ISSN: 2029-4913 / eISSN: 2029-4921

Article in press

TECHNOLOGICAL and ECONOMIC DEVELOPMENT of ECONOMY

https://doi.org/10.3846/tede.2026.28512

EMPIRICAL STUDY REGARDING THE ANALYSIS OF INVESTMENT BEHAVIOR ON HIGH-FREQUENCY TRADING AT THE ROMANIAN STOCK EXCHANGE

Florin Cornel DUMITER [1], Klaus Bruno SCHEBESCH [1], Florin TURCAȘ [2], Marius BOIȚĂ [1]

[1] *Faculty of Economics, IT and Engineering, "Vasile Goldiș" Western University of Arad, Arad, Romania*
[2] *IFB Finwest SA, Arad, Romania*

Article History:

- received 20 March 2026
- accepted 26 August 2026
- first published online 8 October 2026

Abstract. The role and effects of high-frequency trading in financial markets are manifold, spanning from improved (true) price discovery to predatory mechanisms amongst stock market trading agents. The effects of using different types of AI and Machine Learning – while technically capable to react to the accelerated pace of trading opportunities – remain largely unclear. We concentrate on designing potentially useful "generalized prompts" for AI-based trading models, which is the main task from a market analyst's point of view. With a sufficient number of such relevant and often also structurally highly diverse inputs, the computationally intensive modeling and validation part can then be relegated to an otherwise autonomously working AI application. We offer a principled approach of proposing derived time series and diverse filters for time- and resolution-localized information for high-frequency stock market analysis and trading models.

Keywords: high-frequency, financial data, stock exchange, fuzzy decision, emerging economies, sentiment index.

JEL Classification: F32, G15, G18, C63.

Online supplementary material: Supporting information for this paper is available as online supplementary material at https://doi.org/10.3846/tede.2026.28512

Corresponding author. E-mails: ***fdumiter@yahoo.com; dumiter.florin@uvvg.ro***

1. Introduction

Contemporary research into high-frequency trading is deeply rooted in the evolution of market microstructure theory. The fundamental logic of high-frequency trading arises from classic models such as Kyle (1985) and Glosten and Milgrom (1985) which describe order flow, information asymmetry and price information. In modern high-frequency literature, these theoretical underpinnings are executed to account for ultra-low latency order routing, real-time processing, and endogenous liquidity provision. For instance, some empirical recent trends highlight that high-frequency trading participants act both as liquidity providers and informed traders. Their rapid-fire execution allows them to arbitrage away fleeting mispricing, but also creates the potential for adverse selection against slower traders, underpinning current debates on market fairness and efficiency.

Copyright © 2026 The Author(s). Published by Vilnius Gediminas Technical University

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Other trends further debate the equilibrium effects of high-frequency trading: while some posit that these traders contribute to tighter spreads and improved price discovery, others warn that their activities may amplify short-term volatility and encourage predatory behaviors during stressed market conditions.

These theoretical models are now complemented by agent-based simulations and machine learning models, as explored in recent academic debates. More important is the fact that the literature suggest operating the high-frequency trading within a dynamic feedback loop, where strategy adaptation and state dependence can result in nonlinear, sometimes chaotic market outcomes. Here one may face an overwhelming diversity of possble approches, without having any principled guidance of how to proceed from the rather nonstandard data structures of empirical high-frequency stock trading.

Objects of research and research objectives

In recent times, financial trading institutions became increasingly aware of the necessity of storing detailed trading activity at the most basic event level, thus producing very long time series at least by econometric standards. Hence, high-frequency stock market trading data, such as time-stamped trading prices and trading volumes are becoming readily accesible for research. However, owing to the very nature of real life stock market trading events, the entries of these time series appear at very irregular points in physical (or "wall clock") time. There are not just complicated change-of-direction dyanamics (the "truning points"), and associated volume dynamics (contributing to "volatility") but also a peculiar mix of quasi-periodic and seemingly chaotice lengths ("delta-time steps") between succeding single trading events. Furthermore, there is no obvious correlation between the fine structure of the delta-time step dynamics between stocks of different listed firms. The difficulties associated with the latter aspects reveal temselves only if a concrete computational modeling of these time series is attempted. From a maximal time resolution one may proceed to gradually decrease this resolution (in principled steps from millisecond up to monthly or even yearly resolution) thus obtaining more homogenous lengths of succesive time points. In so doing, the amount of available information tends to decrease while the opportunities of transversal comparison or, for example, that of "vector" autoregressive modeling increases. How to do this in a tractable manner (avoiding modelling overkill, etc.) is not at all clear.

This leads to a structural hypothesis, which asserts that in varying the time resolution may offer a key frame to analyzing, interpreting and modeling high-frequency stock market time series.

Research gap

Integrating variation of time resolutions into a procedure by which to to obtain a selection of the types of statistical time series models and other associated information generators, which then to propose as inputs into an "universal" AI application is a neglected aspect in the literature.

The main added value of the paper

The paper offers examples of a multi-view analysis of four high-frequency stock market data sets. These data sets contain stock quotation data for firms of four completely different

industral and service sectors and with different overall trading histories. Such data pose a series of challenges which we overcoome by varying and analyzing different time resolutions, proposing new derived time series and by setting out a program of analyzing replicability, randomness, and extracting distributional as well as correlation-based information of time-located and resolution-dependent time series. These are then cpandidates for problem-adapted inputs ("prompts") to modern, sufficiently powerful AI-modeling applications.

The structure of the article is as follows: The second section presents and analyzes the current state of the art by revealing the most important theoretical and empirical studies in the economic literature regarding high-frequency trading and their impact on the stock market. The third section presents the research methodology and research models that are used in this research, as well as the main rationale behind their usage. The fourth section reveals the empirical results in light of the article's scope, objective, and research methodology, presenting the main important empirical and practical aspects of high-frequency trading. The last section draws the final considerations and important conclusions, and also gives interesting insights regarding the future paths of high-frequency trading as well as some policy implications for traders and brokers.

2. Literature review

High-frequency trading remains one of the most transformative developments in modern financial markets, continously shaping liquidity, volatility, and the very fabric of market infrastructure. Over the past decade,the proliferation of advanced algorithmic strategies, miliseconds execution, and data-driven approaches has not only revolutionized trading tactics but also spured rigorous academic scrutiny. Given its envolving complexity, recent research in the field explores the mutifaced mechanism of high-frequency trading, the role of cutting-edge technology, the consequences for market stability and integrity, and the regulatory environments seeking to acommodate or mitigate its high-speed effects. In this section, it is provided a comprehensive ad current analysis regarding theoretical, technological, empirica land ethical dimensions of high-frequency trading, unveiling consensus and convention while identifying emerging trends and unresolved debates.

Saranj and Zolfaghari (2025) analyze the Turkish stock market in 2 different time frame zones: the first one is the pre-COVID-19 period, and the second one is the actual COVID-19 period. The authors use the stock index fluctuations and the investment group's different trading behavior, and conclude that there is a significant fluctuation trend on the stock market related to which individual investment reacts, highlighting that behavioral aspects of the stock market are very important. The linkages between the stock market and COVID-19 disease were also investigated by authors as Wang and Liu (2022) by linking the Chinese stock market to patients with different COVID-19 in order to test the market reactions; the authors have suggest that the COVID-19 pandemic has caused stock market returns volatility and market fluctuations.

Other studies as Nimalendran et al. (2024), examine the options market liquidity in the context and circumstances of high-frequency trading and reach the conclusion that it can be identified a strong negative correlation between open market liquidity and high-frequency trading strategies and via two channels: first, the arbitrage channel; second, the informed trading

channel. Moreover, deal with the relationship between market stability and high-frequency trading with a special focus on financial cryptocurrencies; the authors have come to the conclusion that there are some complex negative and positive issues between the two aspects, with a strong need oriented for heightened volatility and price manipulations.

Hossain (2022) connects the LOB order book information to the high-frequency trading for the time-stamped intra-day quotes and comes to the conclusion that although there are some counterintuitive connections in developing markets, the narrow and wider quoting can be reached in the high-frequency trading strategies. In the aftermath, Fox et al. (2018) combine aspects of dark pool operators, high-frequency trading strategies, and electronic front running and suggest that the trading systems must have as a background the principal agent problem and adverse selection. Another very interesting study dealing with the problem of flash events and high-frequency trading is the one conducted by Desagre et al. (2025), by introduces a call sequency flash events and finds no evidence of connections between extreme price shocks which are triggered by high-frequency trading strategies.

An interesting study is the one developed by Gomber et al. (2011), analyzing the high-frequency trading system from both the algorithmic trading and high-frequency trading perspectives and reaching the conclusions that there are several advantages as market liquidity, implementing good trading strategies, enhancing market quality, and reducing adverse selection costs. Aquilina et al. (2021) combine high-frequency trading with latency arbitrage, suggesting that there is a high degree of race concentration for a small number of six firms, which are responsible for over 80% of losses and winnings. The financial markets are analyzed in the context of internationalization, developments, and technological shifts, with the identification of common characteristics of high-frequency trading strategies and techniques, especially observed in the regional financial markets from Asia (Kauffman et al., 2015).

The connections between high-frequency trading and IT investments and infrastructure were analyzed by Clapham et al. (2023) in terms of infrastructure and fast connections to exchanges, and they concluded that the quality of security markets can be enhanced by positive economic spillovers generated by high-frequency trading technologies. Subsequently, Menkveld (2011) made an interesting study connecting the high-frequency trading with recent fragmentation in equity; the author has highlighted that the fees are very important for the traders, the low fees are very attractive features, and that the new entry venues are key elements for a successfull market.

Adrian (2016) made an interesting study linking the high-frequency trading to institutional investors' premier information on the stock exchange; the author uses the terminology of front-running to express how traders use the access to information from a superior and privileged angle, and examines the market reaction to such trading practices. Jung et al. (2024) explore the connections between deviations of the stock market prices with high-frequency trading, with a special focus oriented towards intrinsic values; the authors' results suggest that for a small sample of firms, high-frequency trading generates an exogenous shock across the market. Serrano (2020) made an interesting review of the economic literature regarding the soundness of high-frequency trading and pointed out that adverse selection can create systemic risk in the market, and that it is necessary to introduce a certain limitation stop in the speed of trading.

Menkveld (2016) engaged in a qualitative analysis regarding the high-frequency trading literature existing in the economic field in two directions: first, in terms of costs and benefits from the theoretical studies; second, an empirical review analysis in terms of predictions and model assumptions. The high-frequency trading and its impact on the liquidity market from a literature review perspective were analyzed by Ersan et al. (2021) for the last two decades and in the developed countries; the authors conclude that there can be identified two underlying factors for the evolution of the market: first, intraday shocks; second, the liquidity provisions which can generate externalities on the market. Gupta (2022) underscores the important aspects of using artificial intelligence in high-frequency trading strategies by analyzing 24 articles from the economic literature in order to evaluate and assess the impact and regulation aspects of the financial market.

The effects of trade volume on stock price changes from the perspective of high-frequency trading were analyzed by Portnoy (2011) in the context of time series in order to determine if there is any negative or positive correlation of the stock market; the conclusions highlight the correlation between the volume of trade and the degree of volatility regarding the stock market. Ivashchenko (2025) explains that a linkage between trading activity and corporate can be identified, suggesting that ownership of the institutional bond is triggered by liquidity dry-ups. Moreover, Sui et al. (2025) linked the Australian Football League Final to the market quality and investment behavior pattern, and market quality; the authors pointed out that it can be seen that some short-term distortions on the stock market are driven by attention shocks developed by retail investors.

