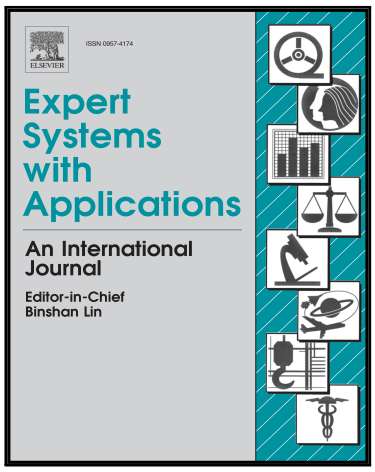

Decision-Making for Financial Trading: A Fusion Approach of Machine Learning and Portfolio Selection

Felipe Dias Paiva , Rodrigo Tomás Nogueira Cardoso , Gustavo Peixoto Hanaoka , Wendel Moreira Duarte

PII: S0957-4174(18)30503-7
DOI: https://doi.org/10.1016/j.eswa.2018.08.003
Reference: ESWA 12132

To appear in: *Expert Systems With Applications*

Received date: 12 February 2018
Revised date: 31 July 2018
Accepted date: 1 August 2018

Please cite this article as: Felipe Dias Paiva , Rodrigo Tomás Nogueira Cardoso , Gustavo Peixoto Hanaoka , Wendel Moreira Duarte , Decision-Making for Financial Trading: A Fusion Approach of Machine Learning and Portfolio Selection, *Expert Systems With Applications* (2018), doi: https://doi.org/10.1016/j.eswa.2018.08.003

Highlights:

- A decision-making model for day trade investments on the stock market is proposed
- The fusion approach between SVM and portfolio selection showed significant results
- The greater the defined target gain, the better discriminatory performance of the SVM
- The results of alternative models were worse than the proposed model
- The brokerage costs can be a strong constraint to feasibility of the proposed model

Decision-Making for Financial Trading: A Fusion Approach of Machine Learning and Portfolio Selection

Felipe Dias Paiva,[a,*] Rodrigo Tomás Nogueira Cardoso,[b] Gustavo Peixoto Hanaoka,[b] Wendel Moreira Duarte[a]

[a] *Finance Research Group (GFin), Centro Federal de Educação Tecnológica de Minas Gerais, Brazil*
[b] *Complex Systems Group (GESC), Centro Federal de Educação Tecnológica de Minas Gerais, Brazil*

Abstract

Forecasting stock returns is an exacting prospect in the context of financial time series. This study proposes a unique decision-making model for day trading investments on the stock market. In this regard, the model was developed using a fusion approach of a classifier based on machine learning, with the support vector machine (SVM) method, and the mean-variance (MV) method for portfolio selection. The model's experimental evaluation was based on assets from the São Paulo Stock Exchange Index (Ibovespa). Monthly rolling windows were used to choose the best-performing parameter sets (the in-sample phase) and testing (the out-of-sample phase). The monthly windows were composed of daily rolling windows, with new training of the classifying algorithm and portfolio optimization. A total of 81 parameter arrangements were formulated. To compare the proposed model's performance, two other models were suggested: (i) SVM + 1/N, which maintained the process of classifying the trends of the assets that reached a certain target of gain and which invested equally in all assets that had positive signals in their classifications, and (ii) Random + MV, which also maintained the selection of those assets with a tendency to reach a certain target of gain, but where the selection was randomly defined. Then, the portfolio's composition was determined using the MV method. Together, the alternative models registered 36 parameter variations. In addition to these two models, the results were also compared with the Ibovespa's performance. The experiments were formulated using historical data for 3,716 trading days for the out-of-sample analysis. Simulations were conducted without including transaction costs and also with the inclusion of a proportion of such costs. We specifically analyzed the effect of brokerage costs on buying and selling stocks on the Brazilian market. This study also evaluated the classifier's performance, portfolios' cardinality, and models' returns and risks. The proposed main model showed significant results, although demand for trading value can be a limiting factor for its implementation. Nonetheless, this study extends the theoretical application of machine learning and offers a potentially practical approach to anticipating stock prices.

Keywords: decision-making, financial trading, support vector machines, portfolio selection, stock market.

* Corresponding author. Address: Centro Federal de Educação Tecnológica de Minas Gerais, Pós-Graduação em Administração, Av. Amazonas, 7675, zip code 30510-000, Belo Horizonte/MG, Brazil. Tel.: +55 31 3319 6740.
Email addresses: fpaiva@cefetmg.br (F. D. Paiva); rodrigocardoso@cefetmg.br (R. T. N. Cardoso); gustavopph@gmail.com (G. P. Hanaoka); wmoreiraduarte@gmail.com (W. M. Duarte).

1. Introduction

Predicting stock returns is considered to be one of the most challenging tasks when dealing with financial time series because the stock market is dynamic, complex, evolutionary, nonlinear, nebulous, nonparametric, and chaotic by nature. Additionally, the stock market is extremely sensitive to political factors, microeconomic and macroeconomic conditions, and investors' expectations and insecurities (Ballings, Van den Poel, Hespeels, & Gryp, 2015; Kara, Boyacioglu, & Baykan, 2011; Tan, Quek, & Ng, 2007).

According to mainstream financial theory, predicting financial asset prices is impossible. The efficient market hypothesis (EMH), which is the literature's main theoretical pillar, suggests that the task of predicting future prices based on financial assets' past behavior cannot achieve abnormal returns. The reason is that the distribution function of a financial time series denotes a Brownian motion, which has random, independent, and Gaussian distribution characteristics.

However, some studies reject the EMH, arguing that the stock market is not actually established at random and that financial time series have long-term memory. For example, recent studies dispute the EMH in the context of different markets and periods. Cervelló-Royo, Guijarro, and Michniuk (2015) showed empirical evidence that compromises the premise of the EMH. The authors researched the intraday markets of the American Dow Jones Industrial Average (DJIA) index, the German Deutscher Aktien 30 Index (DAX), and the British Financial Times Stock Exchange (FTSE) index from 2000 to 2013. By using a methodology based on flag patterns, they tested 96 different configurations. The results were very profitable for all the markets. Chourmouziadis and Chatzoglou (2016) researched daily data from the Athens Stock Exchange from 1996 to 2012. They combined technical analysis and fuzzy logic. The profitability results of the proposed methodology were surprisingly higher than the baselines. Kim and Enke (2016) studied the intraday data of the Korea Composite Stock Price Index (KOSPI) 200 from 2007 to 2014. They formulated a set of rules based on technical analysis combined with genetic algorithms and reported abnormal profits. Chen and Chen (2016) proposed a model based on flag patterns for recognizing bullish reversal patterns. They tested the strategy with the National Association of Securities Dealers Automated Quotation System (NASDAQ) from 1989 to 2004 and with the Taiwan Capitalization Weighted Stock Index (TAIEX) from 1990 to 2004. The results showed high levels of profitability. Kampouridis and Otero (2017) studied the intraday foreign exchange market, analyzing a 10-month period between 2013 and 2014. They proposed a model that combined technical indicators, physical time scales, and genetic algorithms. The results guaranteed higher returns than those of the baselines.

In this sense, there seems to be strong evidence that once a financial time series' pattern of behavior has been identified, it becomes possible to delineate a predictability model (Huang, Yang, & Chuang, 2008; Lo, Mamaysky, & Wang, 2000; Malkiel, 2003; Mandelbrot & Hudson, 2004; Patel, Shah, Thakkar, & Kotecha, 2015). Moreover, consensus exists regarding the complexity and lack of definition of a series' general characteristics. For this reason, it is recommended to use robust and appropriate methods to handle a financial series (Brabazon & O'Neill, 2006).

Given the foregoing, the use of expert systems is consistently increasing in response to environmental characteristics. As such, the challenge is to identify a link between the past and the future with the objective of predicting a stock's price or return. In this regard, the literature has generally considered two research fields (Sheta, Ahmed, & Faris, 2015).

I) Econometric models. These are statistical models such as linear regression, autoregressive (AR), autoregressive moving average (ARMA), autoregressive conditional heteroscedasticity (ARCH), and generalized autoregressive conditional heteroscedasticity (GARCH). A key point in evaluating the use of these models focuses on the assumptions that a financial series must fulfill to guarantee the quality and reliability of the results.

II) Models based on machine learning. These are models based on artificial intelligence methods. They include artificial neural networks, genetic algorithms, fuzzy logic, support vector machines, random forests, and particle swarm optimization. Such options are interesting because of their capacity to treat complex, imprecise, and large amounts of data. The data's characteristics, when applied to other models, tend to obscure the underlying meaning and restrict attempts to obtain useful information. Additionally, artificial intelligence methods enable the use of different types of data (qualitative and quantitative); moreover, such methods are not subject to rigid assumptions such as those imposed on econometric models.

To compare the performance of these two research fields, several comparative studies have been conducted. These studies have highlighted the certain advantages of artificial intelligence methods when addressing problems linked to financial time series. The results have indicated that artificial intelligence methods have a greater capacity to confront problems with nonlinear, nonstationary characteristics than econometric models (Kazem, Sharifi, Hussain, Saberi, & Hussain, 2013; Lu, Lee, & Chiu, 2009; Matías & Reboredo, 2012; Sheta, Ahmed, & Faris, 2015; Teräsvirta, Van Dijk, & Medeiros, 2005; Yu, Wang, & Lai, 2009; Zhang, Cao, & Schniederjans, 2004).

Given the potential of the machine-learning methods that could be utilized, we decided to use the SVM method in the current study. The SVM method emerged in the 1990s with the goal of decreasing structural risks by minimizing empirical risks, thereby resulting in minimizing the upper limit of real risk. Thus, the SVM method minimizes the generalization error rather than the training error. Because of this capability for generalization, it stands out compared with other machine learning methods. Such a characteristic helps overcome overfitting and high-dimensional data problems. Another significantly relevant aspect is that the SVM method's mathematical formulation is restricted to a quadratic programming problem, which guarantees the achievement of a global optimal solution (Coussement & Van den Poel, 2008; Li, Kwok, Zhu, & Wang, 2003). Nevertheless, SVM method also presents disadvantages: it does not present probability prediction values, its sparcity is limited for classification, and its generalization ability can be unsuccessful when data sets have a very large imbalanced ratio (He, Xiao, Wang, Liu, Yang, Lu, Gui, & Sun, 2017; Cervantes, Garcia-Lamont, Rodriguez, López, Castilla, & Trueba, 2017).

We emphasize that the decision to apply the SVM method was influenced by its characteristics and by the results of financial studies that have suggested the method's high potential to address financial problems. Wen, Yang, Song, and Jia (2010) researched 442 assets of the Standard & Poor's (S&P) 500 index from March 2004 to October 2005. The authors proposed an intelligent stock trading model based on market fluctuations and combining stock box theory with the SVM method. The results showed that the model performed significantly better than a buy-and-hold strategy. Ni, Ni, and Gao (2011) studied the assets of the Shanghai Stock Exchange (SSE) from 2000 to 2008. Fractal feature selection was applied to choose the SVM model's features used in the daily price forecast. The results showed that this selection process chooses a number that is slightly lower compared with other selection methods, thereby yielding more accurate values. Zhiqiang, Huaiqing, and Quan (2013) studied the assets listed on the SSE and DJIA from 2000 to 2004. The locality-preserving projection and particle swarm optimization methods were first used to analyze the information. The data were then coupled with the SVM model and comparisons were undertaken. During the out-of-sample analysis, accuracy levels of 61.73% and 57.94% were achieved for the SSE and DJIA, respectively. Thenmozhi and Sarath Chand (2016) studied a sample composed of the indices of global markets (the DJIA, S&P 500, FTSE, National Stock Exchange (NSE) of India, Singapore Exchange (SGX), Hang Seng, and SSE) from 1999 to 2011. The predictions obtained using the SVM method exceeded those of regression models and of the technical analysis indicators in all markets during the full period. Pan, Xiao, Wang, and Yang (2017) studied S&P 500 stocks from June 2008 to April 2015. The authors used an SVM model to predict price fluctuations that considered information in different frequencies. The results indicated that the model outperforms the baselines, thereby representing an interesting option for practical applications.

Consequently, one of the current study's objectives was to build an algorithm that would recognize patterns that arise in the financial time series of stocks part of the São Paulo Stock Exchange Index (Ibovespa). This algorithm would then enable the classification of stocks in accordance with their potential to reach pre-determined daily returns. In other words, the proposal was to delineate a model based on the SVM method that would be able to separate assets into two groups: one that, theoretically, would not reach the stipulated profit goal and another that would meet the condition for success in terms of achieving the estimated profit. As a result of this classification process, we sought to estimate and reduce the set of assets that would be selected for an investment portfolio.

To avoid dividing into daily operations the capital that was to be invested equally (1/N) in the assets that had the potential to successfully reach the defined return, we sought to integrate a portfolio selection method with the SVM method. This procedure aimed to merge the SVM method's results with a method that was able to analyze the assets that had the potential to reach the expected return. Such an approach would help determine which assets should really be part of the investment portfolio and what their respective participation should be. We thus decided to use the classic portfolio selection method developed by Markowitz (1952), who was awarded a Nobel Prize because of his model's contribution to this field.

The formulation of the Markowitz model, which is also called the mean-variance (MV) model, proposes the construction of investment portfolios based on the maximization of expected portfolio returns and the simultaneous minimization of investment risk (Fabozzi, Gupta, & Markowitz, 2002). This model formulates an efficient frontier, which comes from the asset portfolio that reaches a maximum expected return level, given a certain minimum risk. For each desired level of return, the efficiency of the frontier indicates the best investment strategies (Deng, Lin, & Lo, 2012). In this sense, the main objective of the current study was to structure a decision-making model that integrates the machine learning SVM method, which aims to classify the Ibovespa's assets, with the classic MV method to select portfolio investments.

