IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

On the Economic Significance of Stock Market Prediction and the No Free Lunch Theorem

CARLOS BOUSOÑO-CALZÓN[1], JOSUÉ BUSTARVIEJO-MUÑOZ[1], PEDRO ACEITUNO-ACEITUNO[2], AND JOSÉ JOAQUÍN ESCUDERO-GARZÁS.[3]

[1]Department of Signal Theory and Communications, Univ. Carlos III de Madrid, Leganés, Madrid, Spain (e-mail: cbousono,jbustarviejo@tsc.uc3m.es)
[2]Department of Business Administration and Management and Economics, Madrid Open University, Collado Villalba, Madrid, Spain(e-mail: pedro@udima.es)
[3]Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77843 USA (jescugar@tamu.edu)

Corresponding author: Carlos Bousoño-Calzón (e-mail: cbousono@tsc.uc3m.es).

This work was supported in part by the Univ. Carlos III de Madrid under Strategic Action 2013/00199/002.

ABSTRACT Forecasting of stock market returns is a challenging research activity that is now expanding with the availability of new data sources, markets, financial instruments, and algorithms. At its core, the predictability of prices still raises important questions. Here we discuss the economic significance of the prediction accuracy. To develop this question, we collect the daily series prices of almost half of the publicly traded companies around the world over a period of ten years and formulate some trading strategies based on their prediction. Proper visualization of these data together with the use of the No Free Lunch theoretical framework give some unexpected results that show how the *a priori* less accurate algorithms and inefficient strategies can offer better results than the *a priori* best alternatives in some particular subsets of data that have a clear interpretation in terms of economic sectors and regions.

INDEX TERMS Stock market, economic significance, forecasting, prediction algorithm, trading strategies, extended Bayesian framework, no free lunch theorem, support vector machines, big data, visualization.

I. INTRODUCTION

FORECASTING of stock market returns is not only difficult, it may also cause the price generating process to change over time [1], [16]. Interestingly, it seems that the same act of academic publication may interfere with the price of the shares [2]. These facts, together with the availability of new data sources [22], [25], markets [18], [26], financial instruments [8] and algorithms [21], [23] make the predictability of stock returns a hot topic [34].

The econometric approach to forecasting stock market generally proposes mathematical models for the price generation process and then test them with data. A celebrated discussion, usually under the rubric of *Efficient Market Hypothesis* (EMH), is whether the models consistent with the data are predictable or not: It is now widely accepted that the financial series are predictable to a certain extent [5], [6], [11], [17], [18]. Another typical financial objective is the design of *trading strategies* [14]–[16]. A trading strategy is a plan to buy and sell assets that make up a portfolio in order to be profitable. It can use any type of information and traded assets, and it can cover different time horizons [10].

On the other hand, the engineering approach to this topic essentially designs prediction algorithms, mainly based on artificial intelligence and using the time series of prices or other data sources [3], [21]–[26], [30], [31], [34]. Although the typical objective in the design of algorithms lies in the accuracy of the prediction, the final aim should address the corresponding profit. The relationship between the accuracy and profit is not straightforward. Although it is sensible to believe that greater accuracy can generate higher profits, the simple idealized example of an asset with constant prices shows that it may not be like this: since it remains constant, it can be predicted perfectly, however since it does not change its return is none.

In this paper, we address the relationship between the prediction accuracy and the profit that can be obtained in real markets, a relationship that is sometimes referred to as *the economic significance of the stock predictability* [33]. The search for profits generally comes in terms of a trading strategy, while the predictability accuracy is associated with the prediction algorithms. Therefore, we define *prediction strategies* as those trading strategies that are explicitly built

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

on a prediction algorithm. Our purpose is to discuss to what extent the profit can be credited to the prediction algorithm.

To address our objective, we must formulate the problem of stock prediction in a set of very simple prediction strategies, collect real market data and interpret its results in a formal learning theory. As the setting for our prediction strategies, we have considered the problem of forecasting daily stocks individually, taking as the explanatory variables the time series of previous closing prices. As simple as it seems, *it can yield interesting insights*, paraphrasing Campbell [5]. The daily frequency of forecasts is selected because it is a good compromise between the better predictability of short periods and the transaction costs [37] on the one hand, and the resources available to collect and process data on the other. Prediction is made of the binary movements instead of continuous changes; that is, our algorithms predict whether or not the price will rise the next day, instead of the more difficult problem to guess its value.

Our prediction strategies are not designed to find the optimum profit or obtain the best prediction accuracy but to express our results as easily and clearly as possible. We consider four prediction strategies to develop our discussion: the *Support Vector Machine* strategy (SVM), the *Efficient Market* strategy (EMS), the *Buy and Hold* strategy (B&H) and the *Optimal* strategy (OPT). We will refer to their prediction algorithms with the same acronyms when there is no risk of confusion. They all rely on the same data for a fair comparison. The SVM prediction strategy uses a SVM algorithm and decides to buy or sell an asset at the beginning of each day if its SVM predicts that its price will rise or fall, respectively. We use the *Support Vector Machines* as possibly the most fashionable machine learning representative of the years for which we have data and also capable of delivering very good prediction results [3], [30]. The EMS strategy is similar to the SVM strategy but its prediction algorithm uses the day return as the next day's return. The return for a day is defined as the price of that day divided by the price of the previous day so that this algorithm follows a similar idea to the EMH but in trends rather than prices. Therefore, the EMS algorithm is perhaps the simplest possible. When compared to the SVM, we can discuss the effect of the algorithmic complexity on the profits, being the rest of those strategies equal. The other two prediction strategies are references to frame our discussion and do not have prediction algorithms as such, but simply consider the extremes of prediction: the B&H does not predict at all, always buys the asset and keeps it until the end of the considered time horizon; on the other hand, the OPT exhibits a perfect prediction and applies to this prediction the same strategy as the SVM and EMS.

The basis of our discussion is the collection, visualization and interpretation of a large set of asset price data. The lack of a well-established financial data standard [39] and the unstructured nature of these data require special care in their handling as well as the development of customized tools for meaningful visualization. These characteristics allow us to refer to these techniques as *Big Data* [52].We have compiled the daily prices of approximately half of the listed companies worldwide in the period 2007-2016, and some important metadata, such as their industrial classification and their currencies to assign them a regional affiliation.

Finally, we interpret our results against an extension of the Bayesian Learning framework [28]. This theoretical framework allows us to discuss the proposed problem of how the prediction algorithm impacts profit and secondly to show how the *No Free Lunch Theorem* (NFL) applies to the financial stocks prediction strategies [29]. In essence, the NFL says that, under certain conditions, if a prediction algorithm works better in a certain asset class, it will perform worse in the rest. We extend its application to different economic sectors and regions, following the interest of the current literature [9], [11], [13].

Please note that our results do not take into account important issues for the actual trading, such as the spread [32], the trading costs, or the price impact [57], among others.

II. DATA AND THE SVM

The basis for our analysis is a database of listed companies worldwide on top of which we discuss prediction algorithms and some related trading strategies performance. This section presents the details of its collection and the need of filtering in order to make meaningful its economic significance. On the other hand, our prediction strategies are quite simple in their formulation except for the SVM prediction neural network which deserves some previous discussion. Note that the selection of the SVM as a complex type of prediction algorithm is not determinant in the overall discussion and some similar algorithm like Deep Neural Networks might be used. Our main motivation for this selection is that in the time interval corresponding to our data collection, this prediction machine has been the main representative in the literature [30].

A. STOCK MARKET DATABASE

Data are extracted from different online sources with historical closing prices of companies classified by sectors, industries and currencies. Among these sources, there are websites of the main stock exchanges in the world and aggregators such as Yahoo finance.