Acemoglu et al. (2015) connect the financial networks with financial institutions and the contagion risk, revealing that the fragility of the financial system depends on the propagation of the shocks affecting the financial institutions, and financial stability can be enhanced by a more dense financial network. An interesting study dealing with the concentration patterns in the financial market was developed by Metelski and Sobieraj (2024) with a special focus oriented on the impact of major political and social events on the trading volume concentration; the authors conclude that the concentrated portfolios need to be connected with the risk and uncertainty factors. Other authors use Rényi transfer entropy and Diebold-Yilmaz connectedness to measure risk information of the global markets (Jia, 2025); the author suggests that there are significant differences between European and US stock markets, the independence of the Chinese stock maket, and the vulnerability of the Japanese stock market.

A study regarding the connections between emission-regulated enterprises and the carbon emission trading system of China is analyzed by Wang et al. (2025). In the context of overall spillover, the authors conclude that it can be offered a set of guidelines for risk spillovers in terms of characteristics, structural, and dynamics of the firms in order to optimize the investment strategies. Tripathi and Rengifo (2025) present some interesting insights regarding the Nasdaq equity market in terms of commission-free FinTech and fractional trading, suggesting a high degree of volatility and altered liquidity in terms of modern equity markets from the informational and resilience point of view. In subsequent work, the high-frequency trading in the Chinese stock market using a hybrid neural network was developed by Li et al. (2023) for the accurate prediction of time series; the conclusions of the study reveal fast computing due to efficient storage management with a computational sequence of previous trading data and current trading data.

The importance of econophysics and its applications to the stock market was analyzed by Turcaș et al. (2022) in terms of the validation of several underlying stock strategies from different domains like chemistry, mathematics, and physics. The authors have found that the investment strategies can have as a background the foundation of several exact sciences, but in order to be successfully implemented, it is needed that the initial assumptions are validated in the market. In the aftermath, Dumiter et al. (2023) made a triangle assessment between the stock market, technical analysis, and quantitative sentiment indices study on the American stock market, concluding that it manifests a strong correlation between behavioral finance manifested with sentiment indicators, technical analysis applications in practice, and the evolution of the stock market.

The problem of high-frequency finance from the heterogeneous transition was studied by Turiel and Aste (2022) from the propagation of shocks and systemic risk point of view, reaching the conclusions that with the withdrawals of liquidity from the high-frequency trading systems, there are associated large shocks and shock propagation phases. Moreover, Lei et al. (2021) use a deep learning model and a web search index in order to forecast the high-frequency financial data volatility, with the outcome of providing a more equitable and sound method for big data volatility forecasting. Lu et al. (2020) analyze the Chinese stock market in terms of a high-frequency view regarding the trading imbalances phenomenon, reaching out the conclusion that the trading polarity can be estimated in real time using the market sentiment index in a data-driven manner.

Stübinger and Schneider (2019) developed a study to assess the high-frequency data of the S&P500 with a fully-fledged statistical arbitrage strategy, with the outcomes that the developed trading algorithm can register good results in a risk-return based on a multitude of characteristics. Another interesting study was conducted by Donadio et al. (2022), oriented towards the construction and further improvement of high-frequency trading systems using C++ and Java basics. The machine learning approach applied to high-frequency trading was further analyzed by Goudarzi and Bazzana (2023) with intraday precise identifications from proxy indicators, using a probabilistic model and fuzzy logic techniques. Hayashi and Nishide (2024) elaborate on a model using slow and fast traders, and with equation systems and numerical calculations; the authors have come to the conclusion that the fast trader is a liquidity provider and has the advantages of trade frequency.

To subsume the economic literature presented above, it provides a compact overview of the most recent and influential research articles regarding high-frequency trading. Each study and article presented has been selected due to its empirical rigor, relevance to current debates of the theoretical innovation and developments. As can be seen from the current state of the art, the topics gravitate around the market impact, regulatory effectiveness, and technological progress.

3. Research methodology

Our research methodology consists in transforming the raw data of high-frequency financial time series of our four stock market listed companies into inputs suitable for general AI applications for training or successful trading systems that can learn. This includes a close analysis of these time series which contain time-stamps, trading prices and trading volumes

at their maximal time resolution, and, subsequntly, transforming them into aggregate, multi-resolution and also complementarily derived time series, which aid in generating a stepwise process for building increasingly categorical or qualitative features. In order to increase their potential value of inputs for the named AI applications assessment of randomness, replicability (in our study by using wavelets), distributional featurs and time-locatable correlation structures are added. After effectively training such AI models (which is not part of this study) the resulting evaluation of tnput relevance may be used in fuzzy decision models within financial recommender systems.

3.1. The high-frequency data: properties and data models

High-frequency (HF) financial data refers almost as a rule to stock exchange quotations. A variety of views conditioned on different time resolutions are possible and may also be useful to different ends. Various difficulties for comparative data analyses may appear owing to irregularly spaced time intervals of data records. This needs extensive pre-processing, especially if data are collected at a high time resolution or high-frequency. In principle, HF financial data offer manifold trading opportunities, but they can also be highly redundant over time periods. As frequency increases, time spacing, trading volumes, and other non-price information become more important. In the sequel, we therefore describe ways of how to resolve various challenges that irregularly spaced HF data tend to pose to principled data analysis and modeling.

3.2. The primary data at the finest resolution

Our data refer to high-frequency stock exchange transactions from different Romanian stock markets. The single database entries are records of transaction events. They are recorded at the maximal time resolution handled by the data provider, in our case, at the level of milliseconds. In taking this restriction into account, they are recorded at the maximum time resolution. In this respect, financial data are different from most physical process data, where the maximal possible time resolution is either unknown, or – if it is known – it may be inaccessible for any experimental data recording means available. If every recorded point in time contained exactly one transaction, then high-frequency financial data would have an exceptionally solid foundation. Note, however, that in practice, high-frequency financial data may contain more than one recording at the maximal time resolution. In our data, we also find many cases with multiple transactions (trading events) recorded with the same time stamp. Hence, there are more records with different time stamps. The features of such a data record are:

- Company label,
- (Stock-)Market label,
- Time stamp in the sortable long format [*yy:mm*: *dd.hour:minute*: *sec.millisec*],
- Price of the share at the moment of the transaction,
- Trade volume (number of shares transacted), and
- Transaction activity “ticket”.

Hence, a (large) total number of $N > 0$ such data records have the components

- $C_j \in \{s_1, ..., s_n\}$, from the set of company symbols,
- $M_k \in \{l_1, ..., l_m\}$, from the set of market labels,
- t_1 from the set of consecutive time stamps at the technically maximal resolution, with $i \in \{0.1, ..., T_{max} \leq N\}$, the cases with "<" account for multiple trading events at t_i,
- $p_i^{jk} > 0$ the transaction price realized by the company C_j on the market M_k at the trading time t_i,
- $E_i > 0$ the number of transactions at trading time t_i, with $\sum E_i = N$,
- and $A_i = \left\{a_1^l, ..., a_{Ei}^l\right\}$ monotonically increasing trade activity labels ("tickets").

The database has been received from the Bucharest Stock Exchange, Romania, and includes all transactions on the stock market regarding the four companies: ATB (pharmaceuticals), DIGI (telecom), SNP (petrol), and TLV (banking). The total time span covered by the data records is from the trading start of the year 2006 until inclusively July of 2025, more exactly, using the format [yy-mm-dd hh:min: sec.msec], records are formed

2006-01-03 10:15:14.470 until 2025-07-30 17:58:37.547,

with the number after the "seconds" entry being milliseconds. The total number of markets is $m = 9$. The nature of the markets is described by the labels REGS, ODDS, DEALS, POFS, POFA, POF, POFR, POFSV, and BUY-IN. However, starting from the year 2020, only REGS, DEALS, and POF are recorded. In the following, we neglect this information and we highlight filtering by companies: The activity domains of the four companies from the data set are: ATB (pharmaceuticals), DIGI (telecom), SNP (petrol), and TLV (banking). Hence, a randomly extracted list of four consecutive looks like (Symbol refers to the company):

Symbol	Market	Time stamp	Trading price	Trading volume
...				
1 | SNP	REGS	2006-03-30 12:38:24.100	0.57	500
2 | SNP	ODDS	2006-03-30 12:38:26.420	0.56	293
3 | SNP	REGS	2006-03-30 12:38:36.687	0.57	20000
4 | SNP	REGS	2006-03-30 12:38:40.963	0.57	500
...				

Considering the named activity domains of the four companies, the prior expectation is that the four companies should list fairly independently, and, at least during normal trading times, their valuation should not be significantly associated or correlated according to any common measure. However, statistical tests should confirm or reject this in the first place, which is a candidate working hypothesis to be discussed in the sequel.

Figure 1 depicts the transformation of the raw data into more useful, transformed, and standardized data representations. Usefulness is determined by the needs of data series comparisons and time series analyses and modeling in order to serve goals like information extraction, explanation, and, possibly, forecasting. Figure 1 stands for stylized, time-stamped stock quotation data of a single company. Hence, in the sequel, the procedure from the figure will be applied separately for all four companies of the database. The lower three timelines describe the real-time axis t with the irregularly appearing trading events t_i, at the maximal time resolution. However, at each t_i more than one trading event may be registered. In our

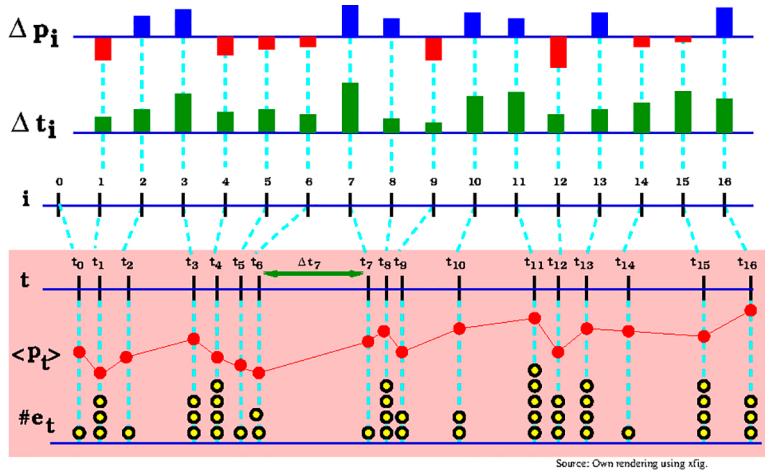


Figure 1. A stylized rendering of raw data as recorded (shaded lower region) on the natural time axis at irregular moments contained in the data base (shaded lower region) and derived time series on a homogeneous time-index axis (the three upper upper series)

data, this occurs quite often, even for any single company. This is accounted for by one or multiple stacked circles in the event timeline e_i. Hence, in the stylized example we have 0, 1, 2, ..., 16, or a total of 17 time stamps and a total of 1 + 3 + 1 + 3 + 4 + 1 + 2 + ... + 3 = N = 44 trading events. Stock prices < p_t > are recorded at the level of time stamps as averages of the respective trading events (variations are rare and very small).

From these data shown in the shaded lower half of Figure 1, we extract for further processing:

- the equally spaced, surrogate timeline $\{i\}$,
- the process of time differences $\{\Delta t_i\}$ between consecutive time stamps, and
- the (stationarized) stock price process $\{\Delta p_i\}$ of the company,

which are depicted in the upper half of Figure 1.

In addition to these derived time series, it is straightforward to add further (complementary) information as needed, for instance

- A trading volume time series $\{V_i\}$: every trading event number e_i corresponds to a set of trading volume counts (number of shares) $\{v_1^i, v_2^i, ..., v_{e_i}^i\}$, which may be used to express a price times quantity volume V_i, with $V_i = p_i \sum_1^{e_i} v_i$, transacted at surrogate time i (or real time t_i).
- A price-change volatility process $\{Z_i\}$: this records the magnitude of price changes, with $Z_i = |\Delta p_i|^\alpha$, with $\alpha > 0$, oftentimes equal 2, or more involved semipositive functions.
- Embedding dimension test using vector sequences $\{(p_i, p_{i-1}, ..., p_{i-p})\}_i$ with $i = p, p+1, p+2, ..., T_{\max}$ for instance, to compute correlation dimension for identifying some underlying dynamical system (Makarov et al., 2023).
- Schemes for value discretization, in our case applied to prices, volumes, coarse-graining time, and trading event time differences for extracting robust information from symbolic dynamics (for recent general approaches to symbolic dynamics for long time series, see Yalavarthi et al., 2025; Yang & Paparrizos, 2025).