We refer to studies that have essentially similar concepts and thereby relate to our main research proposal. Gupta, Mehlawat, and Mittal (2012) presented a model that used the SVM method to categorize 150 assets from the Indian NSE over 36 months. After categorization, the portfolio selection stage was completed with the support of a genetic algorithm. The model yielded good results. Silva, Neves, and Horta (2015) studied assets of the S&P 500 index from 2010 to 2014. The authors developed a genetic algorithm to select and optimize an investment portfolio. The input used for the model originated in fundamental and technical approaches. The simulation yielded returns that were higher than the baselines and with compatible volatility. Huang, Chiou, Wu, and Yang (2015) researched the Taiwanese stock market from 2006 to 2009. Their work consisted of using financial indices to screen the most efficient assets. The authors employed data envelopment analysis in this stage. Based on the preselection of assets, a multi-objective algorithm was applied to select a portfolio. The proposal yielded a result that was higher than the baselines. Machado, Neves, and Horta (2015) researched multiple markets from 2006 to 2014 to preclassify the assets using technical indicators in a system of decision trees and genetic algorithms; only then was the composition of the portfolio determined. This method displayed good profitability, even during times of market turbulence. Petropoulos, Chatzis, Siakoulis, and Vlachogiannakis (2017) studied the foreign exchange market, specifically 10 currency pairs traded against the US dollar, from 2001 to 2015. They proposed an innovative automated investment decision system through the application of several models based on machine learning (the SVM method, random forests, dense-layer neural networks, and naive Bayes) to predict currency movement and logarithmic returns. In the system's next stage, in a process that combined a pair-to-pair correlation study and the predictive results of the machine learning models, the trend markers were expanded starting at five (the number of machine learning methods) up to 50. These decision signals were then added through a majority voting system of trends, genetic algorithms, and regression weighting methods. Tests using a stop–loss mechanism were also conducted. The results yielded by the authors' model were significantly better than those of other benchmarks. Macedo, Godinho, and Alves (2017) studied multiple markets from 2000 to 2015 to verify the performance of a genetic algorithm for portfolio selection, especially when used with signals produced by technical analysis indicators. The authors confirmed the positive effect that resulted from this combination.

In sum, the current study contributes to the literature by establishing a new approach to decision-making models for day trading investments in the stock market. In this regard, the study undertook an experimental evaluation covering a period of 15 years as a robust analysis. Moreover, it provides an innovative model for the integration of methods.

The rest of this study is organized as follows. In Section 2, we review the SVM method and summarize empirical work that has used it to solve problems related to financial series. In Section 3, we discuss a general perspective of the MV method used for portfolio selection. In Section 4, we describe the configuration of the experiments undertaken. In Section 5, we present the results obtained from the experiments and discuss the classifier's performance, portfolio cardinality, the returns and risks of the proposed model, and the baselines. Finally, in Section 6, we state our conclusions and make recommendations for future research.

2. SVM method

The SVM method first appeared in the mid-1990s as a result of the practical use of concepts developed by statistical learning theory. The most prominent researcher in the development of the SVM method is the Russian scientist, Vladimir N. Vapnik. He has been at the forefront of related discussions since the late 1960s when the first studies of statistical learning theory emerged.

The SVM method aims to solve problems of pattern recognition, classification, regression estimation, time series, and density estimation (Huang, Nakamori, & Wang, 2005; Vapnik, 1999). According to

Vapnik (1999), the basic idea of the SVM method is to map input vectors in a large space by means of a defined nonlinear mapping element. Thus, this method applies the concepts of a linear model to separate the *input* from nonlinear mapping in a characteristic high-dimensional space. The linear model constructed in this new space can represent a nonlinear decision limit for the original space.

The following characteristics of the SVM stand out (Awad & Khanna, 2015; Burges, 1998; Kim, 2003; Sands, Tayal, Morris, & Monteiro, 2015; Sheta, Ahmed, & Faris, 2015; Xu, Caramanis, & Mannor, 2009):

- The convexity of the objective function is one of the advantages of the SVM method because the training of this method is equivalent to the solution of a quadratic programming problem, whereby the problem's solution will always be unique and better overall. For example, the artificial neural network method demands nonlinear optimization with the ever-eminent risk of the algorithm being held hostage to local minimums.

- The SVM method has greater precision than other individual forecasting methods. It has superior performance because it is designed to minimize structural risk, while other methods focus on empirical risk minimization. In other words, the SVM method seeks to minimize the upper limit of the generalization error to the detriment of the training error.

- The SVM method can process large volumes of data robustly, without the occurrence of overfitting.

The SVM method also focuses on establishing optimal separation hyperplanes. The training points $(x_i \rightarrow y_i)$ closest to the optimal separation hyperplane are called support vectors and establish the limit of the decision plane. In general cases, where the data are not linearly separated, the SVM method uses nonlinear machines to find the hyperplane that minimizes the number of errors in the training set (Ding, Song, & Zen, 2008).

The core of the SVM method's theory of solving binary classification problems is described below, using the notations of Ding, Song, and Zen (2008), Luo and Chen (2013), and Ni, Ni, and Gao (2011).

Assume we have a set of points for training, $D = \{x_i, y_i\}_{i=1}^{N}$, where the vectors of the input are $x_i = \left(x_i^{(1)}, \dots, x_i^{(n)}\right) \in R^n$ and the vectors of the outputs are $y_i \in \{0, 1\}$ and where n is the amount of training data. Then, to separate two classes of data points, the SVM method seeks to find the optimal hyperplane of separation by solving the following optimization problem:

$$\underset{w,b}{\text{Min}} \left(\frac{1}{2} w^T w\right),$$

Subject: $y_i(w^T \phi(x_i) + b \geq 1), i = 1, \dots, n,$

where w represents the weight vector and b the *bias* variable. The nonlinear function, $\phi(.): R^n \rightarrow R^{nk}$, maps the *input* in the high-dimensional space. However, several classification problems are linearly nonseparable. Thus, we need to introduce gap variables (ξ_i) to allow for misclassification. The optimization problem then becomes

$$\underset{w,b,\xi}{\text{Min}} \left(\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i\right),$$

Subject: $\begin{cases} y_i\Big(\big(w^T \phi(x_i)\big) + b\Big) + \xi_i \geq 1, i = 1, \dots, n \\ \xi_i \geq 0, i = 1, \dots, n, \end{cases}$

where C is the penalty parameter of the error term. The solution of the primary problem is obtained from a Lagrangian construction. Then, the primal problem can be converted into a quadratic optimization problem with bound constraints and a linear equality constraint (Vapnik, 1999):

$$\text{Max}_{\alpha} \left(\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Q_{ij}\right),$$

Subject: $\begin{cases} 0 \leq \alpha_i \leq C, i = 1, \dots, n \\ \sum_{i=1}^{N} \alpha_i y_i = 0, \end{cases}$

where α_i is a Lagrange multiplier, $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$. Because of the required computational effort, the internal product is replaced by the kernel function, which satisfies the Mercer condition, $K_{(x_i,x_j)} = \phi(x_i)^T \phi(x_j)$, and becomes the representation of a measure of similarity or proximity between points. Finally, we reach the nonlinear decision function in the primal space for the linearly nonseparable case:

$$y(x) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_j) + b\right).$$

The *kernel* functions map the input data into a larger dimensional space, where the data can be separated using a hyperplane, meaning that the data become linearly separable (Cristianini & Shawe-Taylor, 2000). The kernel functions used in the SVM method are i) Linear Kernel: $K(x_i, x_j) = x_i^T x_j$; ii) Polynomial Kernel: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$; iii) Radial Kernel: $K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2)$; and iv) Sigmoid Kernel: $K(x_i, x_j) = \tanh(\gamma\, x_i^T x_j + r)$, where $d, r \in N$ and $\gamma \in R^+$ are constants.

The kernel functions play an important role when locating complex decision limits between classes; thus, their selection is critical for the development of the SVM method. When choosing the potential mappings that can be utilized, the first challenge is to identify the best one for a given classification problem in a way that minimizes the generalization error. Mainly, the recommendation is to use the radial basis function (RBF) kernel. This kernel maps the samples in a nonlinear fashion into a high-dimensional space so that it can handle nonlinear problems that cannot be addressed using the linear kernel. The sigmoid kernel behaves like the RBF kernel for a given parameter; however, it is not valid under some parameters. The second challenge is the number of hyperparameters that influence the complexity of model selection. The polynomial kernel has more parameters than the RBF kernel. The RBF kernel has fewer computational difficulties regarding processing. The polynomial kernel can accept values that tend to infinity or zero when the degree is large, which limits its application. As such, the polynomial core requires more time in the training stage and is reported to yield worse results than the RBF kernel (Ding, Song, & Zen, 2008; Huang, Davis, & Townshend, 2002).

3. Portfolio selection model: MV method

In the seminal paper by Markowitz (1952), the forerunner of modern finance theory, a mathematical solution was presented to address the trade-off between maximizing the expected return on investment and minimizing the risk. To quantify securities' returns and risk, Markowitz proposed the mean and variance of returns' distribution as statistical measures (Kolm, Tütüncü, & Fabozzi, 2014). According to Santos and Tessari (2012), the core of this solution is that investors make decisions based on risk and expected returns when determining the best selection for their portfolios. In this regard, investors choose among the lowest variance for portfolios with equal expected returns and for portfolios with the highest expected return in terms of options with the same risk level.

Thus, the MV model can be expressed through a multi-objective optimization formulation (Kolm, Tütüncü, & Fabozzi, 2014). This generation of a set of optimal solutions is called an efficient frontier of investments. The following equations formally describe the model:

$$\underset{w_i, \ldots, w_n}{Min} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i\, w_j \sigma_{ij},$$

$$\underset{w_i, \ldots, w_n}{Max} \sum_{i=1}^{n} w_i \mu_i,$$

$$\text{Subject to: } \begin{cases} \sum_{i=1}^{n} w_i = 1 \\ 0 \leq w_i \leq 1, \forall\, i = 1, \ldots, n, \end{cases}$$

where:
w_i = proportion of the initial value invested in the portfolio for asset i;
w_j = proportion of the initial value invested in the portfolio for asset j;
σ_{ij} = covariance between assets i and j;
μ_i = expected return on asset i.

The same optimal set of assets can be achieved by means of a mono-objective formulation (Jobst, Horniman, Lucas, & Mitra, 2001). In this regard, a variable that represents the investor's risk aversion is introduced into the model as a factor that describes his/her behavior in relation to the risk investment options:

$$\underset{w_i, \ldots, w_n}{Min} \; \lambda \left[\sum_{i=1}^{n} \sum_{j=1}^{n} w_i \, w_j \sigma_{ij}\right] - (1 - \lambda)[\sum_{i=1}^{n} w_i \mu_i],$$

$$\text{Subject to: } \begin{cases} \sum_{i=1}^{n} w_i = 1 \\ 0 \leq w_i \leq 1, \forall \, i = 1, \ldots, n, \end{cases}$$

where λ = the coefficient of risk aversion.

Thus, if, on the one hand, the solution proposed by Markowitz results in the so-called efficient frontier of investments, on the other, it does not indicate an exact point of investment to be considered. Moreover, Markowitz (1952) stated that asset averages, variances, and covariance can be estimated by means of statistical analysis and the use of the analyst's judgment. As a result, a set of MV combinations that match the desired risk–return balance can be derived and presented to the investor. Hence, the investor chooses the point among the possible solutions that accords with his/her risk predisposition (Michaud & Michaud, 2008). The foregoing formulation does not have any risk-free assets; consequently, the portfolio's construction is restricted to risky assets.

4. The experiment

4.1. Data

The assets of companies listed on the Ibovespa were used as the current study's sample. Data from June 2001 to December 2016 were collected for analysis and divided into sets in preparation for the model attributes: in-sample and out-of-sample. During the studied period, 135 assets were listed on the Ibovespa. Of these, between 53 and 73 were members. Appendix B lists the assets utilized in the research. Fig. 1 shows the variations in the number of assets on the Ibovespa.

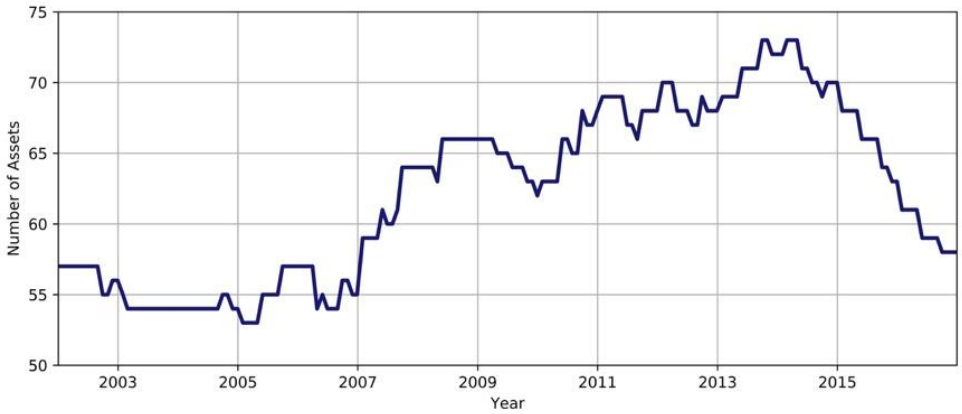

Fig. 1. Number of the Ibovespa's assets

We collected through the Bloomberg terminal the historical series of adjusted opening, closing, maximum, and minimum prices, together with the traded volume of assets, to carry out the experiment.

4.2. Proposed model: SVM + MV

The proposed model is based on technical analysis and anchored on the premise of identifying patterns of behavior in the historical series of asset prices. In this regard, "The basic assumption of all technical theories is that history tends to repeat itself, i.e., past patterns of price behavior in individual securities will tend to recur in the future" (Fama, 1965, 55). Thus, the starting point after data collection was to prepare 22 attributes to be tested as inputs. Among the attributes are return measures based on opening,

closing, maximum, and minimum prices, as well as indicators of momentum, volatility, and volume as follows.