A recursive query method has been used to collect and process all data from these sources in a distributed server system using a DNS-based approach. This system reduces the time in proportion to the number of servers involved [51]. To achieve this, a cloud computing system was implemented with a centralized database optimized for Big Data analysis [50]. In this way, we could get N virtual servers in the cloud, each with L threads, achieving a significant time reduction. Therefore, we translate the computational time problem to an economic cost proportional to the number of virtual servers involved in the calculation [41].

Due to the way in which the operating system treats the threads in each process, there is a maximum number of threads for the efficiency to decrease [49]. In that case, it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

is more convenient to add an additional server instead of continuing to add new threads.

The centralized database, encoded in MySQL, has access control to avoid typical problems derived from multi-threaded computing, such as 'phantom reads', and all database transactions work using a serializable isolation level provided by the *innobd* engine. [48]. However, having a centralized database implies a limitation on the maximum number of read and write operations. To improve the flow of data, we use mirror databases with MySQL replication system [48]. Therefore, we have two data streams: one write-only and some read databases.

An important part of the database is to demonstrate its consistency. During this validation process two problems were found: splits and discontinuous life cycles of the companies. Some price values suffer abrupt changes as a result from the grouping of securities in the stock exchange or their separation, generically known as splits and reverse splits. This puts the consistency of the data at risk, since they can vary very quickly from one day to the next, which leads to the calculation of false returns that can endanger the rigor of the system. For example, Netflix (NFLX - Nasdaq) with a 7-for-1 stock split on July 14 2015 [47] or Apple (AAPL - Nasdaq) on February 18, 2005 with a 2-for-1 split [46]. To avoid this, we work with a model that corrects these prices directly from the source we use. In this way, the algorithm works only with prices corrected for splits and reverse splits. This correction is made as if the price of the entire series was the result of the operation of split or reverse split. On the other hand, the activity of some companies appears and disappears over time, giving rise to discontinuous life cycles. This may be due to temporary or permanent closures of its activity, as well as for regulatory reasons, changes in the name of the company or acquisitions. To achieve temporal coherence in the analysis, we consider companies whose life cycle provides sufficient data to support our analysis. In the case of companies with name changes (e.g.: Alphabet vs Google [45]), it is treated as if it had always been the same company.

There are 35,324 companies in our database with prices compiled for 10 years, making a total of 105,627,027 entries. Historical data are collected until July 19, 2017. Table1 shows their distribution by currency as a proxy of their geographic location, also illustrated in Fig. 1. Table 2 and Fig. 2 show how companies are distributed among the different economic sectors of activity. These sectors are obtained from the Industry Classification Benchmark (ICB) [44], a global standard. Each company is assigned to the sector that most closely represents the nature of its business.

B. DATA SELECTION AND VISUALIZATION

For each company and year in our database, Fig. 3 plots a red dot showing its maximum daily equivalent return versus the percentage of days in that year that has positive returns. However, this figure is quite misleading to draw realistic conclusions: first, the daily return grows higher than is credibly expected for any company in order to translate this return into real profits. Second, the area of a region in the plot does not correspond to the actual number of companies, since they overlap.

To filter out those companies for which the reported equivalent return cannot credibly translate into actual profit, we need to take into account the relationship between the sensitivity of price variation and the market capitalization. Therefore, those companies with very small capitalization or *small*

Currency	Symbol	Number of Companies
US Dollar	USD	7,743
Japanese Yen	JPY	3,597
Euro	EUR	2,758
Canadian Dollar	CAD	1,961
Australian Dollar	AUD	1,766
Hong Kong Dollar	HK	1,762
Indian Rupee	INR	1,711
Pound sterling	GBX	1,462
Chinese Yuan	CNY	1,085
Malaysian Ringgit	MYR	912
Other currencies	-	10,567

TABLE 1. Distribution of collected companies by currency

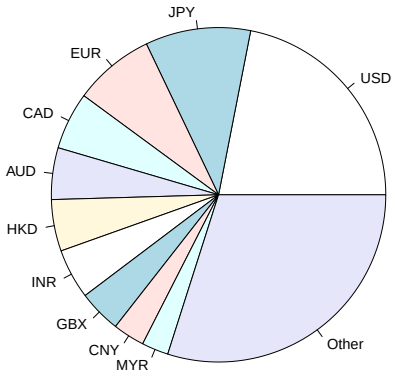

FIGURE 1. Distribution of companies by currency.

Industry	Number of companies
Industrials	7.383
Financials	6.421
Consumer Goods	4.786
Basic Materials	4.357
Consumer Services	3.880
Technology	3.296
Health Care	2.526
Oil and Gas	1.622
Utilities	731
Telecommunications	322

TABLE 2. Distribution of collected companies by industry

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

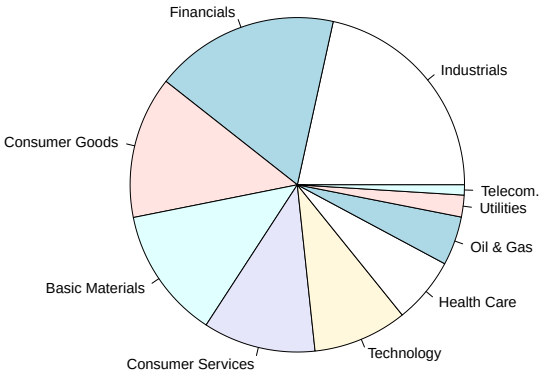

FIGURE 2. Distribution of companies by sectors.

caps exhibit a high return with small volume transaction [35]. An additional relation between market capitalization and average prices can be observed in Fig. 4. As a result of this analysis and considering the relation between the average prices and returns given by Fig. 5, we eliminate those companies with prices below US$ 0.01 for the subsequent discussion.

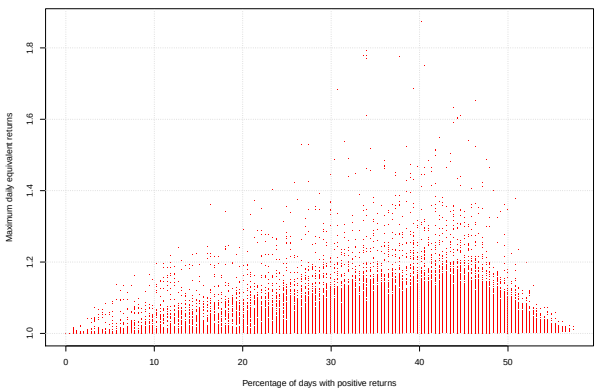

FIGURE 3. A simple illustration of the maximum daily equivalent return versus the percentage of days with positive returns for the total database.

In order to faithfully represent the relationship between prediction accuracy and return, we change the raw points for a statistical aggregate of the companies in accuracy intervals and report the 90% return percentile for that interval, as shown in blue in Fig. 6.

.

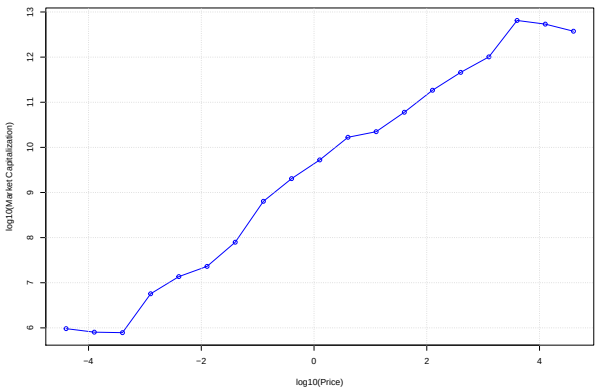

FIGURE 4. The relationship between the market capitalization of companies in our complete database and their prices, both in US$.

