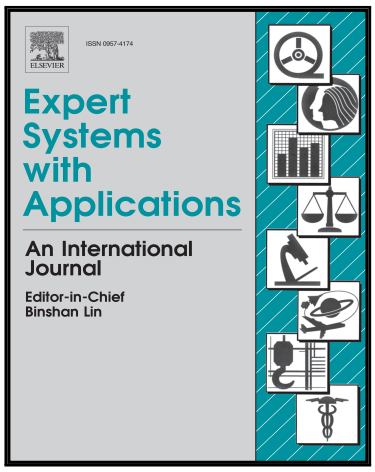

Literature Review: Machine Learning Techniques Applied to Financial Market Prediction

Bruno Miranda Henrique, Vinicius Amorim Sobreiro, Herbert Kimura

PII: S0957-4174(19)30017-X
DOI: https://doi.org/10.1016/j.eswa.2019.01.012
Reference: ESWA 12413

To appear in: *Expert Systems With Applications*

Received date: 15 May 2018
Revised date: 30 August 2018
Accepted date: 4 January 2019

Please cite this article as: Bruno Miranda Henrique, Vinicius Amorim Sobreiro, Herbert Kimura, Literature Review: Machine Learning Techniques Applied to Financial Market Prediction, *Expert Systems With Applications* (2019), doi: https://doi.org/10.1016/j.eswa.2019.01.012

Highlights

- The search for models to predict the prices is still a highly researched topic;
- The prices are financial time series that are difficult to predict; and
- The machine learning area applied to the prediction of financial market prices.

Literature Review: Machine Learning Techniques Applied to Financial Market Prediction☆

Bruno Miranda Henrique[a], Vinicius Amorim Sobreiro[a,*], Herbert Kimura[a]

[a]*University of Brasília, Department of Management, Campus Darcy Ribeiro, Brasília, Federal District, 70910-900, Brazil.*

Abstract

The search for models to predict the prices of financial markets is still a highly researched topic, despite major related challenges. The prices of financial assets are non-linear, dynamic, and chaotic; thus, they are financial time series that are difficult to predict. Among the latest techniques, machine learning models are some of the most researched, given their capabilities for recognizing complex patterns in various applications. With the high productivity in the machine learning area applied to the prediction of financial market prices, objective methods are required for a consistent analysis of the most relevant bibliography on the subject. This article proposes the use of bibliographic survey techniques that highlight the most important texts for an area of research. Specifically, these techniques are applied to the literature about machine learning for predicting financial market values, resulting in a bibliographical review of the most important studies about this topic. Fifty-seven texts were reviewed, and a classification was proposed for markets, assets, methods, and variables. Among the main results, of particular note is the greater number of studies that use data from the North American market. The most commonly used models for prediction involve support vector machines (SVMs) and neural networks. It was concluded that the research theme is still relevant and that the use of data from developing markets is a research opportunity.

Keywords: Financial time series prediction, machine learning, literature review, main path analysis.

☆This document was a collaborative effort.

*Corresponding author

Email addresses: brunomhenrique@hotmail.com (Bruno Miranda Henrique), sobreiro@unb.br (Vinicius Amorim Sobreiro), herbert.kimura@gmail.com (Herbert Kimura)

1. Introduction

The prediction of stock markets is one of the most important and challenging problems involving time series (Chen et al., 2017, p. 340). Despite the establishment of the efficient market hypothesis (EMH) by Malkiel and Fama (1970), which was later revised in Fama (1991), according to which financial markets follow random pathways and therefore are unpredictable, the search for models and profitable systems is still attracting a lot of attention from academia (Weng et al., 2017, p. 153). In the specialized literature, there is evidence contrary to the efficiency of financial markets, as summarized by Kumar et al. (2016), Atsalakis and Valavanis (2009), Malkiel (2003), and Fama (1991). Additionally, a predictive model capable of consistently generating returns above the market indices over time would not only represent strong evidence contrary to the EMH but would enable large profits with financial operations.

The prediction of price time series in financial markets, which have a non-stationary nature, is very difficult (Zhang et al., 2017; Tay and Cao, 2001, p. 161, p. 309). They are dynamic, chaotic, noisy, non-linear series (Bezerra and Albuquerque, 2017; Göçken et al., 2016; Kumar and Thenmozhi, 2014, p. 180, p. 320, p. 285) that are influenced by the general economy, characteristics of the industries, politics, and even the psychology of investors (Zhong and Enke, 2017; Chen et al., 2017, p. 126, p. 340). Thus, the literature about financial market prediction is rich in methods and practical applications regarding historical data for evaluating the profitability of techniques.

Among the classic financial market prediction techniques, of particular note are the following: technical analysis (TA) with standards of support and resistance and indicators calculated from past prices (Chen et al., 2014, pp. 329–330) to indicate bearish or bullish trends (Lahmiri, 2014b, p. 1450013-3) and fundamentalist analysis, which seeks economic factors that have an influence on market trends (Cavalcante et al., 2016, p. 194). Stock prices and market indices are also treated with time series analysis tools. The initial prediction techniques were moving averages, autoregressive models, discriminating analyses, and correlations (Kumar and Thenmozhi, 2014; Wang et al., 2012, p. 285; p. 758). More recently, a promising area of research in the prediction of time series is that of artificial intelligence (Yan et al., 2017; Wang et al., 2012, p. 2266; p. 758), given that the techniques are designed to address chaotic data, randomness, and non-linearity (Chen et al., 2017, pp. 340–341).

Technological advances have made it possible to analyse large historical price databases with computational systems, as introduced by Chiang et al. (2016, p. 195). The intense computational use of intelligent predictive models is commonly studied under the title of machine learning. According to Hsu et al. (2016, p. 215), it is common to test techniques for analysing time series using data from the financial market, given its difficult predictability. Thus, the literature regarding financial market prediction using machine learning is vast, which hinders revisions, systematizations of models and techniques, and searches for material to determine the state of the art. Tools are needed to objectively and quantitatively select the most relevant

works for a literature review covering the most influential articles. Thus, the intention of this article is to present methods for the selection of the main advances in machine learning applied to financial market prediction, in order to present a review of the selected articles, clarifying the knowledge flow that the literature follows, and to propose a classification for the articles. Additionally, this paper brings a summary of the best procedures followed by the literature on applying machine learning to financial time series forecasting.

The selection of the most relevant literature for the proposed review was performed by searching the theme in the *Scopus* database and validating the group of articles selected as a representative sample of the literature. For the review of the articles, objective parameters were proposed as a means of indicating those that are most relevant. Thus, the following were included in the review: the most-cited articles, the articles with the greatest bibliographic coupling, those with the greatest relationship in a co-citation network, those most recently published, and those that are part of the main path of the literature – a technique used to trace the flow of knowledge in a given scientific discipline (Liu et al., 2013, p. 4). The articles were then objectively reviewed and subsequently classified in terms of the following: the markets used as data sources for tests, predictive variables, predicted variables, methods or models, and performance measures used in the comparisons. In all, 54 articles were reviewed and classified, covering the specialized literature from 1991 to 2017. Based on the searches in databases of related articles, no reviews were found with such objective techniques with main path analysis on the theme proposed here.

The literature review is a method for investigating the approaches of a studied topic, as stated by Lage Junior and Godinho Filho (2010, p. 14). The following section briefly presents a review of the main machine learning techniques covered in the articles selected for this study. Subsequently, in Sec. 3, the methods used in the selection of the literature most relevant to this work are described. This concerns seeking the state of the art of a science, systemizing the information, and indicating the challenges for future studies. The objective was to use quantitative methods in the selection of the most important articles about financial market prediction via machine learning, using information from citations and years of publication. In Sec. 4, the results of the bibliometric research are presented, revealing the most important articles in the field under study. The main development path of the theme in the literature is also presented. The researched articles are systematically reviewed and classified in Sec. 5. Finally, based on the reviewed works, the best practices for applying machine learning models to financial time series predictions are outlined in Sec. 6.

2. Brief review of machine learning techniques

Machine learning techniques, which integrate artificial intelligence systems, seek to extract patterns learned from historical data – in a process known as training or learning to subsequently make predictions about new data (Xiao et al., 2013, pp. 99–100). Empirical studies using machine learning commonly have

two main phases. The first one addresses the selection of relevant variables and models for the prediction, separating a portion of the data for the training and validation of the models, thus optimizing them. The second phase applies the optimized models to the data intended for testing, thus measuring predictive performance. The basic techniques used in the literature include the following: artificial neural networks (ANNs), support vector machines (SVMs), and random forests (RFs).

In general, neural networks model biological processes (Adya and Collopy, 1998, p. 481) – specifically the human system of learning and identifying patterns (Tsaih et al., 1998, p. 162). The basic unit of these networks, the neuron, emulates the human equivalent, with dendrites for receiving input variables to emit an output value (Laboissiere et al., 2015, pp. 67–68), which can serve as input for other neurons. The layers of basic processing units of the neural networks are interconnected, attributing weights for each connection (Lahmiri, 2014a, p. 1450008-5), which are adjusted in the learning process of the network (Kumar and Thenmozhi, 2014, p. 291), in the first training phase mentioned in the previous paragraph. This phase optimizes not only the interconnections between the layers of neurons but also the parameters of the transfer functions between one layer and another, thus minimizing the errors. Finally, the last layer of the neural network is responsible for summing all the signals from the previous layer into just one output signal – the network's response to certain input data.

Whereas neural networks seek to minimize the errors of their empirical responses in the training stage, an SVM seeks to minimize the upper threshold of the error of its classifications (Huang et al., 2005, p. 2514). To do so, an SVM takes the training samples and transforms them from their original dimension space to another space, with a greater number of dimensions, in which a linear separation is approximated (Kara et al., 2011, p. 5314) by a hyperplane. This algorithm, which is commonly used to classify data based on input variables in the model, seeks to minimize the margin of the classification hyperplane during the training stage of the model. The transformation of the space of original dimensions to the space in which the classifications are performed is done with the assistance of kernel functions, from estimated parameterization in the training of the model, as detailed by Pai and Lin (2005, pp. 498–499).

Just like ANNs and SVMs, decision trees are often used in the machine learning literature, as reviewed by Barak et al. (2017, p. 91). This method involves subdivision of the data into subsets separated by the values of the input variables until the basic classification unit is obtained, in accordance with the training samples. The consensual classifications of the most accurate trees are combined into a single one, comprising the RF algorithm proposed by Breiman (2001). The combination of decision trees in the RF technique can be used in regressions or classifications, leading to good results for financial market prediction, as demonstrated by Krauss et al. (2017), Kumar et al. (2016), Ballings et al. (2015), Patel et al. (2015), and Kumar and Thenmozhi (2014).

3. Bibliometric analysis methods

Scopus – which is a database of articles and citations from periodicals, of high relevance in the scientific community – was used to survey the most relevant literature about financial market prediction using machine learning. In the system of this database, it is possible to organize a database of citations containing information about the articles, such as title, author, periodical, year of publication, and references cited. The initial bibliometric analysis of the citation database revealed the most-cited articles and the distribution of the articles over the years, as presented in the study by Seuring (2013). The frequency of scientific publications for an area of knowledge obeys Lotka's law of 1926 (Saam and Reiter, 1999, p. 135). Lotka discovered that the productivity of scientists in an area of knowledge follows a power law. Thus, the relative proportion of scientists with n publications is proportional to $1/n^2$ (Saam and Reiter, 1999, p. 137), indicating that many scientists publish little material, while a few have a large number of publications. The law can be generalized to C/n^x, in which C is a constant of proportionality and x has a value of approximately 2. Lotka was unable to explain this law, but other studies interpreted it as being applicable to an area of science, not to individual scientists (Saam and Reiter, 1999, p. 137). In this present study, Lotka's law is used to indicate the validity of the initial search in the *Scopus* database of articles. Thus, a bibliometric search in the area of financial market prediction using machine learning should be sufficiently comprehensive to obey Lotka's law.

One of the products of the bibliometric analysis is the listing of the most-influential authors in a research area. To measure this influence, the h and g indices for the individual performance of each author are used, in accordance with their published works, as performed by Liu et al. (2013). The h index incorporates citations and publications into a single number, following Egghe (2006, p. 132). This index is calculated by ordering the articles of a given author in terms of the respective number of citations and taking the highest h value of articles that have h or more citations (Egghe, 2006, p. 132). Therefore, the articles below this ordination will have no more than h citations. However, Egghe (2006) argued that the h index is insensitive not only to the rarely cited articles of a given author but also to the articles with a very large number of citations. Egghe (2006) also noted that if the number of citations of an article increases over the years, the h index remains unchanged. Thus, Egghe introduced the g index – the most-cited g articles of an author that together have at least g^2 citations (Egghe, 2006, p. 132).

One of the objectives of the bibliometric analysis performed in this study is to indicate the most important articles and journals about financial market prediction using machine learning. For this reason, the following are tabulated: the most-cited articles according to the *Scopus* database and the journals with the largest number of publications – as performed by Mariano et al. (2015, p. 38). In addition to the most-cited articles about the theme in the *Scopus* database, the articles most cited by the entire list searched are also recorded, that is, which articles are the most cited by those surveyed in the bibliometrics. Such a procedure reveals some articles that may not be part of the bibliometric survey but which are important references in the

construction of the basic knowledge about financial market prediction.

In addition to listing the most important bibliography on financial market prediction, this study intends to explain the relationship between the articles. For this reason, direct citations, bibliographic coupling, and a co-citation network are used. Citation analysis is a low-cost evaluation tool used to evaluate the acceptance of an academic article (Liu et al., 2013, p. 4). Bibliographic coupling networks – introduced by Kessler (1963) – relate to articles that use the same set of references. This involves a matrix A, whose a_{ij} element represents how many references the articles i and j have in common. Such networks have the advantage of increasing the chance of representativeness of more current articles, which do not yet have sufficient publication time to reach the level of a large number of citations. Small (1973) defined a co-citation network by the frequency with which authors are cited together. Both networks – bibliographic coupling and co-citation – are means of visualizing the structure of a scientific field. Direct citations, however, can be used to analyse the connectivity between the articles and the path followed by the knowledge, as performed by Hummon and Doreian (1989).

Although widely used, the bibliographic coupling of Kessler (1963) is static, and the similarity between the articles is defined by the bibliography of the authors (Hummon and Doreian, 1989, pp. 57–58). The patterns of the co-citations are more dynamic, changing as the field of knowledge evolves (Small, 1973, p. 265). Hummon and Doreian (1989) used a targeted network composed of the most important discoveries in DNA theory, proposing a new way of analysing the path followed by a science. Representing the network of discoveries as a targeted graph, Hummon and Doreian (1989) proposed to discover the most important path via counting methods in the links between the events. This work constructed the network of direct citations – similar to the network discovered by Hummon and Doreian (1989) – in accordance with Alg. 1, as recommended by Henrique et al. (2018). The result of this algorithm is a graph with vertices representing articles and edges representing direct citations.

Alg. 1 examines the entire list of articles surveyed in the search of the *Scopus* database, seeking the references of each article. When a reference of an article j is identified as an article k in the complete list of articles, there is a direct citation relationship. The edge of the graph originates in the cited work and terminates in the article that references it. This procedure is adopted for the flow of presumed knowledge, as performed by Hummon and Doreian (1989) and Liu et al. (2013); that is, when an article cites an earlier one, the scientific knowledge flows from the reference to the article making the citation (Liu et al., 2013, p. 4). Given that the publication of the cited article is inevitably before the one that cites it, the resulting graph of Alg. 1 is most likely acyclic.

****

Please insert Alg. 1 here.

****

The network of direct citations constructed in this work is used to calculate the paths of scientific

knowledge in the area researched, as in the work of Hummon and Doreian (1989). Thus, the *source* and *sink* – nomenclature used by Liu et al. (2013, pp. 4–5) – vertices are defined. A source vertex is one from which only edges stem; that is, that vertex is only cited by others. Thus, the source vertex does not directly cite any other vertex from the collection of articles studied – it is only cited. In turn, a sink vertex is not cited by any other vertex. It only cites vertices from the collection of articles. A sink vertex therefore has edges only in its direction, and no edge stems from it. Paths are then formed in the literature researched – from source articles and destined to the sink articles, which of course are newer. Thus, one way to survey a bibliography using quantitative methods is to study the paths via which scientific literature evolved until arriving at the current state of the art.

The method used in this work to count the weights of each path in the network of direct citations of the researched literature is known as search path count (SPC) – a method proposed by Batagelj (2003) to complement the techniques proposed by Hummon and Doreian (1989). SPC stems from the definition – according to criteria of the researcher – of the most suitable vertices as sources and sinks in a network of targeted citations. From these definitions, all the paths between the sources and the sinks are computed, recording how many times each of the edges of the graph is traversed. At the end of computing all possible paths, the edges receive weights according to the number of paths that pass through them. From these weights, the main pathways of the literature are calculated. The main path is defined as the one whose sum of the weights of the edges is the highest value among all the paths between source and sink vertices – an approach referred to as Global Search by Liu and Lu (2012). The SPC count method used in this work – with selection of the main path by the approach of that most used overall – is given by Alg. 2, adapted from Henrique et al. (2018).

****

Please insert Alg. 2 here.

****

4. Results of the bibliometric analysis

The search of the *Scopus* database was performed on 13/4/2017. Combinations of the following terms were used: stock market prediction/ forecasting, neural networks, data mining, stock price, classifiers, support vector machine, k-nearest neighbours, and random forest. The survey resulted in 1,478 documents, among which 629 were published articles and 23 were in press. To decrease the number of articles not related to the research theme, the results were restricted to the following areas: economics, econometrics and finance; business, administration, and accounting; social sciences; decision science; engineering; mathematics; and computer science. Thus, 547 articles were selected for this bibliometric work. A description of this database of articles is presented in Tab. 1.