# Attribute	Details	# Attribute	Details
1	$r1 = ln\left(\frac{close\ price_i}{close\ price_{i-1}}\right)$	12	$r12 = ln\left(\frac{low\ price_i}{open\ price_i}\right)$
2	$r2 = ln\left(\frac{close\ price_{i-1}}{close\ price_{i-2}}\right)$	13	$r13 = ln\left(\frac{low\ price_{i-1}}{open\ price_{i-1}}\right)$
3	$r3 = ln\left(\frac{close\ price_{i-2}}{close\ price_{i-3}}\right)$	14	$r14 = ln\left(\frac{low\ price_{i-2}}{open\ price_{i-2}}\right)$
4	$r4 = ln\left(\frac{close\ price_{i-3}}{close\ price_{i-4}}\right)$	15	$r15 = ln\left(\frac{low\ price_{i-3}}{open\ price_{i-3}}\right)$
5	$r5 = ln\left(\frac{high\ price_i}{open\ price_i}\right)$	16	Momentum (close price, period = 10)
6	$r6 = ln\left(\frac{high\ price_i}{open\ price_{i-1}}\right)$	17	Relative strength index (close price, period = 14)
7	$r7 = ln\left(\frac{high\ price_i}{open\ price_{i-2}}\right)$	18	Parabolic SAR (high and low price, acceleration = 0, maximum = 0)
8	$r8 = ln\left(\frac{high\ price_i}{open\ price_{i-3}}\right)$	19	Average true range (high, low and close price, period = 14)
9	$r9 = ln\left(\frac{high\ price_{i-1}}{open\ price_{i-1}}\right)$	20	True range (high, low, and close price)
10	$r10 = ln\left(\frac{high\ price_{i-2}}{open\ price_{i-2}}\right)$	21	Chaikin A/D line (high, low, and close price; volume)
11	$r11 = ln\left(\frac{high\ price_{i-3}}{open\ price_{i-3}}\right)$	22	On balance volume (close price, volume)

The calculated attributes are then used as inputs for the SVM method. The objective of the SVM method here is to perform a binary classification of the return for *day trading* operations. For instance, let us suppose that we wish to identify whether a predetermined asset has the potential to reach a 1% return in t_{+1}. To accomplish this goal, we would collect data from the series of that asset until the closing of the trading session at t_0. Based on the size of the training window, we would complement the data with the series from prior days. The next stage would consist of calculating the attributes and designating them as inputs for the SVM method. The target would then be a 1% return for t_{+1}. Thus, we would establish as a premise that the investor would be able to open a position only at t_{+1}, expecting to successfully close the position on the same day.

Once all the assets are classified, one by one, only those designated as likely to reach the expected return are considered to be eligible to participate in the next stage. The objective of this stage is to define the proportion of capital allocated to each asset. To execute this stage, the Markowitz MV model is used. For simulation purposes, the defined portfolio to which the available resources will be allocated has low variance.

To achieve the optimization and calculation of the minimum-variance portfolio, the measures of the classic Markowitz model are used: mean and covariance matrices. It is worth clarifying that we did not employ simulations of investment decision-making and risk-free assets. As such, the portfolios were composed exclusively of risky assets. Fig. 2 presents the model's scheme.

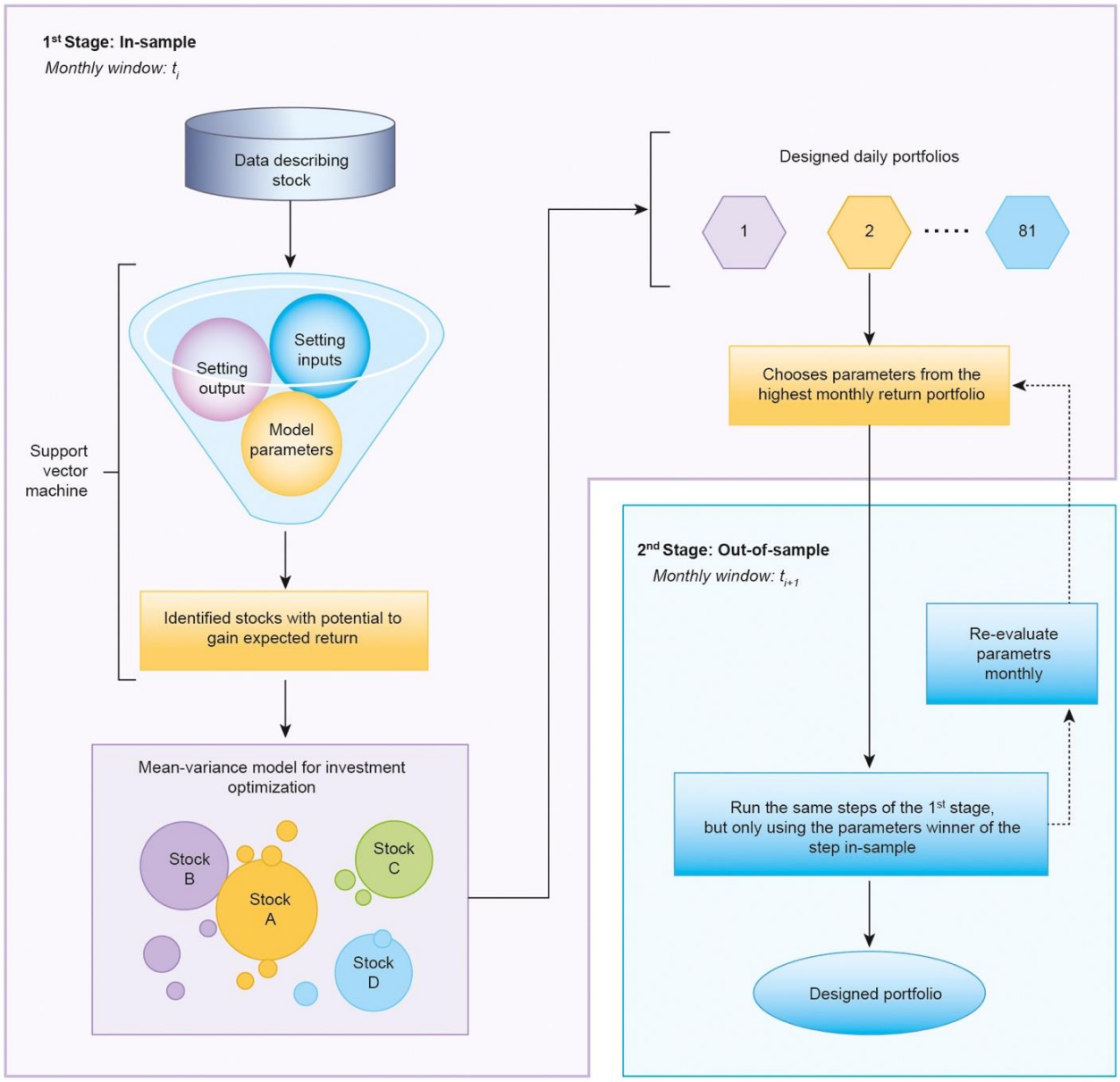

Fig. 2. Proposed model and the experiment's stages

Several important parameters for the execution of the proposed model were varied and tested to define the best overall arrangement. Fig. 3 shows the tested variations. Four parameters are defined; their indicators are then varied over three values. The combinatorial arrangement of these variations results in 81 model structures. There are variations in the number of data instances used in the learning process of the SVM method's algorithm (60, 75, and 90 days), in the set of attributes that compose the SVM[1] method's input (Set 1 = 20 attributes, Set 2 = 22 attributes, and Set 3 = 15 attributes), in the return expected by the investor for a day trading transaction (1.0%, 1.5%, and 2.0%), and in the number of days used to calculate the expected return and the covariance matrix employed for the investment portfolio's optimization process (75, 90, and 105 days).

[1] Attributes of the simulation sets: Set 1 (attributes #1 to #20), Set 2 (attributes #1 to #22), and Set 3 (attributes #1 to #15).

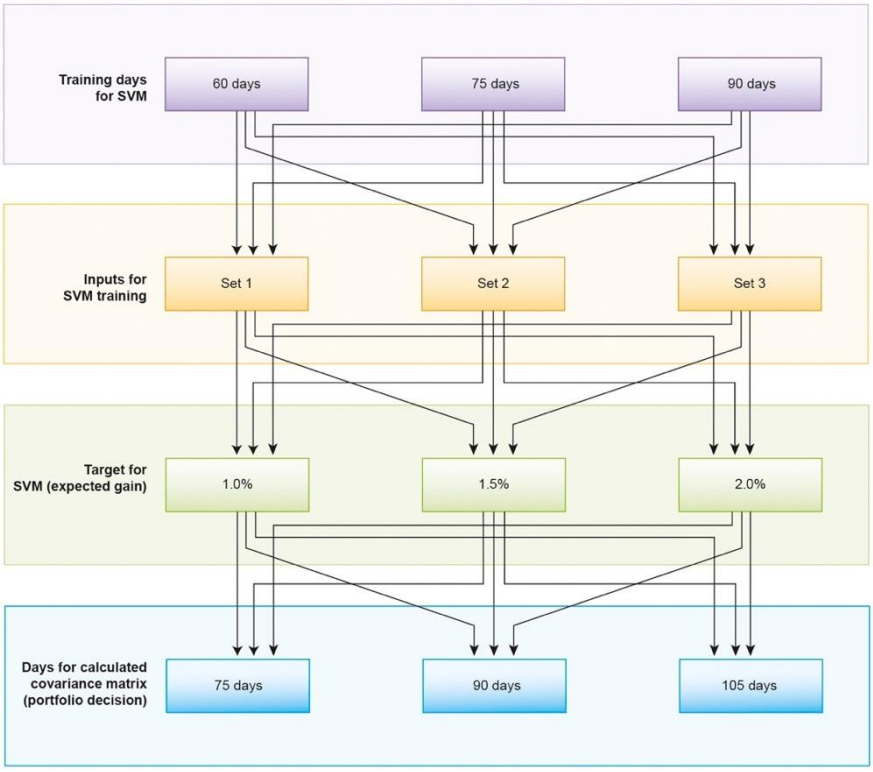

Fig. 3. Set of the experiment's parameters

The simulation used monthly rolling windows to verify the structures of the better performing parameters. The model with the highest cumulative return was defined as the model with the best performance. The monthly windows are the totals of the daily rolling windows for training regarding the classification algorithm and for the optimization of the investment portfolio. Thus, in the in-sample stage, the set of parameters that reached the highest accumulated monthly return was calculated. This set of parameters was then replicated for the following month in the out-of-sample stage. The in-sample stage ran from December 3, 2002 to November 30, 2016 and totaled 180 months of simulation, or 3,713 trading days. The out-of-sample stage ran from January 2, 2002 to December 29, 2016 and was also for 180 months, or 3,716 trading days. Fig. 4 details this scheme.

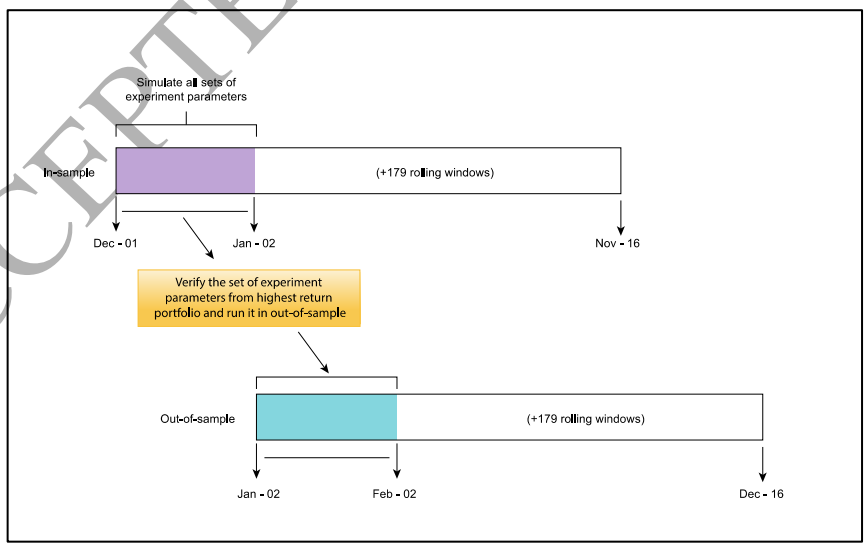

Fig. 4. Rolling windows of the experiment

According to Gerlein, McGinnity, Belatreche, and Coleman (2016), the procedure described here tends to stall the effects of data snooping. White (2000) stated that a good forecasting model is not always the result of real forecasting capacity but of lucky decisions. These lucky decisions can still produce spurious results by using data-mining methods. However, Gerlein, McGinnity, Belatreche, and Coleman (2016) maintained that the adopted procedure guarantees unbiased results because a temporal separation exists between the trained data and testing data. The latter did not allow the selection and optimization of the

parameters. Thus, the model tests received data that were not used in the training process. In theory, testing was considered for future data points, constructing a scenario in which the model would experience real application. This approach avoids the effect of data snooping, namely the use of a future-focused test mechanism as opposed to a back-testing mechanism.

In sum, once the assets and their respective investment ratios had been indicated, the next step was to allocate monetary resources at the opening of the next day. With regard to the exit strategy (the sale of the asset), two possibilities were proposed: first, the asset would be sold during the investment day if it reached the expected gain; second, if the target was not reached, the sale of the asset would occur at the opening of the day subsequent to its purchase.

4.3. Baseline strategies

These were based on the research model proposed in the prior section and used for comparison with this model's performance and its variations.

4.3.1. Alternative model 1: SVM + 1/N

This model's design reproduces the structure of the SVM + MV model, with the exception of the optimization step in the investment portfolio. The proportions of investment in each asset occur in an equitable way; in other words, assets with a true classification signal produced by the SVM method receive the same proportion of investment. As a result of the non-application of the MV model stage for diversification, the parameter variations are as follows: the number of data instances used in the learning process of the SVM method's algorithm, the set of variables for input in the SVM method, and the return expected by the investor for day trading transactions. Thus, there are 27 combinatorial arrangements.

4.3.2. Alternative model 2: Random + MV

The Random + MV model differs from the SVM + MV model in terms of the asset classification phase and features a tendency to reach the expected return. The asset classification is undertaken naively. However, the number of assets chosen randomly must be equal to the number defined by the SVM + MV model. The investment portfolio optimization stage using the Markowitz method is retained. Parameter variations are related to the return expected by the investor for day trading operations, the number of days used to calculate the expected return, and the covariance matrix in the diversification stage, thereby resulting in nine combinatorial arrangements. Additionally, given that the selection of assets during the classification stage is randomized, 51 iterations are performed for each data set. The median is assumed to be the expected result for this model. Hence, the set of parameters to be used in the out-of-sample stage is chosen. Again, 51 iterations are performed and we assume that the median is the expected result.

