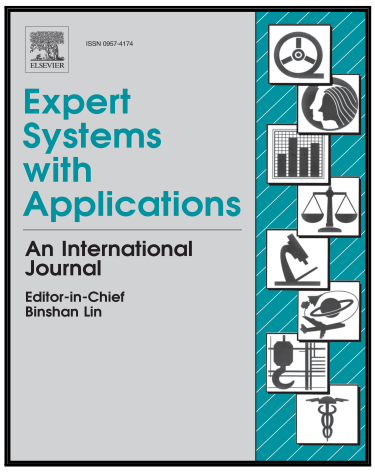

Dynamics of firm financial evolution and bankruptcy prediction

Philippe du Jardin

PII: S0957-4174(17)30025-8
DOI: 10.1016/j.eswa.2017.01.016
Reference: ESWA 11064

To appear in: *Expert Systems With Applications*

Received date: 12 May 2016
Revised date: 22 December 2016
Accepted date: 18 January 2017

Please cite this article as: Philippe du Jardin, Dynamics of firm financial evolution and bankruptcy prediction, *Expert Systems With Applications* (2017), doi: 10.1016/j.eswa.2017.01.016

Highlights

- We estimate the degree of stability of firm financial health over several years
- We typify firms sharing the same level of stability using a quantization method
- We design models that fit each type
- We compare the performance of these models to that of traditional ones
- Our method improves mid-term forecasts when the horizon is higher than 2 years

Dynamics of firm financial evolution and bankruptcy prediction

Philippe du Jardin[a,1]

[a]*Edhec Business School, 393 Promenade des Anglais, BP 3116, 06202 Nice Cedex 3, France*

Abstract

The optimal forecasting horizon of bankruptcy prediction models is usually one year. Beyond this point, their accuracy decreases as the horizon recedes. However, the ability of models to provide good mid-term forecasts is an essential characteristic for financial institutions due to prudential reasons. This is why we have studied a method of improving their forecasts up to a 5-year horizon. For this purpose, we propose to quantize how firm financial health changes over time, typify these changes and design models that fit each type. Our results show that, whatever the modeling technique used to design prediction models, model accuracy can be significantly improved when the horizon exceeds two years. They also show that when our method is used in combination with ensemble-based models, model accuracy is always improved whatever the forecasting horizon, when compared to traditional models used by financial institutions. The method we propose in this article appears to be a reliable solution that makes it possible to solve a real problem most models are unable to overcome, and it can therefore help financial companies comply with the current recommendations made by the Basel Committee on Banking Supervision. It also provides the scientific community (which is interested in designing reliable failure models) with insights about how the evolution of firms' financial situations over time can be modeled and efficiently used to make forecasts.

Keywords: decision support systems, bankruptcy prediction, forecasting horizon

1. Introduction

Bankruptcy prediction models used by financial institutions present a common characteristic: they accurately forecast the fate of firms solely when the horizon of the prediction

*Corresponding author

Email address: philippe.dujardin@edhec.edu (Philippe du Jardin)

[1]Tel.: +33 (0)4 93 18 99 66 - Fax: +33 (0)4 93 83 08 10

does not exceed 1 year on average. Beyond this point, their accuracy worsens all the more as the horizon recedes. All studies that have been conducted on this issue, since that by Altman (1968) to that by Geng et al. (2015), show that this is the case. However, model ability to correctly forecast the fate of companies at a mid-term horizon is essential. Indeed, these models are used by banks to decide whether a loan should be granted based on the probability that a debtor will not reimburse its debt. And most loans that are granted to firms are mid- or long-term loans. In this regard, the European Central Bank estimated, in a recent note (European Central Bank, 2014), that less than 20% of these loans have a maturity lower than 3 years, about 35% have a maturity that ranges between 4 and 5 years and 20% between 6 and 10 years. A model must therefore be able to estimate the robustness of a company, not solely over the first year when it has contracted a loan, but also throughout the whole life of this loan.

The main weakness of traditional models lies in the fact that the maturity of the loans contracted by companies does not match the period during which the risk of default is estimated with very good accuracy. A creditor may decide to lend money based on the estimation of a risk calculated over a short period, although the risk may occur beyond this period with a much higher probability than that estimated by a model. It is precisely for this reason that the Basel Committee on Banking Supervision (Basel Committee on Banking Supervision, 2009) has strongly recommended that banks estimate their risk over the entire duration of their lendings. Nevertheless, attempts to improve model accuracy at a mid-term horizon (5 years) have generally been unsuccessful, except in rare circumstances.