The frequency of the publications per year is shown in Fig. 1, in dark bars, and it can be observed that there was growth, especially from 2007 onward. For comparison, a search in the *Scopus* database using the terms "credit risk" and "machine learning" – under the same conditions described in the preceding paragraph – returned 140 articles. Credit risk is a prominent area of finance research. The distribution of these publications is also shown in Fig. 1 in lighter bars. Examination of this figure suggests opportunities for future research about credit risk using new computational technologies. However, the predominance of the theme of financial market predictions compared to credit risk is clear in the number of publications, when both terms are associated with machine learning, with an annual growth rate of approximately 8.67%. Regarding this theme, the 20 most productive authors in the database of articles addressed are reported in Tab. 2, according to how many articles in this database are produced by each author, even in the case of co-authoring. The vast majority of the 1,151 authors searched (964 or approximately 84%) are authors or co-authors of only one article.

****
Please insert Tab. 1 here.
****

****
Please insert Fig. 1 here.
****

The frequency distribution of the number of publications per author among the articles researched is presented in Tab. 3. In accordance with Lotka's law, it can be observed that most authors are an author or co-author of only one article, and few are responsible for an extensive production. The frequency distribution is shown in Fig. 2 by the circles marked as distribution observed. The frequency distribution theorized by Lotka, C/n^x, is illustrated by the continuous curve in Fig. 2. The circles represent the observed distribution, with coefficient x and constant C estimated to be 2.779123 and 0.567291, respectively. The R^2 of the estimate is 0.9252587. The Kolmogorov-Smirnoff test of significance of the difference between the theoretical and observed distributions of Lotka returned a *p-value* of 0.1640792, thus indicating that there is no significant difference; that is, the bibliographic survey presented here follows Lotka's law, as described previously. Consequently, the comprehensiveness of the results of the bibliometric search is validated as a significant sample of the totality of scientific publications about financial market prediction using machine learning.

****
Please insert Tab. 2 here.
****

In accordance with the previous description, the citation performance of each author can be measured by the h and g indices. The 20 authors with the highest g indices are listed in Tab. 5. However, ordering the authors according to the h index, the authors reported in the list of Tab. 5 are added to the list of Tab. 6. Thus,

the two indices are used to evaluate the contribution and influence of the authors, as performed by Liu et al. (2013, pp. 6–7). As indicated by these authors, the indices are highly correlated, which can be observed in Tabs. 5 and 6. In these tables, a correlation can also be observed between the h and g indices with the number of articles researched in the database surveyed, but not with the number of citations. For example, Lin, C. – author with h and g values of 4 – is cited 362 times in the database surveyed, whereas Wang, J. – author with h and g values of 9 and 16, respectively – is cited 296 times in the database searched. Additionally, studying the 27 authors of these two tables, it can be observed that many of them also appear as authors with the greatest number of articles produced in Tab. 2. Only O, J.; Wang, Y.; Kim, S.; Liu, M.; Chen, T.; Fan, C.Y.; Lu, C.J.; Quek, C.; and Lin, C. do not appear as authors with the most number of articles in Tab. 2. The number of publications per country, in turn, is summarized in Tab. 4. The number of articles and the frequency in the database searched are recorded for the first 20 countries in publications. Together, these countries account for approximately 85% of the database's articles. Similarly, the number of citations per country is listed in Tab. 8. China, Taiwan, and the United States are at the top of the production of articles and citations.

****

Please insert Tab. 3 here.

****

****

Please insert Fig. 2 here.

****

****

Please insert Tab. 4 here.

****

****

Please insert Tab. 5 here.

****

****

Please insert Tab. 6 here.

****

The 20 journals with the highest number of publications in the database of articles searched are listed in Tab. 9. These journals account for approximately 44% of the articles surveyed and therefore are important sources of research in financial market prediction using machine learning. It is worth highlighting that the journal *Expert Systems with Applications* accounts for 13% of the total publications, which indicates not only the number of useful references for the area published in this journal but also that it is a potential target journal for future studies. Listed in Tab. 9, the journals *Neurocomputing*, *Applied Soft Computing Journal*, *Decision Support Systems*, *Neural Computing and Applications*, *Neural Network World*, and *Journal of Forecasting* together account for 17% of the articles surveyed, thus being alternatives for submissions of new studies. For

searches in this area, the 20 most commonly used keywords in the database searched are listed in Tab. 10. Almost 90% of the articles searched use one or more of these keywords. The systems for search and analysis of keywords distinguish terms in singular and plural form, considering them distinct for the purposes of searches.

****
Please insert Tab. 7 here.
****

****
Please insert Tab. 8 here.
****

Upon recording the previous bibliographical statistics, it is then necessary to survey the most relevant literature on financial market prediction using machine learning. An important relationship is that of the most-cited articles – the top 20 are listed in Tab. 11. It is emphasized that the number of citations from each article is related to the entire *Scopus* database of articles. It is also worth listing the references most cited by the articles analysed in the bibliometric research. Such references do not necessarily appear in the database of *Scopus* articles, but they are important sources for the area of financial market prediction. These most-cited articles are listed in Tab. 12. Also listed for review are the 10 most recent articles among those surveyed in the initial bibliometric search, all of which were published in 2017. These articles are listed in Tab. 13.

****
Please insert Tab. 9 here.
****

****
Please insert Tab. 10 here.
****

****
Please insert Tab. 11 here.
****

****
Please insert Tab. 12 here.
****

****
Please insert Tab. 13 here.
****

The network resulting from the bibliographic coupling of Kessler (1963) is shown in Fig. 3 only for the 20 articles with the highest number of relationships between each other. Each node in the figure represents an article, and the arcs, or links, are the relationships between them. As described in Sec. 3, a relationship

represents similarities in the references of the articles. Thus, the articles in Fig. 3 have similar references regarding financial market prediction using machine learning. The review of these articles is therefore a means of summarizing the evolution of the field researched, considering the literature researched by the authors. The data from the articles of Fig. 3 are explained in Tab. 14. This table has more recent articles than the most-cited articles listed in Tab. 11 and 12. The co-citation network for this bibliographic survey – in the model proposed by Small (1973) – is illustrated by Fig. 4 for the 10 greatest relationships. The co-citation network of this figure reveals the articles listed in Tab. 15, also selected for review.

According to the previous description, a network of direct citations is constructed between the 547 articles of the bibliometric research. Alg. 1, which returns an acyclic graph, is used. Isolated vertices are removed; that is, those that do not have direct citations to the others or are not cited. Such vertices will not be considered in the main path calculations because they do not relate to any other. With the network of direct citations at hand, the source vertices and the sink vertices must be selected, in accordance with Liu et al. (2013). Such a choice may be subjective, at the discretion of the researcher of the network; however, this work opted to test all the paths between all possible sources and all possible sinks. The main path selected was the one with the highest sum of weights, as described earlier. Thus, in the network of direct citations among the 547 articles constructed by Alg. 1, 60 articles were identified that are only cited, not referencing any other article from the database. Such articles constitute the sources, from which only edges originate.

****
Please insert Fig. 3 here.
****

****
Please insert Tab. 14 here.
****

****
Please insert Fig. 4 here.
****

****
Please insert Tab. 15 here.
****

The pairing combination then occurs for all 60 source vertices and all 165 sink vertices, calculating all the possible paths between each one. According to Alg. 2, each edge receives a weight according to how many paths pass through it. The main path of the literature – in accordance with the method of Alg. 2 – is shown in Fig. 5, calculated with a total weight of 4686. The articles that constitute the vertices are listed in Tab. 16. It can be observed that half of the articles of the main path calculated were published by the journal *Expert Systems with Applications*, which is consistent with the results presented in Tab. 9.

****
Please insert Fig. 5 here.
****

****
Please insert Tab. 16 here.
****

5. Review of the selected literature

Brief comments on the literature selected via the quantitative methods described in Secs. 3 and 4 follow. Commented upon are the most-cited articles – in accordance with the *Scopus* database – listed in Tab. 11, in addition to those articles most cited by the compiled database of the 547 articles in this work, listed in Tab. 12. The most recent articles selected for the review are shown in Tab. 13. Also selected and reviewed are the articles with the greatest bibliographic coupling – see Tab. 14. Aiming to address the evolution of the state of the art of predictive models of machine learning applied to the financial market, the articles that are part of the main path of the literature are described – see Tab. 16. Finally, the articles reviewed were then classified according to markets, assets, and methods and variables, seeking to highlight some of the main characteristics of the literature.

5.1. Most-cited articles

Among the articles most cited by the bibliographical survey of Sec. 4, the classic work of Malkiel and Fama (1970) deserves attention, given that it established the EMH. According to this theory, financial markets immediately adjust to the information available, and it is impossible to predict their movements. The weak form of the EMH considers available information to be only the past prices of the asset (Malkiel and Fama, 1970, p. 388). By adding other publicly available information, such as annual reports and the issuing of new shares, Malkiel and Fama (1970) addressed the semi-strong form of the EMH. Finally, the strong form of the EMH corresponds to when there is internal information monopolized by some investors. The theory proposed by Malkiel and Fama (1970) is critical for the prediction of financial markets because the construction of consistently profitable systems may mean the existence of evidence contrary to the EMH (Timmermann and Granger, 2004, p. 16).

Also present in Tab. 12, the articles of Engle (1982) and Bollerslev (1986) introduced important econometric models used in financial market prediction. Engle (1982) modelled a time series using a process called autoregressive conditional heteroskedasticity (ARCH). In this model, the conditional variance present depends on the terms of previous errors, keeping the unconditional variance constant. Bollerslev (1986), in turn, generalized the ARCH model, considering its own variance to be an autoregressive process, introducing the generalized autoregressive conditional heteroskedasticity (GARCH) model. Although widely applied in the prediction of time series, ARCH and GARCH assume a linear process of generating the values

of the time series (Cavalcante et al., 2016, p. 197). However, markets are characterized by nonlinearities, interacting with political and economic conditions and the expectations of their operators (Göçken et al., 2016, p. 320), making GARCH assumptions inadequate for many financial time series applications (Lahmiri and Boukadoum, 2015, p. 1550001-2). Thus, other methods are used, such as the one proposed by Elman (1990), which introduced a prediction network with precursor memory of some models of neural networks. Campbell (1987), in turn, sought to document variables that predict stock returns in two distinct periods. However, Campbell (1987, p. 393) concluded that no simple model can anticipate all the variations in return on stock prices.

The most-cited article in the *Scopus* database among those listed in Sec. 4 is that by Kim (2003). As recorded in Tab. 1, this article relies on the total of 546 citations from others in the *Scopus* database and is referenced 39 times on average per year. Kim (2003) addressed the application of SVMs to classify the daily direction of the Korean stock market index (KOSPI), using technical analysis (TA) indicators as predictive variables. The results were compared to those obtained with neural networks and case-based reasoning (CBR), and the support vector machine (SVM) achieved better performance, as measured by the accuracy of the predictions. However, the work of Kim (2003) may be used as a reference with care, for its results contain data-snooping bias: the author selects the best SVM parameters based on test accuracy. Machine learning models should choose the best parameters based on training the model using historical data, given that test data are not known a *priori*. That is precisely what the model seeks to predict. Following the same line of reasoning as Kim (2003), the work of Huang et al. (2005) – present in Tab. 12 – also uses an SVM to classify the direction of the market, comparing its performance with linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and the Elman backpropagation neural network (EBNN). Those authors, however, make use of training data for models' parameters selection. The results indicated greater accuracy in predictions using the SVM alone and combined with the other methods. The tests were performed on the Japanese market index (NIKKEI 225) using weekly quotes. On the other hand, among the articles listed Tab. 12, Pai and Lin (2005) also used an SVM as a prediction method, not for the direction of prices, like Kim (2003) and Huang et al. (2005), but to predict stock values. Additionally, Pai and Lin (2005) combined an SVM with autoregressive integrated moving average (ARIMA) in a hybrid system capable of fewer errors than those obtained using the models separately.

The SVM classification model can be adapted as a regression to predict values in financial time series – in this case it is known as Support Vector Regression (SVR). This model has been used, for example, in the work of Huang and Tsai (2009). The authors combined SVR with a Self-Organizing Feature Map (SOFM) in two stages to predict the value of an index of the Taiwan market (FITX) in daily quotes. The SOFM is a method for spatial mapping of the training samples according to their similarities (Huang and Tsai, 2009, p. 1531). The inputs are indicators of TA, grouped by the SOFM according to their similarities in clusters, which are in turn fed into SVR models. The results achieved by the hybrid model of Huang and Tsai (2009)

are superior to those obtained with only the use of SVR as a predictive model. In their model, Yu et al. (2009) – present in Tab. 12 – used a variation of the SVM known as Least Squares SVM (LSSVM), which has a lower computational cost than the original SVM and good generalization capacity (Yu et al., 2009, p. 88). The authors proposed an evolutionary LSSVM with the use of a Genetic Algorithm (GA) – a process of selecting values optimized for each generation (Yu et al., 2009, p. 88). Thus, in the proposal of Yu et al. (2009), a GA is used at the time of the selection of variables, among the TA values and fundamentalist indices, and in the optimal parameterization of the LSSVM. The results obtained are superior to conventional SVM models, ARIMA, LDA, and neural networks of the Backpropagation Neural Network (BPNN) type. It is worth noting Yu et al. (2009) apply McNemar's tests to formally compare models' performance. Not all works go beyond looking at hit rates. [1]

As observed in Tabs. 11 and 12, the majority of the most-cited articles in the market prediction literature researched apply variations of neural networks. However, another algorithm that is quite present in the texts on prediction is the SVM, which was used by Kim (2003) – the most-cited article from Tab. 11. The work of Kara et al. (2011) – present in Tab. 11 – compares the two basic models of SVMs and ANNs regarding their predictive capabilities for the daily direction in the Turkish market. In addition to using the index values from an emerging market as data, Kara et al. (2011) are also notable because they consider all 10 years of daily prices in the parameterization of the models, ensuring they remain as general as possible. Samples from all 10 years are also considered when training the models. The authors conclude on the predictive superiority of ANNs under their proposed conditions. However, Kara et al. (2011) made predictions for the current direction of daily prices at market close, using TA indicators calculated by those same closing prices as inputs to the models. Obviously, closing prices are available only after the market closes, and the market's direction is already defined. Therefore, predicting the current prices' direction upon market closure using machine learning and those current closing prices, in the form of TA values, as inputs may not make sense for practical use, but this approach enables straightforward model comparisons.

Among the articles compiled in Tab. 12, the earliest reference that effectively addresses the prediction of stock prices is the work of Yoon et al. (1993). The authors applied neural networks to the prediction of the financial performance of stocks relative to the market, and they based their study on the good predictive performances reported in previous works (Yoon et al., 1993, p. 51), demonstrating that neural networks achieve better results than those obtained by discriminant analysis. Neural networks – a model based on the human nervous system (Göçken et al., 2016, p. 322) – have been widely used as prediction methods (Göçken et al., 2016, p. 320). Among the most-cited references about this subject are the works of Hornik et al. (1989) and Hornik (1991), present in Tab. 12. Laying the bases for general applications, such articles rigorously demonstrate the approximation abilities of neural networks for mathematical functions with a certain

[1] McNemar's tests are built pairing the distributions given by the results from each model, applying a χ^2 test on the null hypothesis that the models have the same performance. For more details, refer to Dietterich (1998).

accuracy. This approximation ability of the neural networks is explored in the article by Abu-Mostafa and Atiya (1996), which provides an initial approach regarding financial market predictions before proposing the system based on neural networks and hints. These hints are a learning process joining training data and previous knowledge (Abu-Mostafa and Atiya, 1996, p. 209), such as a known property of any financial asset. However Abu-Mostafa and Atiya (1996) are ultimately vague on formal hints definition and validation.

Neural networks continued to be explored in the articles listed in Tabs. 11 and 12. Of particular note is the review presented by Adya and Collopy (1998), which aimed to summarize the criteria for evaluating predictive work on neural networks. Among the criteria suggested by the authors were validation in test data (out-of-sample) and generalization capacity and stability of the proposed model. Another review study regarding the use of neural networks in financial market predictions – listed in Tab. 12 – is the article by Zhang et al. (1998). The authors presented ANNs as predictive models and commented on previous results from the literature, concluding with the adaptation of neural networks to predictions in the financial market due to their adaptability and ability to handle nonlinearities present in time series (Zhang et al., 1998, p. 55), among other factors.

As observed from the articles listed in Tabs. 11 and 12, the literature about financial market prediction involves the use of many models based on neural networks. For example, the work of Kamstra and Donaldson (1996) used ANNs to combine predictions about the indices of developed markets; for example, the S&P500, NIKKEI, TSEC, and FTSE. The American S&P 500 index was also used to test a hybrid predictive model proposed by Tsaih et al. (1998). These authors constructed their model from variables of the TA and rules given by experts and scholars, using them as inputs for the ANNs in the prediction of the direction of the S&P500. They rely on cases termed "obvious" and "non-obvious", but with neat definitions for each one. The direction of returns was also the dependent variable sought by Fernandez-Rodrıguez et al. (2000) – see Tab. 11. The authors applied ANNs to Madrid's market index, using the returns from nine days earlier as independent variables. The results showed the superiority of the neural networks over buy-and-hold strategies for almost all of the tested periods (Fernandez-Rodrıguez et al., 2000, p. 93). Buy-and-hold means the acquisition and holding of an asset for a given period of time (Chiang et al., 2016, p. 201), thus exposing it to the variations in its market price. A distinguishing characteristic of Fernandez-Rodrıguez et al. (2000) is the use of lagged returns as inputs to the neural networks instead of the conventional TA indicators. The authors, however, do not provide comparisons between both approaches.