4.4. Transaction costs

Two simulation blocks regarding transaction costs were conducted. Section 5 describes the simulations where transaction costs were disregarded and simulations where we added a proportion of the costs. These costs aimed to cover the fees paid to brokers for transactions, including the purchase and sale of stocks, in the Brazilian market.

The official operational costs of the Brazilian stock market, in general terms, are as follows. I) The variable amount: fees and emoluments (day trading = 0.025%, swing trading = 0.0325%) and income tax (day trading = 20%, swing trading = 15%). II) The fixed amount: the brokerage cost, which is a set figure per purchase and sale order, although the amount varies in accordance with the number of orders that are issued. Thus, the variable amount has its values calculated on the basis of the operation (fees and emoluments) and the gain (income tax). With regard to the fixed amount, the unit cost is set regardless of the amount invested or whether the operation is profitable. For example, an investment of BRL 1 million applied for operations involving 1, 10, or 20 shares will have the same expenses in terms of total taxes, fees, and income taxes regardless of the number of orders issued. By contrast, brokerage incurs a direct impact in terms of the number of issued orders. Each brokerage company has a value table for its service. For example, a reasonable value for an average of 20 daily orders would be approximately BRL 2.50 per

order.[2] Consequently, an investor issuing 20 orders (10 for purchase and 10 for sale) at BRL 2.50 each would have a total cost of BRL 50.00. If the total amount invested[3] is BRL 500 thousand, brokerage expenses will represent 0.01%; for an investment of BRL 1 million, the expenses will be 0.005%; for an investment of BRL 5 million, expenses will be 0.001%; etc. Based on these calculations, we simulated transaction costs as 1.00, 0.50, 0.10, and 0.05 basis points (bps)[4] to show the effects produced on return on investment.

We decided to evaluate only the effects of brokerage costs because taxes, emoluments, and income tax have a direct relation with investment performance: the greater the profit and capital invested, the greater such expenses. Moreover, brokerage costs are under the direct control of the investor. Proper management implies minimizing such costs to maximize investment profitability. None of the performed simulations considered bid–ask spread costs.

5. Results

Once the data for the simulation were collected and processed, the formulation of the classifier model based on the SVM method was undertaken to identify assets that had the potential to reach a certain gain target. The daily set of assets classified as having the potential to reach the expected return continued to the optimization stage of the portfolio through the MV method. This process, repeated daily, underwent a monthly evaluation of cumulative returns to identify the list of parameters with the best results to be used as the basis of the out-of-sample test. Table A1 in Appendix A shows the best monthly sets of the model parameters and Fig. 5 summarizes this result, illustrating the division between the configuration options for the SVM + MV model, with the exception of the target, which was chosen as 2%.

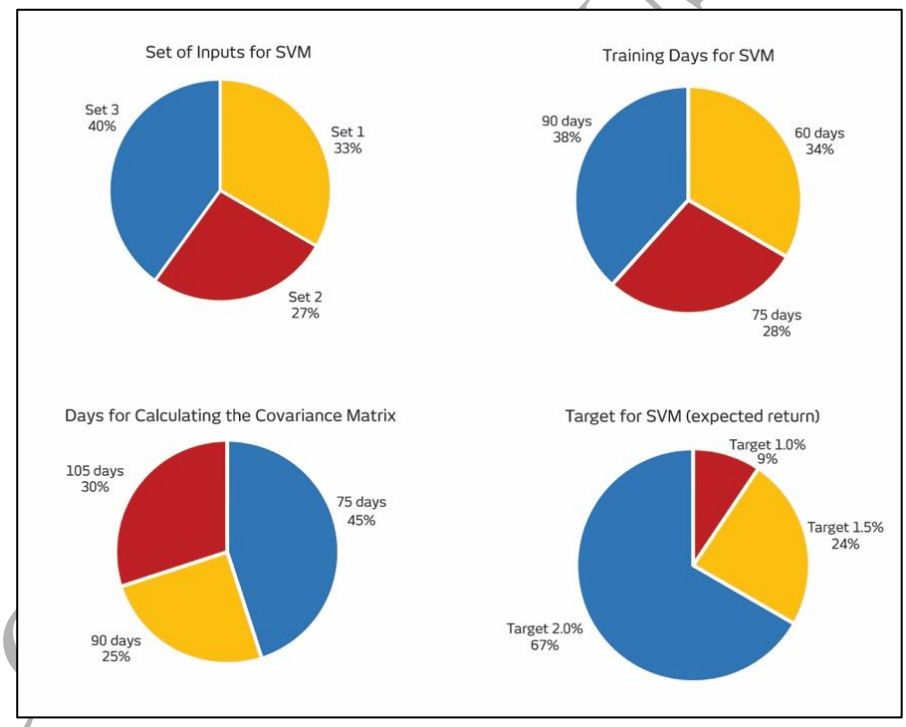

Fig. 5. Results of parameter selection

With regard to the classification algorithm in the in-sample stage, the result of the precision[5] measure was initially presented. To calculate the precision of the historical series of the assets and have a basis of comparison for the SVM method's result, the probability of reaching the target daily was obtained. Table B1 (Appendix B) shows the precision result per asset. On average, 40.70% of assets reach the target of defined gain in the strategy. By using the SVM method, an average precision of 54.97% was achieved. In other words, the use of the SVM method as a resource to classify trades with the potential to reach the

[2] This figure considers operations via a home broker, without the collaboration of the trading desk of the brokerage company that houses the investor's account.

[3] To evaluate the magnitude of these investment values for the Brazilian stock market, see Fig. E1 in Appendix E regarding the average daily trading values of the Ibovespa's assets. For example, in 2016, this average was approximately BRL 5.7 billion.

[4] 1 basis point = 0.01%.

[5] $Precision = \frac{tp}{tp+fp}$, where *tp* and *fp* mean, respectively, *true* and *false positive*. "Positive" and "negative" refer to the forecast and "true" and "false" refer to the judgment of that forecast in relation to the observed value.

target improves the likelihood of success by 35.06%. The Mann–Whitney test was subsequently conducted to assess whether equality existed between the medians of the outcome of the precision measure for the historical series and the SVM method. Given a p-value of 3.12e-37, the null hypothesis of equality was rejected at the 5% significance level.

The specificity[6] measure showed the effectiveness of the model at predicting the days when the expected return would not be achieved. In this regard, the average result was 70.29%. Thus, the classification highlighted the negative signals more effectively than the positive ones with results of 70.29% and 54.97%, respectively.

Table B2 in Appendix B presents the results of the classification phase for the out-of-sample period, which are similar to those for the in-sample window. The data show a 40.42% probability of the assets reaching the target of the gain defined in the strategy for the analyzed trading days. The SVM method achieves an average precision of 54.19%. The Mann–Whitney test showed that we can reject the null hypothesis on the equality of the performance of the SVM method's results and the complete set of assets (p-value = 3.99e-37) at the 5% significance level. The Mann–Whitney test was conducted again for assess the classification equality in the in-sample and out-of-sample periods. The result confirmed the equality of performance for the precision and specificity measures at the 5% significance level, with p-values of 0.08 and 0.50, respectively.

In Table B2, the number of positive signals (the forecast for reaching the target) produced by the SVM method per asset stands out: there are 60,589 positive signals from 231,513 possible signals (the number of trading days × the number of the Ibovespa's assets each day). This result indicates an approximate average of 16 signals per day, or 26.02% of all feasible signals. The active TIMP3 has the highest number of positive signals, 1,361, and a participation rate of 2.26% for the total number of signals. Further, the number of signals was weighted by the number of days during which an asset was in the Ibovespa. Only six assets proved to have signals greater than 50% of the days during which they were present on the index.

In Table 1, we stratify the classification results in accordance with the targets of the test window (out-of-sample). As the target value increases, the precision of the SVM method's classifier improves compared with the observed data and assumes that the target will be reached every day by all assets. In other words, with a target of 1%, the classifier based on the SVM method is 5.49% better than the observed data. With a target of 2%, this figure rises to 44.33%. The relative number of trades also reduces considerably as a higher target is used. Of the possible trades, deals are suggested for 76.20% with a 1% target. With a 2% target, the rate reduces to 15.83% of the investment in available assets.[7]

Table 1
Classification performance and number of signals by target

Descriptive Statistics	Target		
	1%	1.5%	2%
Observed Data (1)	60.16%	44.86%	33.77%
SVM (2)	63.46%	54.95%	48.74%
Δ% (2 - 1)/(1)	5.49%	22.49%	44.33%
Observed Data's Signals	22,298	55,893	15,322
SVM Signals	16,992	19,328	24,269

Table 2 presents the performance of the SVM + MV model in relation to the proposed alternative models, as described in Section 4. All subsequent results refer to the out-of-sample period. The table summarizes the number of deals (the purchase and sale of assets) that each model handled. The SVM + MV model has a much lower number of transactions per day than the others, which in practice would result in lower transaction costs. The Random + MV model has 63.59% more transactions than the SVM + MV model.

Table 2
Trades (buy and sell) per model

Descriptive Statistics	SVM + MV	SVM + 1/N	Random + MV
Total Trades	49,290	121,016	80,632
Daily Minimum	0	0	2
Daily Maximum	78	132	72

[6] $Specificity = \frac{tn}{tn+fp}$, where tn means *true negative*.

[7] Available assets are those part of the Ibovespa on the investment decision date.

Daily Mean	13	33	22
Std. Dev.	7.12	30.00	7.10

Fig. 6 shows the average daily cardinality of the portfolios for each of the investment models and each simulation month. The high dispersion of the SVM + 1/N model is due to the variation in the targets because the number of assets selected to compose the portfolio varies significantly, given the expected target for gain.

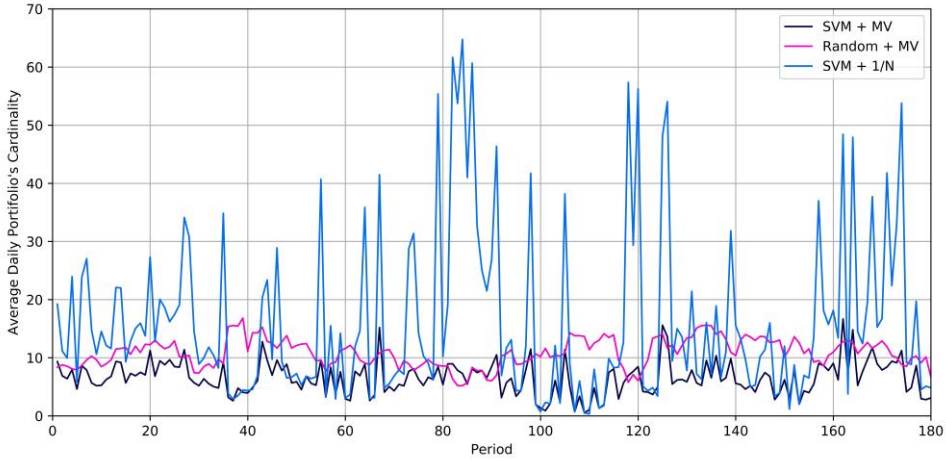

Fig. 6. Average cardinality of portfolios per model

The average cardinality of the portfolios was approximately seven assets for the SVM + MV, 16 for the SVM + 1/N, and 11 for the Random + MV models. According to the Kruskal–Wallis test, the null hypothesis of the equality of strategy distributions was rejected (p-value = 2.19e-30). This test does not indicate the stochastic dominance relation in pairs; thus, Dunn's test was conducted (see Table 3). It can be concluded that there is no equality in the distributions between the Random + MV and SVM + 1/N models at a significance level of 5%.

Table 3
Dunn's test for cardinality (p-value)

	SVM + MV	SVM + 1/N
SVM + 1/N	0.001	-
Random + MV	0.001	0.397

We then analyzed the distribution of the daily returns for the strategies of the investment models. Table 4 presents the results. The SVM + MV model has the highest daily return average: 0.11%. The SVM + 1/N model follows, with a return of 0.06% and a difference of 54.54% between the average returns of the two models. The other two benchmarks, Random + MV and the Ibovespa, have lower daily average returns and greater return dispersions, thus exposing investors to greater risks. The several outlier points featured in all the models are also worth mentioning. At the same time, we emphasize that stop–loss mechanisms were not implemented in any of the models.

Table 4
Descriptive statistics per model

Descriptive Statistics	SVM + MV	SVM + 1/N	Random + MV	Ibovespa
Mean %	0.11	0.06	0.03	0.04
Std. Deviation %	1.42	0.95	1.44	1.78
Pearson Coef. %	13.04	15.44	54.44	45.22
Minimum %	-9.53	-5.14	-9.61	-12.10
Maximum %	3.21	2.22	3.63	13.68

To better describe the results of the simulation that add to the discussion about Table 4, Fig. 7 shows the box plot of the series of daily returns for each model. The SVM + 1/N model least exposes the investor to greater volatility; however, the Ibovespa registers the highest volatility. Despite the lower standard deviation of SVM + 1/N, upon analyzing the relationship between standard deviation and mean and the binomial risk–return relationship, the SVM + MV model has a better outcome. The Kruskal–Wallis test was then conducted and the null hypothesis of the equality of strategy distributions was rejected (p-value = 1.04e-09).

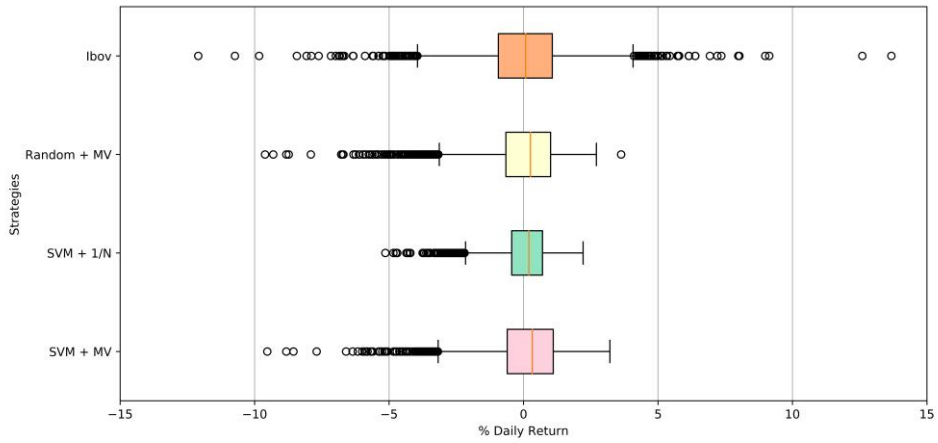

Fig. 7. Box plot of daily returns per model

A Dunn's test was conducted to evaluate the significance of the pairs of differences. According to the results presented in Table 5, we can reject the null hypothesis of equality between the distributions of the returns of the SVM + MV, Random +MV, and SVM +1/N models and the Ibovespa.

