we define a week with inefficiency if the experimental probability is larger than 0.99.

We notice that for the first and the fourth cases the value of the experimental probability before the 0-filtering was >0.98. That is, the experimental probability does not change its value significantly after the 0-filtering but crosses the significance level that is set to be 0.99. For the third case, the 7th week is classified as inefficient before and after the 0-filtering. It has 80 % data in common with week 6, which is classified as inefficient only after the 0-filtering, which makes results more coherent for two adjacent weeks. However, for the second case, a week with detected inefficiency before the 0-filtering is the third, which has only 40 % data in common with week 6.

We conclude that one of the reasons for the appearance of new inefficient weeks is the possibility of the algorithm to detect weeks with inefficiency that were not detected due to a high level of confidence equal to 99 %. On the contrary, as in the example of the second week, sometimes identifying inefficient weeks may be the case of the extreme realization of the test statistic: the estimate of entropy is in the 1 % tail of the entropy distribution associated with the Bernoulli process.

Appendix F. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chaos.2022.112403.

References

[1] Samuelson PA. Proof that properly anticipated prices fluctuate randomly. Ind Manag Rev. 1965;6(2):41–9.

[2] Fama EF. Efficient capital markets: a review of theory and empirical work. J Financ. 1970;25(2):383–417. https://doi.org/10.2307/2325486.

[3] LeRoy SF. Efficient capital markets and martingales. J Econ Lit. 1989;27(4):1583–621.

[4] Ito M, Sugiyama S. Measuring the degree of time varying market inefficiency. EconLett. 2009;103:62–4. https://doi.org/10.1016/j.econlet.2009.01.028.

[5] Cajueiro D, Tabak B. Ranking efficiency for emerging markets. Chaos Solitons Fractals. 2004;22:349–52. https://doi.org/10.1016/j.chaos.2004.02.005.

[6] Cajueiro D, Tabak B. Ranking efficiency for emerging markets ii. Chaos Solitons Fractals. 2005;23:671–5. https://doi.org/10.1016/j.chaos.2004.05.009.

[7] Giglio R, Matsushita R, Figueiredo A, Gleria I, Silva SD. Algorithmic complexity theory and the relative efficiency of financial markets. EPL (Europhysics Letters). 2008;84(4):48005. https://doi.org/10.1209/0295-5075/84/48005.

[8] Shmilovici A, Alon-Brimer Y, Hauser S. Using a stochastic complexity measure to check the efficient market hypothesis. Comput Econ. 2003;22:273–84. https://doi.org/10.1023/A:1026198216929.

[9] Pincus S, Gladstone I, Ehrenkranz R. A regular statistic for medical data analysis. J Clin Monit. 1991;7:335–45. https://doi.org/10.1007/BF01619355.

[10] Alvarez-Ramirez J, Rodriguez E, Alvarez J. A multiscale entropy approach for market efficiency. Int Rev Financ Anal. 2012;21:64–9. https://doi.org/10.1016/j.irfa.2011.12.001.

[11] Pincus S, Kalman R. Irregularity, volatility, risk, and financial market time series. Proc Natl Acad Sci U S A. 2004;101:13709–14. https://doi.org/10.1073/pnas.0405168101.

[12] Duan W-Q, Stanley H. Volatility, irregularity, and predictable degree of accumulative return series. Phys Rev E Stat Nonlinear Soft Matter Phys. 2010;81:066116. https://doi.org/10.1103/PhysRevE.81.066116.

[13] Oh G, Kim S, Eom C. Market efficiency in foreign exchange markets. Physica A. 2007;382(1):209–12. https://doi.org/10.1016/j.physa.2007.02.032.

[14] Molgedey L, Ebeling W. Local order, entropy and predictability of financial time series. Eur Phys J B. 2000;15:733–7. https://doi.org/10.1007/s100510051178.

[15] Risso W. The informational efficiency and the financial crashes. Res Int Bus Financ. 2008;22:396–408. https://doi.org/10.1016/j.ribaf.2008.02.005.

[16] Mensi W, Aloui C, Hamdi M, Nguyen K. Crude oil market efficiency: an empirical investigation via the shannon entropy. Econ Int. 2012;129:119–37. https://doi.org/10.3917/ecoi.129.0119.

[17] Oh G, Kim HYong, Ahn S-W, Kwak W. Analyzing the financial crisis using the entropy density function. Physica A. 2015;419:464–9. https://doi.org/10.1016/j.physa.2014.10.065.

[18] Risso WA. The informational efficiency: the emerging markets versus the developed markets. Appl Econ Lett. 2009;16(5):485–7. https://doi.org/10.1080/17446540802216219.

[19] Ruiz MDC, Guillamón A, Gabaldón A. A new approach to measure volatility in energy markets. Entropy. 2012;14(1):74–91. https://doi.org/10.3390/e14010074.

[20] Ahn K, Lee D, Sohn S, Yang B. Stock market uncertainty and economic fundamentals: an entropy-based approach. Quant Finan. 2019;19(7):1151–63. https://doi.org/10.1080/14697688.2019.1579922.

[21] Pele DT, Lazar E, Dufour A. Information entropy and measures of market risk. Entropy. 2017;19:226. https://doi.org/10.3390/e19050226.

[22] Dionisio A, Menezes R, Mendes D. An econophysics approach to analyse uncertainty in financial markets: an application to the portuguese stock market. Eur Phys J B. 2006;50:161–4. https://doi.org/10.1140/epjb/e2006-00113-2.