Furthermore, data may be down-sampled to lower time resolutions within the [*yy:mm:dd hour:minute:sec.millisec*] convention. In doing so, the need to recalibrate irregular time sampling diminishes (and disappears altogether at low resolutions, such as days, months, years, ...), and one may find causal relations with other economic time series available at these lower resolutions. Such data associations will be described in the next section.

3.3. The modeling pipeline

(1) Analyses for comparing company stock evolutions, be they sporadically observed, down-sampled time series. Possible connections and influences along, for instance, the evolution of correlation structures, become more amenable in periodically observed (calendrical or metrical) time.

(2) Exogenous events (politico-economic decisions, company readjustments, crisis-related situations) may be supplied as "time-localized points of attention" alongside the available stock market trading activity.

(3) Analyses independent of overall time flow, for instance, sequence clustering, based on comparing selected feature combinations. Here, time may enter as a precedence or ordering relation, e.g., to point out possible causalities.

3.4. Finding pockets of dependencies in financial time series

The financial literature was and is emitting hypotheses about stock market "efficiencies", especially information efficiency, where the most extreme being Fama (1970), centering around the description of stochastic price processes. These processes may have various degrees of local regularities or pockets of identifiable dependencies. While such pockets can be spotted ex-post, their ex-ante onset and disappearance may not be more easy to forecast the direction of price change, which is known to be very tough to impossible. The degree of difficulty of the latter may vary with the time resolution of the financial time series under consideration. From an "uninformed" standpoint, but by taking into account the observation of time series, one may expect the following. The extent to which time series appear to be more or less erratic at different time resolutions would influence predictability. Hence, successful higher resolution (HF-)trading would lead to high trading profits (assuming "small enough" transaction costs). However, successful HF forecasts would be hard or impossible to achieve to any acceptable significance level, while low-resolution time series would be less erratic, hence, at least in principle, easier to forecast, but would also strongly limit trading profit potential as compared to the HF resolution cases.

4. Empirical findings

4.1. Description of the sporadically observed data at full resolution

In this section, we emphasize that the stock market data of the four companies with symbols ATB (pharmaceuticals), DIGI (telecoms), SNP (petrol, fossil energy), and TLV (universal banking, finance) do have different periods of overall activity. Thus, their activity over time is sporadic in two respects, namely (1) at the micro-level – as explained in Figure 1, and (2)

at the extended level, in that they show patchy periods of activity over the whole recorded time span.

For all four companies, the time span is always extended to the right, above the last observation of N = 3877601 up to 2^22 = 4194304, in order to accommodate wavelet-based analyses (which require lengths of 2^n). Figures S2–S9 from the supplementary material show these data in detail limited by graphical resolution.

Consecutive trade event time distances are measured in terms of milliseconds (the highest recorded resolution market events). Hence, in order to prevent long distances due to holidays, etc., to absolutely dominate all other values in magnitude, we plot them via ***log***(Delta t(i) + 1). Depending on the view required, these calendar-enforced trading pauses, reflected in the two upper lines of points from the plots, may be cut out (or neglected).

Note that the range of the (log-) distances of 5–15 in all density plots is well out of the range of the calendar-enforced pausing of the trading activities. One common feature of the last Figures 8–9 is the heavy occurrence of zero-time distances between trades. Together with very many zero increments in the price evolution (i.e., first differences, and, consequently, also volatility, to be discussed further below), this underlines the fact that our HF stock market time series of stock quotations from emerging financial markets are highly non-standard in terms of mainstream (econometric) time series analysis. This is mainly caused by sporadically produced (and observed) transactions with irregular periods of intermittency between active markets and periods of calm.

4.2. More data to be associated with down-sampled time resolutions

One should note that there is no naturally stringent inertia in the dynamic stock price process nor in associated trading volume processes, as can certainly be found in many natural or technical physical processes. The latter are often well described by measurements of a time-continuous dynamical system. Pricing and trading events may be totally erratic, but, nevertheless, price trajectories may temporarily behave as if there would represent psychological inertia, herding behavior, etc., of trading participants.

Figure 2 emphasis that this and the following three plots represent the view of filtering out only those minutes for which there is trading activity for all four stocks (therefore, just 8000+). The or numbers in the legend refer to the wavelet coefficients retained for the inverse transform. While 8 coefficients seem to oversimplify but 21 of them capture an acceptably detailed, hi-fidelity reproduction of the original data. The very first and last points of the inverse wavelet transforms are boundary artifacts of the method, present in all such images of the companies, and hence they can be cut out (neglected).

The same procedure is repeated for the processes of

- First differences of the trading prices of the four firms at different time resolutions,
- Time distances between trades of the four firms at different time resolutions, and,
- Trading volumes resp. trading values (price times number of transactions) of the four firms at different time resolutions.

The resulting data and plots are preliminary and not shown in this report.

As down-sampling proceeds (i.e., time resolution becomes successively coarser), the chances of cross-influences between different economic – not just stock-exchange

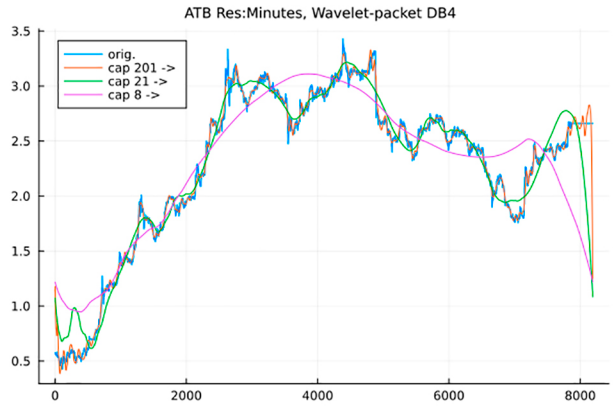


Figure 2. Wavelet smoothing of time series at the minute-resolution level for company ATB

related – variables increase. While at the millisecond level, cross-talk between transaction activities of different stocks is less likely, chances increase as time passes by, and eventually die completely out after some obsolescence levels. The latter may be, for instance, time lag thresholds, which also depend on the relations between stocks, e.g.

Influence (*Firm X* → *Firm Y*) may last longer than influence (*Firm X* → *Firm Z*).

In general, such influences would be determined in the statistical sense, using means or medians and similar.

An interesting approach would therefore be to develop a truly multi-(time)-scale model as was shown to be fruitful (if not essential) for the modeling of time series in other sciences concerned with the explanation and forecasting of dynamical processes.

4.3. First differences of high-frequency time series compared to random references cases

Over the entire trading period recorded in our HF stock market data, namely from the 3rd of January 2006 until the 30th of July 2025 (i.e., from 2006-01-03T10:15:14.470 to 2025-07-30T17:58:37.547 in milliseconds), a series of administrative-institutional decisions, that is, decisions, not necessarily related to market or trading dynamics, lead to several price adjustments for three of the four companies recorded. The prices of the companies ATB and SNP underwent a few minor adjustments, while company TLV saw more, and major such adjustments. Company DIGI had no administrative price adjustments. In the supplementary material we show plots of the adjustments made for the four time series.

First differences of (price-)level time series are a stationarized variant of the original, leveraging the process of directions of change, which is the most relevant for HF trading, that is, for developing a forecast-and-trading strategy, which is beyond the ubiquitous "Buy and Hold".

According to Figure 3, the complexity (degree of irregularity) of these derived time series may be computed and characterized. As will be seen in the sequel, all these profiles also hide very many zero-movements. Hence, additional views, as already proposed are also potentially

informative. Such are, e.g., time differences between successive tradings and time differences between successive non-zero first difference price movements.

Figure 4 shows that the first differences of the ATB price are compared to a normal random distribution with the same number of data points. Such a randomly generated time series would generate a qq-plot as in the upper right image. The actual first differences of the ATB price evolution are depicted in the lower left qq-plot. Note the large number of zero movements, which shapes the regression line of the middle quantile. It therefore seems instructive to show the qq-plot of the first differences without the cases with no price change (lower right). There is a visible departure of the ATB data from the Gaussian (max-uncertainty) case.

The images form Figure 5 highlight that the first differences of the DIGI price are compared to a normal random distribution with the same number of data points. The same logic applies as that from the description of Figure 6. Here, the proportion of zero movements is somewhat lower than that for company ATB. Overall, here too, there are visible differences to the Gaussian case (upper right), but there is also a broader band of agreement between data and the qq-regression, which points to more uncertainty in the price movement (lower right plot).

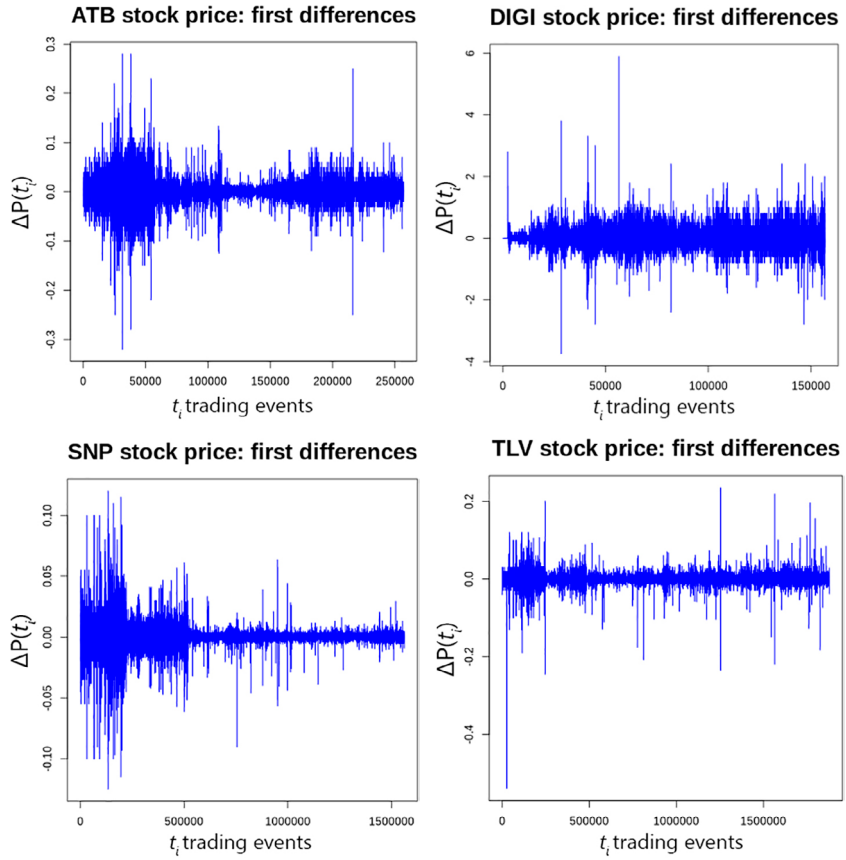


Figure 3. The profiles of the first price differences for the four companies. Implicitly, different volatility clustering patterns are visible