Table 5
Dunn's test for daily returns (p-value)

	SVM + MV	SVM + 1/N	Random + MV
SVM + 1/N	0.015	-	-
Random + MV	0.001	0.001	-
Ibovespa	0.001	0.001	0.001

Subsequently, more detailed information on daily performance, shown in Table 6, was extracted. All models have a similar number of days with a return greater than zero; however, the SVM + MV model is the best option, considering the numerous variables involved in the decision process. With regard to the average daily return on days when the return is equal to or greater than zero, the Ibovespa shows the highest return. By contrast, the Ibovespa has the lowest number of days with a return greater than zero and the highest average value for days with a return below zero. The Random + MV model shows the lowest average loss. In sum, the model that shows the greatest balance between average gain and loss, as well as the number of days of gain and loss, is the SVM + MV.

Table 6
Additional descriptive statistics per model

Descriptive Statistics	SVM + MV	SVM + 1/N	Random + MV	Ibovespa
(1) % days of return >= 0	60.74	60.17	58.40	52.02
(2) % days of return < 0	39.26	39.83	41.60	47.98
(3) Average return on profit days %	1.00	0.92	0.69	1.28
(4) Average return on loss days %	-1.27	-1.32	-0.81	-1.32
Ratio (1)/(2)	1.55	1.51	1.40	1.09
Ratio (4)/(3)	1.27	1.43	1.17	1.02

Fig. 8 shows the cumulative return for each of the models and the Ibovespa. The SVM + MV model has an appreciably higher result and achieves a profitability of 3,809.90% during the analyzed period. The profitability of the Random + MV model follows, with 736.79%, and then the Ibovespa, with 139.55%, and the SVM + 1/N model, with 80.85%.

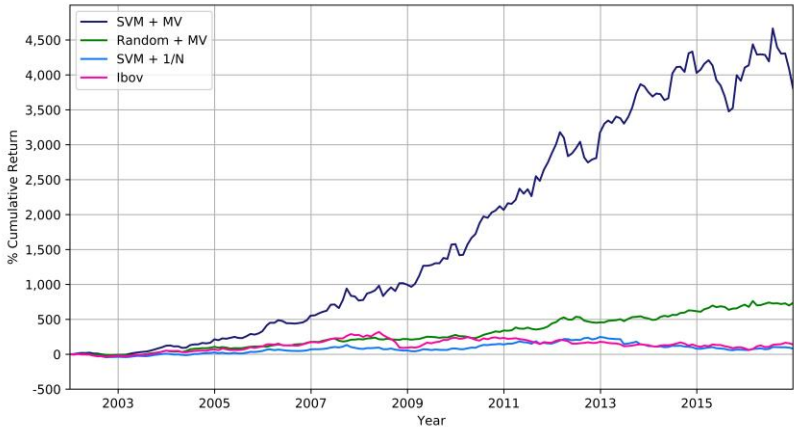

Fig. 8. Cumulative return per model without transaction costs

We already know that the SVM + MV model performs better than the other investment strategies, even after including brokerage costs, since the cumulative return is significantly higher over the studied period (see Fig. 8) and is beneficially associated with the management of a lower number of trades (see Table 2 and Fig. 6). However, it remains pertinent to determine how the performance of the SVM + MV model has a cumulative return higher than the Ibovespa. To this end, we used the Ibovespa as a baseline in the buy-and-hold format. Moreover, it is possible to verify how the SVM + MV and other models behave at different levels of brokerage costs, noting that the weight of these costs is inversely proportional to the invested volume.

In accordance with Figs. 9 and 10, which show the cumulative returns of the simulations involving transaction costs of 0.05 and 0.10 bps, respectively, the SVM + MV model maintains a strong accumulated result. The cumulative return on transactions with a cost of 0.05 bps is 2,956.32%, while for a transaction cost of 0.10 bps it is 2,289.09%. With regard to operations involving transaction costs of 0.50 and 1.00 bps, as shown in Figs. 11 and 12, respectively, the accumulated returns are strongly reduced. Indeed, there is a loss of 61.74% in the simulation with transaction costs of 1.00 bps. Appendix C presents a data set that complements the transaction cost simulations.

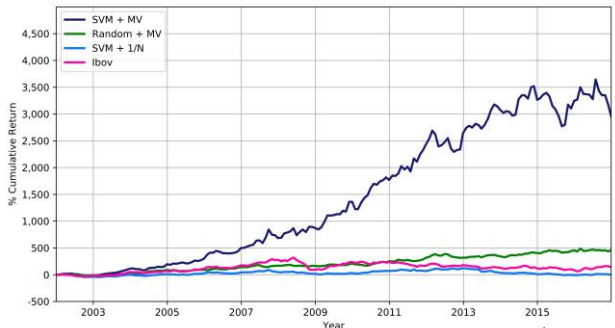

Fig. 9. Cumulative returns including transaction costs (0.05 bps)

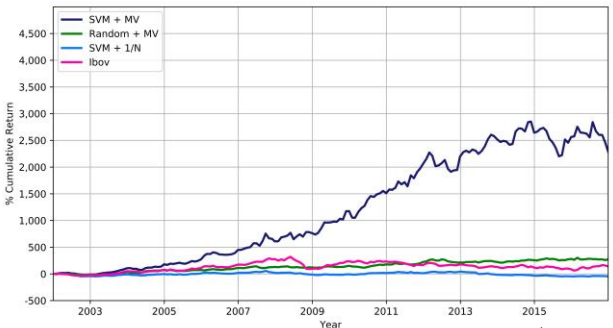

Fig. 10. Cumulative returns including transaction costs (0.10 bps)

Fig. 11. Cumulative returns including transaction costs (0.50 bps)

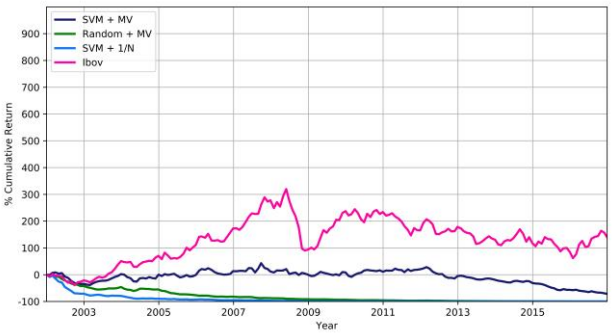

Fig. 12. Cumulative returns including transaction costs (1.00 bps)

The performance of the cumulative returns of the SVM + MV model, compared with the other models and the Ibovespa, triggers the issue of whether performance is really the result of winning operations throughout the studied period or would have occurred only during a specific period. As shown in Fig. 13, we use the performance of the models and the Ibovespa every three years. Thus, we can observe that, of

the 13 surveyed triennia, eight of them show that the cumulative returns of the SVM + MV model perform better than 150% during the respective periods. Moreover, in none of the triennia do the cumulative returns of the other models exceed 150%, except for the Ibovespa, which surpasses the 150% barrier for one period but shows a significant reversal in the following periods. Overall, we find the following average profitability and standard deviations, respectively, for the triennia: 140.08% and 92.43% for the SVM + MV model, 61.41% and 30.73% for the Random + MV model, 29.70% and 62.08% for the Ibovespa, and 27.80% and 51.14% for the SVM + 1/N model.

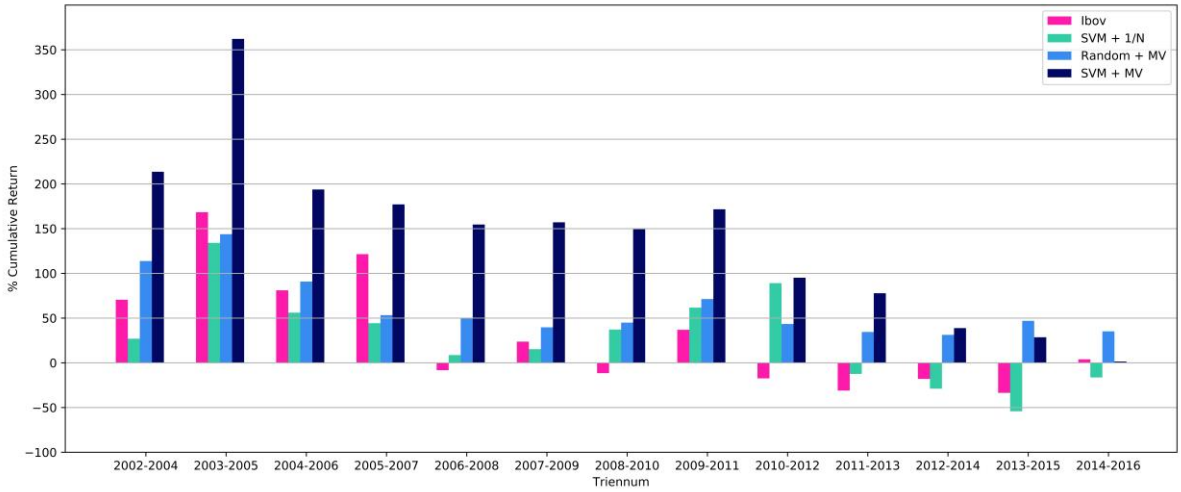

Fig. 13. Cumulative returns of each triennium per model without transaction costs

Appendix D presents the accumulated return data per triennium with the inclusion of transaction costs. The SVM + MV model, with transaction costs of 0.05 and 0.10 bps, behaves similarly to Fig. 13, as shown previously in the cumulative return analysis over the entire studied period. With regard to the simulations with transaction costs of 0.50 and 1.00 bps, the performance of the SVM + MV model makes it unfeasible for adoption.

Fig. 14 shows the result of the calculation for the ratio between return and average monthly risk for each triennium. In addition to having shown a remarkable performance regarding the return metric, the SVM + MV model also obtained the best performance for the return–risk ratio for most periods in the analysis. We find the following average result: 0.40% for the SVM + MV model, 0.32% for the Random + MV model, 0.10% for the SVM + 1/N model, and 0.09% for the Ibovespa. The SVM + MV model stopped having the highest ratio in 2014, which ended up compromising the result for the 2012–2014 triennium and that of the following years. This outcome coincides with the troubled political and economic momentum that Brazil has been dealing with since 2014, given the discovery of various corruption schemes involving public officials and large companies.

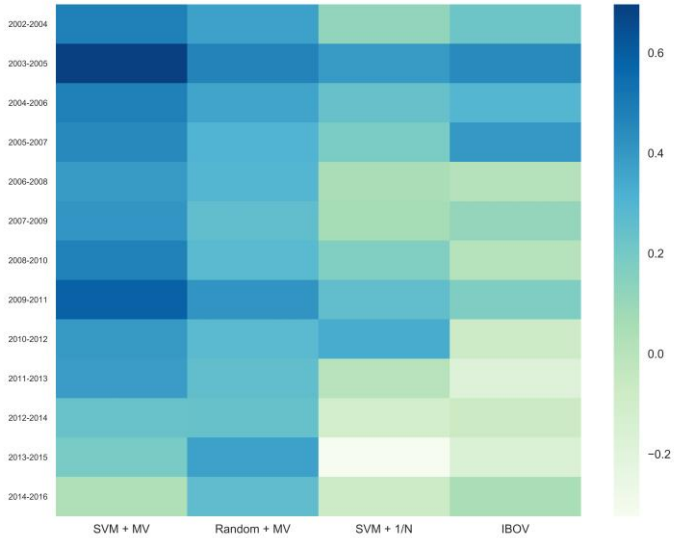

Fig. 14. Monthly average return to the volatility of each triennium per model without transaction costs

Fig. 15 shows the total trades per triennium for each model because this quantification of trades influences models' net profitability. Thus, the Random + MV model, with the second-largest accumulated profitability, would not achieve a net profitability higher than the SVM + MV model even if we considered the transaction costs, given that the average of trades per triennium of the latter is 9,594 as opposed to 16,268 for the former. In other words, the Random + MV model has 69.56% more transactions than the SVM + MV model.

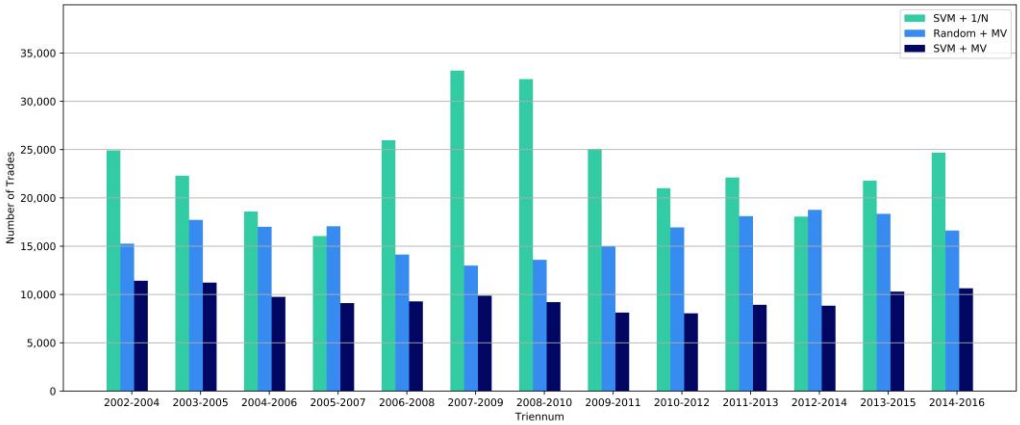

Fig. 15. Total trades of each triennium per model

6. Conclusions

This study extends the theoretical literature on machine learning and stock return prediction. It also provides a practical foundation for a model that can optimize day trading investments. In this regard, the study proposed an investment decision model based on the SVM method to classify assets with a tendency to reach a certain daily return of gain and integrated this classification with the MV diversification method to compose the optimal investment portfolio. The proposed model was referred to as the SVM + MV model. The test period ran from January 2002 to December 2016, totaling 3,716 trading days. Over the studied period, the assets of the Ibovespa were used as the sample. From 135 assets, only 19 were present on all trading days.