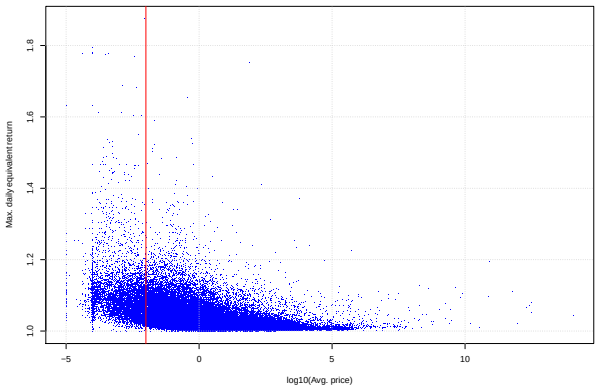

FIGURE 5. Relationship between the maximal equivalent daily returns and the average prices for the complete database. Notice how the 'small caps', that is, those companies with very low prices, show daily equivalent returns that are too large to be credible, as well as high volatility. In red, we show the threshold, 10^{-2} US$, we use to filter our database in order to obtain significant results.

C. THE SVM AND ITS OPTIMIZATION

The SVM is our representative for the best prediction algorithm in our prediction strategies. The full description of the corresponding strategy is provided in the next section, here we describe the SVM parameters optimization in detail. For a description of the SVM and its different classes see [53].

As already discussed in the introduction, our goal is to predict binary outcomes: whether the price will rise or not the next day. Therefore, we evaluate the SVM performance by its prediction accuracy, defined as the percentage of correct predictions in a year of data. Since different companies have different lengths of data in a year, our first challenge is to define *one year of data* for our analysis. We have used the median of the length of data available through the companies, instead of its mean, to get a value insensitive to extreme values. The result is 244 days per year, which is a bit smaller than the number of trading days (255).

In order to find the best predictability, SVMs with different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

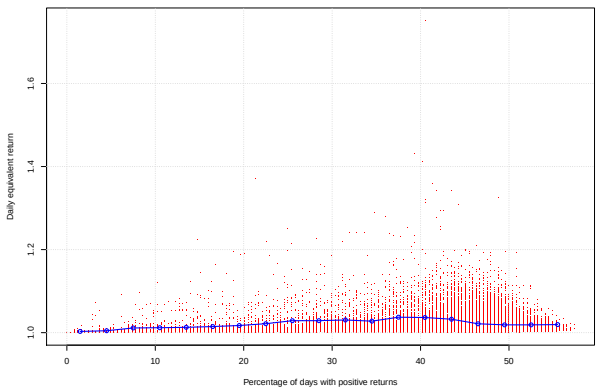

FIGURE 6. In blue, the 90%-percentile of the maximum daily equivalent return in accuracy intervals of the filtered data. The meaning of this percentile is that 10% of the companies have greater return than that shown in the graph. Notice how the basic representation for this relationship, in red, is misleading with respect the actual density of the points

configuration parameters are trained and optimized using the Python implementation of Scikit-learn [43]. The training parameters considered are: the number of vectors, their length, the kernel and type of SVM and the data format, either raw prices or returns. For each company, we search the parameter space for the highest prediction rate. We have considered six vector lengths, six numbers of vectors, three kernels and three SVM types, which means a total of 72,576 simulations per company. We made an initial exploration for optimal parameters with 5% of the companies randomly selected (1,550), making a total of 112,492,800 simulations. To assess the centrality and dispersion of the results, percentiles are taken to avoid the influence of outliers. This is especially useful in the calculation of returns, where the dispersion is greater.

To train the algorithm, future values are not used but only values from the past to the reference date X_0. For each company, in order to predict the return obtained from day X_1, we build n training vectors, where $n \in \{1, 5, 19, 61, 122, 244\}$, by rotating a time window of length W:

$$
\begin{aligned}
&Vector_1 : [X_0, X_{-1}, \ldots, X_{-W+1}] \\
&Vector_2 : [X_{-1}, X_{-2}, \ldots, X_{-W}] \\
&\ldots \\
&Vector_n : [X_{-n+1}, X_{-n}, \ldots, X_{-n-W+2}]
\end{aligned}
$$

Table 3 shows that a vector of length $n = 5$ (days) provides slightly greater percentiles than the rest of the lengths.

For the sample length, W, we have tested the following alternatives: one day (1 closing price), one week (5 closing prices), one month (19 closing prices), one quarter (61 closing prices), half a year (122 closing prices) and one year (244 closing prices). Table 4 shows that the size of the vector does not have much impact on the results, so we use single-day samples, $W = 1$, since it accelerates data processing.

Q	1	5	19	61	122	244
0%	0.3483	0.3319	0.3442	0.2540	0.1188	0.0081
25%	0.5040	0.5204	0.5081	0.5040	0.5081	0.5040
50%	0.5450	0.5778	0.5696	0.57377	0.5737	0.5737
75%	0.6106	0.6885	0.6844	0.6721	0.6721	0.6680
100%	0.9959	0.9959	0.9959	1.0000	1.0000	1.0000

TABLE 3. Accuracy percentiles for the number of training vectors

Q	1 day	1 week	1 month	1 quarter	1/2 year	1 year
0%	0.0122	0.0081	0.0081	0.1106	0.1844	0.1639
25%	0.5081	0.5081	0.5081	0.5081	0.5081	0.5081
50%	0.5737	0.5737	0.5655	0.5614	0.5614	0.5655
75%	0.6680	0.6680	0.6680	0.6639	0.6639	0.6680
100%	0.9959	0.9959	0.9959	1.0000	1.0000	1.0000

TABLE 4. Accuracy percentiles for the training vector lengths

We have considered three different kernels for the SVM: the linear, the radial basis function (RBF) and the sigmoidal. As shown in Table 5, there is a small predictive improvement of the RBF and sigmoidal kernels over the linear ones. Between these two kernels, we have opted for the RBF because its calculation is the fastest.

Finally, two different SVM configurations are tested for prediction optimality: the regression (SVR) and classifier (SVC) modes. Although the classifier configuration adapts naturally to a binary forecast, the regression configuration output can be considered a *soft* decision to be later quantized into two *hard* binary values using a convenient threshold. Additionally, two different data formats are tested: the price closing data and the return. Table 6 shows the results of our tests; since the SVC gets the same results for both data formats, only one column is shown. There, it can be observed how the SVR working on returns offers the best performance.

In summary, our SVM algorithm is selected to work with a RBF kernel in the regression mode, over the returns, and with 5 training vectors of length 1.

Q	Linear	RBF	Sigmoid
0%	0.008	0.0081	0.0163
25%	0.5081	0.5122	0.5122
50%	0.5532	0.5737	0.5778
75%	0.6557	0.6762	0.6762
100%	1.0000	1.0000	1.0000

TABLE 5. Accuracy percentiles for the SVM kernels

Q	SVC (prices/return)	SVR (prices)	SVR (return)
0%	0.3483	0.1639	0.0081
25%	0.5081	0.4790	0.6106
50%	0.5573	0.5163	0.6844
75%	0.6229	0.5614	0.7172
100%	1.0000	1.0000	0.9959

TABLE 6. Accuracy percentiles for both SVM types and data formats

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

III. PREDICTION, TRADING STRATEGIES AND THE NO FREE LUNCH THEOREM

The interpretation of the data to find the aimed relationships among the algorithms predictive power and their related trading strategies' returns requires a theoretical framework which may trigger further research questions. We first define some very simple trading strategies whose main purpose is to translate the predictive power of the algorithms they are based on directly to their returns. Please note that our objective is not to get an optimum trading strategy but to prepare the tools for our discussion. The *Extended Bayesian Framework* (EBF), which gives the complete real data a predominant role in the relationship between the algorithmic prediction and the obtained results, is therefore a good candidate for getting the most out of a data set which is more representative of the whole real world data than a statistical sample.