ANNs were also used in comparison with other models in the work of Leung et al. (2000) – an important article present in both Tabs. 11 and 12. The authors contrasted predictions of values and monthly direction of the S&P500, FTSE, and NIKKEI indices using neural networks, LDA, and regressions. Among the input variables used in the models are interest rates, indices of industrial production and consumer prices, and previous returns. The main conclusion of the study is that the models for direction prediction have superior performance to the models for predicting value (Leung et al., 2000, p. 188), measured not only by the

successful predictions but also by the return obtained in operations strategies. Chen et al. (2003) – another important study present in Tabs. 11 and 12 – applied ANNs in the prediction of returns. The authors conducted their predictions in the emerging market of Taiwan, using for this purpose a Probabilistic Neural Network (PNN). Such networks use Bayesian probability, deal better with the effects of outliers, and are faster in the learning process (Chen et al., 2003, p. 906). Kim and Han (2000), in turn, addressed the reduction in dimensionality of variables to obtain the input variables of the ANNs through a GA. This algorithm is used to discretize the continuous values of the TA indicators and optimize the weights in the network connections, which leads to better predictive performance. Leigh et al. (2002) – present in Tab. 11 – also discussed the use of GAs for optimizing neural networks, comparing their predictive performance with a visual pattern of the TA called "bull flag". Although that pattern finds definition in the text of Leigh et al. (2002), results may vary as how other researchers build their template for a pattern recognition tool. Such is a big challenge of TA visual patterns recognition for they are generally based on loose or ambiguous definitions.

Thawornwong and Enke (2004) and Enke and Thawornwong (2005) – presented in Tab. 12 – investigated how the predictions of neural networks vary according to the selection of the input variables. To do so, they proposed a measure of the relevance of each input variable according to how much information is added to the model with their use. The authors concluded that the models that dynamically select variables with the period are more profitable and less risky (Thawornwong and Enke, 2004, pp. 226–227). Armano et al. (2005), in turn, proposed the use of a layer – prior to the neural network predictors – responsible for the selection of the predictors, taking indicators of TA as inputs. Only the best predictors are selected by means of a GA. Despite using a limited number of TA indicators, results outperform buy-and-hold strategies, even considering trading costs, which is a differential feature of Armano et al. (2005). Hassan et al. (2007) also used a GA; however, it was to optimize the parameters of a hidden Markov model (HMM), using an ANN to transform the input variables of the general model. The HMM is based on transition matrices and probabilities used in general predictions, such as DNA sequencing and voice recognition (Hassan et al., 2007, p. 171). The combined use of an ANN, a GA, and an HMM proposed by Hassan et al. (2007) was tested on the next day's closing price forecasting. However, the proposed model has only been applied to three stocks in the computing area, with the reported results being very close to an ARIMA model. The historical data was also very limited, revealing roughly 2 years of daily prices.

A good review of studies involving neural networks in financial market prediction is the article by Atsalakis and Valavanis (2009) – see Tab. 12. It is worth noting that this article is also one of the most cited among the articles compiled in the bibliographical survey of Sec. 4, listed in Tab. 11. In addition, the article by Atsalakis and Valavanis (2009) is in the list of the articles with the most bibliographic coupling – as indicated in Tab. 14 – and is part of the main path of the literature about financial market prediction, in accordance with the method presented in Sec. 3. This important work analysed 100 articles, classifying them

regarding markets analysed, input variables of each model, sample size, and performance and comparative measurements. Among the findings of the review, of note are the following: the average use of 4 to 10 variables; an estimated total of 30% of the articles apply indicators on closing prices; and 20% use TA indicators as input variables (Atsalakis and Valavanis, 2009, pp. 5933 –5936). However, most of the articles resort to a combination of technical and fundamentalist indicators as inputs for their models (Atsalakis and Valavanis, 2009, p. 5936).

In addition to ANNs and SVMs, the most-cited articles among those surveyed in Sec. 4 also address other prediction methods. Chiu (1994) – listed in Tab. 12 – used fuzzy logic for grouping data and classification. The purpose of grouping into clusters is to create larger groups, thus representing the behaviour of the system (Chiu, 1994, p. 267). The article by Chiu (1994) did not address classification in the financial market, but the method can be used in the prediction of the direction of stock prices and indices. Fuzzy logic was also used by Wang (2002) and Wang (2003) in the construction of predictor systems for Taiwan's market. Another method for predicting stock values is the combination of representation algorithms and textual relevance of news to machine learning, as in Schumaker and Chen (2009, p. 20). These authors reported prediction results superior to those obtained by the SVM and ARIMA hybrid used by Pai and Lin (2005).

5.2. *Articles with the greatest bibliographic coupling*

As discussed earlier, the method of bibliographic survey by coupling enables the listing of more recent literature, even without a large number of citations. The articles with greater bibliographic coupling – as described in Sec. 3 – are listed in Tab. 14 and can be observed in Fig. 3. Some of these works have already been commented in Sec. 5.1, which addresses the most cited articles within the bibliographical survey performed: the article by Atsalakis and Valavanis (2009), which is a review of works about predictions with neural networks, and the articles by Huang and Tsai (2009), Yu et al. (2009), and Kara et al. (2011), which effectively applied neural networks and SVMs in predictions in the financial market. This current section is dedicated to the other articles listed in Tab. 14.

As noted in the articles studied in Sec. 5.1, neural networks are mainly applied to the prediction of prices and movements in the financial market. For example, Thawornwong et al. (2003) used only neural network algorithms to predict the direction of three American stocks in daily quotes, using the indicators from the TA as inputs. The results indicated that TA may be inconsistent in the prediction of short-term trends and that the use of neural networks and their indicators may increase predictive performance (Thawornwong et al., 2003, p. 323). Additionally, TA and neural networks result in better strategies than the simple buy-and-hold (Thawornwong et al., 2003, pp. 320–321), which is evaluated not only by accuracy of predictions but by profitability in simulations. As another example, the work of Rodríguez-González et al. (2011) applied neural networks to the relative strength indicator (RSI), achieving better predictive performance than the simple use of such an indicator. In turn, exploring the predictive effects of neural networks in developing markets,

Cao et al. (2005) applied univariate and multivariate ANNs to Chinese stocks, measuring performance through the mean absolute percentage error (MAPE). The authors concluded that the neural networks surpassed the predictive performance of linear models, such as the capital asset pricing model (CAPM). This model assumes that the return of an asset is a linear function of its risk in relation to the market (Cao et al., 2005, p. 2501). In their conclusions, the authors pointed out that neural networks may be more effective in predictions of developing markets' prices.

Neural networks have also been applied to hybrid models. Wang et al. (2012) combined predictions using the Exponential Smoothing Model (ESM) and ARIMA, which are capable of capturing linear characteristics of time series, with BPNN – a neural network for addressing the non-linear characteristics of these series. Using for tests the monthly closing prices of the Chinese SZII index, and the monthly opening prices of the American DJIA index, the authors concluded that the predictive performance of the hybrid model is superior to that obtained using the models individually. Wang et al. (2012) could, however, explicitly state which variables are used as input and briefly justify choosing opening prices for an index and closing prices for the other. Kumar and Thenmozhi (2014) also worked with hybrid models in a developing market, combining ARIMA, ANN, SVM, and RF to predict the daily returns of an index of the Indian market. The authors argue that financial time series are not absolutely linear or non-linear (Kumar and Thenmozhi, 2014, p. 288), which justifies the combination of the two types of predictive models. The article indicates superiority in prediction and profitability with the use of the ARIMA and SVM hybrid. Other than traditional error measures, like mean absolute error and root mean square error, the hybrid models are also compared in terms of return, volatility, maximum drawdown and percentage of winning up and down periods, which could be incorporated as performance measurements in future articles.

The article by Tsai and Hsiao (2010) uses methods of selecting variables before processing the data by ANN to predict the direction of prices. Principal Component Analysis (PCA), GA, decision trees and their combinations are used, demonstrating that the previous selection of variables increases the predictive performance of neural networks. PCA is a multivariate statistical method that extracts a reduced number of factors, or components, of highly correlated elements, from the original variables (Tsai and Hsiao, 2010, p. 260). This method is presented in variants, some of which are examined by Zhong and Enke (2017) in the selection of variables before applying them to an ANN. Although it is concluded that pre-processing in the selection of variables increases the predictive performance of neural networks, the work of Zhong and Enke (2017, p. 137) indicates traditional PCA as a simpler and more efficient method than its variants in use combined with ANN. Both Tsai and Hsiao (2010) and Zhong and Enke (2017) make respective results more robust by using t-tests to assure differences in performance measures from each model is statistically significant. Similarly, Chiang et al. (2016) selected variables based on the information gain inherent to each one, before applying them to the ANN. In their approach, Chiang et al. (2016) apply variable selection in a set smaller than Tsai and Hsiao (2010) and Zhong and Enke (2017), but increment results by smoothing data

with wavelet transformations.

This paragraph was improved. Pre-processing data with wavelets prior to the application of neural networks was also explored by Li and Kuo (2008), who used a technique – known as the Discrete Wavelet Transform (DWT) – that involves decomposition of digital signals into their components. The components are then processed by a special class of neural network known as the Self-Organizing Map (SOM) to generate short- and long-term purchase and sale signals. The work done by Li and Kuo (2008) is remarkable for its pattern labeling scheme of long- and short-term trading signals as well as a defined, reproducible and profitable trading strategy. Chang and Fan (2008) also used decomposition into wavelets as data pre-processing, but they grouped them by homogeneous characteristics into clusters that are mapped into fuzzy logic rules. Subsequently, Chang and Fan (2008) applied a system proposed for the interpretation of these rules in the generation of predictions, using also the k-Nearest Neighbours (kNN) algorithm to reduce errors. The authors highlighted results superior to other models such as BPNN, although they used a small 2 years dataset of daily prices from a Taiwanese index. Transformations involving wavelets have also been used to mitigate short-term noise in index prices; for example, in Chiang et al. (2016). The authors showed larger returns when the data are smoothed through transformations with wavelets before applying them to neural networks (Chiang et al., 2016, p. 205), as stated before.

Widely applied to time series, the fuzzy logic theory was developed in human linguistic terms (Chen et al., 2014, p. 330). For example, the work of Chen et al. (2014) uses fuzzy logic in time series, aiming to overcome limitations regarding linearity assumptions in other models, such as ARIMA and GARCH. In a previous study, Ang and Quek (2006) combined neural networks with fuzzy logic to obtain daily predictions of stock prices better than other traditional neural network models. The system proposed by Ang and Quek (2006) also provides interpretability of rules – a property that is often absent in traditional ANN and SVM approaches (Al Nasseri et al., 2015; Yu et al., 2009).

As noted in the articles listed in Sec. 5.1, many works compare predictions obtained with various methods. In this context, the article by Ballings et al. (2015) – listed in Tab. 14 – compares the combination of predictions of multiple classifiers, among them ANN, SVM, kNN, and Random Forest (RF). The results indicated better accuracy in the prediction of the direction of stock prices in a year using RF (Ballings et al., 2015, p. 7051). Distinguishing features of Ballings et al. (2015) are the wide list of classifiers selected by the authors and their numerous input variables, drawn from fundamentals of 5.767 European companies. Therefore, the authors are limited to yearly predictions, for that is a very low frequency type of data, typically released monthly or yearly. Gorenc Novak and Velušček (2016, p. 793) worked with the prediction of the stock price direction for the following day, but applying only the SVM. The work of these authors stands out due to using the daily maximums of the assets, as opposed to the traditional closing price. Gorenc Novak and Velušček (2016, p. 793) observed that the volatility of the maximums is less than that of the prices at the end of the negotiation session, at closing. Therefore, the maximums would be easier to predict, as stated by the

authors.

Finally, the articles (in Tab. 14) that address financial market predictions using textual analysis deserve to be highlighted. Hájek et al. (2013) analysed sentiment about the annual reports of companies, processing terms that exert positive or negative influence on asset prices. Corporate reports are tools for communication with investors, which contain terms loaded with qualitative data (Hájek et al., 2013, p. 294). The authors processed these terms through previously constructed dictionaries, and the resulting categorizations serve as inputs for both neural networks and SVR models. The method proposed by Hájek et al. (2013) proved to be capable of predicting returns for one year in advance of the data used in the tests. Al Nasseri et al. (2015) also analysed sentiments, but in publications of a blog specializing in stock markets. The authors concluded that variations in the terms used in the texts of the blog predict trends of the American DJIA index.

5.3. *Articles with greater co-citation relationships*

According to Small (1973), a co-citation occurs when two articles are quoted by a third party in the same work. The more frequent these joint citations, the stronger the relationship between the two articles. Sec. 4 lists the 10 articles with the greatest co-citation relationships in Tab. 15. Some of these articles were reviewed earlier, in Sec. 5.1, and therefore are not commented on in this section. The articles are as follows: Atsalakis and Valavanis (2009), Chen et al. (2003), Hornik et al. (1989), Huang et al. (2005), Kim and Han (2000), and Zhang et al. (1998). Thus, the next few paragraphs briefly comment only on the other articles of greater co-citation frequency from Tab. 15.

The study by Kimoto et al. (1990) used combined neural networks to form a single prediction of weekly buying or selling for a Japanese stock market index. It applied the basic ANN model to six economic indicators as input variables, with more lucrative results than the basic buy-and-hold strategy. The major innovation by Kimoto et al. (1990) is the notion that prediction rules must change with time, according to market conditions. Therefore, the neural networks are re-trained as time passes, in a fixed moving window fashion. A prediction period is defined and as new data arrives, the training period if forward shifted. Chang et al. (2009) also worked with the classic neural network model; however, they increased the returns obtained in simulations combining the predictions of the networks with CBR. The authors also used a stock selection model based on financial health indicators of the respective companies. The returns obtained with the combined ANN and CBR model of Chang et al. (2009) were greater than the individual returns of each model for the nine stocks selected in that work. In that paper, the innovation rests on previously selecting potentially profitable stocks and reusing only price pattern cases in which the system had previous success.

Contrary to the models used by Kimoto et al. (1990) and Chang et al. (2009), the article by Tay and Cao (2001) obtained superior predictions with the use of SVM. The authors compared their results to those obtained by neural networks, and they concluded that the best performance of the SVM is due to: the minimization of the structural risk, to a fewer number of parameters to be optimized by the SVM; and

the possibility of the neural networks converging for local solutions (Tay and Cao, 2001, p. 316). Finally, also present in Tab. 15, the book of Box et al. (2015) is an ample introduction to the subject of prediction of time series for general applications that is not restricted to only financial series. The book covers linear techniques, correlations, moving averages, and autoregressive models. Being a general compendium of time series theory, Box et al. (2015) did not specifically address machine learning techniques, but even so, it is still listed as one of the works with the highest number of co-citations among the references surveyed in this present study.

5.4. *Main path*

Below is a brief review of the main path of the literature on financial market prediction using machine learning. As stated earlier, it is a chronological survey of the main articles published on the subject, which are described in detail in Sec. 3. The main path of the literature addressed in this work – illustrated in Fig. 5 – suggests the most important works for the review of methods, experiments, findings, and scientific conclusions regarding the proposed theme. Thus, this section is dedicated to detailing the main aspects of the articles listed in Tab. 16, exploring the state of the art of this literature.

The main path of Fig. 5 begins with the article by Kamstra and Donaldson (1996). This article is also listed in Tab. 11 and has already been commented on in Sec. 5.1. It addresses the use of neural networks in the prediction of daily volatility of the S&P500, NIKKEI, TSEC, and FTSE indices, compared with the popular linear model GARCH, based on out-of-sample test data. As the ANN is a collection of non-linear transfers that relate the output variables to the inputs (Kamstra and Donaldson, 1996, p. 51), it is a more suitable proposal for potentially non-linear data. In fact, the empirical tests of Kamstra and Donaldson (1996) indicated that the predictions of the volatility of market indices using GARCH have deviations from the actual values that are higher than those obtained with ANN. The results are confirmed by the second article of the main path, which is by the same authors as the first article. Thus, Donaldson and Kamstra (1999) concluded that combination of predictions with ANNs can provide significant improvements when compared to the linear combinations approach. It is important to notice Donaldson and Kamstra (1999) aim to predict returns, differently from Kamstra and Donaldson (1996), which work to predict volatility. Moreover, both papers confront linear and non-linear models, although some authors seem to agree that non-linear models are more suitable for financial time series predictions.

The authors of the first two articles of the main path did not yet consider the SVM classification model, which the initial publication is credited to Vapnik (1995). Approximately five years after Donaldson and Kamstra (1999), the work of Pai and Lin (2005) considered the combination of SVM with a linear model in the prediction of stock prices. At the time, one of the linear models most commonly used in predictions was the ARIMA (Pai and Lin, 2005, p. 498), which has limitations for capturing non-linear characteristics of time series. Thus, Pai and Lin (2005) stand out due to combining this model with an SVM, which is based

on minimizing structural risk through limitations in the error thresholds. Despite the limited amount of data used by the authors – just over a year of daily closing data (Pai and Lin, 2005, pp. 499–500), the work concluded that the proposed hybrid model can overcome the individual use of its components but suggests optimization of parameters to achieve the best results. As a very well written paper and a great reference on early models hybridization, Pai and Lin (2005) could state more clearly the input variables to their system. It should be noted that the manuscript by Pai and Lin (2005) is also one of the most-cited articles according to the survey of Sec. 4 – see Tabs. 11 and 12.

Reviewing 100 articles about predictive models in the financial market using computational techniques, Atsalakis and Valavanis (2009) provided a classification of the studies regarding markets, variables, prediction methods, and performance measures. The importance of this article for the main path is demonstrated by its presence in all the results of the bibliographical survey of Sec. 4, listed, therefore, in Tabs. 11, 12, and 14. In general, the studies used four to ten predictive variables (Atsalakis and Valavanis, 2009, pp. 5933–5936), with the most common being the opening and closing prices of market indices. Also according to the authors, 30% of the proposed models in the articles analysed use closing prices, and 20% use TA variables, most of which combine them with statistical and fundamentalist data. Atsalakis and Valavanis (2009) highlighted ANN and SVM, with data pre-processing using normalization or PCA, among other methods. The most common performance measures used by the authors are also listed; for example, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), profitability, and annual return. The authors concluded that neural and neuro-fuzzy networks are suitable predictive algorithms for the stock market, but there is still no definition for the structures of such networks – they are determined by trial and error (Atsalakis and Valavanis, 2009, p. 5938).