First, the classification process using the SVM method presented significant results. The classification performance of the proposed model was higher than the probability of the occurrence of the events delineated and calculated using our sample. The findings confirmed that the snooping data effect was outweighed by the rolling window strategy because there were no statistical differences between the in-sample and out-of-sample results. According to the results, the classifier has a greater discriminatory power when it is required to work with higher targets. Additionally, it proportionately reduces the number of trades among potential businesses.

The cardinality of the daily portfolios in the SVM + MV model also presented an interesting result: a daily average of seven assets per portfolio. The SVM + 1/N model used a higher number of assets (129%), while in the Random + MV model this number was 57.14% higher. The lower the cardinality of the portfolio, the lower are the transaction costs. The two alternative models were derived from the fragmentation of the SVM + MV model: one only maintains a similar classification method and the other only the optimization method. The merger of the classification and optimization processes seemed to converge positively to lower portfolio cardinality. To further portray this process of bottlenecking in the number of assets that composed, on average, the portfolio of the SVM + MV model during the studied period, the Ibovespa was given a theoretical portfolio of 61 assets on average. In the classification stage, these 61 assets were reduced, on average, to 16 assets with the potential to reach the expected gain. Then, the next step of portfolio optimization restricted the average number of assets to seven (i.e., a reduction of approximately 89% in the number of assets for investment).

With regard to cumulative return performance without the inclusion of transaction costs, the SVM + MV model was satisfactorily better than the other models and the Ibovespa. A three-year cumulative return verification confirmed the better performance of the SVM + MV model. Following the inclusion of transaction costs, the SVM + MV model still showed a better result than the other models up to a certain level of such costs, specifically brokerage costs. Thus, the feasibility of the model's implementation is directly related to the amount of financial resources that the investor is willing to apply. Such implementation is also related to demand for a more significant contribution of financial resources to

dilute the proportion of brokerage costs and thereby make the model viable. This situation may, however, find a barrier in the form of the market's liquidity.

The highest performance for the SVM + MV model when faced with all baselines might be checked by different forms of comparison: less portfolio cardinality, higher returns, and/or a better return–risk ratio. Taking account of these results, it is thus important to indicate that there is a side gain given by using the SVM method for asset selection, followed by the MV method for portfolio selection (SVM + MV), since the simple SVM technique selection and equally weighted investment distribution (SVM + 1/N) or the random selection of assets, followed by the MV method (Random + MV) application have shown significantly lower results.

Such a potential barrier means that the study has a possible limitation. The study is also limited because it uses only the Ibovespa's assets. Thus, future research could compare the model with regard to market liquidity indicators and develop simulations on other stock markets. In addition, there is significant scope for research on the application of other categorical predictive methods of machine learning, implementation of stop–loss mechanisms, expansion of the parameter sets for modeling, and use of other investment portfolio optimization techniques.

Acknowledgments

This study is a collaborative effort. The authors are grateful to the Companhia Energética de Minas Gerais (CEMIG) for providing the financial database of Brazilian companies. The authors would also like to thank Mr. Joaquim Dias for his support.

Funding

This work was supported by the Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (Process number 460048/2014-7).

References

Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Berkeley, CA: Apress.

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*, 7046–7056.

Brabazon, A., & O'Neill, M. (2006). *Biologically inspired algorithms for financial modelling*. Berlin: Springer.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Cervantes, J., Garcia-Lamont, F., Rodriguez, L., López, A., Castilla, J. R., Trueba, A. (2017). PSO-based method for SVM classification on skewed data sets. *Neurocomputing*, *228*, 187-197.

Cervelló-Royo, R., Guijarro, F., & Michniuk, K. (2015). Stock market trading rule based on pattern recognition and technical analysis: Forecasting the DJIA index with intraday data. *Expert Systems with Applications, 42*, 5963–5975.

Chen, T.-L., & Chen, F.-Y. (2016). An intelligent pattern recognition model for supporting investment decisions in stock market. *Information Sciences, 346–347*, 261–274.

Chourmouziadis, K., & Chatzoglou, P. D. (2016). An intelligent short term stock trading fuzzy system for assisting investors in portfolio management. *Expert Systems with Applications, 43*, 298–311.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*, 313–327.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machine and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

Deng, G.-F., Lin, W.-T., & Lo, C.-C. (2012). Markowitz-based portfolio selection with cardinality constraints using improved particle swarm optimization. *Expert Systems with Applications, 39*,

4558–4566.

Ding, Y., Song, X., & Zen, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Systems with Applications*, *34*, 3081–3089.

Fabozzi, F. J., Gupta, F., & Markowitz, H. M. (2002). The legacy of modern portfolio theory. *Journal of Investing*, *11*, 7–22.

Fama, E. F. (1965). The behavior of stock-market prices. *Journal of Business*, *38*, 34–105.

Gerlein, E. A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, *54*, 193–207.

Gupta, P., Mehlawat, M. K., & Mittal, G. (2012). Asset portfolio optimization using vector machines and real-coded genetic algorithm. *Journal of Global Optimization, 53,* 297–315.

He, S., Xiao, L., Wang, Y., Liu, X., Yang, C., Lu, J., Gui, W., Sun, Y. (2017). A novel fault diagnosis method based on optimal relevance vector machine. *Neurocomputing, 267,* 651-663.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*, 725–749.

Huang, C.-J., Yang, D.-X., & Chuang, Y-T. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications*, *34*, 2870–2878.

Huang, C.-Y., Chiou, C.-C., Wu, T.-H., & Yang, S.-C. (2015). An integrated DEA-MODM methodology for portfolio optimization. *Operational Research, 15*, 115–136.

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, *32*, 2513–2522.

Jobst N. J., Horniman, M. D., Lucas, C. A., & Mitra G. (2001). Computational aspects of alternative portfolio selection models in the presence of discrete asset choice constraints. *Quantitative Finance, 1*, 489–501

Kampouridis, M., & Otero, F. E. B. (2017). Evolving trading strategies using directional changes. *Expert Systems with Applications, 73*, 145–160.

Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange. *Expert Systems with Applications*, *38*, 5311–5319.

Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing, 13*, 947–958.

Kim, K.-J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*, 307–319.

Kim, Y., & Enke, D. (2016). Developing a rule change trading system for the futures market using rough set analysis. *Expert Systems with Applications*, *59*, 165–173.

Kolm, P.N., Tütüncü, R., & Fabozzi, F. J. (2014). 60 years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, *234*, 356–371.

Li, S., Kwok, J. T., Zhu, H., & Wang, Y. (2003). Texture classification using the support vector machines. *Pattern Recognition, 36*, 2883–2893.

Lo, A. W., Mamaysky, H., & Wang, J. (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, *55*, 1705–1765.

Lu, C.-J., Lee, T.-S., & Chiu, C.-C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems, 47*, 115–125.

Luo, L., & Chen, X. (2013). Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction. *Applied Soft Computing*, *13*, 806–816.

Macedo, L. L., Godinho, P., & Alves, M. J. (2017). Mean-semivariance portfolio optimization with multiobjective evolutionary algorithms and technical analysis rules. *Expert Systems with Applications, 79*, 33–43.

Machado, J., Neves, R. F., & Horta, N. (2015). Developing multi-time frame trading rules with a trend following strategy, using GP. *GECCO*.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, *17*, 59–82.

Mandelbrot, B., & Hudson, R. L. (2004). *The (mis)behavior of markets: A fractal view of financial turbulence*. New York: Basic Books.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*, 77–91.

Matías, J. M., & Reboredo, J. C. (2012). Forecasting performance of nonlinear models for intraday stock returns. *Journal of Forecasting, 31*, 172–188.

Michaud, R. O., & Michaud, R. O. (2008). *Efficient asset management: A practical guide to stock portfolio optimization and asset allocation*. New York: Oxford University Press.

Ni, L.-P., Ni, Z.-W., & Gao, Y.-Z. (2011). Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications*, *38*, 5569–5576.

Pan, Y., Xiao, Z., Wang, X., & Yang, D. (2017). A multiple support vector machine approach to stock index forecasting with mixed frequency sampling. *Knowledge-Based Systems*, *122*, 90–102.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*, 259–268.

Petropoulos, A., Chatzis, S. P., Siakoulis, V., & Vlachogiannakis, N. (2017). A stacked generalization system for automated FOREX portfolio trading. *Expert Systems with Applications, 90,* 290–302.

Sands, T. M., Tayal, D., Morris, M. E., & Monteiro, S. T. (2015). Robust stock value prediction using support vector machines with particle swarm optimization. *2015 IEEE Congress on Evolutionary Computation (CEC)*, 3327–3331.

Santos, A. A. P., & Tessari, C. (2012). Técnicas quantitativas de otimização de carteiras aplicadas ao mercado de ações Brasileiro (Quantitative portfolio optimization techniques applied to the Brazilian stock market). *Revista Brasileira de Finanças, 10*, 369–394.

Sheta, A. F., Ahmed, S. E. M., & Faris, H. (2015). A comparison between regression, artificial neural networks and support vector machines for predicting stock market index. *International Journal of Advanced Research in Artificial Intelligence*, *4*, 55–63.

Silva, A., Neves, R., & Horta, N. (2015). A hybrid approach to portfolio composition based on fundamental and technical indicators. *Expert Systems with Applications, 42*, 2036–2048.

Tan, T. Z., Quek, C., & Ng, G. S. (2007). Biological brain-inspired genetic complementary learning for stock market and bank failure prediction. *Computational Intelligence*, *23*, 236–261.

Teräsvirta, T., Van Dijk, D., & Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: a re-examination. *International Journal of Forecasting*, *21*, 755–774.

Thenmozhi, M., & Sarath Chand, G. (2016). Forecasting stock returns based on information transmission across global markets using support vector machines. *Neural Computing and Applications*, *27*, 805–824.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*, 988–999.

Wen, Q., Yang, Z., Song, Y., & Jia, P. (2010). Automatic stock decision support system based on box theory and SVM algorithm. *Expert Systems with Applications*, *37*, 1015–1022.

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*, 1097–1126.

Xu, H., Caramanis, C., & Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of Machine Learning Research, 10*, 1485–1510.

Yu, L., Wang, S., & Lai, K. K. (2009). A neural-network-based nonlinear metamodeling approach to financial time series forecasting. *Applied Soft Computing*, 9, 563–574.

Zhang, W., Cao, Q., & Schniederjans, M. J. (2004) Neural network earnings per share forecasting models: A comparative analysis of alternative methods. *Decision Sciences, 35*, 205–237.

Zhiqiang, G., Huaiqing, W., & Quan, L. (2013). Financial time series forecasting using LPP and SVM optimized by PSO. *Soft Computing*, *17*, 805–818.

APPENDIX A. Selected sets for simulation

Table A1
Best performance of the simulation parameters over the in-sample window

Month	2001	2002	2003	2004
January		Set2_75d_90d_200p	Set1_90d_90d_150p	Set1_90d_90d_200p
February		Set2_75d_75d_200p	Set1_60d_90d_200p	Set1_60d_75d_150p
March		Set2_90d_75d_150p	Set2_90d_90d_200p	Set1_90d_75d_150p
April		Set2_75d_90d_200p	Set1_90d_75d_200p	Set2_90d_90d_200p
May		Set1_75d_105d_150p	Set3_90d_75d_200p	Set1_60d_75d_200p
June		Set3_75d_105d_150p	Set1_90d_105d_200p	Set2_90d_75d_200p
July		Set1_75d_105d_200p	Set3_90d_90d_150p	Set3_75d_90d_200p
August		Set3_90d_90d_200p	Set3_75d_105d_200p	Set1_60d_75d_200p
September		Set2_75d_75d_200p	Set1_60d_105d_200p	Set3_75d_75d_200p
October		Set3_60d_75d_200p	Set3_75d_75d_200p	Set2_60d_90d_100p
November		Set1_60d_90d_200p	Set3_90d_90d_200p	Set3_90d_105d_200p
December	Set2_60d_105d_200p	Set3_90d_105d_150p	Set1_90d_75d_200p	Set3_75d_105d_200p

Month	2005	2006	2007	2008
January	Set2_90d_75d_200p	Set1_75d_90d_200p	Set1_60d_105d_150p	Set1_60d_75d_150p
February	Set3_90d_75d_200p	Set3_90d_90d_200p	Set3_90d_75d_150p	Set2_75d_105d_200p
March	Set3_90d_75d_200p	Set2_60d_90d_200p	Set2_75d_75d_100p	Set3_90d_105d_200p
April	Set1_90d_90d_200p	Set1_60d_105d_200p	Set3_90d_75d_200p	Set3_60d_75d_200p
May	Set2_60d_105d_200p	Set3_75d_75d_200p	Set3_60d_75d_200p	Set3_60d_75d_200p
June	Set2_60d_90d_150p	Set2_90d_90d_100p	Set1_60d_75d_100p	Set3_90d_75d_100p
July	Set3_90d_75d_150p	Set1_90d_75d_200p	Set1_60d_75d_200p	Set3_60d_75d_200p
August	Set3_75d_90d_200p	Set2_75d_105d_150p	Set1_60d_75d_200p	Set3_90d_105d_200p
September	Set3_90d_105d_150p	Set3_75d_90d_200p	Set3_60d_75d_200p	Set3_75d_75d_150p
October	Set2_75d_90d_200p	Set2_60d_105d_150p	Set2_90d_75d_200p	Set3_60d_105d_200p
November	Set3_60d_105d_200p	Set2_60d_90d_200p	Set2_75d_75d_200p	Set1_90d_75d_100p
December	Set1_60d_75d_200p	Set2_60d_90d_200p	Set1_90d_75d_150p	Set1_60d_75d_200p