A. PREDICTION STRATEGIES

We define a *prediction strategy* as a trading strategy that uses a prediction algorithm. While a precision can be attributed to the prediction algorithm, the final performance obtained must be credited to the trading strategy. An important objective in this paper is to analyze the impact that the algorithmic accuracy has on the final return, and to delimit what part of the credit it has in the entire strategy. We will refer to the whole trading strategy as the *prediction strategy* when we want to emphasize this mixture and use the term *prediction algorithm* when we want to emphasize the impact of the predictive algorithm. To elaborate our discussion, we have defined four prediction strategies: the SVM strategy (SVM), the Efficient Market strategy (EM), the Buy and Hold strategy (B&H), and the Optimal strategy (OPT).

The SVM strategy uses the prediction capabilities of a Support Vector Machine. This artificial neural network has been the most celebrated among practitioners in the prediction of stock prices for many years, see [30]. We use it to predict from the previous daily returns weather the next day return will be above or below one, which will result directly in a purchase or not purchase order, respectively. This strategy is our representative of the machine learning ingenuity towards trading.

The EM strategy is inspired in the Efficient Market Hypothesis. It projects the return of one day to the next: if it is greater than one, the action is to buy. The generalization of the EMH to returns implies that, instead of prices, we consider that efficiency also entails price derivatives. Let us clarify that we are not claiming any kind of statement, simply using this projection as a prediction strategy adequately close in terms of previous returns.

The B&H strategy is a typical reference for the other strategies, see for instance [42]. The asset in this strategy is always bought at the opening price of the day and sold at the closing price of that day. Obviously, it does not take into account the changes of price between the closing price of a day and the opening price of the following, but in this way its comparison with the other strategies is fairer in terms of earnings due to predictability. Note also that its predictability is zero, so the *the accuracy* of 'this prediction' is the actual percentage of days with returns greater than one for each asset.

The OPT is also a reference strategy for our discussion purpose. It implies a perfect predictability, so it reflects the highest possible return that a stock can have.

For each company in our database and each strategy, we have calculated their accuracy and return: Fig. 7 shows their return distribution and Fig. 8 their accuracy distribution. The compounded daily return in a year grows exponentially, so to summarize the annual result in a daily return figure, we consider the *equivalent daily return* r that is calculated from the total annual return of a company AR considering 255 days of trading in a year as $r = AR^{1/255}$.

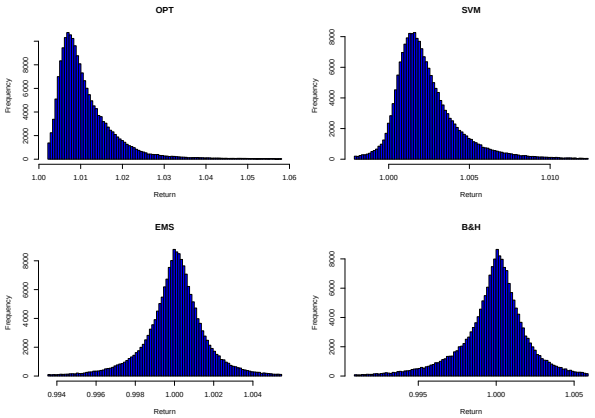

FIGURE 7. Return histograms for the prediction strategies applied to our filtered database.Please note that the ranges shown for the different strategies are different.

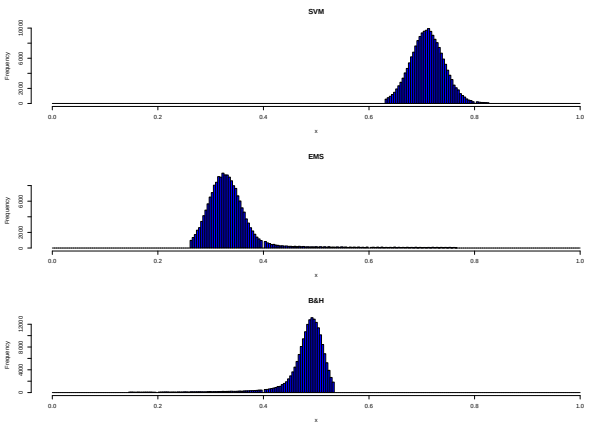

FIGURE 8. Accuracy histograms for the prediction strategies.We do not include the OPT strategy since its accuracy is 1 by definition. For the B&H strategy, the accuracy is the number of days with returns greater than or equal to 1.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

B. THE NO FREE LUNCH THEOREM INTERPRETATION

First we summarize the theoretical framework related to the NFL theorem and then proceed to interpret the data by defining heuristics that conceptually coincide with the theoretical concepts in the aforementioned framework.

The Extended Bayesian Framework (EBF) is formalized following [28]. Let X be an input space and Y an output space. The unknown function from X to Y to be learned is called *target* function, f. We are given a set of samples of the target function called the *training* set, $d = \{(x_i, f(x_i))\}_{i=1..m}$, for some $m > 0$. The output of the learning algorithm is called *hypothesis* function, h. We assume that the set of all f's is F, and the set of all h's is H. All information about the relationships between these elements is specified by a probability distribution that depends on the learning algorithm i, $P_i(f, h, d)$ over an appropriate event space. For example, the dependence of the hypothesis function on the training data is given by $P_i(h|d) = P_i(h, d)/P_i(d)$. The measure of how well a learning algorithm performs is given by a *cost* function associated with a discrepancy between h and f that can also dependend on d, $c(h, f, d)$. This cost in the Wolpert model can be written as a inner product between their distributions: $E_i(C|d) = \sum_{h,f} c(h, f, d)P_i(h|d)P(f|d)$. The key point in this probabilistic framework is the explicit distinction between the two spaces F and H. While the researcher has total control of H, F is *set by the physical universe and outside the researcher's control*. If $H = F$, we are back to the conventional *Bayesian Framework*. Discussing their difference allows to encompass and compare different learning frameworks like the *PAC* or the Vapnik's*VC* [29].

Within the EBF, the *No Free Lunch Theorem* for supervised learning can be expressed as follows [29]: Let i and j be two learning algorithms, then *uniformly* averaging over all possible fs, and for any training set d, $E_i(C|f, d) = E_j(C|f, d)$, that is, their performances are equal. This theorem only states that there is no best algorithm for everything, it must be adapted to its role, formally given by the projection of $P(F|d)$ on $P(H|d)$ in the Wolpert's model.

To interpret the data, we propose an intuitive mapping between the data distributions and the concepts in the EBF. To do that, we introduce an additional independent variable in the EBF terms, the algorithmic prediction accuracy α. As a rigorous mathematical characterization is beyond the scope of this paper, we will refer generically to them as *heuristics*. First, we would like to characterize the predictive algorithm i with the projection of the real data $P(f|d)$ to the hypothesis functions $P_i(h, \alpha|d)$. A reasonable candidate heuristic could be the distribution on the accuracy of this algorithm, see Fig. 8. Intuitively, this distribution measures how well the algorithm matches the actual function. However, the economic significance of this match is actually given by the return delivered by the prediction, so this term should be taken into account. As a simple example already given in the introduction, consider a fixed price that gives a flat return of 1. Taken the previous day price as today's would offer the maximum accuracy, but its benefit is none. To provide a meaningful heuristic, the gain obtained by the strategy must be included in this heuristic and the *cost* term $c(h, f, d)$ in the EBF can do the job. To emphasize the new role for this formal term, we will refer to it as the *gain* term, $g(h, \alpha, f, d)$. However, its full inclusion would take into account how optimal the strategy is above the prediction value. Therefore, we divide this gain into a term that explains the *efficiency* of the strategy, $eff(h, \alpha, f, d)$, and another related to the *prediction matching*, $pm(h, \alpha, f, d)$, such that: $g(h, \alpha, f, d) = pm(h, \alpha, f, d)\ eff(h, \alpha, f, d)$.