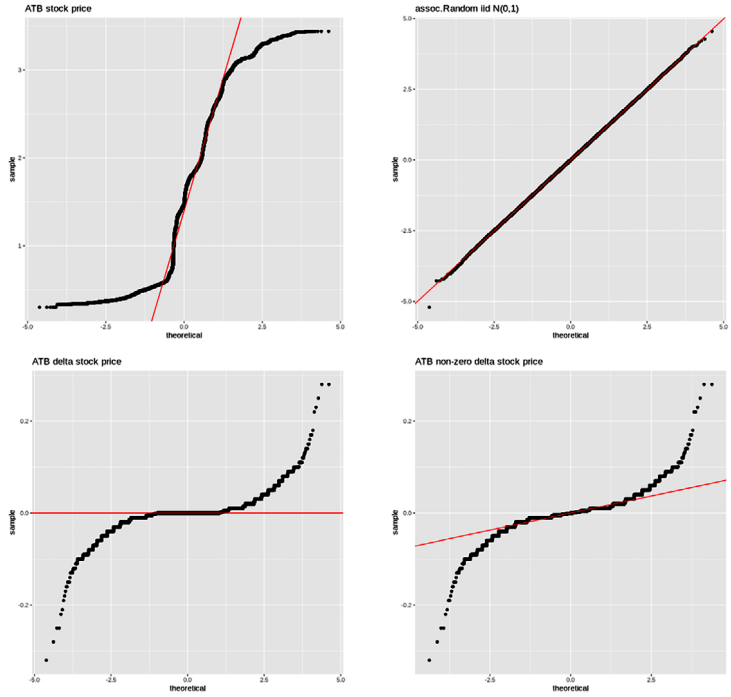


Figure 4. The historically recorded ATB stock price in a qq-plot (upper left)

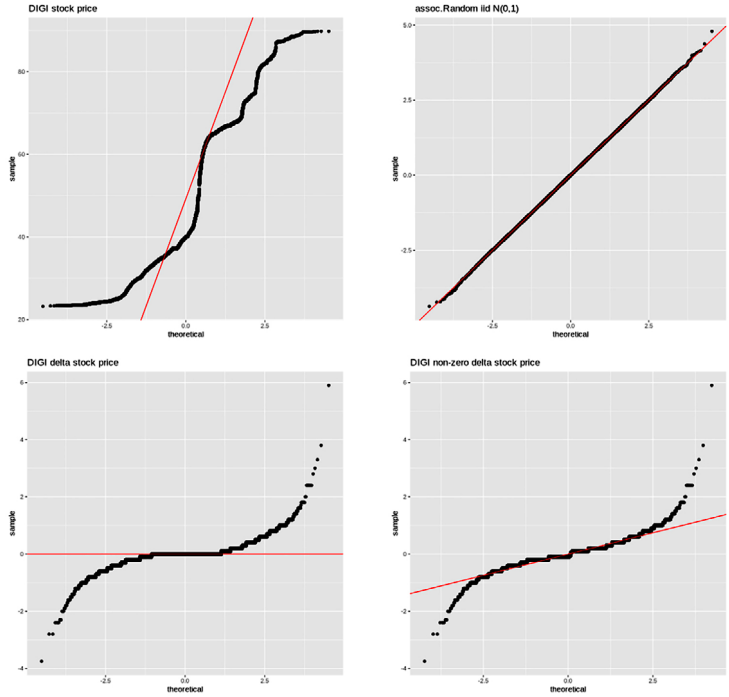


Figure 5. The historically recorded DIGI stock price in a qq-plot (upper left)

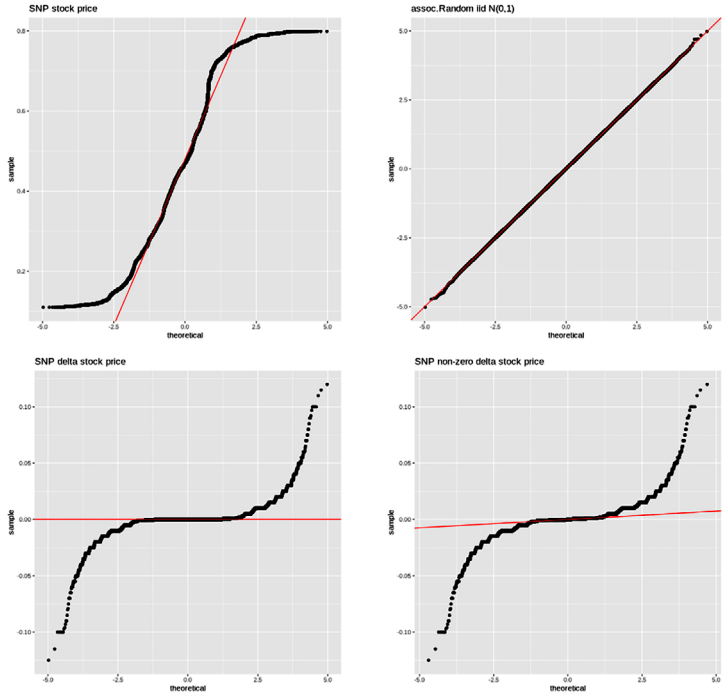


Figure 6. The historically recorded SNP stock price in a qq-plot (upper left)

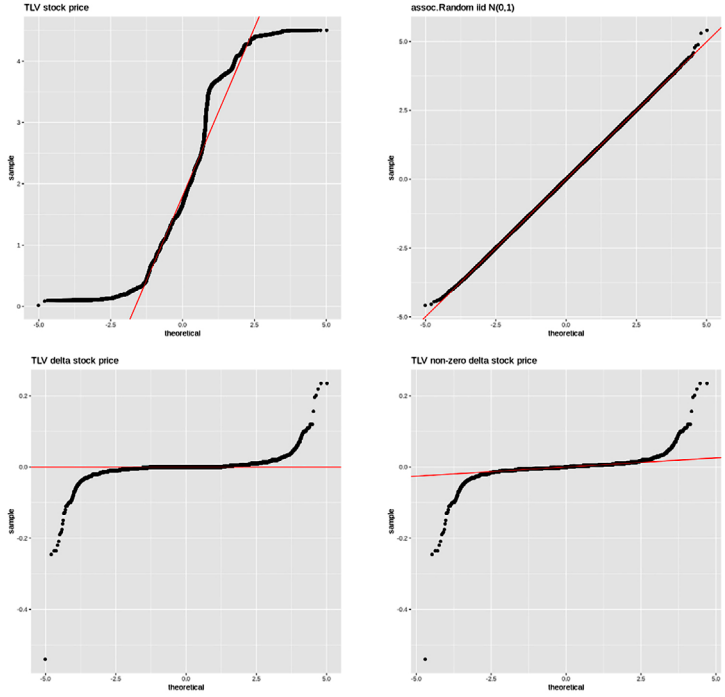


Figure 7. The adjusted TLV stock price in a qq-plot (upper left)

According to Figure 6, there are many more trading events of the company's stock (1567727 in total) than for company stocks ATB and DIGI (262999 and 162881, respectively), traversing more economic regimes, which shows up in a more Gaussian-like overall behavior in the price qq-plot. However, the first differences of the lower line plots show a strong departure from the Gaussian case (upper right), indicating that there might be some more detectable structure.

As in the case of company SNP, Figure 7 shows that there are also many trading events of the company's stock (1883994 in total). Overall, especially compared to SNP, there is more asymmetry in the data, both in the price-level and in the first difference price qq-plots. This also points to some detectable structure.

4.4. Characterizing the high-frequency event data by prediction potential

After analyzing the entire high-frequency (HF) data set from a descriptive perspective, one should attempt to develop a more prescriptive point of view. To this end, a simplification with regard to the tough time resolution problem will be adopted. For the time being, we define a portfolio of all four available companies, which is composed of the sum of the prevailing company prices p times their trading volume V, which is henceforth called the **portfolio value** *PoV*:

$$PoV_t = \sum\nolimits_{i\in\{1,2,3,4\}} p_{i,k(i,t)} V_{i,r(i,t)},$$

with $k(i, t)$, $r(i, t)$ times of the last prevailing price and volume.

This ensures that such a portfolio always exists, irrespective of the active market presence of all firms at the event time t. Hence, *PoV* will serve as a predictor (when lagged) and also as prediction target. The second feature (and restriction) of the time series models is that they use no other information than that available or derived from the trading event records of the Bucharest stock exchange. A typical input for the time series model (in the training and in the test-sets)

$$\left[PoV_t : \left(PoV_{t-1}, PoV_{t-2}, \ldots, PoV_{t-M}\right); \left(U_{t-1}, U_{t-2} \ldots\right)\right],$$

where *PoV* at t (left) is the prediction target, followed by a vector of lagged *PoV* inputs (middle), and possibly adjoint by other lagged variables (U's at right). In the sequel, the U's used contain the time difference between the present and the last trade. As both *PoV* and U are both positive or zero ("simultaneous" trades), with very high occasional peaks, we use $\log(PoV+1)$ and $\log(U+1)$ instead.

Training sets are chosen to start at random inside the whole time trading-time range, providing for sufficient time margins and last 500 or 600 trading events. The associated test sets start next to the end of their training set and last for the same number of trading events. Three batches of model runs are performed, producing 300, 2000, and 5000 prediction models each. The input structures have all incrementally lagged variables, namely (1) 3 *PoV* and 3 *U variables*, *followed by* (2) 10 *PoV* and 10 *U variables*, and (3) 30 *PoV* and 6 *U* variables. Batch (1) and (2) use 500 such data vectors, and (3) uses 600 data vectors, respectively.

This large number of models is produced by the least resource-consuming model-types, namely generalized linear models and logistic regression as a sub-category thereof. In the

sequel, we display the typical outcomes of a single model, and finally, we discuss some evaluation of the collection of all models. In Figure 8, we display the transformed data of both a training and a test set.

The overall schedule for training and for evaluating of each single model is:

- [A] generate training and test data sets for each model;
- [B] run a generalized linear model (glm) with gaussian distribution family (as R function);
- [C] run a Logistic regression, that is a glm, with binomial distribution family (as R function); it uses an above/below mean binary variable as prediction target, making it a (0,1)-classifier;
- [D] evaluating the in-sample (training set) and the out-of-sample (test set) predictions along the ROC curves and their AUC value (Theodoridis & Koutroumbas, 2009, p. 275ff). They are the most telling performance characteristics for binary classification models, as with increasing AUC (area under the ROC curve, which should be above 0.5), the prediction performance increases as well. The shape of ROC-curves themselves may be exploited too; this will be a task of future work.

Both sets of Figure 8 show a complex pattern of time clustering and expose a variable duration of overall trading activity. The outlier points (above 15 on the log scale) denote the passage to the next day, as the pass from 2nd to the 3rd day (training) and from the 3rd to the 4th day of the month (see plot titles).

In Figure 9, we show in the two upper panels the resulting phase plots of training and testing the model [B]. The ideal situation would be a narrow band of points around a positive slope indication that the training set has been learned to minimal errors. The horizontal axis indicated the data point to be predicted, and the vertical one the actual predictions. As expected, the glm does not have the power to predict these highly irregular data. Hence, their prediction performance is weak. This may also be an advantage, since more powerful model types like neural nets tend to overtrain, which requires expensive in-sample pruning strategies. Nevertheless, the two scatter plots have a positive linear tendency (dashed line obtained by linear regression), with the train-set "prediction" slightly better than that of the test-set.

Figure 9 (lower panels) shows the rather successful pairing of training and test results measured by AUC. As should be expected from predictive data modeling, the AUC-value of the training set is higher than that of the test set. Furthermore, the shape of the ROC curve is more uniform, indicating a better prediction quality (stability) if, for instance, the cutoff-value (the separation value for class membership of the targets) changes. The run #147 shown here is from the batch (1) with (3,3)-time lagged input vectors. The other run batches with 2000 and 5000 models, respectively, show similar pairs, some with still higher AUC over their test sets.

As it turns out from the evaluation of all run batches, this convenient situation of inferring that training AUC-values and their associated, time-adjacent, test AUC-values are positively correlated is, in general, not true. Oftentimes, one finds inferior test-set AUC values, where the ROC curve oscillates closely around the diagonal (the dashed line in the AUC panels of Figure X.2), meaning that the underlying model has no predictive value. Furthermore, one also finds "inversely" predictive models, meaning that they indeed capture structure but confuse the binary classes (AUC values significantly below 0.5). An evaluation of these aspects is summarized in the following for all batch 3 models.