Month	2009	2010	2011	2012
January	Set1_60d_105d_100p	Set2_75d_90d_100p	Set1_75d_75d_200p	Set1_60d_90d_200p
February	Set3_90d_90d_200p	Set3_90d_90d_200p	Set3_75d_90d_150p	Set3_60d_105d_200p
March	Set2_60d_90d_200p	Set3_75d_90d_200p	Set1_60d_75d_200p	Set1_75d_105d_200p
April	Set1_75d_105d_200p	Set3_75d_105d_200p	Set1_60d_90d_200p	Set3_90d_90d_100p
May	Set2_60d_105d_150p	Set3_90d_75d_200p	Set1_60d_90d_150p	Set3_90d_75d_100p
June	Set3_60d_75d_100p	Set3_90d_75d_150p	Set1_75d_105d_150p	Set1_90d_105d_200p
July	Set1_75d_75d_200p	Set3_90d_105d_200p	Set2_60d_75d_200p	Set1_75d_105d_200p
August	Set2_90d_105d_150p	Set2_60d_105d_100p	Set2_90d_90d_200p	Set3_60d_75d_200p
September	Set1_60d_105d_150p	Set1_75d_75d_150p	Set3_90d_75d_100p	Set1_90d_105d_200p
October	Set1_90d_75d_200p	Set1_90d_75d_200p	Set1_90d_75d_150p	Set1_75d_105d_150p
November	Set1_75d_90d_200p	Set2_90d_105d_150p	Set2_90d_90d_100p	Set1_75d_90d_200p
December	Set3_90d_75d_150p	Set3_75d_75d_200p	Set3_75d_90d_200p	Set1_90d_105d_200p

Month	2013	2014	2015	2016
January	Set2_60d_75d_150p	Set1_75d_105d_150p	Set2_90d_75d_200p	Set1_90d_90d_200p
February	Set2_75d_105d_200p	Set2_75d_75d_150p	Set3_75d_75d_200p	Set2_60d_90d_150p
March	Set3_75d_105d_150p	Set3_90d_75d_200p	Set1_60d_105d_200p	Set3_60d_90d_200p
April	Set1_90d_90d_200p	Set3_60d_75d_200p	Set2_90d_75d_200p	Set3_90d_75d_150p
May	Set2_60d_75d_200p	Set3_90d_75d_150p	Set2_90d_105d_100p	Set1_90d_75d_100p
June	Set2_60d_75d_150p	Set3_60d_75d_200p	Set3_60d_105d_200p	Set2_60d_105d_200p
July	Set3_60d_105d_200p	Set1_75d_105d_150p	Set3_60d_75d_100p	Set1_75d_75d_200p
August	Set3_75d_75d_200p	Set3_90d_105d_200p	Set2_60d_105d_200p	Set3_75d_90d_150p
September	Set1_75d_105d_200p	Set1_60d_105d_200p	Set3_75d_75d_200p	Set3_90d_105d_200p
October	Set3_60d_75d_200p	Set3_90d_75d_200p	Set2_60d_90d_200p	Set3_75d_90d_200p
November	Set3_90d_105d_200p	Set2_60d_75d_200p	Set3_90d_75d_150p	Set3_90d_75d_200p
December	Set2_90d_75d_150p	Set1_75d_75d_150p	Set2_90d_75d_200p	

Note: The simulation coding is (Set of input for SVM)_(Training days for SVM)_(Days for calculating the covariance matrix)_(Target for SVM)

APPENDIX B. Results of the performance metrics for the SVM classifier

Table B1
Classification metrics of the in-sample period

Ticker	Precision		Specif.	Signals Total	Ticker	Precision		Specif.	Signals Total	Ticker	Precision		Specif.	Signals Total
	Observed	SVM				Observed	SVM				Observed	SVM		
ABEV3	26.41	56.31	76.47	325	CTIP3	20.52	46.00	87.05	50	OIBR4	41.53	54.61	61.26	954
ACES4	25.90	45.40	90.69	163	CYRE3	46.00	59.72	73.55	916	PCAR4	33.27	58.53	71.31	422
AEDU3	33.50	58.33	64.83	24	DASA3	37.22	56.76	73.49	111	PDGR3	45.94	55.22	70.40	536
AELP3	56.41	53.33	70.73	15	DTEX3	39.58	58.58	78.73	338	PETR3	35.81	54.83	69.26	777
AGEI3	66.67	100.00	56.00	13	DURA4	48.54	58.82	34.69	221	PETR4	30.93	54.61	73.04	531
ALLL11	44.68	58.23	43.44	316	EBTP3	43.32	57.38	79.77	359	POMO4	44.98	53.75	66.06	80
ALLL3	35.04	52.41	73.89	145	EBTP4	43.32	55.13	76.72	448	PRML3	45.75	54.76	75.36	420
ARCE3	34.86	56.52	92.26	92	ECOR3	37.95	57.32	87.26	164	PRTX3	33.33	76.92	95.60	13
ARCZ6	37.80	54.38	85.04	480	EGIE3	42.94	56.95	65.56	518	PTIP4	43.07	56.06	20.29	462
BBAS3	39.30	52.51	65.07	918	ELET3	39.03	50.74	63.30	875	QUAL3	43.63	56.25	89.58	240
BBAS4	45.78	43.55	98.65	62	ELET6	39.66	51.67	61.99	838	RADL3	34.95	52.63	76.72	57
BBDC3	36.56	51.55	72.41	161	ELPL4	35.05	51.00	76.30	400	RDCD3	40.67	62.33	89.89	361
BBDC4	29.85	48.83	72.36	385	ELPL5	50.34	56.03	85.35	589	RENT3	37.91	55.51	69.43	254
BBSE3	33.24	38.98	93.12	118	LBR3	35.56	57.12	69.07	653	RLOG3	59.68	58.49	92.83	53
BISA3	42.67	59.84	68.58	366	LBR4	35.67	46.96	81.28	115	RSID3	48.97	55.68	57.15	792
BNCA3	32.78	58.76	45.88	177	ENBR3	40.10	48.31	74.67	236	RUMO3	51.57	60.98	96.77	164
BRAP4	37.94	54.59	66.86	828	EQTL3	38.19	60.38	84.27	53	SANB11	38.01	58.06	85.93	341
BRDT4	37.27	43.33	87.99	30	ESTC3	45.71	55.47	80.17	256	SBSP3	38.89	53.42	65.15	803
BRFS3	34.34	53.18	69.40	440	EVEN3	37.08	46.15	74.02	39	SDIA4	45.04	57.93	32.08	347
BRKM5	40.97	58.28	65.52	1,009	FIBR3	38.15	56.45	85.59	372	SMLE3	44.78	52.59	91.54	135
BRML3	38.21	55.08	76.22	305	GFSA3	48.13	58.31	73.61	902	SUZB5	34.54	51.80	67.07	139
BRPR3	34.80	48.94	85.89	94	GGBR4	39.19	58.04	66.68	877	TAMM4	45.17	57.04	84.28	540
BRTP3	46.42	53.80	83.96	790	GOAU4	42.34	57.66	67.50	907	TCOC4	43.06	48.48	12.89	328
BRTP4	45.64	56.70	25.45	709	GOLL4	46.58	58.03	66.91	903	TCSL4	41.91	48.20	78.26	724
BTOW3	46.59	60.30	68.15	728	HGTX3	40.96	57.81	73.42	301	TDBH4	54.82	53.10	59.10	113
BVMF3	37.17	54.75	83.92	442	HYPE3	37.27	54.14	80.51	314	TESA3	53.08	70.75	75.39	465
CCPR3	45.86	59.18	85.40	98	INEP4	51.00	56.69	55.02	127	TIMP3	44.57	54.22	61.06	1,363
CCRO3	38.96	53.85	68.41	624	ITSA4	29.85	42.99	71.43	328	TLCP4	52.97	57.40	7.20	655
CESP5	48.28	53.33	74.37	465	ITUB4	30.07	50.84	72.57	417	TMAR5	39.28	54.15	40.90	663
CESP6	42.55	57.23	78.29	788	JBSS3	46.10	56.07	77.64	815	TMCP4	40.11	48.06	82.58	412
CGAS5	42.01	53.86	73.33	583	KLBN11	27.00	42.86	95.19	35	TNEP4	47.10	54.65	7.36	269
CIEL3	29.78	58.06	87.38	186	KLBN4	39.62	51.97	66.37	787	TNLP3	43.76	51.77	35.59	902
CLSC4	40.29	53.46	73.29	621	KROT3	43.03	55.51	78.87	236	TNLP4	32.78	51.93	46.04	441
CMET4	32.12	40.48	82.51	42	LAME4	38.25	54.24	64.23	483	TRPL4	38.21	53.00	67.74	800
CMIG3	42.00	46.81	65.69	282	LIGH3	44.59	49.03	11.72	359	UBBR11	41.44	54.21	85.30	273
CMIG4	35.53	51.57	68.42	731	LIGT3	37.95	56.61	76.19	507	UGPA3	24.42	52.59	74.89	135
CPFE3	32.26	57.01	78.48	428	LREN3	41.97	54.65	74.05	763	UGPA4	33.56	57.36	84.16	129
CPLE6	39.59	51.90	64.65	948	MMXM3	49.22	58.66	71.47	554	USIM3	47.15	53.82	66.45	589
CRTP5	37.96	50.79	16.09	189	MRFG3	45.64	56.38	77.52	603	USIM5	43.90	57.07	62.79	1,244
CRUZ3	36.96	54.83	71.32	808	MRVE3	49.24	55.42	74.03	803	VALE3	34.27	56.61	70.41	620
CSAN3	40.25	57.18	75.55	689	MULT3	42.62	58.33	77.60	132	VALE5	28.34	58.39	75.43	411
CSNA3	43.57	60.16	63.85	1,142	NATU3	39.43	56.61	72.51	643	VCPA4	39.17	56.75	84.75	437
CSTB4	44.57	50.14	92.28	353	NETC4	48.82	62.13	76.73	1,014	VIVO4	44.47	54.36	62.84	802
CTAX3	50.20	58.04	76.81	112	OGXP3	48.00	64.90	71.80	433	VIVT4	32.96	49.77	70.62	649
CTAX4	40.16	49.25	72.08	67	OIBR3	50.00	55.52	66.70	317	WEGE3	39.91	57.14	72.80	63

Table B2
Classification metrics of the out-of-sample period

Ticker	Precision		Specif.	Signals Total	Ticker	Precision		Specif.	Signals Total	Ticker	Precision		Specif.	Signals Total
	Observed	SVM				Observed	SVM				Observed	SVM		
ABEV3	25.30	48.41	76.61	283	CTIP3	19.55	21.43	86.71	14	OIBR4	41.47	56.31	61.89	927
ACES4	25.26	45.74	90.69	129	CYRE3	45.45	58.78	73.65	934	PCAR4	32.76	56.15	72.07	447
AEDU3	33.00	65.00	64.37	20	DASA3	37.22	54.08	72.94	98	PDGR3	45.72	57.44	70.82	484
AELP3	53.85	55.00	71.43	20	DTEX3	38.96	52.82	78.46	337	PETR3	35.84	57.72	70.31	816
AGEI3	59.52	100.00	55.65	7	DURA4	46.59	58.38	34.47	197	PETR4	31.59	57.73	73.07	563
ALLL11	44.49	58.57	43.39	321	EBTP3	43.72	57.49	80.11	367	POMO4	44.18	47.62	65.92	63
ALLL3	34.77	51.40	74.44	179	EBTP4	42.66	56.18	77.45	461	PRML3	45.66	56.50	75.03	400
ARCE3	34.59	50.00	92.12	76	ECOR3	39.70	54.60	87.10	174	PRTX3	34.96	37.50	95.47	8
ARCZ6	38.31	53.18	85.29	534	EGIE3	42.34	53.40	64.88	530	PTIP4	44.33	54.06	19.24	468
BBAS3	39.37	52.84	65.04	916	ELET3	39.46	52.24	63.64	894	QUAL3	42.40	53.60	89.44	250
BBAS4	45.18	43.08	98.71	65	ELET6	39.63	50.79	62.17	827	RADL3	34.24	51.92	76.97	52
BBDC3	36.59	56.32	72.56	174	ELPL4	35.25	50.50	75.74	398	RDCD3	41.36	62.93	89.40	348
BBDC4	29.79	51.36	72.85	405	ELPL5	50.93	56.07	85.37	585	RENT3	36.54	52.31	69.59	260
BBSE3	34.72	45.83	92.99	120	LBR3	34.82	50.48	68.34	624	RLOG3	58.06	59.18	92.47	49
BISA3	42.87	58.19	68.18	397	LBR4	35.58	44.30	81.70	158	RSID3	49.33	57.97	58.00	790
BNCA3	33.23	59.68	45.67	186	ENBR3	41.37	53.16	75.25	237	RUMO3	51.38	60.22	96.83	186
BRAP4	37.89	56.51	67.72	860	EQTL3	40.30	60.78	83.79	51	SANB11	38.46	56.23	85.84	361
BRDT4	35.79	46.88	88.27	32	ESTC3	45.63	54.15	79.82	277	SBSP3	39.10	53.18	65.07	850
BRFS3	34.60	55.73	69.71	454	EVEN3	37.08	54.35	74.17	46	SDIA4	44.74	58.18	31.99	318
BRKM5	40.93	53.92	64.42	1,085	FIBR3	37.76	52.56	85.42	390	SMLE3	44.20	54.11	91.68	146
BRML3	38.29	54.35	75.40	322	GFSA3	47.32	59.41	74.18	877	SUZB5	35.45	49.34	67.03	152
BRPR3	35.85	48.60	85.90	107	GGBR4	39.29	56.54	66.24	902	TAMM4	44.89	59.44	84.37	498
BRTP3	46.10	55.05	85.11	841	GOAU4	41.86	56.20	67.49	911	TCOC4	42.87	46.15	12.36	351
BRTP4	45.49	57.26	25.43	723	GOLL4	45.82	56.94	66.89	850	TCSL4	42.84	51.80	78.25	695
BTOW3	45.26	56.70	68.06	739	HGTX3	40.26	56.85	73.39	292	TDBH4	54.22	53.78	58.83	119
BVMF3	36.79	55.48	84.29	456	HYPE3	35.94	53.53	80.73	312	TESA3	51.43	69.43	75.45	458
CCPR3	48.12	62.50	85.33	96	INEP4	48.59	60.94	55.55	128	TIMP3	44.48	54.15	61.10	1,361
CCRO3	37.55	52.85	69.58	632	ITSA4	30.19	48.08	71.75	364	TLCP4	52.60	58.09	7.15	680
CESP5	48.63	53.19	74.80	470	ITUB4	30.14	51.27	72.65	433	TMAR5	39.55	52.98	40.22	638
CESP6	42.43	54.77	77.84	796	JBSS3	45.92	58.06	77.79	794	TMCP4	40.05	49.64	83.16	421
CGAS5	42.21	51.22	71.98	572	KLBN11	26.63	42.86	95.21	42	TNEP4	46.39	51.61	7.10	279
CIEL3	29.52	53.71	86.87	175	KLBN4	39.36	53.04	67.00	807	TNLP3	43.60	51.54	35.64	846
CLSC4	40.19	51.62	72.40	585	KROT3	42.30	58.23	79.15	249	TNLP4	32.47	55.40	46.44	417
CMET4	33.33	48.15	82.80	54	LAME4	37.21	53.80	64.95	461	TRPL4	37.81	52.28	67.87	790
CMIG3	42.72	48.76	65.65	322	LIGH3	44.69	51.92	11.87	364	UBBR11	41.96	58.94	85.43	263
CMIG4	36.11	49.50	67.55	798	LIGT3	37.91	55.87	76.16	537	UGPA3	23.20	44.85	75.31	136
CPFE3	32.11	55.36	78.07	392	LREN3	41.93	56.14	73.92	725	UGPA4	32.77	49.58	84.01	119
CPLE6	39.80	51.12	64.01	935	MMXM3	50.00	60.49	71.29	529	USIM3	47.01	56.63	66.54	581
CRTP5	38.05	49.47	15.88	190	MRFG3	45.85	57.88	77.89	641	USIM5	43.35	55.62	62.77	1,237
CRUZ3	36.99	54.36	71.15	780	MRVE3	47.14	56.60	75.32	795	VALE3	34.53	54.77	69.88	608
CSAN3	40.07	57.26	75.99	709	MULT3	40.61	48.55	77.75	138	VALE5	28.28	53.94	74.98	419
CSNA3	43.06	58.04	63.93	1,182	NATU3	38.99	53.56	72.65	674	VCPA4	38.17	57.52	85.29	419
CSTB4	44.26	50.67	92.76	375	NETC4	48.50	60.50	76.44	995	VIVO4	44.04	52.98	62.45	821
CTAX3	46.59	54.46	76.87	101	OGXP3	47.79	63.38	71.27	396	VIVT4	32.29	48.54	71.14	649
CTAX4	38.96	50.00	71.30	54	OIBR3	51.01	56.04	66.44	298	WEGE3	38.15	50.00	72.62	64