With these definitions, we propose our first heuristic $H1$ as the distribution on the accuracy of the maximum return a prediction algorithm can have, as shown in Fig.9. This distribution is interpreted as the EBF term $\sum_{h,f} pm(h, \alpha, f, d)P_i(h, \alpha|d)P(f|d)$. Since we reflect in this heuristic the maximum gain for an algorithm, therefore, the efficiency term of the strategy is bounded by 1. The second heuristic, $H2$, is the distribution of the real return by a prediction strategy that takes into account both the matching of the prediction algorithm and the efficiency of the strategy, see Fig.10. In relation to the EBF, $H2$ corresponds to the total gain, G: $E_i(G, \alpha|d) = \sum_{h,f} g(h, \alpha, f, d)P_i(h, \alpha|d)P(f|d) = \sum_{h,f} pm(h, \alpha, f, d)eff(h, \alpha, f, d)P_i(h, \alpha|d)P(f|d)$.

These definitions can be supported intuitively by carefully comparing their graphs. To appreciate it better, we use a logarithmic scale and we weight the real returns so that its maximum is adjusted to the maximum returns, see Fig. 11. It can be observed how the maximum potential return basically follows the real returns obtained by the strategy, so that the efficiency term is independent of the accuracy to a large extent, that is $eff(h, \alpha, f, d) \sim eff(h, f, d)$. Therefore, the accuracy of the algorithmic prediction is mainly related to $H1$. These heuristics make precise the difference between the return credited to the prediction algorithm and that attributed to the strategy using it, although this conceptual partition may not be easy or even possible to apply in real strategies.

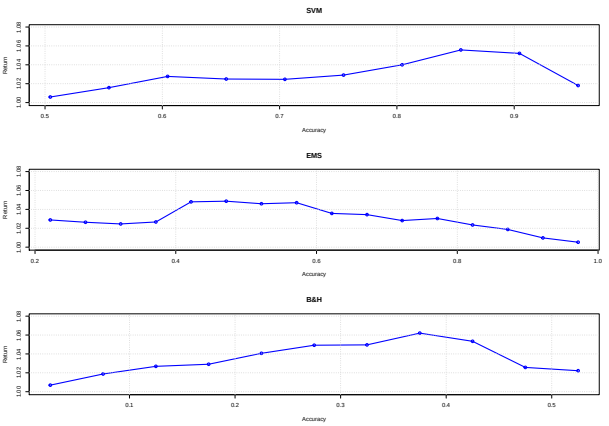

FIGURE 9. Maximal returns are plotted versus accuracy for the different prediction strategies. These distributions define the $H1$ heuristic.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

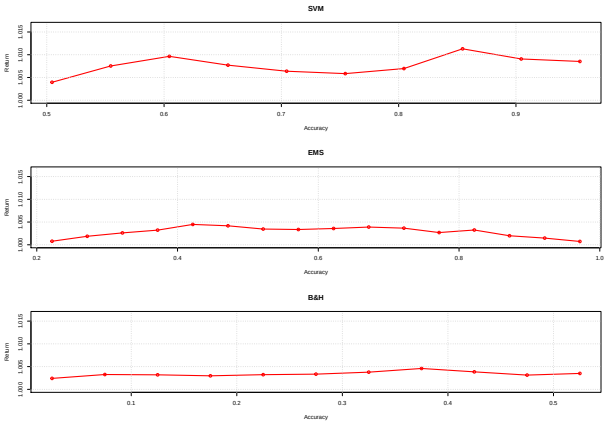

FIGURE 10. Actual returns are plotted versus accuracy for the different prediction strategies. These distributions define the $H2$ heuristic.

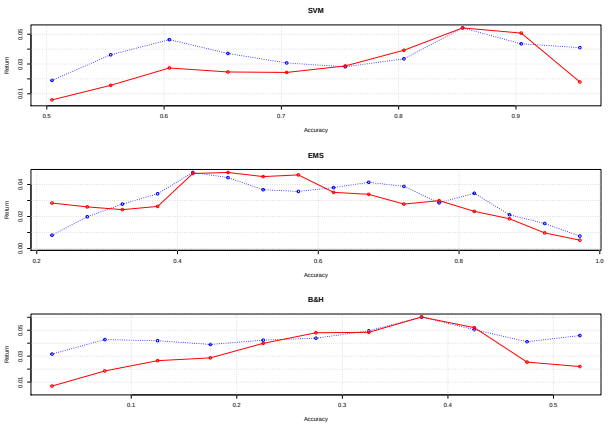

FIGURE 11. The maximum and weighed real returns are plotted on a logarithmic scale versus accuracy for the different prediction strategies. This visualization allows us to appreciate how the real returns essentially follows the maximum returns for different values of accuracy.

$H1$ offers interesting information about the economic significance of the algorithmic accuracy. Looking at Fig.9, it can be seen that the financial assets with higher returns have different accuracies due to use of different algorithms. Therefore, the simple conclusion that a lower accuracy is related to a lower return is not correct. Nor can it be expected that a prediction strategy with maximum return at a low accuracy can offer sufficient efficiency to obtain better returns than prediction strategies with high accuracy at their maximal returns. However, the NFL theorem will point out that, for different subsets of data, that may indeed be possible.

A first illustration of the NFL is provided for the SVM and EMS prediction algorithms in Figs.12 and Fig. 13, which show for each company the accuracies obtained with these prediction algorithms and their corresponding returns, respectively. It is important to bear in mind that these basic representations are a little misleading regarding with respect to the number of companies in each region of the plots: the percentage of companies with the highest accuracy using the SVM is 99% and the corresponding percentage with the highest return is 89%. For this 11% of the companies, their EMS's return is greater than their SVM's, although their accuracy is lower. This fact emphasizes the fundamental mismatch between the accuracy of the prediction and the return, that we attribute to the term $pm(h, \alpha, f, d)$.

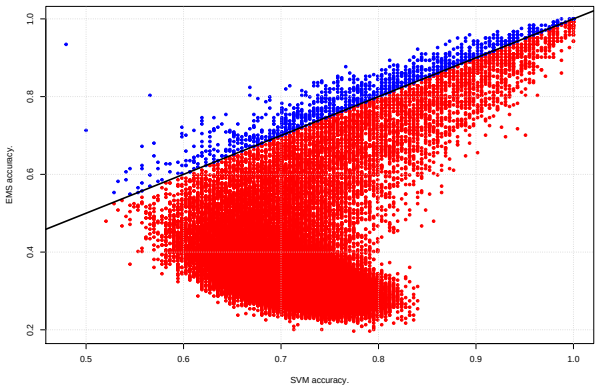

FIGURE 12. The SVM accuracy for each company is shown versus its EMS accuracy. For one 1% of the companies, the EMS accuracy is greater than the SVM's (in blue).