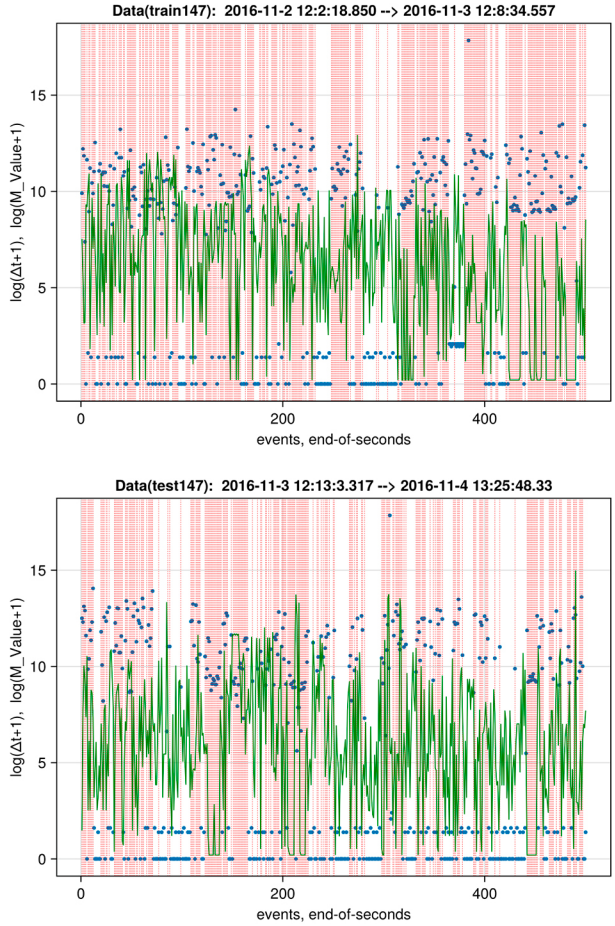


Figure 8. Training data (above) and test data (below) of model 147. Points are the log-duration from the last trading event. Points at zero indicate high intensity trading. The continuous dark line is the evolution of the prediction target *PoV*. Vertical dashed lines represent one second (source: own computation)

This evaluation is presented for brevity by discussing the features visible in Figure 10. The first goal is to verify if there is any significant dependence of the AUC values in the test-set on those of the training-set. An affirmative answer would imply that a forecaster (or trader, etc.) could rely on such a directed correlation: observing higher AUC-values in training on a data-subset would reasonably imply finding higher AUC-values in an associated forecasting period (in the associated test set) as well. Such a situation would be immediately visible by a positively sloped data cloud in the increasingly sorted training-set AUC values. Unfortunately, at least for the input structures selected for the batch of 5000 LogReg models (presented in Figure 10), there is no way to rely on in-sample AUC information! This comes, in fact, with little surprise for mature capital markets, but it seems to be true for the portfolio of our four Romanian stocks as well. However, there is a significant relation between increasing AUC on the training set (in-sample) and the variance of the AUC on the test-set (out-of-sample): If there is a high in-sample AUC, then expect a high risk of obtaining, respectively,

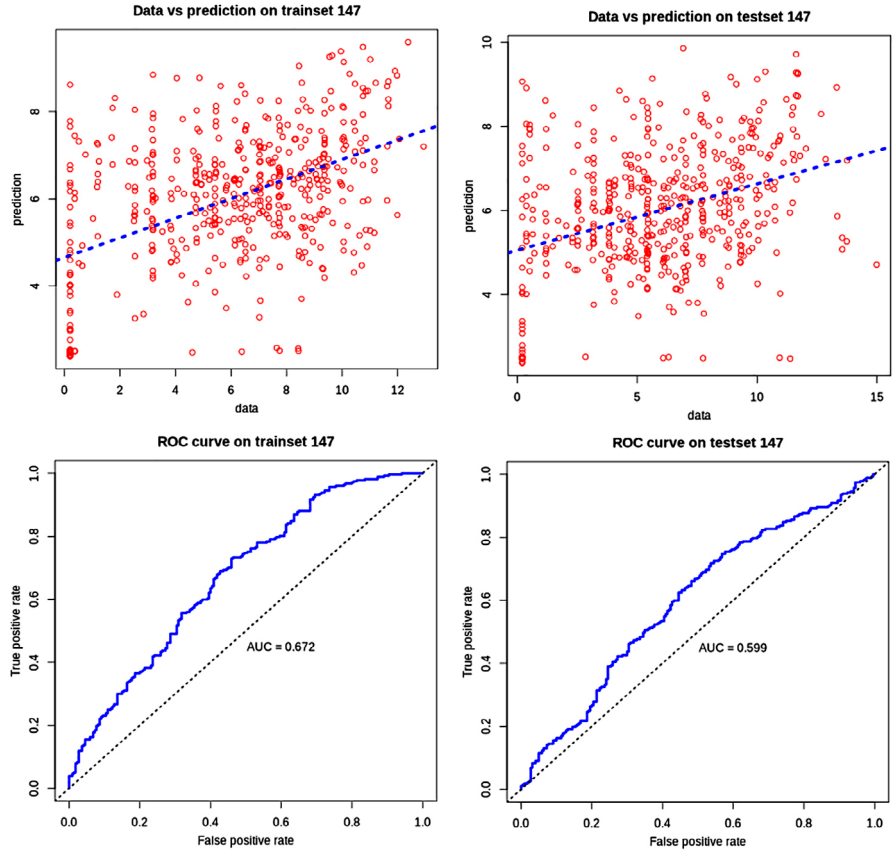


Figure 9. Upper panels: prediction performance of the glm model in-sample (left) and out-of-sample (right). Lower panels: ROC-curves and their AUC of the binary classification problem (LogReg) for both data sets (source: own computations)

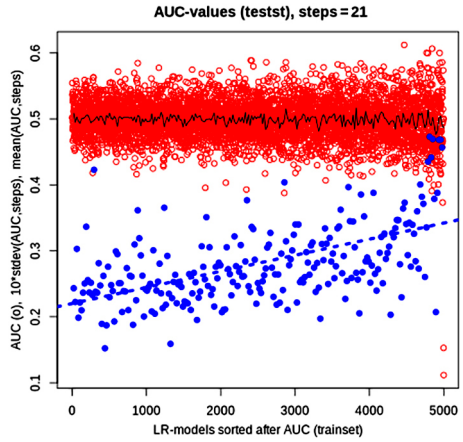


Figure 10. A comparison of what may be inferred from the AUC values on both, training- and test-sets from the batch of 5000 Logistic Regression models as described in the main text above (source: own computations)

both high-gain and high-loss predictions. These are visible by means of the best linear fit to the variances of the models (the filled points) and by reading the summary of this fit from Table 1. Both the step-wise mean (the line around AUC value of 0.5, which is exactly the case of "no signal over noise") and the step-wise variance are computed repeatedly over succeeding 21 single models. Overall, the stepwise mean is not changing much, but does fluctuate around the 0.5 AUC-level. These model-to-model fluctuations, which continue over the test-set, may be partly responsible for a relatively high overall discrepancy between in-sample AUC (between 0.6 and approaching 1.0, not shown) and the out-of-sample AUC, which reaches more than 0.6 (significant for 600 test examples), respectively less than 0.4 for those models which are "inversely" significant. Note that the single models differ solely by behavior and not by structure – apart from individual parametrization. Hence, their behavioral variation is entirely determined by the different train- and test-sets. Note also from Figure 16, a quite visible outlier problem, both in test AUC and test AUC-variance.

Table 1. The statistical summary of the regression leading to the dashed line in Figure 16

Call: lm(formula = 10*a_sd ~ a_seq) # "10*" to comply with image size

Residuals:

Min	1Q	Median	3Q	Max
–0.13249	–0.03668	–0.00508	0.02521	0.53466

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.201e-01	7.930e-03	27.751	<2e-16***
a_seq	2.441e-05	2.739e-06	8.913	<2e-16***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.06096 on 236 degrees of freedom
Multiple R-squared: 0.2518, Adjusted R-squared: 0.2487
F-statistic: 79.44 on 1 and 236 DF, **p-value: <2.2e-16**

Table 1 indicates a significant (linear) dependence of variation of test-set AUC over the 5000 Logistic Regression models.

As a consequence of comparing the 5000 different LogReg models, we do not find clear evidence that the choice of different time-lagged inputs from variables composed out of the event-by-event trading records is changing the overall picture. This may well change if using external covariates. To accelerate the choice of potentially informative covariates for financial HF-predictions, one may well use recommendations from some AI-based dialog system, which efficiently scans through pertinent opinions. As (lagged) covariates and time resolutions must be jointly identified, this seems much of an open-ended process – at least so under pragmatic time constraints. Nevertheless, one may deduce a recipe for investors with innate risk propensity so far: Risk-prone traders should act if the in-sample AUC is high, which may pay off after many repeated investments. The converse would then apply for the risk-averse actors.

4.5. Evaluation of some Machine Learning approaches for characterizing the predictive power of high-frequency time series

Data-driven Machine Learning, which includes neural networks, are important part of today's AI applied to long-time series analyses and predictions. In general, they are very expressive; hence, if large enough, they never undertrain but rather *overtrain* quite easily. The challenge of over-training is especially acute if they are also models may help as are, for instance, neural networks.

For exploratory purposes alone, such heavily parameterized models may help in discovering predictive potential in time series, but for actual production – or, out-of-sample real-life predictions – the less- or non-parametric methods may be favored.

The length of the time series, that is, the number of (consecutive) observations, will indicate if it makes sense to use a model with strongly nonlinear expressive power. Generalized linear models, possibly with a regularization term added to the error function, in order to identify the most relevant explanatory variables, do often suffice for economic time series.

Financial HF time series, with their high time resolution, may justify the use of more expressive models. However, this is only the case if the time series itself is not highly redundant or "pathological" in other ways, just as big data (a large number of observations and/or features) does not necessarily imply that there also is a large amount of content to be found!

4.5.1. From market value at high-frequency time resolution to prediction of stock price changes

Predicting stock price changes is a fundamental ingredient for capital market trading systems. While time series analysis can be done in a completely exploratory and descriptive manner (see previous sections), realistic prediction tasks must invariably be effected on a split of data, namely by using a training set and a test set, respectively. Traders need models that are normative, which are able to recommend (1) ***buy***, (2) ***sell,*** and in general also (3) ***wait*** actions. Especially, for prediction model building, one cannot use the entire data for training (for learning a hopefully good model for later predictions).

Besides predicting composites, or transformed stock price changes (like risk, volatilizes, etc.), the most demanding prediction targets are ***price changes of single stocks***, especially the direction of change.

In the context of using even-driven (max-time-resolution) data, as are ours, this amounts to learn time series models with a variable lag structure. Even if a formal lag structure reads to predict some

x_t with a model $f(x_{t-1}, x_{t-2}, x_{t-3})$, say, then the real time duration between a t_k to a t_{k+1} if effectively a random variable, the distribution of which may not even be similar for different single stock price evolutions. Therefore, it may be a sensible approach for extracting structure from the highly irregular footprint of HF time series by constant lag structures, for instance, with a formal lag sequence of [1, 2, 3], which over time means something effectively different for every single time series.

Stock price changes recorded by transaction events of single firms come with a series of issues:

(1) Their transactions are fewer than the numbers of recorded transactions in the market as a whole.

(2) Prices of single firms do not even change in most of their own observed transactions.

We first used different simple as well as Neural Network-based methods on the unfiltered time series of the single stocks (firms). Independent of potential model expressiveness, all of them predict an unconditional mean! More expressive models just tended to adapt somewhat better on the training sets, while their predictions on the test sets were heavily deteriorating and almost worthless. In the sequel, we refrain from further reporting on these results.