APPENDIX C. Analysis of daily returns for strategies with transaction costs

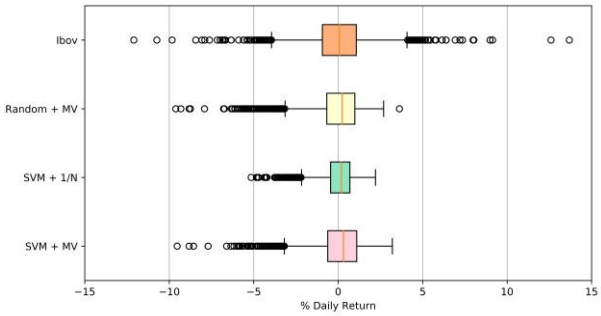

Fig. C1. Daily returns including transaction costs (0.05 bps)

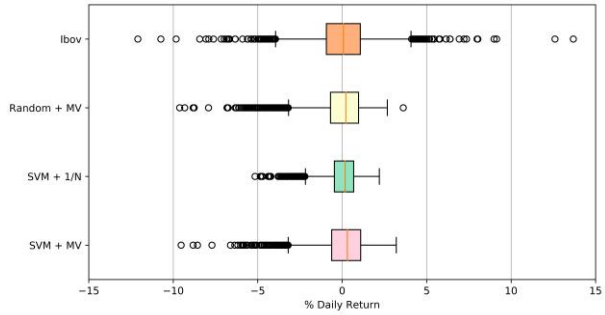

Fig. C2. Daily returns including transaction costs (0.10 bps)

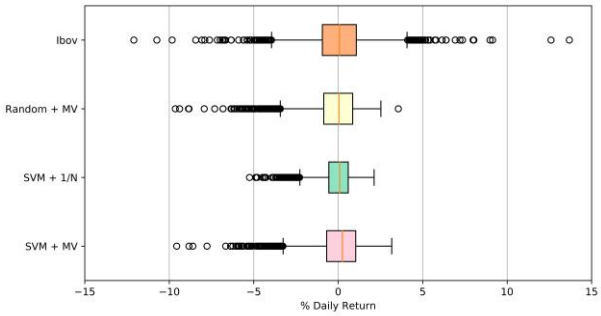

Fig. C3. Daily returns including transaction costs (0.50 bps)

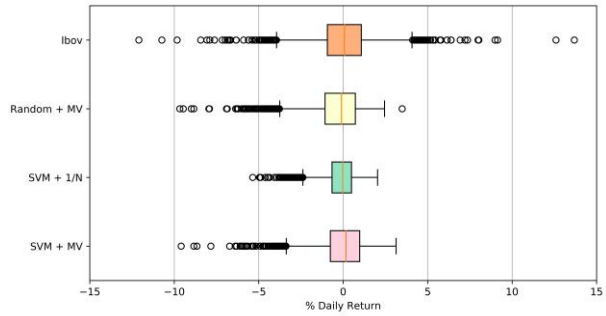

Fig. C4. Daily returns including transaction costs (1.00 bps)

Table C1
Average daily return for each strategy (%)

Characteristic	SVM + MV	SVM + 1/N	Random + MV	Ibovespa
Without transaction cost	0.11	0.03	0.06	0.04
Transaction cost (0.05 bps)	0.10	0.01	0.05	0.04
Transaction cost (0.10 bps)	0.10	-0.01	0.04	0.04
Transaction cost (0.50 bps)	0.04	-0.05	-0.14	0.04
Transaction cost (1.00 bps)	-0.02	-0.30	-0.16	0.04

Table C2
Normality test for daily returns (p-value)

Characteristic	SVM + MV	SVM + 1/N	Random + MV	Ibovespa
Without transaction cost	0.00	0.00	0.00	0.00
Transaction cost (0.05 bps)	0.00	0.00	0.00	0.00
Transaction cost (0.10 bps)	0.00	0.00	0.00	0.00
Transaction cost (0.50 bps)	0.00	0.00	0.00	0.00
Transaction cost (1.00 bps)	0.00	0.00	0.00	0.00

Table C3
Kruskal–Wallis test for daily returns

Characteristic	p-value
Without transaction cost	1.04e-09
Transaction cost (0.05 bps)	2.30e-09
Transaction cost (0.10 bps)	4.05e-09
Transaction cost (0.50 bps)	3.66e-12
Transaction cost (1.00 bps)	1.42e-27

Table C4
Dunn's test for daily returns including a transaction cost of 0.05 bps (p-value)

	SVM + MV	SVM + 1/N	Random + MV
SVM + 1/N	0.001	-	-
Random + MV	0.001	0.001	-
Ibovespa	0.001	0.001	0.983

Table C5
Dunn's test for daily returns including a transaction cost of 0.10 bps (p-value)

	SVM + MV	SVM + 1/N	Random + MV
SVM + 1/N	0.001	-	-
Random + MV	0.001	0.001	-
Ibovespa	0.001	0.012	0.764

Table C6
Dunn's test for daily returns including a transaction cost of 0.50 bps (p-value)

	SVM + MV	SVM + 1/N	Random + MV
SVM + 1/N	0.001	-	-
Random + MV	0.001	0.387	-
Ibovespa	0.001	0.043	0.001

Table C7
Dunn's test for daily returns including a transaction cost of 1.00 bps (p-value)

	SVM + MV	SVM + 1/N	Random + MV
SVM + 1/N	0.001	-	-
Random + MV	0.001	0.180	-
Ibovespa	0.044	0.001	0.001

APPENDIX D. Cumulative returns for strategies with transaction costs

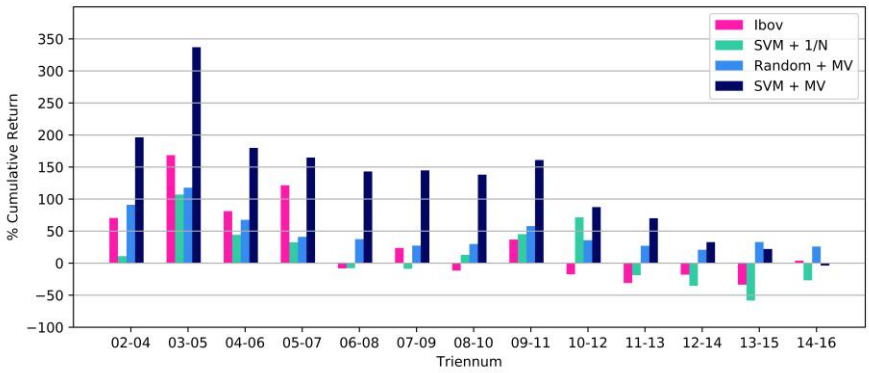

Fig. D1. Cumulative returns of each triennium per model including transaction costs (0.05 bps)

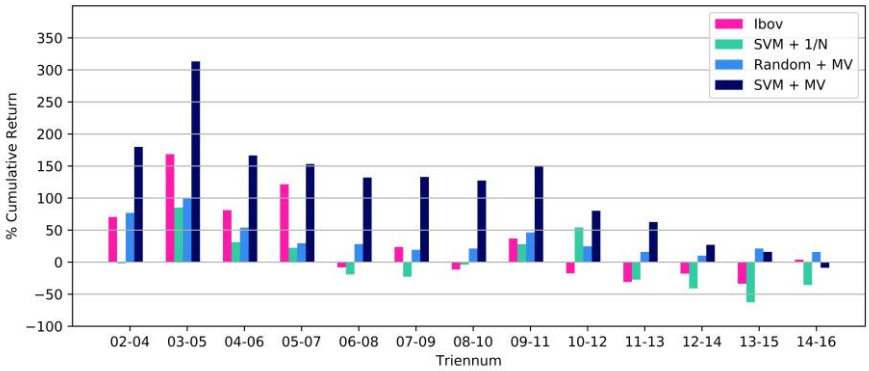

Fig. D2. Cumulative returns of each triennium per model including transaction costs (0.01 bps)

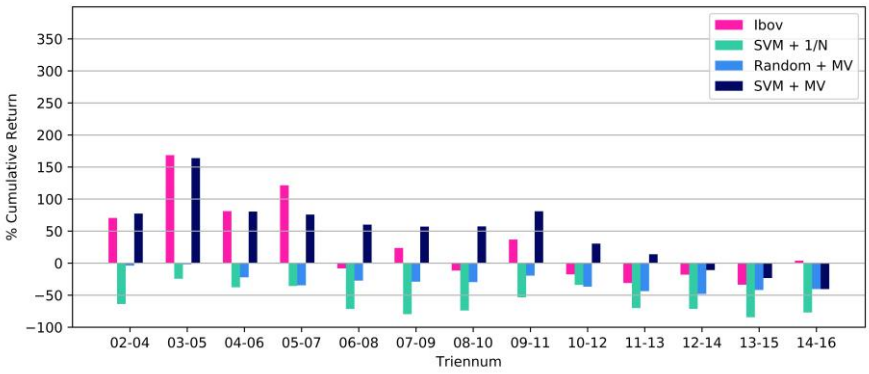

Fig. D3. Cumulative returns of each triennium per model including transaction costs (0.50 bps)

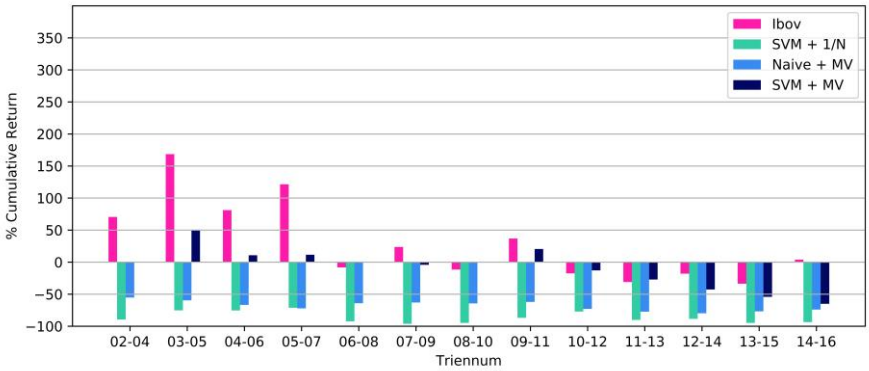

Fig. D4. Cumulative returns of each triennium per model including transaction costs (1.00 bps)

APPENDIX E. Average daily trading value for the Ibovespa's assets

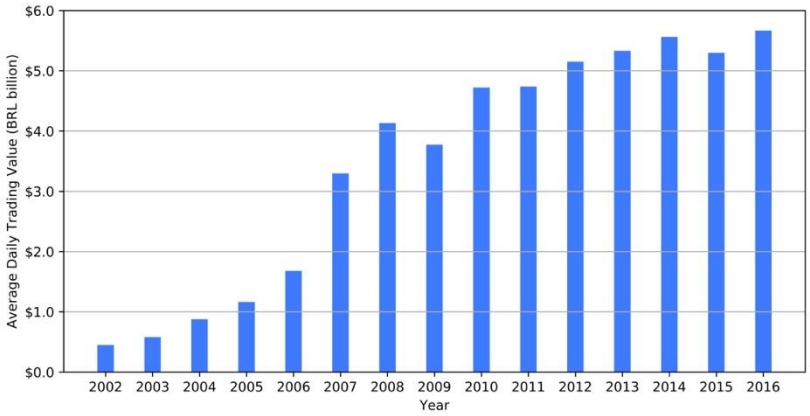

Fig. E1. Average daily trading value for the Ibovespa's assets (Brazilian real (BRL))