A few years after the review of Atsalakis and Valavanis (2009), the main path continued with the work of Kara et al. (2011) – an article from the most-cited group, as observed in Sec. 5.1, and listed as having the greatest bibliographic coupling, as addressed in Sec. 5.2. Kara et al. (2011) is an excellent reference for comparing predictions using ANN and SVM in their basic forms. To compare the models, Kara et al. (2011) used 10 years of daily market prices from the Istanbul index, practically balanced on days of high days and lows in the prices. It is worth highlighting the procedure for selecting the samples for the sets of parameters, training, and tests of the authors, which ensures the presence of samples of each year in each set. Ten indicators were calculated from TA – they were pre-processed only with normalization of values before their use in the models. ANN had slightly better performance than the SVM, with statistical significance calculated via *t-tests*. Following the same methods of Kara et al. (2011), Patel et al. (2015) included in the comparisons – in addition to ANN and SVM – the RF and Naïve Bayes (NB) models. These authors also used TA indicators as predictive variables, but they innovated by considering – in addition to the continuous values of the indicators – the price trend indicated by each variable. The overall result indicated that this approach of the indicators – referred to as discrete – improves the predictions (Patel et al., 2015, p. 268).

In an article contemporaneous to that of Patel et al. (2015), Laboissiere et al. (2015) sought the prediction of stock prices in the Brazilian market using the basic ANN form. These authors differed from previous approaches by focusing on a specific sector of the market – electric energy – and using as input variables not only the prices but also the index of the São Paulo stock exchange (BOVESPA), a specific index for the electric energy market (IEE), and the US dollar. In addition, Laboissiere et al. (2015) focused on the prediction of daily maximums and minimums, seeking to define thresholds for operations with the stocks studied. The authors opted for pre-processing of the prices with the Weighted Moving Average (WMA), filtering noisy fluctuations and highlighting trends (Laboissiere et al., 2015, p. 68). Additionally, they also used correlation analysis between stock prices and the indices to select the most important ones as inputs for the ANN model. Finally, it is important to record that the articles of Patel et al. (2015) and Laboissiere et al. (2015) delimit the most recent publications of the main path of the literature. Thus, such articles and their successors in the main path of the literature do not have sufficient publication time to reach the number of citations of the previous articles, present in Tabs. 11 and 12.

After the article by Laboissiere et al. (2015), the main path of the literature about financial market prediction continued with an example of pattern recognition. This study by Chen and Chen (2016) specified an algorithm for operations based on a visual signal of a high in the prices of a stock or index. As a means of reducing dimensionality, the authors used a method of weighting the most important points of a time series (Chen and Chen, 2016, p. 262), based on a visual pattern used in the TA known as a bull-flag. To avoid incurring subjectivities, a specific definition of the bull-flag pattern is given and parameterized (Chen and Chen, 2016, p. 264). Based on the pattern sought, a model was calculated dynamically, seeking to assess how well the pattern adapted to the daily prices of the NASDAQ and TAIEX indices. Thus, the pattern was recognized computationally, and an operation initiated, varying as a parameter the time until its liquidity. Chen and Chen (2016) evaluated their algorithm through the return generated, comparing it with more advanced models, such as Genetic Algorithms (GAs).

At the end of the main path of the literature is the article by Zhong and Enke (2017), which was already commented on in Sec. 5.2. Zhong and Enke (2017) sought to predict the daily direction of a fund based on the American index S&P500 using the basic ANN algorithm but differing in the selection of the predictive variables. For this purpose, the authors used PCA and its variations to select the most significant variables for predictions among economic indicators, such as American Treasury rates, foreign exchange, indices of international markets, and returns for companies with large capitalization. The performance of each method of variable selection – known as dimensionality reduction – allied to ANN was measured by the MSE, confusion matrices, and profitability in a simple strategy of following purchase signals and investing in American Treasury securities in the case of selling signals. Zhong and Enke (2017) concluded that there is no statistically significant difference in performances using the different variations of PCA. However, the profitability measured is slightly higher using the traditional PCA for dimensionality reduction (Zhong and

Enke, 2017, p. 135).

Finally, the article by Chen et al. (2017) represents the sink node in the main path of the literature shown in Fig. 5; that is, where the state of the art of financial market prediction applying machine learning converges. Chen et al. (2017, p. 341) argued that most of the previous research considers that the predictive variables make equal contributions to the value obtained by the predictive market model. Chen et al. (2017) innovated by applying a measure of information gain in the weighting of the variables, taken from indicators of the TA. The variables subsequently weighted are used as inputs for the SVM and kNN models, comparing the results to the use of the classifiers without the weighting of variables. Subsequently, a hybrid model – combining SVM and kNN – was proposed in a manner similar to that of Nayak et al. (2015). However, the model described by Chen et al. (2017) considered the weighting of the variables, obtaining better performance measures than those reported by Nayak et al. (2015).

As observed in the preceding paragraphs, the initial studies of the main literature about financial market prediction using machine learning contrast linear models of prediction with non-linear ones. The literature seems to agree that the former are surpassed by the latter and are therefore only used as a benchmark in the most recent works. Thus, the main predictive models explored in the main path of Fig. 5 are ANNs and SVMs. The applications of both prediction approaches have variations in pre-processing, the selection of variables, and, more recently, the hybridization of the models. Most of the studies – especially the most recent ones of the main path – present a comparison of models and combinations between them, thus representing the state-of-the-art aspects of the theme.

5.5. *Most recent articles*

The following paragraphs are dedicated to reviewing the most recent publications among the articles researched in the bibliometrics of this study. The 10 most recent articles are listed in Tab. 12. These articles summarize the most modern approaches in the use of machine learning for financial market prediction. For example, Weng et al. (2017) used algorithms that have been widely studied in previous literature; however, they used input data originating from public sources of knowledge via the Internet. Specifically, Weng et al. (2017) derived news indicators from *Google News* and from visits to the *Apple*® page on *Wikipedia*, in addition to the products of this company, combining traditional indicators from the TA and predicting the direction of the following day's prices using ANNs, SVMs and decision trees. This approach attained accuracies higher than 80%, allowing the authors to claim a higher performance than the use of SVMs in the manner proposed by Kim (2003). It should be noted that Weng et al. (2017) explores crowd-sourced information bases freely available on the Internet, leveraging their prediction system with theoretically other potential traders' opinions. Naturally, how to obtain, interpret and apply all of the relevant data are rich, open research questions.

As stated in Sec. 5.4, the most recent approaches to predicting stocks and indices of financial markets

use data pre-processing and hybridization of classifiers. For example, in the work of Zhang et al. (2017), time series of four market indices were decomposed into individual components before the application of a multidimensional variant of kNN for predictions of closing prices and maximums. The results obtained by Zhang et al. (2017) were better than those obtained using traditional ARIMA and kNN. However, Zhang et al. (2017) do not compare their unique method to the more popular machine learning approaches of SVM or ANN. The approach of Barak et al. (2017) involved the diversification of classifiers regarding the data used for training, using methods of multiple partitioning of these data, as well as sampling techniques. Barak et al. (2017) used the selection of variables among fundamentalist indices and finally merged the classifications of the methods of greater accuracy into a single prediction of returns and risk. Barak et al. (2017) concluded with the superiority of combined predictions, reporting accuracies exceeding 80%. A fundamental base for the work of Barak et al. (2017) rests on the assumption that diversity of results by different classifiers may be combined into a superior prediction. Supposedly, diversification schemes could balance each machine learning classifiers' weakness, while strengthening overall accuracy. Therefore, combining results of individual machine learning models may be a research area as prolific as the fusion of those models themselves.

Still concerning the hybridization of classification models for predictions in financial markets, Krauss et al. (2017) worked with the combination of variants of neural networks and random trees and random forests by comparing stock portfolios created with these techniques. The authors verified that the models used in combinations have better performance than when they are used individually (Krauss et al., 2017, p. 694) – the individual model used as the basis, which had the best predictive performance regarding the data used by the authors, was the RF. Working with a long period of S&P500 daily prices history, the paper from Krauss et al. (2017) accounts for survivorship bias in their portfolio building proposal, that is, not all stocks have always been constituent of the index. Portfolio building schemes must take that into consideration when testing for return using historical prices. Pan et al. (2017), in turn, applied an SVM to multi-frequency independent variables to obtain weekly price predictions of the S&P 500 index – they reported results better than those of the models using single frequency variables. Pan et al. (2017) stand out from other works that use SVM to predict financial time series because the independent variables are not necessarily sampled at the same rate as the model's output. It would be interesting for future works to explore economic implications from this approach. Bezerra and Albuquerque (2017) also applied an SVM, but in its regressor mode, combined with GARCH, to predict the volatility of the returns on financial assets. The greatest innovation of the SVR – GARCH model proposed by Bezerra and Albuquerque (2017) was the combination of Gaussian functions as the kernel function of the SVR. The predictions were compared to those obtained using traditional GARCH models, among others, and, for most of the tests, they achieved results with fewer MAE and RMSE errors.

Following the line of research of Schumaker and Chen (2009) and Al Nasseri et al. (2015), which have

already been reviewed in previous paragraphs, the work of Oliveira et al. (2017) proposed the prediction of return, volume, and volatility of multiple portfolios using textual analysis and the sentiments of specialized blogs. Listed among the most recent of Tab. 13, Oliveira et al. (2017) created indicators of sentiments with textual data, and they applied SVMs, RF and neural networks, among other techniques, to evaluate the contribution that the information from blogs has on financial market predictions. Among the various results, of particular note were the superior accuracies when taking into account the data of the blogs and SVM classifiers in the prediction of returns, especially for companies with lower capitalization, in addition to the technology, energy, and telecommunications sectors. However, the textual data did not significantly increase the accuracy of the predictions of volatility and volume. Researchers exploring sentiment analysis combined with machine learning for stock prices prediction will find a rich review of previous literature and a well developed framework example in Oliveira et al. (2017).

Neural networks continue to be researched for use in predicting financial market prices. Yan et al. (2017) used neural networks combined with Bayesian probability theory to obtain predictions better than those obtained via SVMs and traditional neural networks. The errors are further reduced by the authors with the application of Particle Swarm Optimization (PSO), which is an optimization technique based on grouping and migration of artificial life forms (Yan et al., 2017, p. 2278), in which each state, or particle, is a candidate for an optimal solution, adjusting itself according to the others. Pei et al. (2017) modified traditional neural networks by applying Legendre [2] polynomials in the internal layers of the networks, in addition to a special time function in the method for updating the weights of the connections between the layers. One difference in the work of Pei et al. (2017) is that the authors sought to predict the moving averages of different periods for the prices, not the prices directly or their direction. According to the authors, this approach removes the influence of accidental factors in identifying the direction of the trend (Pei et al., 2017, p. 1694). Finally, Mo and Wang (2017) also proposed neural networks modified by time functions; however, they applied them in the prediction cross-correlations between Chinese and American market indices. The work of Mo and Wang (2017) can therefore be used in the optimization of asset portfolios.

5.6. *Classification of the articles*

This section is dedicated to a classification of the articles reviewed in Secs. 5.1, 5.2, 5.3, 5.4, and 5.5. The articles are classified according to the markets addressed, the assets used in the empirical analyses, the types of predictive variables used as inputs, the dependent variables of the predictions, the main predictive methods used in the models, and the performance measures considered in each author's evaluations. The

[2] Denoted by $L_p(X)$, the Legendre polynomials are a set of orthogonal polynomials of order p that are solutions for the differential equation $\frac{d}{dx}\left[(1-x^2)\frac{dy}{dx}\right] + p(p+1)y = 0$ Legendre polynomials can be used to expand the coefficients of the hidden layers of neural networks. For more details, see Dash (2017).

812 classification subsequently proposed is in Tab. 17, with the following being sought: the most commonly
813 used methods, methods of measuring performance considered, and markets addressed.

814 Tab. 17 lists 57 articles reviewed in the previous sections. The review works and those that did not
815 directly consider financial market prediction were excluded from the classifications, as well as those used
816 as basic references to the methods. Thus, the articles of Adya and Collopy (1998), Zhang et al. (1998), and
817 Atsalakis and Valavanis (2009) were not classified because they are review papers. The article by Malkiel
818 and Fama (1970), which addresses EMH, was also excluded from the classification, in addition to Engle
819 (1982) and Bollerslev (1986), who introduced ARCH and GARCH, respectively. Although Campbell (1987)
820 addressed the prediction of stock prices, this author's methods are not directly related to machine learning,
821 and therefore, it was excluded from the classification. The article by Elman (1990) and the book of Box
822 et al. (2015) are used as bases for the construction of models and are therefore outside the scope of the
823 classification proposed in this section. Finally, the following articles were not classified: Chiu (1994), which
824 introduced fuzzy logic in predictions but did not address financial markets; and Hornik et al. (1989) and
825 Hornik (1991), which generically demonstrated the predictive capabilities of neural networks but did not
826 directly address financial markets.

827 Analysing the articles listed in Tab. 17 regarding the markets addressed, it can be observed that almost
828 half of the studies used North American data (approximately 47%), and one sixth of them (approximately
829 17%) refer to data from Taiwan. This is expected, given the economic hegemony of the USA and the vast
830 academic production of Taiwan, as quantified in Tab. 4. Another interesting fact is that despite the large
831 Chinese academic productivity, only six articles from Tab. 16 use data from China in their predictions
832 (approximately 10% of the articles). Likewise, studies are recorded using data from Brazil, Russia, India,
833 China, and South Africa (BRICS) – 10 articles or approximately 17%. As for the assets for which the
834 predictions are calculated, most articles in Tab. 17 focus on stock market indices (more than 60%). In addition,
835 only two studies applied prediction models simultaneously to indices and stocks.

836 Regarding the variables used as inputs in the models of financial market prediction, the TA indicators are
837 the most popular in Tab. 17 – they were used in approximately 37% of the studies, followed by fundamentalist
838 information, which was used in 26% of the studies. Only two studies explicitly applied both types of variables
839 in their predictive models. Also of note are some studies that applied the prices of the assets themselves –
840 or lagged prices – as inputs in their models. Regarding the prediction sought by the articles of Tab. 17, the
841 models are basically divided between those that seek future prices or returns, and those that seek only the
842 future direction or trend of the market analysed. In this respect, the articles whose dependent variable is the
843 direction of the markets are predominant – approximately 42% of the articles aimed to predict the direction
844 of selected indices or stocks, while 31% of the articles predicted prices.

****

Please insert Tab. 17 here.

****

Among the prediction methods used by the articles in Tab. 17, it can be observed that approximately 70% of the studies used at least some type of neural network; therefore, it was the classification method most used among those of machine learning. The second most used model was SVMs/SVR – a more recent approach than neural networks, which was used in approximately 37 % of the articles reviewed. The hegemony of the ANN and SVM models has already been observed in the main path of literature, in accordance with the reviews of Sec. 5.4. Few studies use other models, such as kNN, RF, or NB. Thus, the articles that only used techniques of classification or regression different from ANNs and SVMs/SVR accounted for approximately 14% of the total researched.

Finally, it should be noted that the method for measuring performance varies with the type of prediction – direction or price – sought by the articles. Thus, articles that seek to predict the direction of the market tend to measure the performance of their models by means of accuracy. Similarly, articles that seek to predict prices, verify their performance by calculating prediction errors. Specifically, MAE and RMSE measure the average magnitude of the error, as given in Bezerra and Albuquerque (2017, p. 188), being the RMSE only a square root applied to the MSE. The MAE is also known as Mean Absolute Deviation (MAD) by some authors, such as Cao et al. (2005, p. 2506). The MAPE measure, in turn, is a percentage measure of the error. MAE, MSE, and MAPE are given by Eqs. 1, 2, and 3, respectively (Bezerra and Albuquerque, 2017; Cao et al., 2005, p. 188, p. 2506), in which N is the number of observations, y is the class or real value of an observation, and $\hat{y}$ is the class or value estimated by the model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2 \tag{2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

Some articles in Tab. 17 that involve classification of the direction of financial markets measure the performance of their respective models via a measure denominated Area Under the Curve (AUC). This is the area under a curve known as Receiver Operating Characteristics (ROC), which is constructed from the variation of the classification thresholds of a model, denoting the rates of false positives on the axis of the abscissa and the ratios of the true positives on the ordinate axis (Zweig and Campbell, 1993, pp. 564–565). There are also many articles that measure the financial returns made possible by the use of suggested

strategies based on their respective predictive models. One of these measures is the Sharpe rate, which is calculated as the rate between average return from an operations strategy and its standard deviation (Fernandez-Rodrıguez et al., 2000, p. 92).

For future reference, the main forecasting and optimization methods considered in machine learning applied to financial markets forecasting, according to the research conducted in this paper, are listed in Tab. 18 and Tab. 19, respectively. References are grouped by each applied method, making future citations of those papers easier. It should be noted, however, that much of the machine learning literature consists of customization and tuning of base models. Therefore, models are listed as their basic form in Tab. 18 and Tab. 19, even though some authors call them by different names once they alter or hybridize them according to their applications.

****
Please insert Tab. 18 here.
****

****
Please insert Tab. 19 here.
****

6. Robust research structure

Based on the reviewed works on the previous section, the next paragraphs summarize the research structure followed by the selected literature on financial markets prediction using machine learning. This section describes in detail the steps to a robust research on predicting markets variables applying machine learning, the desired results and their validation. It aims to be a future reference for the best practices established by the most prominent papers in the field.