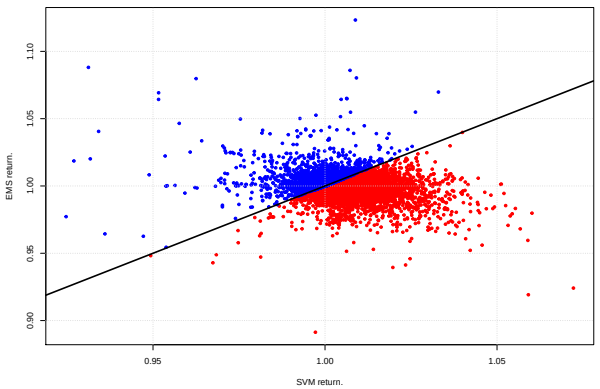

FIGURE 13. The SVM real return for each company is shown versus its EMS actual return. For $11\%$of the companies, the EMS return is greater than the SVM's (in blue).

The comparison between the B&H and EMS strategies, makes this effect more striking, see Figs. 14 and 15. While the percentage of companies for which the accuracy obtained by B&H is higher than that obtained with the EMS is 93%, the percentage of companies that get more return with B& H is only 49%. This can be explained because the strategy B& H does not reject those days with losses while the EMS gets rid of them. Therefore, the efficiency of the EMS is much greater than that of the B&H, which does not actually use any prediction capability as already mentioned.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

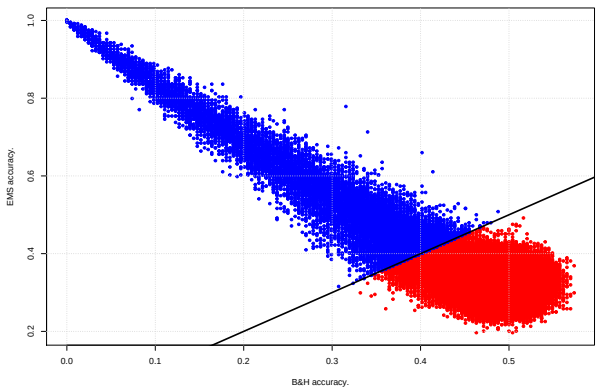

FIGURE 14. The EMS accuracy for each company is shown versus its B&H accuracy. For 7% of the companies, the EMS accuracy is greater than the B&H's (in blue).

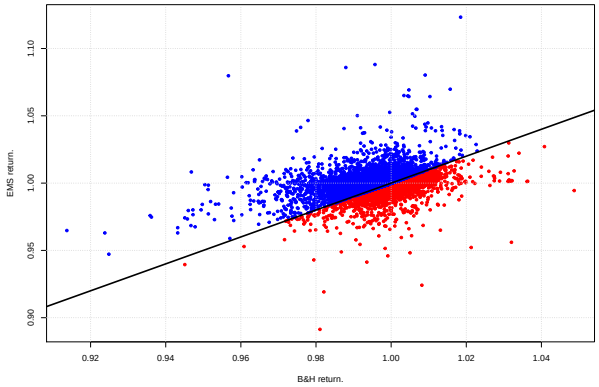

FIGURE 15. The EMS real return for each company is shown versus its B&H actual return. For $51\%$of the companies, the EMS return is greater than the B&H's (in blue).

C. ECONOMIC SECTORS AND REGIONS AND PREDICTION PERSISTENCY THROUGH TIME

As suggested in the previous section, our $H2$ heuristic allows us to deepen empirically in the relationship of the prediction algorithms and the different real data partitions, in sectors, geographical regions, and years, under the perspective of the NFL theorem. To emphasize the difference in the returns, we have selected the companies with the highest returns, setting the percentile of the representation at 99.0% for the plots in this section.

Fig. 16 shows the real returns delivered by the prediction strategies for different economic sectors. Notice how the worst *a priori* strategy, the B&H, has obtained highest return peaks than the SVM, the *a priori* best strategy, in the sectors of Utilities, Technology, Finance and Health Care in the period considered. Even the EMS has got a best peak in the sector of Consumer Services.

Fig. 17 shows similar illustrations of the NFL theorem in the geopolitical regions as show by coin proxies. It is particularly interesting to note that, in the USA market, the best prediction strategy seems to be not predict at all, since it is the B&H that gets the best results.

Finally, Fig. 18 allows us to discuss the NFL theorem during the ten years covered by the collected data. When considering the set of companies worldwide, the performance of the prediction algorithms seems quite persistent in their returns versus accuracy. The best peaks typically belong to the SVM, although the B&H strategy takes the lead every three or four years.

IV. CONCLUSIONS AND FURTHER WORK

We have used the Extended Bayesian Framework and the No Free Lunch theorem together with a big set of daily prices for almost half of the publicly traded companies around the world to discuss the relationship between the accuracy of prediction algorithms and their use by means of trading strategies. Through the definition of a pair of heuristics related to the theoretical terms, some unexpected results show how the *a priori* less accurate algorithms and inefficient strategies can offer better results than the *a priori* best alternatives in some particular subsets of data that have a clear interpretation in terms of economic sectors and regions.

The proposed link between the theoretical concepts in the Extended Bayesian Framework and our Big Data heuristics allows some additional research questions. For example, we have shown how the efficiency of a strategy, $eff(h, \alpha, f, d)$, is independent of the accuracy of the prediction algorithm α to a large extent. However, it is not independent totally, and this fact may suggest a deeper relationship between the prediction algorithm and the trading strategy that uses it. Given that many trading strategies can be formalized as optimization algorithms, and given that the No Free Lunch Theorem was formulated in that framework [27], unfolding such a relationship would result in a fruitful discussion.

REFERENCES

[1] A. Timmermann, "Elusive return predictability," *Intl. J. of Forecasting*., vol. 24, no. 1, pp. 1–18, 2008.

[2] R.D. McLean, J. Pontiff, "Does academic research destroy stock return predictability?," *The Journal of Finance*., vol. 71, no 1, p. 5-32,2016.

[3] M.-W. Hsu et al.,"Bridging the divide in financial market forecasting: machine learners vs. financial economists", *Expert Systems with Applications*., vol. 61, p. 215–234, 2016.

[4] J. Y. Campbell, A. W. Lo and A. C. MacKinlay, "Introduction," in *The Econometrics of Financial Markets*.Princeton, (New Jersey), USA: Princeton Univ. Press, 1997, ch. 1, sec. 1.5, pp. *20–25.*

[5] J. Y. Campbell, A. W. Lo and A. C. MacKinlay, "The Predictability of Asset Returns," in *The Econometrics of Financial Markets*.Princeton, (New Jersey), USA: Princeton Univ. Press, 1997, ch. 2, pp. *27–82.*

[6] A.W. Lo, A.C. MacKinlay, "Stock market prices do not follow random walks: Evidence from a simple specification test", *The review of financial studies*, vol. 1, no 1, p. 41-66, 1988.

[7] H. Jang, J. Lee,"An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information",*IEEE Access*, vol. 6, p. 5427–5437, 2018.

[8] A.Detzel, et al.,"Bitcoin: predictability and profitability via technical analysis". *SSRN Electronic Journal, DOI*, vol. 10, 2018.

[9] D. Bannigidamath, P.K. Narayan, "Stock return predictability and determinants of predictability and profits",*Emerging Markets Review*, vol. 26, p. 153-173, 2016.

[10] J. Conrad, G. Kaul, "An anatomy of trading strategies", *The Review of Financial Studies*, vol. 11, no 3, p. 489-519, 1998.

[11] A. Charles,O. Darné,J.H. Kim, "International stock return predictability: Evidence from new statistical tests",*International Review of Financial Analysis*, vol. 54, p. 97–113, 2017.

[12] V. Poti, et al., "Predictability, trading rule profitability and learning in currency markets", *International Review of Financial Analysis*, vol. 33, p. 117–129, 2014.

[13] C.R.Harvey, "Predictable risk and returns in emerging markets", *The review of financial studies*, vol. 8, no 3, p. 773–816, 1995.