[23] London M, Evans A, Turner M. Conditional entropy and randomness in financial time series. Quant Finan. 2001;1(4):414–26. https://doi.org/10.1088/1469-7688/1/4/302.

[24] Tsallis C. Possible generalization of boltzmann-gibbs statistics. J Stat Phys. 1988;52:479–87. https://doi.org/10.1007/BF01016429.

[25] Gradojevic N, Caric M. Predicting systemic risk with entropic indicators. J Forecast. 2017;36:16–25. https://doi.org/10.1002/for.2411.

[26] Gençay R, Gradojevic N. The tale of two financial crises: an entropic perspective. Entropy. 2017;19(6):244. https://doi.org/10.3390/e19060244.

[27] Zhao X, Ji M, Zhang N, Shang P. Permutation transition entropy: measuring the dynamical complexity of financial time series. Chaos Solitons Fractals. 2020;139:109962. https://doi.org/10.1016/j. chaos.2020.109962.

[28] Marschinski R, Kantz H. Analysing the information flow between financial time series. Eur Phys J B. 2002;30:275–81. https://doi.org/10.1140/epjb/e2002-00379-2.

[29] Kwon O, Yang J-S. Information flow between composite stock index and individual stocks. Physica A. 2008;387(12):2851–6. https://doi.org/10.1016/j.physa.2008.01.007.

[30] Schreiber T. Measuring information transfer. Phys Rev Lett. 2000;85:461–4. https://doi.org/10.1103/PhysRevLett.85.461.

[31] Liu A, Chen J, Yang SY, Hawkes AG. The flow of information in trading: an entropy approach to market regimes. Entropy. 2020;22(9). https://doi.org/10.3390/e22091064.

[32] Calcagnile L, Corsi F, Marmi S. Entropy and efficiency of the etf market. Comput Econ. 2020;55:143–84. https://doi.org/10.1007/s10614-019-09885-z.

[33] Bandi FM, Kolokolov A, Pirino D, Renò R. Zeros. Manag Sci. 2020;66(8):3466–79. https://doi.org/10.1287/mnsc.2019.3527.

[34] Kolokolov A, Livieri G, Pirino D. Statistical inferences for price staleness. J Econ. 2020; 218(1):32–81. https://doi.org/10.1016/j.jeconom.2020.01.021.

[35] Brownlees C, Gallo G. Financial econometric analysis at ultra-high frequency: data handling concerns. Comput Stat Data Anal. 2006;51(4):2232–45. https://doi.org/10.1016/j.csda.2006.09.030.

[36] Cont R. Empirical properties of asset returns: stylized facts and statistical issues. Quant Finan. 2001;1(2):223–36. https://doi.org/10.1080/713665670.

[37] Wood RA, McInish TH, Ord JK. An investigation of transactions data for nyse stocks. J Financ. 1985;40(3):723–39. https://doi.org/10.2307/2327796.

[38] Bulla J, Bulla I. Stylized facts of financial time series and hidden semi-markov models. Comput Stat Data Anal. 2006;51(4):2192–209. https://doi.org/10.1016/j.csda.2006.07.021. nonlinear Modelling and Financial Econometrics.

[39] Sucarrat G, Grønneberg S. Risk estimation with a time-varying probability of zero returns. J Financ Economet. 2020:1–32. https://doi.org/10.1093/jjfinec/nbaa014.

[40] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B Methodol. 1977;39(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[41] Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[42] Marton K, Shields PC. Entropy and the consistent estimation of joint distributions. Ann Probab. 1994;22(2):960–77. https://doi.org/10.1214/aop/1176988736.

[43] Shields PC. The ergodic theory of discrete sample paths. American Mathematical Society; 1996.

[44] Grassberger P. Entropy estimates from insufficient samplings, arXiv: data analysis, statistics and probability. URL. https://arxiv.org/abs/physics/0307138v2; 2008.

[45] Fama EF, Blume ME. Filter rules and stock-market trading. J Bus. 1966;39(1):226–41. https://doi.org/10.1086/294849.

[46] Tsutsui Y, Hirayama K, Tanaka T, Uesugi N. Can we make money with fifth-order autocorrelation in Japanese stock prices?, ISER discussion paper 0639. Institute of Social and Economic Research, Osaka University; 2005.https://doi.org/10.2139/ssrn.770565.

[47] Lo AW, MacKinlay AC. Stock market prices do not follow random walks: evidence from a simple specification test, working paper 2168. National Bureau of Economic Research; February 1987.https://doi.org/10.3386/w2168.

[48] Bandi FM, Pirino D, Ren´o R. Excess idle time. Econometrica. 2017;85(6):1793–846. https://doi.org/10.2139/ssrn.2199468.

[49] Glosten LR, Milgrom PR. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. J Financ Econ. 1985;14(1):71–100. https://doi.org/10.1016/0304-405X(85)90044-3.

[50] Kyle AS. Continuous auctions and insider trading. Econometrica. 1985;53(6): 1315–35. https://doi.org/10.2307/1913210.

[51] Jacod J, Li Y, Zheng X. Statistical properties of microstructure noise. Econometrica. 2017;85(4):1133–74. https://doi.org/10.3982/ECTA13085.

[52] Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4. https://doi.org/10.1214/aos/1176344136.

[53] Jones RH. Maximum likelihood fitting of Arma models to time series with missing observations. Technometrics. 1980;22(3):389–95. https://doi.org/10.1080/00401706.1980.10486171.