Most financial markets prediction papers first acknowledge the difficult task at hand. As commented on Sec. 1, financial markets prices are influenced by a myriad of factors and there has been a large number of proposals for their prediction. Among many markets values, researchers begin by setting their target for prediction, for example, returns and prices, as Zhong and Enke (2017) and Chen et al. (2017), direction, as Ballings et al. (2015) and Patel et al. (2015) or volatility, as Oliveira et al. (2017). The selected dependent variable to be predicted is generally some future, unknown value of a time series, supposedly useful for trading or hedging strategies. As examples, Pan et al. (2017) aim to predict USA stock index for weeks ahead and Gorenc Novak and Velušček (2016) set the next day prices of USA stocks as the dependent variables.

Although not common, present market values may also be chosen as variables of interest to forecast. For instance, Patel et al. (2015) and Kara et al. (2011) evaluate prediction models' performance on closing prices directions forecasting using those very same closing prices for calculating their independent variables. As argued in Sec. 5, for practical implementations, no models would be necessary to forecast market closing

direction if the closing prices are already available at the time of the forecast. Arguably, this approach is more suited to explore models' capabilities than to guide the building of profitable strategies. However, as financial markets future values prediction is the main interest of the academy and practitioners, best practices would dictate the choosing of a future market variable as the forecasting target.

According to the dependent variable chosen for prediction, a performance measure is more adequate. For instance, accuracy definitions, like F-measure, apply to works predicting prices direction classification (up or down), as Weng et al. (2017) and Patel et al. (2015), while error measurements, like RMSE or MAE, are more suited for prices and returns predictions, as applied by Yan et al. (2017) and Göçken et al. (2016). Overall, financial return may or may not be used as performance measure, because not all research focus on building trading strategies. In fact, some works classified on Tab. 17 propose and test models for predicting financial market variables, regardless as how those predictions should be used for profit. Some models obviously translate to trading strategies. For example, direction prediction can surely be interpreted as a buy/sell signals strategy, depending on the direction of the forecast. However, some authors refrain from advising a specific practical usage of their models, concentrating on the report of results using measures other than returns. In case financial returns are used to evaluate models and strategies, it is advisable to include other practical measures, such as drawdown and volatility as risk parameters.

After specifying the prediction target and the measurement of how close the results are from the real outcome, the researcher moves to the methods of prediction, arguably the richest and most creative part of the work. At this point, models are conceived, created, modified, tuned or even hybridized, as commented in Sec. 5. Tab. 18 and Tab. 19 bring the main machine learning algorithms and optimizations compiled by the papers considered in this review. They are only a small number of possibilities, although promising, for market forecasting. There is a large number of new methods being developed and studied, some tuned to other areas of forecasting, but plainly customizable for financial market time-series analysis. Visual recognition models, for example, could greatly enhance predictions based on given patterns, as in the work of Chen and Chen (2016). As another example, semantics and sentiment recognition innovations are potentially applicable to forecasting, as in Oliveira et al. (2017).

Machine learning model building does not necessarily involve the development of an algorithm entirely new. Customizing and tweaking well known models can lead to improved prediction results. Even pre-processing data before running the model is subject to research innovation, as exemplified by the PCA variations applied by Zhong and Enke (2017). Other examples of optimization techniques are found in Tab. 19. As for model customizations examples, basic neural networks present many possibilities throughout specialized literature. For instance, Yan et al. (2017) combine traditional ANN with Bayesian probability theory and Pei et al. (2017) modify internal layers of neural networks with Legendre polynomials. SVM also presents opportunities for innovations through modifications of the traditional approach and the vast possibilities for kernel functions. Combining methods can also result in new and improved algorithms for

predictions, as exemplified by Krauss et al. (2017), Zhang et al. (2017) and Chen et al. (2017).

The next task at researching financial time series predictions using machine learning is obtaining data. However independent variables are chosen by the researcher, historical data must be gathered and somehow processed. All of the above reviewed papers apply their models to real, past data, being them stock prices (Gorenc Novak and Velušček, 2016), market indexes (Pei et al., 2017), fundamental company data (Barak et al., 2017) or text (Oliveira et al., 2017). Few authors rely exclusively on simulated data and, therefore, robust research in the area should always include real data. In fact, independent variables planning can be conducted simultaneously with data gathering, mainly because of availability concerns. Economic data, for example, are not frequently released, which may hinder higher frequency works, and involve reading or processing many company reports. High frequency prices, or even tick data, are commonly not freely available and its processing may require high computing power.

Data gathering may incur in high expenses and should be considered in the budget of the research. Some stock markets provide prices and transaction information free of costs. Brazilian BMF&Bovespa, for example, hosts a database of tick-by-tick level information, with some time constraints, free of costs at the Internet. Yahoo!Finance, a free online quotes database, is a common choice for data, as in Weng et al. (2017), Bezerra and Albuquerque (2017), Chiang et al. (2016) and Gorenc Novak and Velušček (2016). Another aspect that should be considered in selecting an information database is data quality. Few authors dedicate time to the treatment of data problems, such as outliers and missing data, as done by Zhong and Enke (2017). Data problems can be harmful to research results, specially for data-intense algorithms of machine learning, and should be dealt with by a well specified treatment process.

The variables to be used by machine learning models are intimately related to the available data. As exemplified by the papers listed in Tab. 17, predictive or independent variables may be chosen from TA, fundamentalist data, text from news, Internet blogs or prices themselves. Predictive variables choosing may in fact be the research question pursued by some works, as in Chen et al. (2017) and Göçken et al. (2016). Regardless of the type or number of independent variables selected, researchers must be cautious about only using data that would be available at the moment of the forecast. Making use of information prior to its release in the prediction procedure imply in a problem known as data snooping, which harms the validity of results. For example, using closing prices of a given period for calculating TA indicators and using them to predict the closing prices for that same period surely raises questions about validity, which can be observed in the work of Patel et al. (2015) and Kara et al. (2011). Results would be more robust if authors explicitly reported avoidance of snooping problems, as done by Chen et al. (2003, pp. 905–906), Tsai and Hsiao (2010, p. 264) and Gerlein et al. (2016, p. 198).

In machine learning applications, the whole dataset is divided into subsets for training and testing the chosen models. Training data, also called in-sample (Krauss et al., 2017, p. 692), is commonly used in a supervised manner, that is, the real past outcomes are known by the models so they can be optimized to the

training data, expecting the testing data will hold the same underlying characteristics. A good practice is to avoid reusing data, which can lead to data snooping bias, by further dividing the training set into subgroups for model parametrization, as done by Bezerra and Albuquerque (2017). For testing purposes and measuring performance, test subset, called out-of-sample (Huang et al., 2005, p. 2518) or holdout dataset (Kim and Han, 2000, p. 192), is used. No observation from this set should be used in the training or optimization of the models, under the risk of snooping. Rigidly separating the subsets ensures validity of results, as the testing of the models are conducted on new data, unknown by the algorithms, simulating a real situation of forecast.

According to the reviewed literature of Tab. 17, training machine learning models is necessary to select the best parameters for future predictions. Values commented on Sec. 2, such as interconnections' weights in neural networks layers, SVM kernel function parameters and an optimal number of decision trees in a RF model, are obtained in the training phase. Assuming the test dataset behaves similarly to the training dataset, those optimized parameters are used on the models to predict the target variable using test data samples. The best parameters' values must always be chosen by in-sample training, without knowledge of test data, for validating results. That mimics practical implementations because out-of-sample data are not available on the moment of the forecast. Comparison results presented by Kim (2003), for example, may be confronted simply by observing that the author selects the best SVM parameters based on accuracy results from the test set data, incurring in another example of data snooping bias.

Another research decision is whether the machine learning model is periodically updated or it remains unchanged through the testing phase. It is not possible to observe a best practice in this regard, for both approaches are valid in the reviewed literature of Tab. 17. Keeping the models unchanged once they are optimized has the advantage of low computational costs. Once the models are trained, they are promptly used as each testing sample becomes available. Krauss et al. (2017), Pan et al. (2017), Yan et al. (2017), Pei et al. (2017) and Bezerra and Albuquerque (2017) are examples of papers which consider fixed, unchanged models once they are trained. On the other hand, the machine learning models may be periodically updated once new data become available, a procedure called sliding window (Gerlein et al., 2016, p. 198). Computational costs are higher because the models have to be retrained every time the sliding window moves. However, they are constantly adapting to new market conditions. Examples of papers which follow this last approach are Chiang et al. (2016), Gorenc Novak and Velušček (2016), Tsai and Hsiao (2010), Li and Kuo (2008) and Thawornwong and Enke (2004).

Finally, after comparing predictions with the real, test samples, the selected performance measures are calculated. Some authors further apply statistical tests to evaluate the significance of results. For instance, Zhong and Enke (2017) and Kara et al. (2011) apply t-tests while Kim and Han (2000) and Yu et al. (2009) apply McNemar's tests. Relying on statistical tests well established in the scientific literature improves robustness of the research, given the authors correctly interpret the results. Although many works still

report results without them, statistical tests should be incorporated as best practices in the field of machine learning financial time series prediction for significance and robustness.

7. Conclusions about the review

This article provided previously available quantitative and objective methods for selecting the relevant literature on a particular theme of scientific research. The literature available regarding any topic can be broad, and a complete coverage of all the published documents can be challenging or even impossible. A systematic selection of the literature most relevant to a study is thus necessary, taking into account not only the history of an area but its state of the art. Thus, this review described bibliographic survey techniques and used them in the systematic review of financial market predictions that use machine learning techniques. This review led to the summarization of the best procedures established by the scientific literature on the field in order to achieve robust results when researching financial markets prediction using machine learning.

As this literature review addressed machine learning techniques, Sec. 2 briefly commented on popular models; for example, ANN, SVM, and RF. In relation to the broader survey of the literature, Sec. 3 described how to proceed to a search of databases using keywords and filters by subject. It is emphasized that the quality of this initial search for articles determined the final quality of the results obtained by the bibliographic survey. Sec. 4, in turn, presented the results of the database of articles surveyed, validating it objectively with the use of Lotka's distribution, in addition to analysing the results regarding the most productive authors and countries, the most-cited periodicals, and the potential targets for submitting new works for publication.

Moving on to the actual review of the most important articles about financial market prediction using machine learning, this work commented on the following: the most-cited articles, those with the greatest bibliometric coupling and the highest co-citation frequencies, the most recently published articles, and those that are part of the main path of the knowledge flow of the literature studied. It is emphasized that these were objective and clear methods of surveys, independent of the experience of the researcher, serving not only for initial studies in research but also as validation of knowledge for experienced specialists.

Finally, this work proposed a classification of the 57 articles reviewed, based on the markets addressed, the type of index predicted, the variables used as inputs for the models, and the type of prediction sought. Additionally, the prediction methods used and the main performance measures used by each article were summarized. There is extensive use of data from the North American market, in addition to the application of neural and SVM networks. Similarly, most of the predictions relate to market indices. Among the possible conclusions about the classification proposed here, it is to be expected that new proposed models will be compared to the benchmarks of neural and SVM networks, as is the use of data from the North American

market. The use of new models in financial market prediction continues providing research opportunities, as does the exploration of the behaviour of predictions in developing markets, such as those of the BRICS.

References

Abu-Mostafa, Y. S., Atiya, A. F., 1996. Introduction to financial forecasting. Applied Intelligence 6 (3), 205–213.

Adya, M., Collopy, F., 1998. How efective are neural networks at forecasting and prediction? A review and evaluation. Journal of Forecasting 17 (1), 481–495.

Al Nasseri, A., Tucker, A., de Cesare, S., 2015. Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. Expert Systems with Applications 42 (23), 9192–9210.

Ang, K. K., Quek, C., 2006. Stock trading using RSPOP: A novel rough set-based neuro-fuzzy approach. IEEE Transactions on Neural Networks 17 (5), 1301–1315.

Armano, G., Marchesi, M., Murru, A., 2005. A hybrid genetic-neural architecture for stock indexes forecasting. Information Sciences 170 (1), 3–33.

Atsalakis, G. S., Valavanis, K. P., 2009. Surveying stock market forecasting techniques–Part II: Soft computing methods. Expert Systems with Applications 36 (3), 5932–5941.

Ballings, M., den Poel, D. V., Hespeels, N., Gryp, R., 2015. Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications 42 (20), 7046–7056.

Barak, S., Arjmand, A., Ortobelli, S., 2017. Fusion of multiple diverse predictors in stock market. Information Fusion 36 (1), 90–102.

Batagelj, V., 2003. Efficient algorithms for citation network analysis. arXiv preprint cs/0309023.

Bezerra, P. C. S., Albuquerque, P. H. M., 2017. Volatility forecasting via SVR–GARCH with mixture of Gaussian kernels. Computational Management Science 14 (2), 179–196.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31 (3), 307–327.

Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M., 2015. Time series analysis: forecasting and control, 3rd Edition. Vol. 1. John Wiley & Sons, Hoboken, New Jersey.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Campbell, J. Y., 1987. Stock returns and the term structure. Journal of Financial Economics 18 (2), 373–399.

Cao, L., 2003. Support vector machines experts for time series forecasting. Neurocomputing 51 (1), 321–339.

Cao, Q., Leggio, K. B., Schniederjans, M. J., 2005. A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. Computers & Operations Research 32 (10), 2499–2512.

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., Oliveira, A. L., 2016. Computational intelligence and financial markets: A survey and future directions. Expert Systems with Applications 55 (1), 194–211.

Chang, P.-C., Fan, C.-Y., 2008. A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38 (6), 802–815.

Chang, P.-C., Liu, C.-H., Lin, J.-L., Fan, C.-Y., Ng, C. S., 2009. A neural network with a case based dynamic window for stock trading prediction. Expert Systems with Applications 36 (3, Part 2), 6889–6898.

Chen, A.-S., Leung, M. T., Daouk, H., 2003. Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index. Computers & Operations Research 30 (6), 901–923.

Chen, H., Xiao, K., Sun, J., Wu, S., 2017. A Double-Layer Neural Network Framework for High-Frequency Forecasting. ACM Transactions on Management Information Systems (TMIS) 7 (4), 11:2–11:17.

Chen, T.-l., Chen, F.-y., 2016. An intelligent pattern recognition model for supporting investment decisions in stock market. Information Sciences 346 (1), 261–274.

Chen, Y.-S., Cheng, C.-H., Tsai, W.-L., 2014. Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting. Applied Intelligence 41 (2), 327–347.

Chiang, W.-C., Enke, D., Wu, T., Wang, R., 2016. An adaptive stock index trading decision support system. Expert Systems with Applications 59 (1), 195–207.

Chiu, S. L., 1994. Fuzzy model identification based on cluster estimation. Journal of Intelligent & Fuzzy Systems 2 (3), 267–278.

Dash, R., 2017. Performance analysis of an evolutionary recurrent Legendre Polynomial Neural Network in application to FOREX prediction. Journal of King Saud University-Computer and Information SciencesIn Press.

Dietterich, T. G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10 (7), 1895–1923.

Donaldson, R. G., Kamstra, M., 1999. Neural network forecast combining with interaction effects. Journal of the Franklin Institute 336 (2), 227–236.

Egghe, L., 2006. Theory and practise of the *g*-index. Scientometrics 69 (1), 131–152.

Elman, J. L., 1990. Finding structure in time. Cognitive Science 14 (2), 179–211.

Engle, R. F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica: Journal of the Econometric Society 50 (4), 987–1007.

Enke, D., Thawornwong, S., 2005. The use of data mining and neural networks for forecasting stock market returns. Expert Systems with Applications 29 (4), 927–940.

Fama, E. F., 1991. Efficient capital markets: II. The Journal of Finance 46 (5), 1575–1617.

Fernandez-Rodrıguez, F., Gonzalez-Martel, C., Sosvilla-Rivero, S., 2000. On the profitability of technical trading rules based on artificial neural networks: Evidence from the Madrid stock market. Economics Letters 69 (1), 89–94.

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 (1), 193–207.

Göçken, M., Özçalıcı, M., Boru, A., Dosdoğru, A. T., 2016. Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction. Expert Systems with Applications 44 (1), 320–331.

Gorenc Novak, M., Velušček, D., 2016. Prediction of stock price movement based on daily high prices. Quantitative Finance 16 (5), 793–826.

Hájek, P., Olej, V., Myskova, R., 2013. Forecasting stock prices using sentiment information in annual reports–a neural network and support vector regression approach. WSEAS Transactions on Business and Economics 10 (4), 293–305.

Hassan, M. R., Nath, B., Kirley, M., 2007. A fusion model of HMM, ANN and GA for stock market forecasting. Expert Systems with Applications 33 (1), 171–180.

Henrique, B. M., Sobreiro, V. A., Kimura, H., 2018. Building direct citation networks. Scientometrics 115 (2), 817–832.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251–257.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., Johnson, J. E., 2016. Bridging the divide in financial market forecasting: Machine learners *vs.* financial economists. Expert Systems with Applications 61 (1), 215–234.

Huang, C.-L., Tsai, C.-Y., 2009. A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert Systems with Applications 36 (2), 1529–1539.

Huang, W., Nakamori, Y., Wang, S.-Y., 2005. Forecasting stock market movement direction with support vector machine. Computers & Operations Research 32 (10), 2513–2522.

Hummon, N. P., Doreian, P., 1989. Connectivity in a citation network: The development of DNA theory. Social Networks 11 (1), 39–63.

Kamstra, M., Donaldson, G., 1996. Forecasting combined with neural networks. Journal of Forecast 15 (1), 49–61.

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 (5), 5311–5319.

Kessler, M. M., 1963. Bibliographic coupling between scientific papers. Journal of the Association for Information Science and Technology 14 (1), 10–25.

Kim, K., 2003. Financial time series forecasting using support vector machines. Neurocomputing 55 (1-2), 307–319.

Kim, K.-j., Han, I., 2000. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Systems with Applications 19 (2), 125–132.

Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M., 1990. Stock market prediction system with modular neural networks. In: Neural Networks, 1990., 1990 IJCNN International Joint Conference on. IEEE, pp. 1–6.

Krauss, C., Do, X. A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research 259 (2), 689–702.