In order to remedy these problems from our stock-specific increment $\Delta X_t^i := x_t^i$ time series, which have sporadic activity and very long and irregular periods of "silence", we cut out all entries with zero or very small increments, resulting in a much shorter time series y_k:

$$y_k = \emptyset;\ k = 0;\ t = 0;$$

FOR ($t \le T$) IF abs(x_t) > $\varepsilon \ge 0$ THEN append x_t to y_k $k = k + 1$; ELSE skip; $t = t + 1$ **END**,

where $T > 0$ is the length of the unfiltered time series, *and* > 0 may vary over the stocks (firms). In what follows, this filtered time series we be labeled non-***silent first differences***. The advantage of obtaining a more tractable time series for prediction modeling purposes, stands against some potential disadvantages, namely (1) the remaining activity points do not coincide in the real time sense between different firms, that is, for any two firms, the same *k*-index value may mean records at totally different natural time points, hence making comparisons between firms more demanding or unfeasible; (2) the next price movement happens after a variable time span in natural time, which is a potential problem for trading on out-of-sample predictions, especially in the case of $\varepsilon > 0$ (how long to wait upon observing the small movements?).

In view of the non-*silent first differences* filter and its implied challenges, we restrict further analyses to univariate modeling, by using the same kind of Machine Learning (ML) models on a 3-lag data input sequence for the price variation (first differences) of all four stocks. Note that 3-lag may mean here non-near entries in natural time and also occurring in different time patterns across firms.

Both a training and test example are hence given by the structure $[y_k \mid y_{k-1}, y_{k-2}, y_{k-3}]$ followed by

$[y_{k+1} \mid y_k, y_{k-1}, y_{k-2}]$, etc., where the respective targets are the left outermost entries and the model inputs the rest. Note that, in this representation, the order of presenting the examples to the model and, subsequently, to the predictor does not matter. The input order is also permutation invariant if changed in all examples in the same way. This leads to the assumption that fully featured neural networks are over-engineered for the task at hand, since their capacity is best exploited for more complex data structures, with non-interchangeable spatial correlations, say. As our data are technically simple regression tables, a class of ML methods that is expressive, suitable for large data sets, and is training much faster than (deep) neural networks is preferable.

In this situation, Random Forests (Athey et al., 2019; Mentch & Zhou, 2020), henceforth RF, seem to be the right choice. They come in many flavors, but their essential ingredients are proposing many decision trees and up-voting mechanisms as a dynamic selector of consecutively better models. In general, trees are much easier to produce, to traverse, and to purposefully modify than networks. Then we follow the standard procedure:

(1) For every stock, choose sets of training and of test data, with roughly an equal number of entries; However, the number of data entries differs widely across firms.

(2) For every stock build a RF-model ***RF***$[y_{k-1}, y_{k-2}, y_{k-3}] \rightarrow y_k$ on the filtered training data.

(3) For every stock, predict the price movements on the filtered test data using its RF-model.

(4) Evaluate the trading potential of all models on the filtered test data by using sign(***RF***$[y_{k-1}, y_{k-2}, y_{k-3}]) * y_k$ as the basic gain (profit) contribution in period k.

If the sign of the predicted value coincides with the sign of the target, then the amount of change in the target will be the gain, otherwise it will be accounted for as a loss. The sum over these period-wise contributions will be compared to the hypothetical Max-abs-gain (all directions of change correctly predicted).

Figure 11 shows the alignment of target training data with predictions. We use two types of RF modeling software, namely (1) ***randomForest()*** and (2) ***ranger()***, both as implemented in ***R***. Training time for (1) takes between 15 seconds for the models on shorter filtered training sets (around 20000 recorded transactions) and more than 2 minutes for the larger ones (more than 120000 recorded transactions), while for (2) a few seconds. Variant (1) produces better training alignment but does not translate into better out-of-sample (test set) predictions. Hence, in Figure 11 and in the sequel, we only show the results using the RF variant (2). The expressive power of the RF models is obvious from the near-resulting diagonal like shapes of the alignments. Note oval-like and more dispersed point structures.

Figure 12 depicts the time evolution of prices over the filtered test data time series. Consecutive time entries are not always also nearby time successors in natural time. However, the shown time series follows in filtered time directly after the training set. Their different lengths reflect both the total data available and the number of silent first differences. DIGI (telecom) and SNP (petrol) show both a stronger overall trend, while ATB (pharma) and TLV (bank) exhibit a mixed overall evolution. Evolutions, which may be relatable to herding behavior, can be observed in the early portions of the test set for ATB, DIGI, and TLV. The dotted Max-gain curve may be a rough indicator for herding, especially for TLV and SNP, the two longer test time series. The dashed curve stands illustrates the buy-and-hold gain.

Figure 13 contrasts the alignments of predictions with targets on the test sets with those on the training sets (Figure 11). As expected, the alignment is here weaker than that found for the training sets. However, this does not have a proportional detrimental influence on the gains obtainable by trading, as will be reported next in Table 2. The alignments on test sets have a distinctly different shape, ranging from concentrated, oval-like, to more dispersed and disconnected (i.e. lowe left inset for TVL). Future research will attempt to further interpret these and related data shapes.

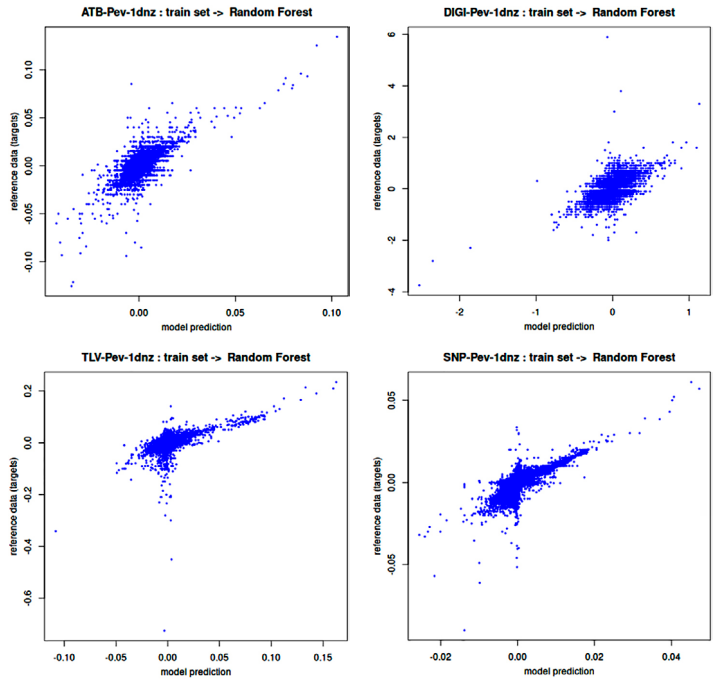


Figure 11. Results of training the four stocks (ATB, DIGI, TLV, and SNP, see main text) with Random Forest (RF) models on the non-silent first differences data

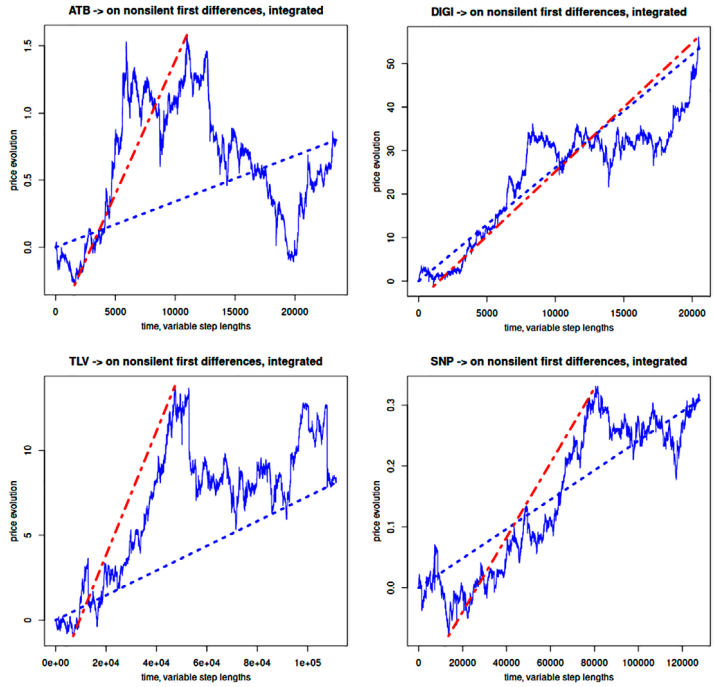


Figure 12. Integrated time series of the non-silent first differences of the four stock price evolutions

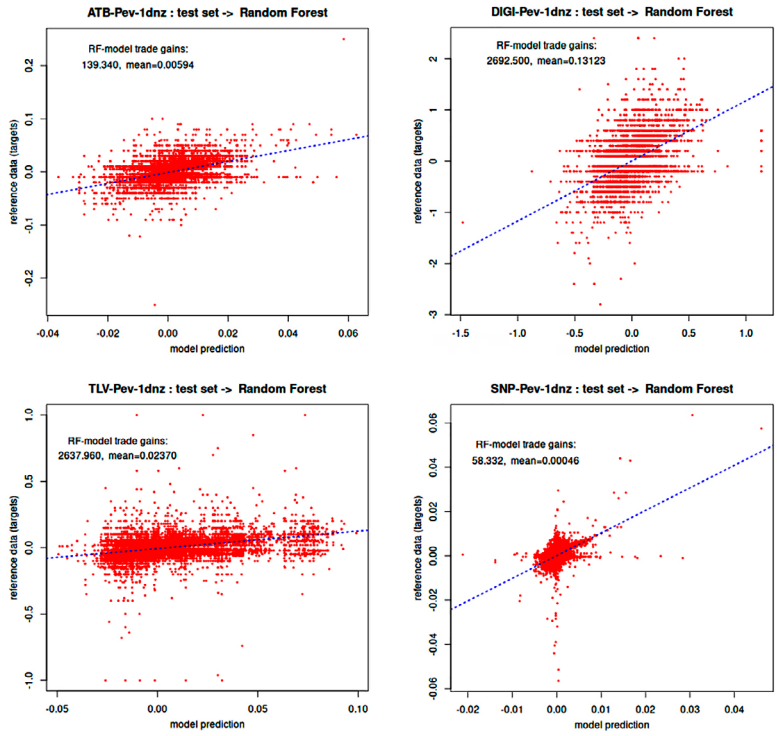


Figure 13. The target variable of the test set of the four listed companies plotted against the prediction of the respective trained RF-models. In the lower left panel (for TLV), the few very large price changes are remapped to –1 and +1, respectively, in order to magnify the display of the many remaining points. All evaluations are made on the original data

Table 2. Comparing trade gains when using different trading mechanisms (source: own computations)

Stock of company	ATB (pharma)	DIGI (telecom)	SNP (petrol)	TLV (bank)
Max. absolute gain**	299.63	5222.1	114.65	4531.0
Max. gain/loss*	1.86	57.4	0.41	14.7
Buy and Hold	0.80	53.5	0.308	8.10
Model (Rand. Forest)	**139.34**	**2692.5**	**58.33**	**2637.96**

Note: **) impossible to achieve in practice; *) improbable to realize.

According to Table 2, the four stock evolutions over non-silent first differences may be characterized by a vector of measures concerning the gain over various time periods and, hence, also relating to trading mechanisms of different complexity. The first two of them, (1) *Maximal absolute gain*, and (2) *Maximal gain/loss*, *can* never be achieved in practice as they would require 100% correct prediction of the directions of change, and (2) knowledge about when to enter and when to leave a buy-and-hold (respectively a sell-and-hold) position (as is obvious by looking at the dotted lines of Figure 18). Remain the simple (3) *Buy and Hold* and the (4) *Prediction model* (our RF models).

It is obvious that gains pertaining to criterion (1) must be highest for every stock. Also gain of (2) must be greater than or equal to gains of (3). The interesting message refers to the fraction of (4) over (1), not their absolute values. Note that, in the context of non-silent first difference time series, high-frequency trading gains are nominally high, perhaps unexpectedly so. As Figure 17 shows, this is not solely due to trends, as there are extended intervals of *sidewalks* (no significant trends).