[14] R. Gencay, "Optimization of technical trading strategies and the profitability in security markets", *Economics Letters*, vol. 59, no 2, p. 249–254,, 1998.

[15] A.C.Szakmary, Q. Shen, S.C. Sharma, "Trend-following trading strategies in commodity futures: A re-examination", *Journal of Banking & Finance*, vol. 34, no 2, p. 409–426, 2010.

[16] K.-Y. Kwon, R.J. Kish, "Technical trading strategies and return predictability: NYSE", *Applied Financial Economics*, vol. 12, no 9, p. 639–653, 2002.

[17] J. Westerlund, P. Narayan, "Testing for predictability in conditionally heteroskedastic stock returns", *Journal of Financial Econometrics*, vol. 13, no 2, p. 342–375, 2014.

[18] E.J. Chang, E.J.A. Lima, B.M. Tabak, "Testing for predictability in emerging equity markets", *Emerging Markets Review*, vol. 5, no 3, p. 295–316, 2004.

[19] J.S. Armstrong, "Forecasting with econometric methods: Folklore versus fact", *Journal of Business*, p. 549–564, 1978.

[20] A.W. Lo,A.C. MacKinlay, "Stock market prices do not follow random walks: Evidence from a simple specification test", *The review of financial studies*, vol. 1, no 1, p. 41–66, 1988.

[21] X. Zhang, et al., "Stock market prediction via multi-source multiple instance learning", *IEEE Access*, vol. 6, p. 50720–50728, 2018.

[22] X. Li, et al., "Stock Prediction via Sentimental Transfer Learning", *IEEE Access*, vol. 6, p. 73110–73118, 2018.

[23] M. Prause, J. Weigand, "Market Model Benchmark Suite for Machine Learning Techniques", *IEEE Computational Intelligence Magazine*, vol. 13, no 4, p. 14–24, 2018.

[24] L. Chen, et al., "Which artificial intelligence algorithm better predicts the chinese stock market?", *IEEE Access*, vol. 6, p. 48625–48633, 2018.

[25] G. Liu, X. Wang, "A Numerical-Based Attention Method for Stock Market Prediction With Dual Information", *IEEE Access*, vol. 7, p. 7357–7367, 2019.

[26] P. Raña, J. Vilar, G. Aneiros, "On the Use of Functional Additive Models for Electricity Demand and Price Prediction", *IEEE Access*, vol. 6, p. 9603–9613, 2018.

[27] D.H. Wolpert, et al., "No free lunch theorems for optimization", *IEEE transactions on evolutionary computation*, 1997, vol. 1, no 1, p. 67-82, 1997.

[28] D.H. Wolpert, "On the connection between in-sample testing and generalization error", *Complex Systems*, vol. 6, no 1, p. 47–94, 1992.

[29] D.H. Wolpert,"The supervised learning no-free-lunch theorems", in *Soft computing and industry*, Springer, London, p. 25-42, 2002.

[30] J. Jaramillo, J.D. Velasquez, C.J. Franco, "Research in financial time series forecasting with SVM: Contributions from literature", *IEEE Latin America Transactions*, vol. 15, no 1, p. 145–153, 2017.

[31] B. Weng et al., "Predicting short-term stock prices using ensemble methods and online data sources", *Expert Systems with Applications*, vol. 112, p. 258–273, 2018.

[32] P.J. Knez, M.J. Ready, "Estimating the profits from trading strategies", *The Review of Financial Studies*, vol. 9, no 4, p. 1121–1163, 1996.

[33] M.H. Pesaran, A. Timmermann, "Predictability of stock returns: Robustness and economic significance", *The Journal of Finance*, vol. 50, no 4, p. 1201-1228, 1995.

[34] T.P. Michalak, M. Wooldridge, "AI and Economics" [Guest editors' introduction], *IEEE Intelligent Systems*, vol. 32, no 1, p. 5-7, 2017.

[35] W. De Groot, J. Huij, W. Zhou, "Another look at trading costs and short-term reversal profits", *Journal of Banking & Finance*, vol. 36, no 2, p. 371-382, 2012.

[36] B.S. Abbey, J.A. Doukas, "Do individual currency traders make money?", *Journal of International Money and Finance*, 2015, vol. 56, p. 158-177, 2015.

[37] B. Foltice, T. Langer, "Profitable momentum trading strategies for individual investors", *Financial Markets and Portfolio Management*, vol. 29, no 2, p. 85-113, 2015.

[38] W.Small, H.-H. Hsieh, "Style Influences And JSE Sector Returns: Evidence From The South African Stock Market", *Journal of Applied Business Research*, vol. 33, no 5, p. 863-872, 2017.

[39] S. Yang et al., "Framework formation of financial data classification standard in the era of the big data", *Procedia Computer Science*, vol. 30, p. 88-96, 2014.

[40] F. Lillo, J.D. Farmer, R.N. Mantegna, "Econophysics: Master curve for price-impact function", *Nature*, vol. 421, no 6919, p. 129-130, 2003.

[41] P. Kazenin, 'Optimal number of threads in parallel computing', *Pavel Kazenin's blog*., Aug 2, 2014 [Online]. Available: https://pavelkazenin.wordpress.com/2014/08/02/optimal-number-of-threads-in-parallel-computing/.

[42] F. Fernández-Rodríguez, C. González-Martel, S. Sosvilla-Rivero,"On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market", *Economics letters*., vol. 69, no 1, pp. 89–94, 2000.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research 12*., pp. 2825–2830, 2011.

[44] 'ICB Structure And Definitions' *FTSE Russell*, 2012. [Online]. Available: https://www.ftse.com/products/downloads/ICBStructure-Eng.pdf.

[45] J. D'Onfro, 'Google is now Alphabet,' *Business Insider*, Oct 2, 2015. [Online]. Available: https://www.businessinsider.com/google-officially-becomes-alphabet-today-2015-10.

[46] Apple, 'Apple Announces Two-for-One Stock Split,' *Apple press release*, Cupertino (California), USA: Apple Newsroom, Feb. 11, 2005. [Online]. Available: https://www.apple.com/newsroom/2005/02/11Apple-Announces-Two-for-One-Stock-Split/.

[47] J. Pramuk, "Netflix splits stock 7 for 1',' *CNBC*., Jun. 23, 2015. [Online]. Available: https://www.cnbc.com/2015/06/23/netflix-splits-stock-7-for-1.html.

[48] B. Schwartz et al. "High Performance MySQL" *O'Reilly Media* 2nd ed., Sebastopol (California), USA: O'Reilly Media, 2008, pp. 7–31, 447–581.

[49] Saavedra-Barrera, R. H., Culler, D. E., and Von Eicken, T. "Analysis of multithreaded architectures for parallel computing" *2nd Annual ACM Symposium on Parallel Algorithms and Architectures*, *ACM*, (New York), USA: pp. 2–8, 1990.

[50] Ji et al, "Big Data Processing in Cloud Computing Environments," *International Symposium on Pervasive Systems, Algorithms and Networks*., San Marcos (Texas), USA: pp. 17–21, 1999.

[51] V. Cardellini, "Dynamic Load Balancing on Web-server Systems," *Iee Explore*., vol. 3, no. 3, pp. 5–9, 1999.

[52] E.Y. Gorodov, V.V. Gubarev, "Analytical review of data visualization methods in application to big data", *Journal of Electrical and Computer Engineering*, vol. 2013, p. 22-29, 2013.

[53] B. Scholkopf, A.J.Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond",*MIT press*, 2001.

[54] N.K. Ahmed, et al., "An empirical comparison of machine learning models for time series forecasting", *Econometric Reviews*, vol. 29, no 5-6, p. 594-621, 2010.