The very first models, that were designed after that by Altman (1968), were all based on the use of a single classification rule and data that characterized companies over a single moment of their life. They relied on the assumption that firm history has no influence on their future and that a single measure of their financial health is sufficient to forecast their possible bankruptcy. However, this assumption does not hold. We know that some firms that may appear to be in a very bad situation have accumulated, over time, a sort of resilience that makes them able to survive while nothing suggests they will. We also know that some others may appear to be in a very good shape at a given moment, but their history shows that they have many weaknesses that will eventually lead them to bankruptcy (Miller &

Friesen, 1977; D'Aveni, 1989). It is precisely for all these reasons that the accuracy of single models decreases as the horizon of the prediction increases.

To overcome such limitations, different ways have been explored. Some authors used data measured over several consecutive years in conjunction with either traditional classification methods (Altman et al., 1977; Betts & Belhoul, 1987; Dambolena & Khoury, 1980) or survival methods (Gepp & Kumar, 2008). Some used methods where the model no longer relies on a unique classification rule, but on a set of rules designed using techniques such as bagging or boosting (Sun et al., 2011; Huang et al., 2012) or techniques where rules are built with different classification techniques (Geng et al., 2015). Others used multi-rule models where each rule is *a priori* specialized in a certain region of the decision space (du Jardin, 2015). All these works show that multi-period models tend to offer short-term predictions (1 year) that are more accurate than those achieved with single-period models, but without providing better mid-term predictions. The same conclusion holds for models that rely on a set of rules: at a short-term horizon, they are more accurate than single-rule models, but not at a mid-term horizon. However, models that are *a priori* specialized by region of the decision space provide better mid-term forecasts than traditional single-rule models do, but not beyond a 3-year horizon and without improving model accuracy at a 1-year horizon.

One conclusion can be drawn from these results: a correct estimation of the decision boundary between firms to be classified relies both on the way firm history is taken into account and on the diversity of classification rules. This is the reason why we studied a method to improve mid-term forecasts based on a new way of producing model diversity. It relies on results that were drawn from research works carried out in the field of organizational ecology and that showed that the more a firm experienced organizational changes, the higher its probability of bankruptcy (Amburgey et al., 1993). The method we propose in this article aims to measure the degree of organizational perturbation such changes may incur, typify firms depending on their degree of perturbation and design models that fit each type of firm. To assess the performance of these models, their accuracy is estimated using different samples and it is compared to that of traditional models.

2. Literature review

The accuracy of failure models at a horizon that exceeds 1 year can be considered a real challenge for financial institutions, especially for prudential reasons, as underlined in a note by the Basel Committee on Banking Supervision (Basel Committee on Banking Supervision, 2009) mentioned previously. The literature that has analyzed model ability to offer accurate forecasts up to a 3-year horizon is relatively wide. Beyond this point, it is rather thin. Table 1 gives an overview of the main studies that have calculated model accuracy up to 4 or 5 year horizons. When we estimate the average correct classification rates at a 1-year horizon and at a 5-year horizon, using the sample size of each study, we notice that this rate ranges from 85% to 69.5%, that is to say a difference of more than 15%. This not only illustrates how difficult it is to provide reliable mid-term forecasts, but also how small the discrepancy is between the very first models based on single rules (which were designed in the 1960s and 1970s), and the latest models based on ensemble techniques. So as to analyze the real weaknesses of past and current models, we have grouped them into four categories. These categories are illustrated in Tables 2, 3, 4 and 5 using the studies that are presented in Table 1. Table 2 presents the issue of each study and their possible theoretical foundations; Table 3 depicts the different modeling methods and the type of models and variables that were used, but also the way explanatory variables were chosen and selected; Table 4 shows the characteristics of the data used to estimate model parameters; and Table 5 shows the methods and criteria used to estimate model performance. Within each table, studies are grouped into categories, one per panel.