Gain refers to the theoretical profit, which does not take into account transaction costs of any kind. There is also no information about when exactly in the future the next change in stock price will occur. Hence, the rather large gains resulting technically from the high-frequency model predictions may turn out to be lower in practice.

Concerning investors' herding behavior, the high-frequency univariate time series approach followed in this research allows just for hypothesizing herding for extended periods of price increase, where price change volatility is relatively minor with regard to a temporary trend (depicted for our four stock price evolutions for various sub-periods in Figure 11). As herding is a kind of imitative behavior, it has complex psychological and social determinants (Rahayu et al., 2020), including herding sub-types like *spurious herding* and *intentional herding* and it heavily draws on network and spatial modeling approaches (Eberhard et al., 2019), requiring information about covariates which are not in our data base and, in many cases, even cannot be part of a HF financial data base.

In the wider context of stock price prediction itself, recent literature points to the importance of considering (or learning) "... the long-term relative stability of stock relation structures while retaining the flexibility to respond to sudden market shifts ..." (Chen et al., 2025), which clearly requires a large number of commensurate stock market time series for modeling of inter-stock relationships. Hence, a possible solution would be to partially retreat from the pure HF data view and seek to combine some HF-grounded information with covariates available at lower time resolutions.

4.5.2. Portfolios build by non silent first difference data models

Time series such as ours, running on different natural time rhythms owing to the use of non-silent first difference data filters, hence on individual "synthetic" time scales each, cannot be combined in the classical straightforward way in order to evaluate portfolios. Alternatively, one may use the different trained models in order to predict a given time series. While for each time series, there is the ***original model***, one may evaluate for each of our four HF time series the trading gains obtainable from the prediction of these same test data by all the other remaining models. Hence, in the sequel, we will display for every model five trading gain measures resulting from the original and the three respective cross-predictions using the gain formula explained in the previous section. The fifth entry contains the portfolio gain. A portfolio may contain all or a selection of predictors as evaluated by running a generalized linear model on the combination of predictors previously generated by the respective Random Forest models discussed in the previous section. Note that this approach is nonstandard as it cross-predicts by means of models built with other intentions. Technically, this works as all models predict the same test data for every one of the four experiments.

The most surprising result from Table 3 is that ATB by SNP and ATB by TLV outperform ATB by ATB, the original. This means that the reaction to all information patterns is present in

the cross-predicting models, SNP and TLV. The superior performance of the portfolio is to be expected on purely technical grounds; a somewhat unusual situation is the highly significant increase in trading gain.

The ATB portfolio is built by linearly combining ATB, SNP, and TLV model predictions Table 4).

The DIGI Portfolio (Table 5) is effectively replicating the original by resulting in a gain level just slightly above that of predictions made by the RF model trained on the DIGI data. The information patterns in the models trained on data from SNP and TLV data produce trading gain levels rather close to the original.

The DIGI portfolio is built by linearly combining DIGI, ATB, and TLV model predictions (Table 6).

Table 3. HF prediction gains on the ATB test set (source: own computation)

ATB by ATB	139.340	Original
ATB by DIGI	69.526	Cross-prediction
ATB by SNP	144.982	<< Remarkable cross-prediction
ATB by TLV	152.854	<<< Remarkable cross-prediction
ATB Portfolio	**156.042**	Significantly superior to the original

Table 4. The glm() coefficients of the portfolio weights for the ATB prediction (source: own computation)

	Estimate	Std. Error	t value	Pr(>|t|)	Select
(Intercept)	–0.001244	0.000109	11.410	<2e-16***	X
pred_test_ATB	0.336608	0.022003	15.299	<2e-16***	X
pred_test_AbD	–0.013863	0.008979	–1.544	0.122637	
pred_test_AbS	0.101271	0.029845	3.393	0.000692***	X
pred_test_AbT	0.732544	0.025777	28.419	<2e-16***	X

Note: --- Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table 5. HF prediction gains on the DIGI test set (source: own computation)

DIGI by DIGI	2692.5	Original
DIGI by ATB	2001.9	Cross-prediction
DIGI by SNP	2618.7	Remarkable cross-prediction
DIGI by TLV	2618.7	Remarkable cross-prediction
DIGI Portfolio	2700.3	Slightly superior to the original

Table 6. The glm() coefficients of the portfolio weights for the DIGI prediction (source: own computation)

	Estimate	Std. Error	t value	Pr(>|t|)	Select
(Intercept)	–0.008264	0.002558	–3.230	0.00124**	
pred_test_DIGI	1.012052	0.018904	53.537	<2e-16***	X
pred_test_DbA	0.957868	0.211531	4.528	5.98e-06***	X
pred_test_DbS	0.198824	0.517931	0.384	0.70107	
pred_test_DbT	0.325907	0.121830	2.675	0.00748**	X

Note: --- Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Table 7. HF prediction gains on the SNP test set (source: own computation)

SNP by SNP	58.3324	Original
SNP by ATB	24.6412	Poor cross-prediction
SNP by DIGI	26.7770	Poor cross-prediction
SNP by TLV	8.1230	Very poor cross-prediction
SNP Portfolio	**58.7300**	Slightly superior to the original

The SNP Portfolio (Table 7) also effectively replicates the original by resulting in a gain level just slightly above that of predictions made by the RF model trained on the SNP data. Overall, this may not be very surprising, since the filtered SNP test set is a much longer time series than those of ATB and DIGI. In contradistinction to the DIGI portfolio, information patterns in the models trained on ATB, DIGI, and TLV data do not gain levels anywhere close to the original. Somewhat surprisingly, is the poor performance revealed with regard to the model trained on filtered TLV data (i.e., SNP by TLV), the latter also being a long test time series.

Table 8. The glm() coefficients of the portfolio weights for the SNP prediction (source: own computation)

	Estimate	Std. Error	t value	Pr(>|t|)	Select
(Intercept)	1.370e-04	7.532e-06	18.185	<2e-16***	X
pred_test_SNP	8.454e-01	6.986e-03	121.004	<2e-16***	X
pred_test_SbA	–1.479e-02	2.677e-03	–5.524	3.32e-08***	X
pred_test_SbD	7.549e-03	2.196e-04	34.369	<2e-16***	X
pred_test_SbT	2.073e-01	8.315e-03	24.926	<2e-16***	X

Note: --- Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

The SNP portfolio model is built by linearly combining all model predictions (Table 8). Note that we do not use convex combinations; hence, prediction vectors with negative coefficients may also contribute to improving a portfolio.

Table 9. HF prediction gains on the TLV test set (source: own computation)

TLV by TLV	2637.96	Original
TLV by ATB	2324.26	Remarkable cross-prediction
TLV by DIGI	1331.66	Poor cross-prediction
TLV by SNP	2410.96	Remarkable cross-prediction
TLV Portfolio	**2685.24**	Slightly superior to the original

The TLV Portfolio (Table 9) also effectively replicates the original by resulting in a gain level slightly above that of predictions made by the RF model trained on the TLV data. TLV by ATB and TLV by SNP have remarkable performance, while the opposite is found for SNP by TLV from Table 6.

Table 10. The glm() coefficients of the portfolio weights for the TLV prediction (source: own computation)

	Estimate	Std. Error	t value	Pr(>|t|)	Select
(Intercept)	–0.0083636	0.0001631	–51.288	<2e-16***	X
pred_test_TLV	0.9273616	0.0122347	75.798	<2e-16***	X
pred_test_TbA	0.3260112	0.0168763	19.318	<2e-16***	X
pred_test_TbD	0.3114824	0.0105297	29.581	<2e-16***	X
pred_test_TbS	0.0834593	0.0350841	2.379	0.0174*	X-

Note: --- Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

The TLV portfolio model (Table 10) is built by linearly combining all model predictions. Leaving out, and the training on SNP data (pred_test_TbS, the fifth row entry of Table 10) produces a negligible change in portfolio gain.

To conclude, the non-silent first difference time series approach makes it difficult to directly compare model outputs of different time series in a classical way (by, e.g., convex combinations of prediction vectors), in principle owing to a non-existent time synchronicity between these predictions. However, it is possible to study the information similarity of model reactions by letting models predict test time series on which they were not trained on. This indirectly shows that – perhaps contrary to ex ante expectation – there are indeed cases of shared information patterns between what models effectively learn. While we find situations with a substantial deficit between potential trading gains (based on correct prediction of the direction of the next price change) of models trained on the stock data they are supposed to predict, in many remaining cases, the models trained on other data can be used to quite succesfully predict the original. In all but in the somewhat surprising case of the ATB, models trained on other data do not reach or surpass the original trading gains. As expected, on purely statistical all are contributing to portfolios that at least replicate the respective original model. A tentative semi-ordering of the information pattern cascade may be attributed to models as follows

SNP > (TLV or DIGI) > ATB, with ">" meaning more useful information patterns.

Although one also finds some reverse flow (example: TLV by ATB from the previous tables). A general lesson from these numerical experiments is that predictions for financial time series may be replicable to a sufficient degree without training models on past original data (which might not be available), but by constructing these predictions from models trained on data that may be just faintly related to the original. Another possibility might be to use models trained on surrogate data, which are simulations generated via adequate statistical distributions. This would reflect the opinion of some AI developers, who claim that any (practically relevant – meaning mathematically non-exotic) data series can be predicted by using a large pool of real-life examples.

5. Conclusions, discussion, and policy recommendations

5.1. Towards implementing the derived information into AI models

Regarding the use of AI or Machine Learning for an effective trading model based on high-frequency data, nowadays it is much less relevant to search for some "intelligent" mathematical models or to dive into the details of the AI per se. The used AI should simply be powerful enough to learn from very many input-output examples of complicated (composed) structures. Think of giving the AI application all the information we derived in section 4 from the high-frequency data as a kind of "generalized prompt". They contain selected time series of different resolutions, time-resolution-localized distributions, randomness-degrees, correlation structures, and complexity measures (like truncated the wavelet replicas) as a kind of domain knowlege.

The objectives of a pertinent AI-model are straightforward to state. An example would be: *Propose time points for the "buy" and "sell" orders*, *such that the cumulated profit is maximized*. Such orders may indeed be of "high frequency," but oftentimes they will occur much rarer over time, as transaction cost and other constraints come into play.

Thereafter, today's powerful AI applications know how to translate these requirements into working time series models (e.g. large recurrent neural nets, with additional constraints suitably translated from the "prompts") and will also statistically cross-validate the trading results in-sample. Such trained AI-models can then be subjected to a "meta-learning" loop, which will output the most relevant input prompts (somewhat analogous to relevant variable detection of structurally much simpler classical econometric models). Such validated input prompts may then be combined into fuzzy recommendations.

If there is a low (or unknown) signal-to-noise ratio – as one expects to find in HF stock exchange time series, neural networks may be employed to learn an involved statistical distribution (its density), which can be adopted to explain the (random) generators of such time series. This may improve predicting certain aspects (events like major turning points), but will probably not improve predicting HF time series at their high time resolution,

To this end, we propose to use a model setting that learns the extreme-value (unusual peaks) distribution of a long time series. While not originating from a financial application, in (Pasche & Engelke, 2024), a recurrent neural network is used for a similar rare event prediction purpose.

5.2. A final general critical discussion

The multifaceted and complex scientific field of high-frequency trading represents a dynamic phenomenon in itself. Analyzing the economic literature presented in the literature review section has provided several insights into high-frequency trading technological innovations, theoretical backgrounds, market impact, and policy dilemmas. It can be seen that the dichotomy of high-frequency trading lies between greater efficiency and market liquidity under normal conditions, and systemic risk and high volatility, which need to reshape and rewrite risk management strategies.