Kumar, D., Meghwani, S. S., Thakur, M., 2016. Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. Journal of Computational Science 17 (1), 1–13.

Kumar, M., Thenmozhi, M., 2014. Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models. International Journal of Banking, Accounting and Finance 5 (3), 284–308.

Laboissiere, L. A., Fernandes, R. A., Lage, G. G., 2015. Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks. Applied Soft Computing 35 (1), 66–74.

Lage Junior, M., Godinho Filho, M., 2010. Variations of the kanban system: Literature review and classification. International Journal of Production Economics 125 (1), 13–21.

Lahmiri, S., 2014a. Improving forecasting accuracy of the S&P500 intra-day price direction using both wavelet low and high frequency coefficients. Fluctuation and Noise Letters 13 (01), 1450008.

Lahmiri, S., 2014b. Entropy-based technical analysis indicators selection for international stock markets fluctuations prediction using support vector machines. Fluctuation and Noise Letters 13 (02), 1450013.

Lahmiri, S., Boukadoum, M., 2015. An Ensemble System Based on Hybrid EGARCH-ANN with Different Distributional Assumptions to Predict S&P 500 Intraday Volatility. Fluctuation and Noise Letters 14 (01), 1550001.

Leigh, W., Purvis, R., Ragusa, J. M., 2002. Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. Decision Support Systems 32 (4), 361–377.

Leung, M. T., Daouk, H., Chen, A.-S., 2000. Forecasting stock indices: a comparison of classification and level estimation models. International Journal of Forecasting 16 (2), 173–190.

Li, S.-T., Kuo, S.-C., 2008. Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks. Expert Systems with Applications 34 (2), 935–951.

Liu, J. S., Lu, L. Y., 2012. An integrated approach for main path analysis: Development of the Hirsch index as an example. Journal of the American Society for Information Science and Technology 63 (3), 528–542.

Liu, J. S., Lu, L. Y., Lu, W.-M., Lin, B. J., 2013. Data envelopment analysis 1978–2010: A citation-based literature survey. Omega 41 (1), 3–15.

Malkiel, B. G., 2003. The Efficient Market Hypothesis and Its Critics. Journal of Economic Perspectives 17 (1), 59–82.

Malkiel, B. G., Fama, E. F., 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance 25 (2), 383–417.

Mariano, E. B., Sobreiro, V. A., Rebelatto, D. A. N., 2015. Human development and data envelopment analysis: A structured literature review. Omega 54 (1), 33–49.

Mo, H., Wang, J., 2017. Return scaling cross-correlation forecasting by stochastic time strength neural network in financial market dynamics. Soft Computing 1 (1), 1–13.

Nayak, R. K., Mishra, D., Rath, A. K., 2015. A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices. Applied Soft Computing 35 (1), 670–680.

Oliveira, N., Cortez, P., Areal, N., 2017. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications 73 (1), 125–144.

Ortega, L., Khashanah, K., 2014. A Neuro-wavelet Model for the Short-Term Forecasting of High-Frequency Time Series of Stock Returns. Journal of Forecasting 33 (2), 134–146.

Pai, P.-F., Lin, C.-S., 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33 (6), 497–505.

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 (1), 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 (1), 259–268.

Pei, A., Wang, J., Fang, W., 2017. Predicting agent-based financial time series model on lattice fractal with random Legendre neural network. Soft Computing 21 (7), 1693–1708.

Rodríguez-González, A., García-Crespo, Á., Colomo-Palacios, R., Iglesias, F. G., Gómez-Berbís, J. M., 2011. CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator. Expert systems with Applications 38 (9), 11489–11500.

Saam, N., Reiter, L., 1999. Lotka's law reconsidered: The evolution of publication and citation distributions in scientific fields. Scientometrics 44 (2), 135–155.

Schumaker, R. P., Chen, H., 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS) 27 (2), 12.

Seuring, S., 2013. A review of modeling approaches for sustainable supply chain management. Decision Support Systems 54 (4), 1513 – 1520, rapid Modeling for Sustainability.

Small, H., 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the Association for Information Science and Technology 24 (4), 265–269.

Son, Y., Noh, D.-j., Lee, J., 2012. Forecasting trends of high-frequency KOSPI200 index data using learning classifiers. Expert Systems with Applications 39 (14), 11607–11615.

Tay, F. E., Cao, L., 2001. Application of support vector machines in financial time series forecasting. Omega 29 (4), 309–317.

Thawornwong, S., Enke, D., 2004. The adaptive selection of financial and economic variables for use with artificial neural networks. Neurocomputing 56 (1), 205–232.

Thawornwong, S., Enke, D., Dagli, C., 2003. Neural networks as a decision maker for stock trading: A technical analysis approach. International Journal of Smart Engineering System Design 5 (4), 313–325.

Timmermann, A., Granger, C. W., 2004. Efficient market hypothesis and forecasting. International Journal of Forecasting 20 (1), 15–27.

Tsai, C.-F., Hsiao, Y.-C., 2010. Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems 50 (1), 258–269.

Tsaih, R., Hsu, Y., Lai, C. C., 1998. Forecasting S&P 500 stock index futures with a hybrid AI system. Decision Support Systems 23 (2), 161–174.

Vapnik, V., 1995. The nature of statistical learning theory. Springer Heidelberg, New York, New York.

Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., Guo, S.-P., 2012. Stock index forecasting based on a hybrid model. Omega 40 (6), 758–766.

Wang, Y.-F., 2002. Predicting stock price using fuzzy grey prediction system. Expert Systems with Applications 22 (1), 33–38.

Wang, Y.-F., 2003. Mining stock price using fuzzy rough set system. Expert Systems with Applications 24 (1), 13–23.

Weng, B., Ahmed, M. A., Megahed, F. M., 2017. Stock market one-day ahead movement prediction using disparate data sources. Expert Systems with Applications 79 (1), 153–163.

Xiao, Y., Xiao, J., Lu, F., Wang, S., 2013. Ensemble ANNs-PSO-GA Approach for Day-ahead Stock E-exchange Prices Forecasting. International Journal of Computational Intelligence Systems 6 (1), 96–114.

Yan, D., Zhou, Q., Wang, J., Zhang, N., 2017. Bayesian regularisation neural network based on artificial intelligence optimisation.

1207 International Journal of Production Research 55 (8), 2266–2287.
1208 Yoon, Y., Swales Jr, G., Margavio, T. M., 1993. A comparison of discriminant analysis versus artificial neural networks. Journal of the
1209 Operational Research Society 44 (1), 51–60.
1210 Yu, L., Chen, H., Wang, S., Lai, K. K., 2009. Evolving least squares support vector machines for stock market trend mining. IEEE
1211 Transactions on Evolutionary Computation 13 (1), 87–102.
1212 Zhang, G., Patuwo, B. E., Hu, M. Y., 1998. Forecasting with artificial neural networks: The state of the art. International Journal of
1213 Forecasting 14 (1), 35–62.
1214 Zhang, N., Lin, A., Shang, P., 2017. Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting.
1215 Physica A: Statistical Mechanics and its Applications 477 (1), 161–173.
1216 Zhong, X., Enke, D., 2017. Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications
1217 67 (1), 126–139.
1218 Zweig, M. H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine.
1219 Clinical Chemistry 39 (4), 561–577.

Algorithm 1: Algorithm for the construction of networks of direct citations.

```
1  Initialize list_articles;
2  for i ← 1 to Total (list_articles) do
3  |   article ← list_articles[i];
4  |   list_references ← References (article);
5  |   for j ← 1 to Total (list_references) do
6  |   |   for k ← 1 to Total (list_articles) do
7  |   |   |   if list_references[j] = Title (list_articles[k]) then
8  |   |   |   |   Trace an edge starting at list_articles[k] and finishing at list_articles[i];
9  |   |   |   end
10 |   |   end
11 |   end
12 end
```

Algorithm 2: Algorithm for finding the main path of the literature via SPC.

```
1  Initialize list_sources;
2  Initialize list_sinks;
3  for i ← 1 to Total (list_sources) do
4  |   source ← list_sources[i];
5  |   for j ← 1 to Total (list_sinks) do
6  |   |   sink ← list_sinks[j];
7  |   |   paths ← Paths (source, sink);
8  |   |   foreach (path in paths) do
9  |   |   |   Add 1 to the weight of each esge that is part of path;
10 |   |   end
11 |   end
12 end
13 Return the path between the source and the sink with the largest sum of weights.
```

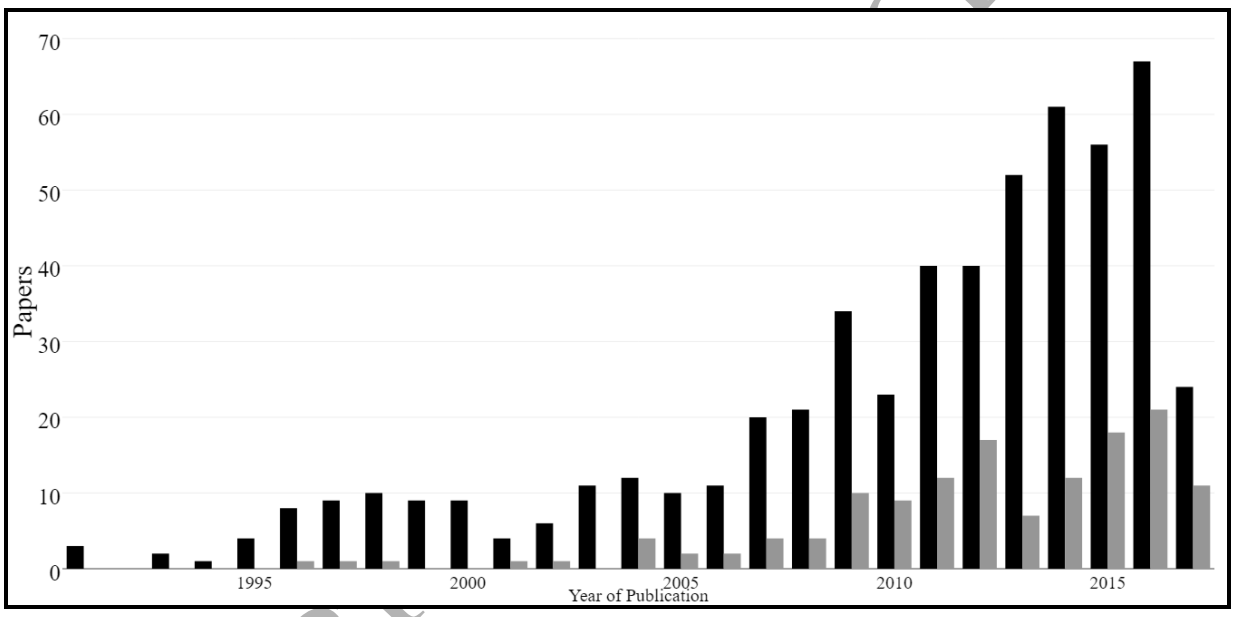

Figure 1: Frequency of publication of the articles considered in the bibliometrics, between 1991 and 2017, in dark bars. The lighter bars, for comparison, illustrate the publications about credit risk that apply machine learning.

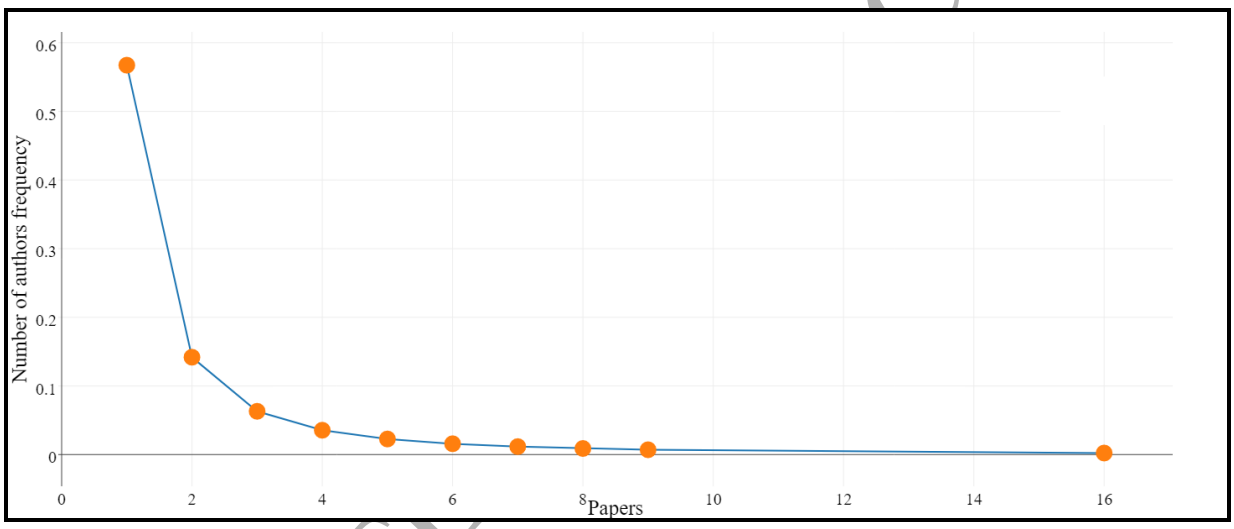

Figure 2: Frequency of publication, by author, in the database of articles considered in the bibliometrics. The continuous line represents Lotka's theoretical distribution, whereas the circles mark the distribution observed in the bibliographic survey of this present work.

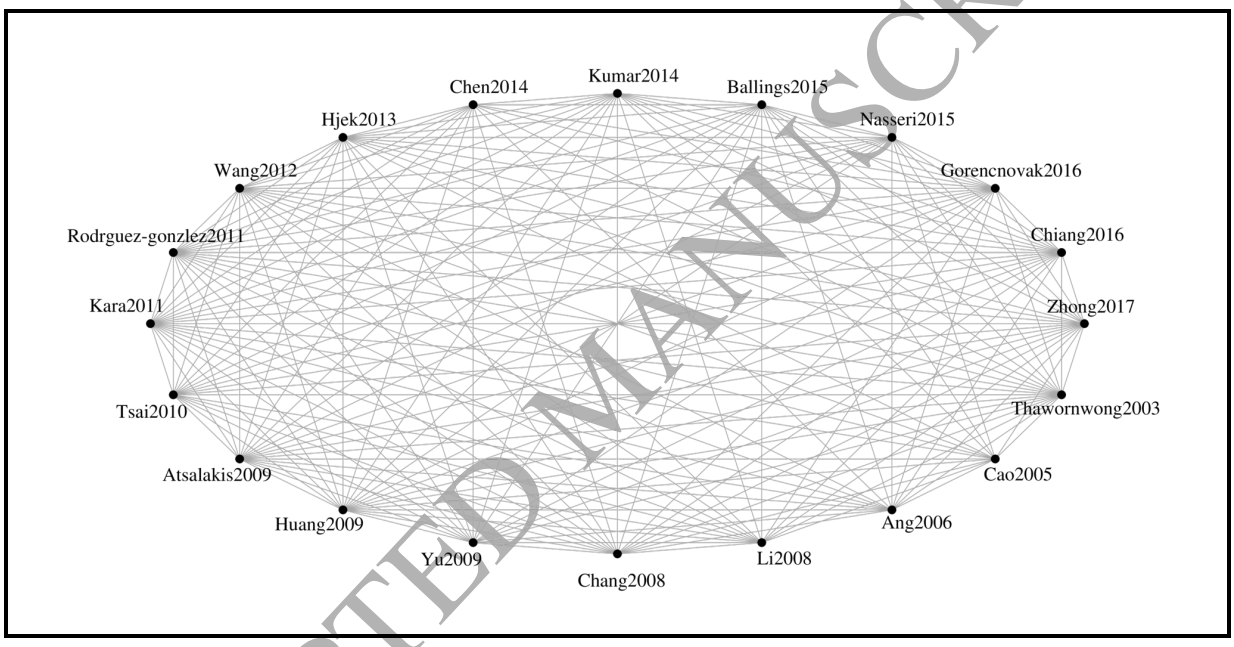

Figure 3: Bibliographic coupling of the 20 articles with the highest degrees of relationship. Each line represents a coupling relationship between the articles. The interconnections make a dense, almost fully-mashed network of papers with similar references.