The first category (Panels A in Tables 2, 3, 4 and 5) is made up of the historical form of failure models where the classification rule is unique and where the variables that characterize firm activity are measured solely over one period of time (single-period single-rule models). This was the method used by Altman to design his very first model, nearly 50 years ago (Altman, 1968). All these works lie at the root of current banking models and have largely contributed to shape the history of bankruptcy models. They made it possible to assess the conditions under which models should be estimated and validated but also the criteria that should be used to assess their generalization ability and their robustness. They empirically grounded the usefulness of discrimination techniques and assessed the strengths and weak-

nesses of different criteria used to evaluate model performance. They demonstrated that, on the whole, linear models are rather accurate, but often less accurate than non-linear ones. Even if the former are less effective than the latter, they are nonetheless more robust, as shown by different studies conducted by Banque de France (Bardos, 2007), the French central bank. They highlighted that financial ratios are by far the best bankruptcy predictors: they are indeed well standardized, within a given accounting framework, easily accessible and make it possible to design reliable and accurate models. Nevertheless, financial ratios do not embody all causes or symptoms of financial failure and, for this reason, they can sometimes be well complemented by other types of variables that represent dimensions of firm management other than the financial one. They also indicated that most models do not rely on any conceptual or theoretical justification but solely on empirical findings; as a consequence, the robustness and generalization ability of many models are often unstable (Balcaen & Ooghe, 2006). Finally, they suggested that the traditional way of designing models, where a model is solely based on a single classification rule and single-period data, represents, by construction, their main limit. Due to the uniqueness of the rule, a model captures a probability of bankruptcy using a sort of distance between the financial situation of a given company and a standard financial bankruptcy situation. And because of the uniqueness of the period, a model relies on data that characterize firms over a single moment of their existence. However, bankruptcy cannot be embodied in a single reference state and is rarely the consequence of a sudden event; it is often the result of a long and dynamic process that occurs over time (Dimitras et al., 1996) and that results in firms presenting very different financial situations before failure. Indeed, many firms that, at a given moment of their life, may appear to be in a very fragile situation, are finally able to survive, while others that seem to be in rather good shape eventually fail, and sometimes very quickly (D'Aveni, 1989). We know that history explains that two firms with similar instantaneous probabilities of failure calculated using a single-period model may have very different real probabilities of going bankrupt. The fact that models are not able to take into account the temporal dimension of bankruptcy lies partly at the root of their inability to offer accurate forecasts at a 5-year horizon.

The second category (Panels B in Tables 2, 3, 4 and 5) corresponds to single-rule models that

try to overcome the weaknesses of the latter using multi-period data, that is to say data that are measured over several years. Studies which are at the root of these models rely on the widespread idea, grounded in organizational theory, that the temporal dimension is a fundamental explanatory variable of bankruptcy. This is the reason why they try to find ways to embody this dimension using historical data, that is to say data that characterized the evolution of firms' financial situations over time, and also using particular modeling methods. As financial ratios were long considered to be good predictors of failure, most research focused on calculating indicators that were supposed to properly capture variations of financial variables over time. And since traditional modeling methods were ill-suited to taking time into account, new techniques such as survival methods or hazard models were then tested. The main contribution of such works lies precisely in these attempts. The results, however, were not entirely conclusive. Taking time into account, both through historical data and ad hoc modeling methods, helped to improve forecasts compared to those achieved with previous methods, but did not efficiently reduce the difference between short-term (1 year) and mid-term (5 years) predictions. Presumably, if the idea that time is a good explanatory factor seems perfectly right, the way this dimension is embodied within models does not appear to be relevant. Indeed, variables used by models essentially represent measures of ratios over several years, or measures of variations of such ratios over a given period of time, but also statistical measures (mean, standard deviation, standard error...) of ratios estimated over different periods; however, all these measures lack robustness because their discriminating power is rather unstable over time (Bardos, 2008). Moreover, the modeling techniques that were used do not fundamentally change the way bankruptcy is represented; the classification rule remains unique and as a consequence is not able to account for the complexity of the failure process. Therefore, the principles that lie at the root of all modeling attempts are rather inappropriate. Incidentally, what we have just noticed is partly confirmed by some research works which have shown that models that do not directly use measures of firm financial evolution, but rather employ them after a process where these measures have been typified through the concept of "trajectories", manage to provide forecasts that are nearly

as accurate at a 1-year horizon as they are at a 3-year horizon (du Jardin & Séverin, 2011)[2]. This clearly shows that the way time is embodied in a model is not neutral and has an influence on model performance. But, here as well, this method has a limit because beyond a 3-year horizon, model accuracy does not exceed that of traditional models.