The question is whether market integrity can be safeguarded by the arms race, the introduction of unknown risks by machine learning, and ultra-fast finance can be regulated by

the existing judicial framework in the context of emerging markets. In this context, a holistic approach needs to be adopted by practitioners, practitioners and regulators in order to make a trade-off between innovation, efficiency, and the global realities.

Regarding the evolution of the Bucharest Stock Exchange, it can be seen that, because absolutely all trades are available, some special events can be tracked that only daily charts may not accurately render.

We are referring here to events that brokers certainly felt as a game-changer, but maybe the charts didn't surprise them. Such days were:

1. The days of price collapse during the 2008 world financial crisis, when on some days the stock market only had pre-opening and pre-closing to prevent the massive losses of intraday speculators;
2. 21.12.2018 – issuance of the Greed Ordinance, which established an additional tax for companies with a high turnover, especially affecting banks and companies in the energy field;
3. 08.03.2020 – COVID lockdown – The prohibition of the free movement of people implied significant uncertainty in the market, which led to massive sales;
4. 24.11.2024 – posting of the results of the first round of the presidential elections, which brought a huge surprise being won with an anti-European and anti-NATO candidate;
5. 06.12.2024 – annulment of the presidential elections – An important and controversial decision of the Constitutional Court that annulled the entire process of electing the president, leading to uncertainty regarding Romania's European future.

The effect of these events on stock market transactions can only be highlighted by following the principles of HFT analysis.

Other elements that may be relevant if viewed from an HFT perspective should also be studied:

1. Japanese candlestick.
2. Long body candle days.
3. Study of variations (first derivative) – comparison of statistics covering all transactions vs. daily, weekly, and monthly.
4. Fractal study – the possibility of predicting variations according to patterns highlighted in the past.
5. Tracking the days with large divergences between the evolution of companies and conclusions on the reasons.

These conclusions can be useful on "hectic" or high volatility days, since on "quiet" days, fast trading and speculation are not relevant.

Acknowledgements

We appreciate the comments and suggestions from the editor(s) and anonymous reviewers. A special merit deserves the Bucharest Stock Exchange of Romania, which provided the data for this study, especially Alin Barbu, Alexandru Ștefan, and Octavian Molnar.

Funding

The authors received no funding.

Author contributions

The authors contributed equally. All authors approved and read the final manuscript.

Disclosure statement

There are no competing interests in this study.

References

Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, *105*(2), 564–608. https://doi.org/10.1257/aer.20130456

Adrian, J. (2016). Informational inequality: How high-frequency traders use premier access to information to pray to institutional investors. *Duke Law and Technology Review*, *14*, 156–279.

Aquilina, M., Budish, E., & O'Neill, P. (2021). *Quantifying the high-frequency trading "arms race"* (BIS Working Paper No. 29011). National Bureau of Economic Research. https://doi.org/10.3386/w29011

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Chen, L., Bai, H., Wang, M., Huang, X., Guan, Z., & Zhao, J. (2025). *Stock price prediction via a dual relation fusion network incorporating static and dynamic relations*. arXiv. https://arxiv.org/pdf/2510.10695

Clapham, B., Haferkorn, M., & Zimmermann, K. (2023). The impact of high-frequency trading on modern securities markets. *Business & Information Systems Engineering*, *65*(1), 7–24. https://doi.org/10.1007/s12599-022-00768-6

Desagre, C., Laly, F., & Petitjean, M. (2025). Revisiting the trading activity of high-frequency trading firms around ultra-fast flash events. *Financial Innovation*, *11*, Article 68. https://doi.org/10.1186/s40854-024-00726-z

Donadio, S., Ghosh, S., & Rossier, R. (2022). *Developing high-frequency trading systems*: *Learn how to implement high-frequency trading from scratch with C++ or Java basics*. Packt Publishing.

Dumiter, F. C., Turcaș, F., Nicoară, Ș. A., Bențe, C., & Boiță, M. (2023). The impact of sentiment indices on the stock exchange–The connections between qualitative sentiment indicators, technical analysis and stock market. *Mathematics*, *11*(14), Article 3128. https://doi.org/10.3390/math11143128

Eberhard, J., Lavín, J. F., Montecinos-Pearce, A., & Arenas, J. (2019). Analyzing stock brokers' trading pattern: A network decomposition and spatial econometrics approach. *Complexity*, *2019*, Article 490640. https://doi.org/10.1155/2019/7490640

Ersan, O., Dalgic, N., Ekinci, C., & Bodur, M. (2021). High-frequency trading and its impact on market liquidity: A review of literature. *Alanya Academic Review Journal*, *5*(1), 345–368. https://doi.org/10.29023/alanyaakademik.799039

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.2307/2325486

Fox, M. B., Glosten, L. R., & Rauterberg, G. V. (2018). High-frequency trading and the new stock market: Sense and nonsense. *Journal of Applied Corporate Finance*, *29*(4), 30–44. https://doi.org/10.1111/jacf.12260

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100. https://doi.org/10.1016/0304-405X(85)90044-3

Gomber, P., Arndt, B., Lutat, M., & Uhle, T. (2011). *High-frequency trading*. SSRN. https://doi.org/10.2139/ssrn.1858626

Goudarzi, M., & Bazzana, F. (2023). Identification of high-frequency trading: A machine learning approach. *Research in International Business and Finance*, *66*, Article 102078. https://doi.org/10.1016/j.ribaf.2023.102078

Gupta, M. (2022). High-frequency trading in stock market. *International Research Journal of Engineering and Technology*, *9*(10), 494–499.

Hayashi, T., & Nishide, K. (2024). Strategic liquidity provision in high-frequency trading. *International Review of Financial Analysis*, *93*, Article 103168. https://doi.org/10.1016/j.irfa.2024.103168

Hossain, S. (2022). High-frequency trading (HFT) and market quality research: An evaluation of the alternative HFT proxies. *Journal of Risk and Financial Management*, *15*, Article 54. https://doi.org/10.3390/jrfm15020054

Ivashchenko, A. (2025). (In)frequently traded corporate bonds and pricing implications liquidity dry-ups. *Finance Research Letters*, *75*, Article 106880. https://doi.org/10.1016/j.frl.2025.106880

Jia, J. (2025). Tail risk spillover between global stock markets based on effective Rényi transfer entropy and wavelet analysis. *Entropy*, *27*, Article 523. https://doi.org/10.3390/e27050523

Jung, M. J., Kwon, K. Y., & Park, H. (2024). Does high-frequency trading cause stock prices to deviate from fundamental values? *Accounting and Business Research*, *54*(5), 580–613. https://doi.org/10.1080/00014788.2023.2258787

Kauffman, R., Hu, Y., & Ma, D. (2015). Will high-frequency trading practices transform the financial markets in the Asia Pacific Region? *Financial Innovation*, *1*(4), 1–27. https://doi.org/10.1186/s40854-015-0003-8

Kyle, A. S. (1985). Continuous auctions and inside trading. *Econometrica*, *53*(6), 1315–1335. https://doi.org/10.2307/1913210

Lei, B., Zhang, B., & Song, Y. (2021). Volatility forecasting for high-frequency financial data based on web search index and deep learning model. *Mathematics*, *9*, Article 320. https://doi.org/10.3390/math9040320

Li, C., Shen, L., & Qian, G. (2023). Online hybrid neural network for stock prediction: A case study of high-frequency stock trading in the Chinese market. *Econometrics*, *11*, Article 13. https://doi.org/10.3390/econometrics11020013

Lu, S., Zhao, J., & Wang, H. (2020). Trading imbalance in Chinese stock market – A high-frequency view. *Entropy*, *20*, Article 897. https://doi.org/10.3390/e22080897

Makarov, V. A., Muñoz-Arnaiz, R., Herreras, O., & Makarova, J. (2023). Correlation dimension of high-dimensional and high-definition experimental time series. *Chaos*, *33*, Article 123114. https://doi.org/10.1063/5.0168400

Menkveld, A. J. (2011). *High-frequency trading and the new market makers* (Tinbergen Institute Discussion Paper No. TI 2011-076/2/DSF 21). Duisenberg School of Finance.

Menkveld, A. J. (2016). The economis of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, *8*(1), 1–24. https://doi.org/10.1146/annurev-financial-121415-033010

Mentch, L., & Zhou, S. (2020). Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, *21*(171), 1–36.

Metelski, D., & Sobieraj, J. (2024). Trading volume concentration across S&P 500 Index constituents – A Gini-based analysis and concentration-driven (daily rebalanced) portfolio performance evaluation: Is chasing concentration profitable? *Journal of Risk and Financial Management*, *17*, Article 325. https://doi.org/10.3390/jrfm17080325

Nimalendran, M., Rzaye, K., & Sagade, S. (2024). High-frequency trading in stock market and the costs of options market making. *Journal of Financial Economics*, *159*, Article 103900. https://doi.org/10.1016/j.jfineco.2024.103900

Pasche, O. C., & Engelke, S. (2024). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics*, *18*(4), 2818–2839. https://doi.org/10.1214/24-AOAS1907

Portnoy, K. (2011). High-frequency trading and the stock market: A look at the effects of trade volume on stock price changes. *The Park Place Economist*, *19*(1), 69–75.

Rahayu, A. D., Putra, A., Chiata, C., & Ningtyas, R. A. (2020). Herding behavior in the stock market: A literature review. *International Journal of Social Sciences Review*, *1*(2), 8–24.

Saranj, A., & Zolfaghari, M. (2025). Trading behavior-stock market volatility nexus among institutional and individual investors. *Financial Innovation*, *11*, Article 98. https://doi.org/10.1186/s40854-024-00717-0

Serrano, A. S. (2020). High-frequency trading and systemic risk: A structured review of findings and policies. *Review of Economics*, *71*(3), 169–195. https://doi.org/10.1515/roe-2020-0028

Stübinger, J., & Schneider, L. (2019). Statistical arbitrage with mean-reverting overnight price gaps on high-frequency data of the S&P 500. *Journal of Risk and Financial Management*, *12*, Article 51. https://doi.org/10.3390/jrfm12020051

Sui, Y., Tian, X., Zhong, A., & Chiah, M. (2025). Beyond the final whistle: AFL grand final and retail investor trading. *Finance Research Letters*, *85*, Article 108137. https://doi.org/10.1016/j.frl.2025.108137

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern recognition* (4th ed.). Elsevier Academic Press.

Tripathi, J. S., & Rengifo, E. W. (2025). FinTech, fractional trading and order book dynamics: A study of US equity markets. *Fintech*, *4*(2), Article 16. https://doi.org/10.3390/fintech4020016

Turcaș, F., Dumiter, F. C., & Boiță, M. (2022). Econophysics techniques and their applications on the stock market. *Mathematics*, *10*, Article 860. https://doi.org/10.3390/math10060860

Turiel, J. D., & Aste, T. (2022). Heterogeneous criticality in high frequency finance: A phase transition in flash crashes. *Entropy*, *24*, Article 257. https://doi.org/10.3390/e24020257

Wang, Q., & Liu, L. (2022). Pandemic or panic? A firm-level study on the psychological and industrial impacts of COVID-19 on the Chinese stock market. *Financial Innovation*, *8*, Article 36. https://doi.org/10.1186/s40854-022-00335-8

Wang, Y., Zeng, Y., & Wu, Z. (2025). Risk spillover in the carbon-stock system and sustainability transition: Empirical evidence from China's ETS pilots and a-share emission-regulated firms. *Sustainability*, *17*, Article 4274. https://doi.org/10.3390/su17104274

Yalavarthi, V. K., Burchert, J., & Schmidt-Thieme, L. (2025). *DCSF*: *Deep convolutional set functions for classification of asynchronous time series*. Information Systems and Machine Learning Lab, University of Hildesheim.

Yang, F., & Paparrizos, J. (2025). SPARTAN: Data-adaptive symbolic time-series approximation. *Proceedings of the ACM on Management of Data*, *3*(3), Article 220. https://doi.org/10.1145/3725357