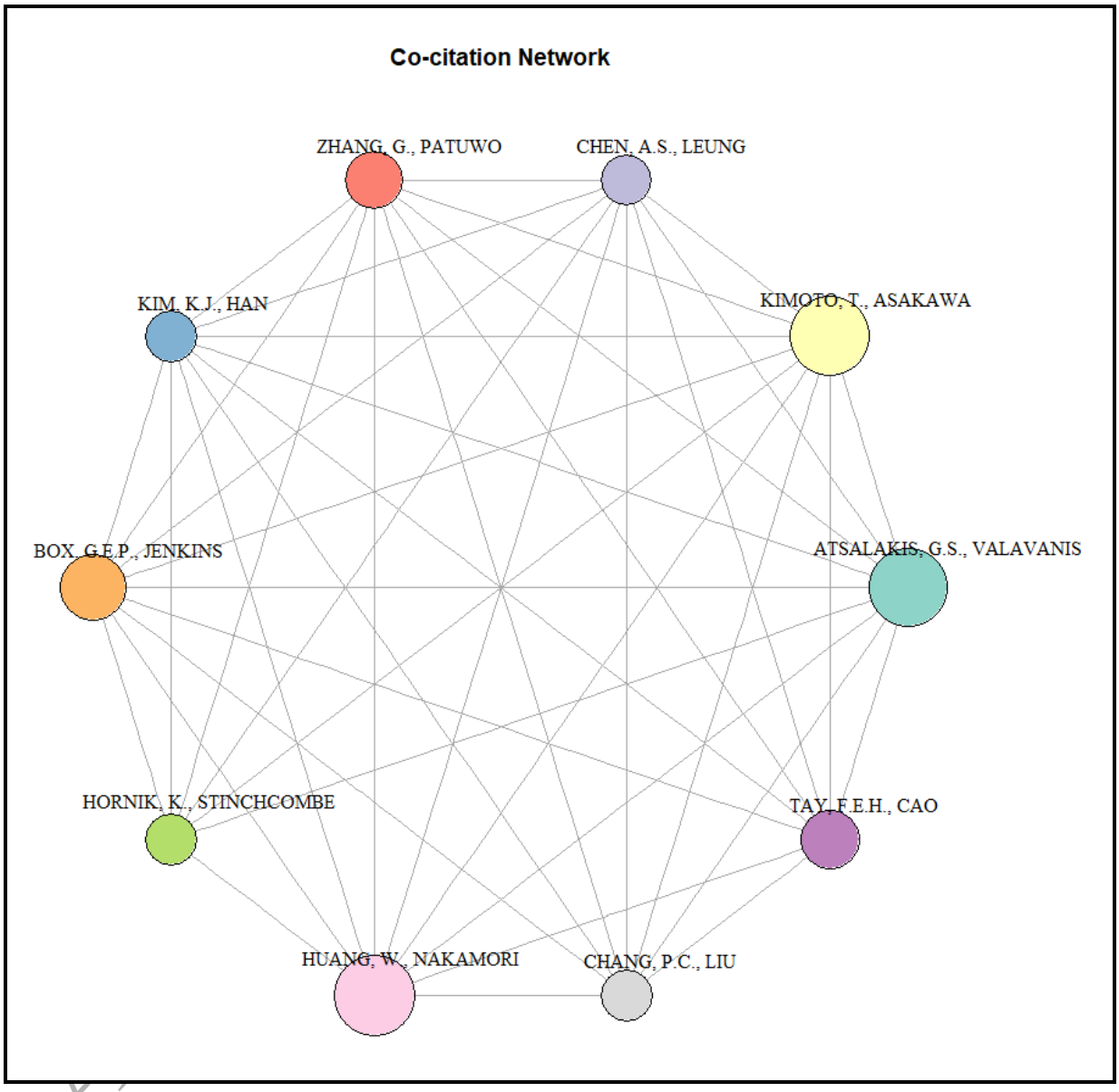

Figure 4: Network of co-citations for the 10 most-related authors and co-authors.

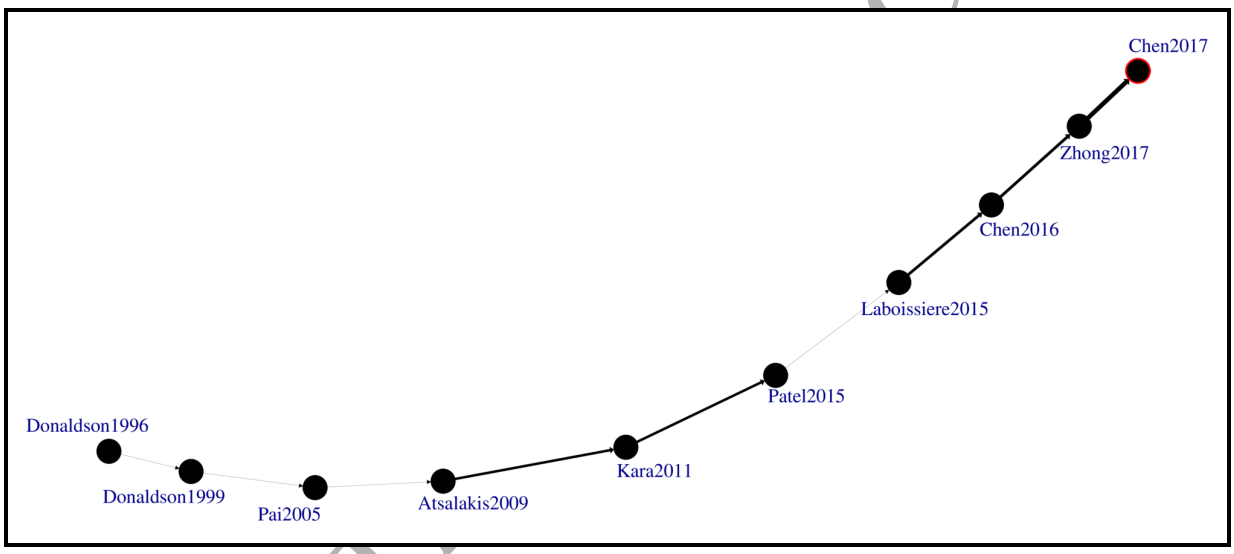

Figure 5: Main path followed by the literature. The edges have a thickness proportional to the weight assigned by Alg. 2.
Source: Henrique et al. (2018).

Characteristic	Value
Number of articles.	547
Periodicals.	243
Number of keywords.	1336
Period of the publications.	1991 – 2017
Average number of citations per article.	17.63
Authors.	1.151
Authors with single-author articles.	43
Articles per author.	0.475
Authors per article.	2.1

Table 1: Description of the database of articles used in the bibliometrics.

Author	Articles in the Database
Wang, J.	16
Cheng, C-H.	9
Wei, L-Y.	9
Enke, D.	8
Dash, PK.	7
Chang, P-C.	6
Chen, H.	6
Lahmiri, S.	6
Sun, J.	6
Dhar, J.	5
Hu, Y.	5
Kamstra, M.	5
Li, H.	5
Zhang, Y.	5
Zhang, Z.	5
Bekiros, SD.	4
Bisoi, R.	4
Chen, T-L.	4
Chen, Y.	4
Donaldson, RG.	4

Table 2: The 20 authors with the highest number of published articles in the database of articles searched.

Number of Articles	Authors	Distribution of Frequency
1	964	0.837533
2	126	0.109470
3	35	0.030408
4	11	0.009557
5	6	0.005213
6	4	0.003475
7	1	0.000869
8	1	0.000869
9	2	0.001738
16	1	0.000869

Table 3: Authors added to the most productive according to the *h* index.

Country	Number of Articles	Frequency
China.	88	0.16635
Taiwan.	72	0.13611
India.	60	0.11342
USA.	55	0.10397
Korea.	24	0.04537
Iran.	18	0.03403
United Kingdom.	16	0.03025
Spain.	15	0.02836
Greece.	14	0.02647
Singapore.	14	0.02647
Italy.	12	0.02268
Australia.	11	0.02079
Brazil.	11	0.02079
Hong Kong.	10	0.01890
Canada.	9	0.01701
Germany.	9	0.01701
Japan.	8	0.01512
Turkey.	8	0.01512
Lithuania.	6	0.01134
Malaysia.	5	0.00945

Table 4: The 20 countries with the highest number of articles according to the database of articles searched.

Author	*h* index	*g* index	Citations	Articles
Wang, J.	9	16	296	16
O, J.	6	11	122	13
Cheng, C-H.	6	9	243	9
Wei, L-Y.	7	9	231	9
Chen, Y.	3	9	86	9
Enke, D.	5	8	295	8
Chen, H.	5	8	438	8
Zhang, Z.	3	7	121	7
Dash, PK.	3	6	41	7
Chang, P-C.	6	6	214	6
Sun, J.	5	6	184	6
Wang, Y.	3	6	292	6
Kim, S.	5	6	89	6
Hu, Y.	3	5	30	5
Kamstra, M.	5	5	311	5
Li, H.	5	5	184	5
Zhang, Y.	2	5	60	5
Liu, M.	3	5	35	5
Bekiros, S.	3	5	27	5
Chen, T.	5	5	187	5

Table 5: Authors added to the most productive according to the *g* index.

Author	h index	g index	Citation	Articles
Chen, T-L.	4	4	182	4
Donaldson, RG.	4	4	270	4
Fan, C-Y.	4	4	177	4
Lu, C-J.	4	4	104	4
Quek, C.	4	4	101	4
Lin, C.	4	4	362	4
Lahmiri, S.	3	4	20	6

Table 6: Authors added to the most productive according to the h index.

Country	Number of Articles	Frequency
China.	88	0.16635
Taiwan.	72	0.13611
India.	60	0.11342
USA.	55	0.10397
Korea.	24	0.04537
Iran.	18	0.03403
United Kingdom.	16	0.03025
Spain.	15	0.02836
Greece.	14	0.02647
Singapore.	14	0.02647
Italy.	12	0.02268
Australia.	11	0.02079
Brazil.	11	0.02079
Hong Kong.	10	0.01890
Canada.	9	0.01701
Germany.	9	0.01701
Japan.	8	0.01512
Turkey.	8	0.01512
Lithuania.	6	0.01134
Malaysia.	5	0.00945

Table 7: The 20 countries with the highest number of articles according to the database of articles searched.

Country	Number of Citations	Mean Number of Citations per Article
Taiwan.	2429	33.736
USA.	1913	34.782
Korea.	1211	50.458
China.	900	10.227
Greece.	340	24.286
Singapore.	336	24.000
Canada.	282	31.333
India.	239	3.983
Spain.	234	15.600
Italy.	226	18.833
Australia.	190	17.273
United Kingdom.	173	10.812
Turkey.	152	19.000
Iran.	142	7.889
Brazil.	141	12.818
Germany.	133	14.778
Hong Kong.	106	10.600
Thailand.	49	9.800
Norway.	39	39.000
Slovenia.	35	17.500

Table 8: The 20 countries with the highest number of citations according to the database searched.

Periodical	Number of Articles
Expert Systems with Applications.	76
Neurocomputing.	21
Applied Soft Computing Journal.	20
Decision Support Systems.	14
Neural Computing and Applications.	11
Neural Network World.	11
Journal of Forecasting.	8
Studies in Computational Intelligence.	8
International Journal of Applied Engineering Research.	7
Journal of Theoretical and Applied Information Technology.	7
Mathematical Problems in Engineering.	7
Soft Computing.	7
Information Sciences.	6
Journal of Information and Computational Science.	6
Knowledge-Based Systems.	6
Lecture Notes in Computer Science.	6
Physica A: Statistical Mechanics and its Applications.	6
Applied Intelligence.	5
Fluctuation and Noise Letters.	5
Computational Economics.	4

Table 9: The 20 journals with the highest number of articles in the database searched.

Keyword	Articles that Use the Keyword
Neural networks.	59
Forecasting.	45
Data mining.	36
Stock market.	34
Artificial neural networks.	30
Neural network.	29
Artificial neural network.	28
Prediction.	25
Genetic algorithm.	22
Machine learning.	21
Stock price forecasting.	19
Time series.	19
Feature selection.	17
Support vector machine.	17
Technical analysis.	17
Support vector machines.	16
Stock market prediction.	15
Support vector regression.	15
Stock prediction.	14
Genetic algorithms.	13

Table 10: The 20 keywords most used in the database of articles searched.

References	Title	Journal	Citations	Citations per Year
Kim (2003).	*Financial Time Series Forecasting Using Support Vector Machines.*	*Neurocomputing.*	546	39.00
Kim and Han (2000).	*Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index.*	*Expert Systems with Applications.*	279	16.41
Pai and Lin (2005).	*A Hybrid ARIMA And Support Vector Machines Model in Stock Price Forecasting.*	*Omega.*	278	23.17
Atsalakis and Valavanis (2009).	*Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods.*	*Expert Systems with Applications.*	2 24	28.00
Schumaker and Chen (2009).	*Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System.*	*ACM Transactions on Information Systems.*	186	23.25
Chen et al. (2003).	*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index.*	*Computers and Operations Research.*	184	13.14
Wang (2002).	*Predicting Stock Price Using Fuzzy Grey Prediction System.*	*Expert Systems with Applications.*	163	10.87
Enke and Thawornwong (2005).	*The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns.*	*Expert Systems with Applications.*	152	12.67
Armano et al. (2005).	*A Hybrid Genetic-Neural Architecture for Stock Indexes Forecasting.*	*Information Sciences.*	130	10.83
Leigh et al. (2002).	*Forecasting the NYSE Composite Index with Technical Analysis, Pattern Recognizer, Neural Network, and Genetic Algorithm: A Case Study in Romantic Decision Support.*	*Decision Support Systems.*	130	8.67
Hassan et al. (2007).	*A Fusion Model Of HMM, ANN and GA for Stock Market Forecasting.*	*Expert Systems with Applications.*	124	12.40
Tsaih et al. (1998).	*Forecasting S&P 500 Stock Index Futures with a Hybrid AI System.*	*Decision Support Systems.*	121	6.37
Leung et al. (2000).	*Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models.*	*International Journal of Forecasting.*	118	6.94
Wang (2003).	*Mining Stock Price Using Fuzzy Rough Set System.*	*Expert Systems with Applications*	117	8.36
Kamstra and Donaldson (1996).	*Forecast Combining with Neural Networks.*	*Journal of Forecasting*	115	5.48
Kara et al. (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.*	104	17.33
Yu et al. (2009).	*Evolving Least Squares Support Vector Machines for Stock Market Trend Mining.*	*IEEE Transactions on Evolutionary Computation.*	104	13.00
Fernandez-Rodrıguez et al. (2000).	*On the Profitability of Technical Trading Rules Based on Artificial Neural Networks: Evidence from the Madrid Stock Market.*	*Economics Letters.*	100	5.88
Huang and Tsai (2009).	*A Hybrid SOFM-SVR with a Filter-Based Feature Selection for Stock Market Forecasting.*	*Expert Systems with Applications.*	99	12.38
Yoon et al. (1993).	*A Comparison of Discriminant Analysis Versus Artificial Neural Networks.*	*Journal of the Operational Research Society.*	99	4.12

Table 11: The 20 articles most cited in the compiled database. The number of citations refers to the citations in the entire *Scopus* database.

References	Title	Journal	Citations
Bollerslev (1986).	*Generalized Autoregressive Conditional Heteroscedasticity.*	*Journal of Econometrics.*	23
Kim (2003).	*Financial Time Series Forecasting Using Support Vector Machines.*	*Neurocomputing.*	18
Enke and Thawornwong (2005).	*The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns.*	*Expert Systems with Applications.*	12
Elman (1990).	*Finding Structure in Time.*	*Cognitive Science.*	10
Engle (1982).	*Autoregressive Conditional Heteroscedasticity with Estimator of the Variance of United Kingdom Inflation.*	*Econometrica.*	9
Huang et al. (2005).	*Forecasting Stock Market Movement Direction with Support Vector Machine.*	*Computers and Operations Research.*	9
Thawornwong and Enke (2004).	*The Adaptive Selection of Financial and Economic Variables for Use with Artificial Neural Networks.*	*Neurocomputing.*	9
Campbell (1987).	*Stock Returns and the Term Structure.*	*Journal of Financial Economics.*	8
Chen et al. (2003).	*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index.*	*Computers and Operations Research.*	8
Pai and Lin (2005).	*A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting.*	*Omega.*	8
Malkiel and Fama (1970).	*Efficient Capital Markets: A Review of Theory and Empirical Work*	*Journal of Finance.*	7
Hornik et al. (1989).	*Multilayer Feedforward Networks are Universal Approximators.*	*Neural Networks*	7
Leung et al. (2000).	*Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models.*	*International Journal of Forecasting.*	7
Tsaih et al. (1998).	*Forecasting S&P 500 Stock Index Futures with a Hybrid AI System.*	*Decision Support Systems.*	7
Zhang et al. (1998).	*Forecasting with Artificial Neural Networks: The State of the Art*	*International Journal of Forecasting.*	7
Abu-Mostafa and Atiya (1996).	*Introduction to Financial Forecasting.*	*Applied Intelligence.*	6
Adya and Collopy (1998).	*How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation.*	*Journal of Forecasting.*	6
Atsalakis and Valavanis (2009).	*Surveying Stock Market Forecasting Techniques-Part II: Soft Computing Methods.*	*Expert Systems with Applications.*	6
Chiu (1994).	*Fuzzy Model Identification Based on Cluster Estimation.*	*Journal of Intelligent and Fuzzy Systems.*	6
Hornik (1991).	*Approximation Capabilities of Multilayer Feedforward Networks.*	*Neural Networks.*	6

Table 12: The 20 articles most cited by the articles in the compiled database.
Note: It should be noted that the articles in this table may not be part of the initially compiled database.

Reference	Title	Journal
Weng et al. (2017).	*Stock Market one-day ahead Movement Prediction using Disparate Data Sources.*	*Expert Systems with Applications.*
Zhang et al. (2017).	*Multidimensional k-Nearest Neighbor Model based on EEMD for Financial Time Series Forecasting.*	*Physica A: Statistical Mechanics and its Applications.*
Barak et al. (2017).	*Fusion of Multiple Diverse Predictors in Stock Market.*	*Information Fusion.*
Krauss et al. (2017).	*Deep Neural Networks, Gradient-boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500.*	*European Journal of Operational Research.*
Oliveira et al. (2017).	*The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices.*	*Expert Systems with Applications.*
Yan et al. (2017).	*Bayesian Regularisation Neural Network Based on Artificial Intelligence Optimisation.*	*International Journal of Production Research.*
Pan et al. (2017).	*A Multiple Support Vector Machine Approach to Stock Index Forecasting with Mixed Frequency Sampling.*	*Knowledge- Based Systems.*
Pei et al. (2017).	*Predicting Agent-based Financial Time Series Model on Lattice Fractal with Random Legendre Neural Network.*	*Soft Computing.*
Bezerra and Albuquerque (2017).	*Volatility Forecasting via SVR-GARCH With Mixture of Gaussian Kernels.*	*Computational Management Science.*
Mo and Wang (2017).	*Return Scaling Cross-Correlation Forecasting by Stochastic Time Strength Neural Network in Financial Market Dynamics.*	*Soft Computing.*

Table 13: The 10 most recent articles in the compiled database.