[55] S.Kandel,R.F. Stambaugh, "On the predictability of stock returns: an asset-allocation perspective", *The Journal of Finance*, vol. 51, no 2, p. 385–424, 1996.

[56] K. Joseph, M.B. Wintoki, Z. Zhang, "Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search",*International Journal of Forecasting*, vol. 27, no 4, p. 1116-1127, 2011.

[57] A. Dufour, R.F. Engle, "Time and the price impact of a trade", *The Journal of Finance*, vol. 55, no 6, p. 2467-2498, 2000.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

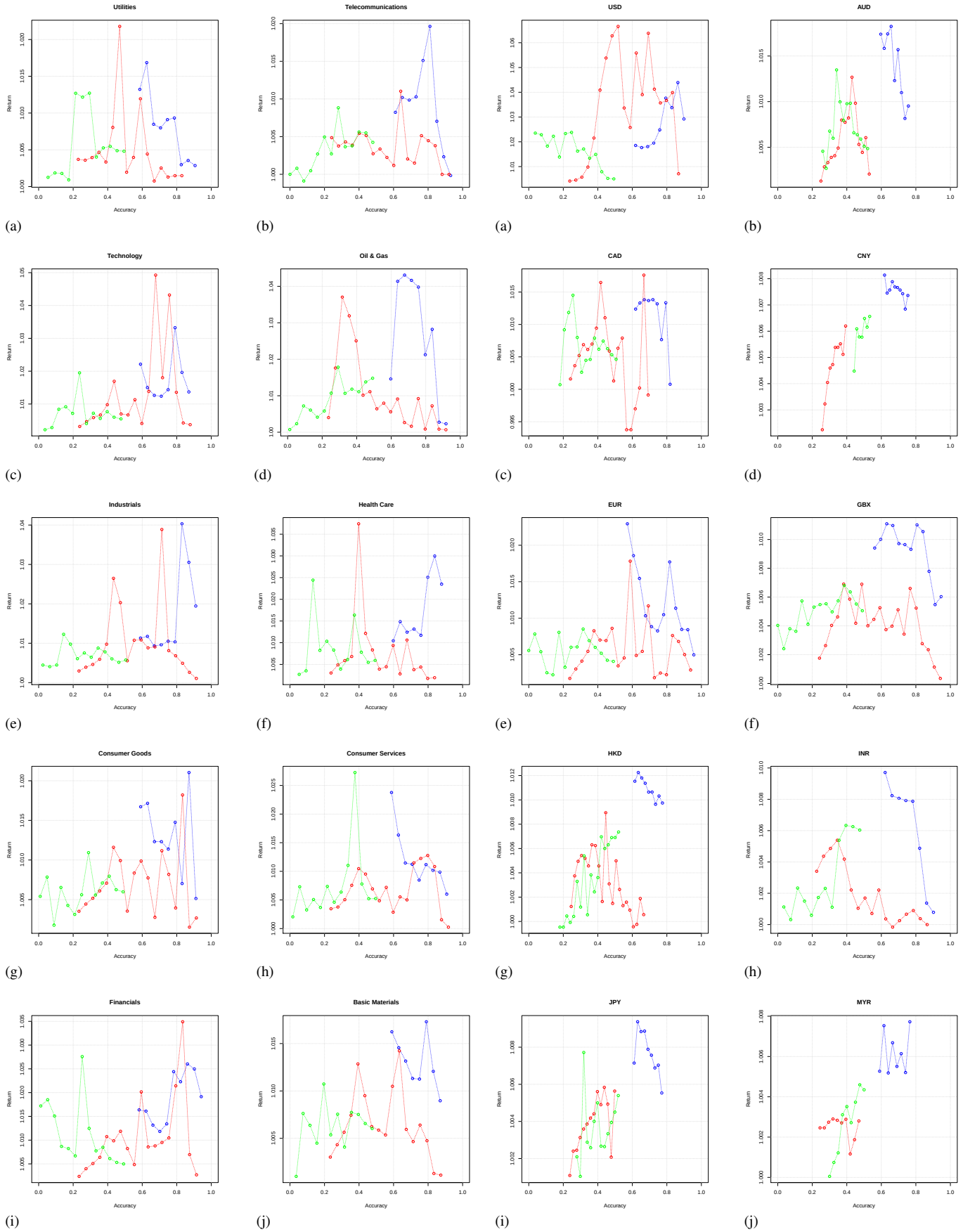

FIGURE 16. Return vs. accuracy for economic sectors: SVM, EMS and B&H.

FIGURE 17. Return vs. accuracy for economic regions: SVM, EMS and B&H.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2921092, IEEE Access

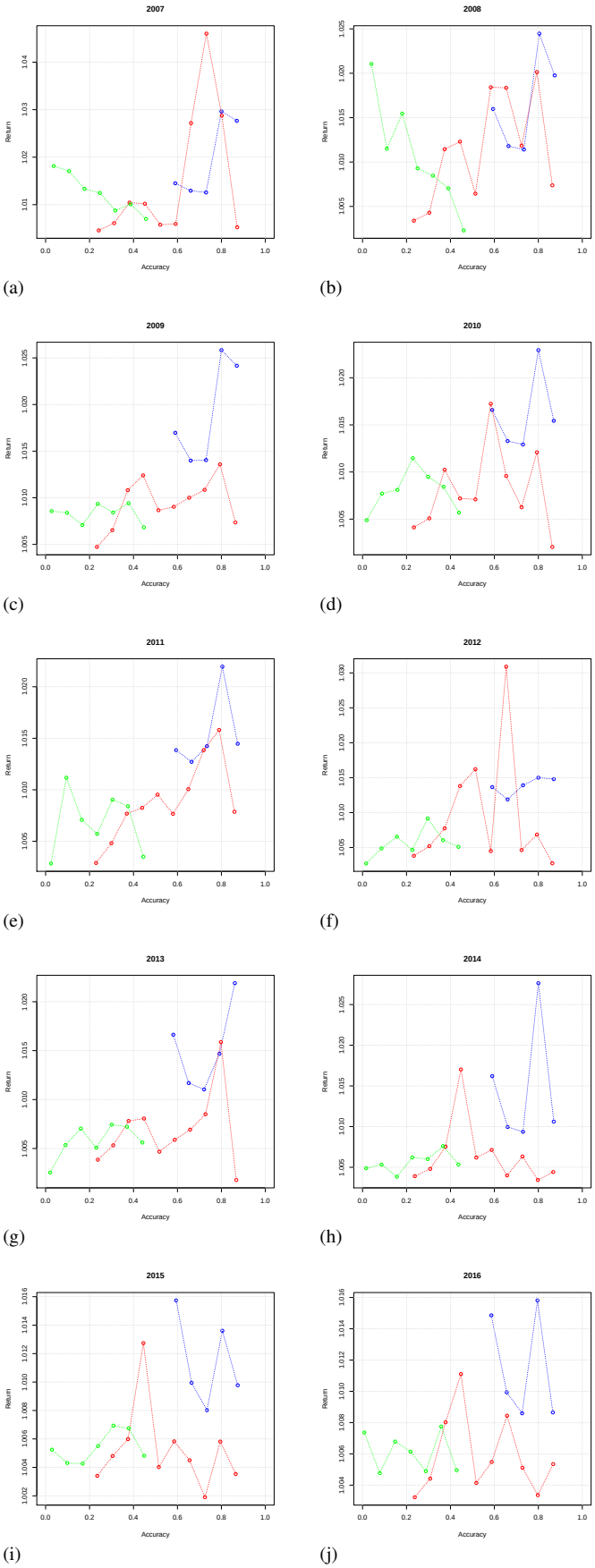

FIGURE 18. Return vs. accuracy for years: SVM, EMS and B&H