Reference	Title	Journal
Kumar and Thenmozhi (2014).	*Forecasting Stock Index Returns Using ARIMA-SVM, ARIMA-ANN, and ARIMA-Random Forest Hybrid Models.*	*International Journal of Banking, Accounting and Finance.*
Ballings et al. (2015).	*Evaluating Multiple Classifiers for Stock Price Direction Prediction.*	*Expert Systems with Applications.*
Al Nasseri et al. (2015).	*Quantifying Stocktwits Semantic Terms' Trading Behavior in Financial Markets: An Effective Application of Decision Tree Algorithms.*	*Expert Systems with Applications.*
Gorenc Novak and Velušček (2016).	*Prediction of Stock Price Movement Based on Daily High Prices.*	*Quantitative Finance.*
Chiang et al. (2016).	*An Adaptive Stock Index Trading Decision Support System.*	*Expert Systems with Applications*
Zhong and Enke (2017).	*Forecasting Daily Stock Market Return Using Dimensionality Reduction.*	*Expert Systems with Applications.*
Thawornwong et al. (2003).	*Neural Networks as a Decision Maker for Stock Trading: A Technical Analysis Approach.*	*International Journal of Smart Engineering System Design.*
Cao et al. (2005).	*A Comparison Between Fama and French's Model and Artificial Neural Networks in Predicting the Chinese Stock Market.*	*Computers and Operations Research.*
Ang and Quek (2006).	*Stock Trading Using RSPOP: A Novel Rough Set-Based Neuro-Fuzzy Approach.*	*IEEE Transactions on Neural Networks.*
Li and Kuo (2008).	*Knowledge Discovery in Financial Investment for Forecasting and Trading Strategy Through Wavelet-Based SOM Networks.*	*Expert Systems with Applications.*
Chang and Fan (2008).	*A Hybrid System Integrating a Wavelet and TSK Fuzzy Rules for Stock Price Forecasting.*	*IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews.*
Yu et al. (2009).	*Evolving Least Squares Support Vector Machines for Stock Market Trend Mining.*	*IEEE Transactions on Evolutionary Computation.*
Huang and Tsai (2009).	*A Hybrid SOFM-SVR with a Filter-Based Feature Selection for Stock Market Forecasting.*	*Expert Systems with Applications.*
Atsalakis and Valavanis (2009).	*Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods.*	*Expert Systems with Applications.*
Tsai and Hsiao (2010).	*Combining Multiple Feature Selection Methods for Stock Prediction: Union, Intersection, and Multi-Intersection Approaches.*	*Decision Support Systems.*
Kara et al. (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.*
Rodríguez-González et al. (2011).	*CAST: Using Neural Networks to Improve Trading Systems Based on Technical Analysis by Means of the RSI Financial Indicator.*	*Expert Systems with Applications.*
Wang et al. (2012).	*Stock Index Forecasting Based on a Hybrid Model.*	*Omega.*
Hájek et al. (2013).	*Forecasting Stock Prices Using Sentiment Information in Annual Reports - A Neural Network and Support Vector Regression Approach.*	*WSEAS Transactions on Business and Economics.*
Chen et al. (2014).	*Modeling Fitting-Function-Based Fuzzy Time Series Patterns for Evolving Stock Index Forecasting.*	*Applied Intelligence.*

Table 14: The 20 articles with the greatest bibliographic coupling among all the articles searched.

Reference	Title	Journal
Atsalakis and Valavanis (2009).	*Surveying Stock Market Forecasting Techniques–Part II: Soft Computing Methods.*	*Expert Systems with Applications.*
Box et al. (2015).	*Time Series Analysis: Forecasting and Control* (Book).	*John Wiley & Sons (Publisher).*
Chang et al. (2009).	*A Neural Network with a Case Based Dynamic Window for Stock Trading Prediction.*	*Expert Systems with Applications.*
Chen et al. (2003).	*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index.*	*Computers & Operations Research.*
Hornik et al. (1989).	*Multilayer Feedforward Networks are Universal Approximators.*	*Neural Networks.*
Huang et al. (2005).	*Forecasting Stock Market Movement Direction with Support Vector Machine.*	*Computers & Operations Research.*
Kim and Han (2000).	*Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index.*	*Expert Systems with Applications.*
Kimoto et al. (1990).	*Stock Market Prediction System with Modular Neural Networks.*	*International Joint Conference on Neural Networks.*
Tay and Cao (2001).	*Application of Support Vector Machines in Financial Time Series Forecasting.*	*Omega.*
Zhang et al. (1998).	*Forecasting with Artificial Neural Networks: The State of the Art.*	*International Journal of Forecasting.*

Table 15: The 10 articles with the greatest co-citation relationship among those searched.

Reference	Title	Journal
Kamstra and Donaldson (1996).	*Forecast Combining with Neural Networks.*	*Journal of Forecasting.*
Donaldson and Kamstra (1999).	*Neural Network Forecast Combining with Interaction Effects.*	*Journal of the Franklin Institute.*
Pai and Lin (2005).	*A Hybrid ARIMA and Support Vector Machines Model in Stock Price Forecasting.*	*Omega.*
Atsalakis and Valavanis (2009).	*Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods.*	*Expert Systems with Applications.*
Kara et al. (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.*
Patel et al. (2015).	*Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques.*	*Expert Systems with Applications.*
Laboissiere et al. (2015).	*Maximum and Minimum Stock Price Forecasting of Brazilian Power Distribution Companies Based on Artificial Neural Networks.*	*Applied Soft Computing Journal.*
Chen and Chen (2016).	*An Intelligent Pattern Recognition Model for Supporting Investment Decisions in Stock Market.*	*Information Sciences.*
Zhong and Enke (2017).	*Forecasting Daily Stock Market Return Using Dimensionality Reduction.*	*Expert Systems with Applications.*
Chen et al. (2017).	*A Feature Weighted Support Vector Machine and K- Nearest Neighbor Algorithm for Stock Market Indices Prediction.*	*Expert Systems with Applications.*

Table 16: Articles that are part of the main path of the literature searched.

Reference	Market/s	Asset/s	Predictive Variable/s	Prediction/s	Main Method/s	Performance Measure/s
Al Nasseri et al. (2015).	USA.	Index.	Text.	Direction.	Analysis of sentiment.	Return.
Ang and Quek (2006).	Singapore.	Stocks.	TA.	Prices.	Neural networks.	Return.
Armano et al. (2005).	USA, Italy.	Indices.	TA.	Prices.	Neural networks, GA.	Sharpe rate.
Ballings et al. (2015).	Europe.	Stocks.	Fundamentalist.	Direction.	Neural networks, SVM, kNN, and RF.	AUC.
Barak et al. (2017).	Iran	Stocks.	Fundamentalist.	Return and risk.	Neural networks, SVM, decision trees.	Accuracy.
Bezerra and Albuquerque (2017).	Brazil, Japan.	Indices.	Prices.	Volatility.	SVR, GARCH.	MAE, RMSE.
Cao et al. (2005).	China.	Stocks.	Fundamentalist.	Return.	Neural networks, CAPM.	MAD, MAPE, MSE.
Chang and Fan (2008).	Taiwan.	Index.	TA.	Prices.	kNN, DWT, fuzzy logic.	MAPE.
Chang et al. (2009).	Taiwan.	Stocks.	TA.	Direction.	Neural networks, CBR.	Return.
Chen et al. (2003).	Taiwan.	Index.	Fundamentalist.	Return.	Neural networks, GMM.	Return.
Chen et al. (2014).	Taiwan, Hong Kong.	Indices.	TA.	Prices.	Fuzzy logic.	RMSE.
Chen and Chen (2016).	USA, Taiwan.	Indices.	TA.	Return.	Pattern recognition.	Return.
Chen et al. (2017).	China.	Indices.	TA.	Direction.	SVM, kNN.	MAPE, RMSE, AUC.
Chiang et al. (2016).	Multiple.	Indices.	TA.	Direction.	Neural networks.	Accuracy, return.
Enke and Thawornwong (2005).	USA.	Index.	Fundamentalist.	Direction.	Neural networks.	RMSE.
Fernandez-Rodrıguez et al. (2000).	Spain.	Index.	Prices.	Direction.	Neural networks.	Accuracy, Sharpe rate.
Gorenc Novak and Velušček (2016).	USA.	Stocks.	TA.	Direction.	SVM.	Return, Sharpe rate.
Hájek et al. (2013).	USA.	Stocks.	Fundamentalist.	Return.	Neural networks, SVR, analysis of sentiment.	MSE.
Hassan et al. (2007).	USA.	Stocks.	Prices.	Prices.	Neural networks, GA.	MAPE.

to be continued.

Reference	Market/s	Asset/s	Predictive Variable/s	Prediction/s	Main Method/s	Performance Measure/s
Huang and Tsai (2009).	Taiwan.	Index.	TA.	Prices.	SVR.	MSE, MAE, MAPE.
Kara et al. (2011).	Turkey.	Index.	TA.	Direction.	Neural networks, SVM.	Accuracy.
Kamstra and Donaldson (1996).	Multiple.	Indices.	Prices.	Volatility.	Neural Networks.	MSE, MAE.
Kim and Han (2000).	Korea.	Index.	TA.	Direction.	Neural networks, GA.	Accuracy.
Kimoto et al. (1990).	Japan.	Index.	Fundamentalist.	Direction.	Neural networks.	MAPE.
Krauss et al. (2017).	USA.	Stocks.	Prices.	Returns.	Neural networks, RF, decision trees.	Return, Sharpe rate.
Laboissiere et al. (2015).	Brazil.	Stocks.	Indices.	Maximums, minimums.	Neural networks.	MAE, MAPE, RMSE.
Leigh et al. (2002).	USA.	Index.	Prices, volume.	Prices.	Neural networks, GA.	Return.
Leung et al. (2000).	USA, United Kingdom, Japan.	Indices.	Fundamentalist.	Return.	Neural networks, LDA, regressions.	Return.
Li and Kuo (2008).	Taiwan.	Indices.	Prices.	Prices.	DWT, SOM.	MSE, MAE.
Mo and Wang (2017).	China, USA.	Indices.	Prices.	Correlation.	Neural networks.	MAE, RMSE, MAPE.
Oliveira et al. (2017).	Multiple.	Indices.	Text.	Return, volume, volatility.	Neural networks, SVM, RF.	MAE.
Pai and Lin (2005).	USA.	Stocks.	Prices.	Prices.	SVM.	MAE, MAPE, MSE, RMSE.
Pan et al. (2017).	USA.	Index.	Fundamentalist, prices.	Prices.	SVM.	RMSE, MAE.
Patel et al. (2015).	India.	Indices, stocks.	TA.	Direction.	Neural networks, SVM, RF, NB.	Accuracy.
Pei et al. (2017).	China.	Index.	Prices.	Mean of the prices.	Neural networks.	RMSE, MAE, MAPE.
Rodríguez-González et al. (2011).	Spain.	Index, stocks.	TA.	Direction.	Neural networks.	Accuracy.
Schumaker and Chen (2009).	USA.	Stocks.	News.	Prices.	Textual analysis, SVR.	MSE.
Thawornwong et al. (2003).	USA.	Stocks.	TA.	Direction.	Neural networks.	Return, Sharpe rate.
Tsai and Hsiao (2010).	Taiwan.	Stocks.	Fundamentalist.	Direction.	Neural networks.	Accuracy.
Tsaih et al. (1998).	USA.	Index.	TA.	Direction.	Neural networks.	Accuracy.

to be continued.

Reference	Market/s	Asset/s	Predictive Variable/s	Prediction/s	Main Method/s	Performance Measure/s
Wang (2002).	Taiwan.	Stocks.	Prices.	Prices.	Fuzzy logic.	Accuracy.
Wang (2003).	Taiwan.	Stocks.	Prices.	Prices.	Fuzzy logic.	Accuracy.
Wang et al. (2012).	China, USA.	Indices.	Prices.	Prices.	Neural networks, GA.	Accuracy, MAE, RMSE, MAPE.
Weng et al. (2017).	USA.	Stocks.	TA, text.	Direction.	Neural networks, SVM, decision trees.	Accuracy, AUC, F-measure.
Yan et al. (2017).	China.	Index.	Prices.	Prices.	Neural networks.	MAE, MAPE, MSE.
Yoon et al. (1993).	USA.	Stocks.	Fundamentalist.	Return.	Neural networks, LDA.	Accuracy.
Yu et al. (2009).	USA.	Indices.	Fundamentalist, TA.	Direction.	SVM, GA.	Accuracy.
Zhang et al. (2017).	USA.	Indices.	Prices.	Prices.	kNN, ARIMA.	MAPE, MSE.
Zhong and Enke (2017).	USA.	Index.	Fundamentalist.	Direction.	Neural networks.	MSE, Sharpe rate.
Abu-Mostafa and Atiya (1996).	FOREX.	Currency.	Fundamentalist, "hints", TA.	Direction.	Neural networks.	Return.
Donaldson and Kamstra (1999).	USA.	Index.	Prices.	Return.	Neural networks, GARCH.	RMSE, MAE.
Enke and Thawornwong (2005).	USA.	Index.	Fundamentalist.	Direction.	Neural networks.	RMSE.
Huang et al. (2005).	USA, Japan.	Indices, currency.	Indices, currency	Direction.	SVM, neural networks, LDA.	Accuracy.
Kim (2003).	Korea.	Index.	TA.	Direction.	SVM, neural networks.	Accuracy.
Kumar and Thenmozhi (2014).	India.	Index.	Returns.	Return.	Neural networks, SVM, RF, ARIMA.	Accuracy, MAE, RMSE.
Tay and Cao (2001).	USA.	Indices.	Prices, TA.	Prices.	Neural networks, SVM.	MAE, MSE.
Thawornwong and Enke (2004).	USA.	Index.	Fundamentalist.	Direction.	Neural networks.	RMSE.

Table 17: Classification of the reviewed articles about financial market prediction using machine learning techniques.
Note: AUC: Area Under the receiver operating characteristic Curve; GMM: Generalized Methods of Moments; MAD: Mean Absolute Deviation.

Main Method	Number of References	References
Neural Networks.	42	Ang and Quek (2006), Armano et al. (2005), Ballings et al. (2015), Barak et al. (2017), Cao et al. (2005), Chang et al. (2009), Chen et al. (2003), Chiang et al. (2016), Enke and Thawornwong (2005), Fernandez-Rodrıguez et al. (2000), Hájek et al. (2013), Hassan et al. (2007), Kara et al. (2011), Kamstra and Donaldson (1996), Kim and Han (2000), Kimoto et al. (1990), Krauss et al. (2017), Laboissiere et al. (2015), Leigh et al. (2002), Leung et al. (2000), Mo and Wang (2017), Oliveira et al. (2017), Patel et al. (2015), Pei et al. (2017), Rodríguez-González et al. (2011), Thawornwong et al. (2003), Tsai and Hsiao (2010), Tsaih et al. (1998), Wang et al. (2012), Weng et al. (2017), Yan et al. (2017), Yoon et al. (1993), Zhong and Enke (2017), Abu-Mostafa and Atiya (1996), Donaldson and Kamstra (1999), Enke and Thawornwong (2005), Huang et al. (2005), Kim (2003), Kumar and Thenmozhi (2014), Tay and Cao (2001), Thawornwong and Enke (2004), Lahmiri (2014a), Lahmiri and Boukadoum (2015).
SVM/SVR.	20	Ballings et al. (2015), Barak et al. (2017), Bezerra and Albuquerque (2017), Chen et al. (2017), Gorenc Novak and Velušček (2016), Hájek et al. (2013), Huang and Tsai (2009), Kara et al. (2011), Oliveira et al. (2017), Pai and Lin (2005), Pan et al. (2017), Patel et al. (2015), Schumaker and Chen (2009), Weng et al. (2017), Yu et al. (2009), Huang et al. (2005), Kim (2003), Kumar and Thenmozhi (2014), Tay and Cao (2001), Lahmiri (2014b).
RF/Decision Trees.	7	Ballings et al. (2015), Barak et al. (2017), Krauss et al. (2017), Oliveira et al. (2017), Patel et al. (2015), Weng et al. (2017), Kumar and Thenmozhi (2014).
Sentiment/Text Analysis.	5	Al Nasseri et al. (2015), Hájek et al. (2013), Schumaker and Chen (2009), Weng et al. (2017), Oliveira et al. (2017).
kNN.	4	Ballings et al. (2015), Chang and Fan (2008), Chen et al. (2017), Zhang et al. (2017).
ARIMA/GARCH.	4	Bezerra and Albuquerque (2017), Donaldson and Kamstra (1999), Zhang et al. (2017), Kumar and Thenmozhi (2014).
Fuzzy Logic.	4	Chang and Fan (2008), Chen et al. (2014), Wang (2002), Wang (2003).
LDA.	3	Leung et al. (2000), Yoon et al. (1993), Huang et al. (2005).
NB.	1	Patel et al. (2015).

Table 18: Main forecasting techniques applied by each reviewed reference.

Main Method	Number of References	References
GA.	11	Armano et al. (2005), Bezerra and Albuquerque (2017), Hassan et al. (2007), Kim and Han (2000), Leigh et al. (2002), Wang et al. (2012), Yu et al. (2009), Donaldson and Kamstra (1999), Göçken et al. (2016), Chen and Chen (2016), Tsai and Hsiao (2010).
DWT.	6	Li and Kuo (2008), Chang and Fan (2008), Chiang et al. (2016), Ortega and Khashanah (2014), Xiao et al. (2013), Lahmiri (2014a).
PCA.	3	Zhong and Enke (2017), Son et al. (2012), Tsai and Hsiao (2010).
PSO.	3	Yan et al. (2017), Chiang et al. (2016), Xiao et al. (2013).
SOM.	2	Li and Kuo (2008), Cao (2003).

Table 19: Main optimization techniques applied by each reviewed reference.

Author Contributions Section

Bruno Miranda Henrique (BMH), Vinicius Amorim Sobreiro (VAS) and Herbert Kimura (HK) participated in the development of the research. The first author conducted the study and the results were discussed initially with VAS and HK. Following the three authors developed the initial version of the manuscript. Then, VAS revised and improvement in the paper. Finally, all authors read and approved the final manuscript.