The third includes multi-rule single-period models where a set of models are combined so as to make forecasts (Panels C in Tables 2, 3, 4 and 5). They rely on techniques such as bagging or boosting, where classifiers are designed with the same modeling method, or on techniques where classifiers are built using different methods. Unlike previous techniques, where rules are unique, such techniques rely on a variety of rules so as to better embody the decision boundary between classes. Thus, instead of relying on a single model that has a global expertise on this boundary, they rely on a set of models where each of them has a particular expertise on a portion of the boundary. Their results are rather good in certain circumstances, depending especially on the sensitivity of a model to sampling variations or to sample size. Their main weakness, from a practical point of view, remains their complexity as they are absolutely not understandable by a financial analyst, as are single-rule models estimated using non-linear methods (neural networks, support vector machines, etc.). Indeed, some legal frameworks require financial institutions to provide explanations when they decide not to grant a loan to one of their customers. Therefore, if an analyst uses a bankruptcy models as a decision-making tool, the model must be interpretable. However, ensemble models are not. Moreover, on the whole, such models are more accurate at a short-term horizon than those presented above (Alfaro et al., 2008; Marques et al., 2012), but their accuracy also worsens as the horizon of a prediction increases (Sun et al., 2011; Huang et al., 2012). These results indicate that ensemble techniques do not lead to forecasts at horizons between 3 and 5 years that are more accurate than those achieved with traditional single-period models.

[2]In this study, firms in a sample are individually characterized by a measure of evolution of their financial health over time: this measure is called "trajectory" and represents the way each company moves in a space at risk over six consecutive years. Then, all individual trajectories are typified and grouped within a few meta-trajectories. Once these meta-trajectories are estimated, each of them is given a label of the class (failed or non-failed) for which a meta-trajectory is considered a prototype. Finally, forecasts are performed by comparing the individual trajectory of a given company to all meta-trajectories: a firm is then classified in the same group as that of the meta-trajectory that is the closest to its individual trajectory.

The fourth and last category (Panels D in Tables 2, 3, 4 and 5) is made up of multi-rule multi-period models that are estimated with conventional (Geng et al., 2015) or specific methods such as that proposed by du Jardin (2015)[3], where models are built based *a priori* on a partitioning of the decision space. In the first case, the error decreases as the horizon of the prediction recedes, and the mid-term error is far higher than the short-term one, and in the second case, the error is rather stable up to a 3-year horizon, but not beyond. Such models share the same strengths and weaknesses as those of all models previously presented, and the fact that they use multi-period data in conjunction with a set of classification rules does not especially improve mid-term forecasts. However, these models provide some indications about the way the time dimension might be embodied. They reinforce the idea that usual multi-period data are not a good means to measure the influence of firm history on its probability of failure. And as a consequence, they suggest that if one changes the way these data are used, as that proposed by du Jardin (2015), then one may improve model performance, even if this improvement does not exceed 3 years.

The results achieved with the four categories of models that have just been presented show that a good estimation of the boundary between classes relies on model multiplicity, regardless of the method used to embody this multiplicity, hence on a certain way of approximating the different financial situations firms may experience before failing. However, the contribution of data multiplicity seems to depend more on the method used to account for such multiplicity than on the data; indeed, it seems to be more efficient to use multi-period data so as to design groups and then design models that fit each group, than to use such data to directly design models.

Nevertheless, none of these methods are able to provide accurate results at a mid-term horizon (5 years) or results between a 1 and a 5-year horizon that are not too different. This is why we studied a way to improve single and multi-rule models in an attempt to take advantage of data that characterize firm history, but using such data in a completely different way than that used so far. For this purpose, our method relies on some research works that have

[3]This study is based on that by du Jardin & Séverin (2011). Firms are also categorized using trajectories, but this time they are calculated over a short period of 3 years. A few meta-trajectories are then estimated, and for each meta-trajectory, a forecasting model is designed.

demonstrated that there is a relationship between organizational change and firm failure. Some studies have shown that, over time, firms experience periods of organizational stability followed by periods of organizational change, and the more a firm experiences changes, the higher its probability of failure and the shorter the elapsed time since the occurrence of other changes (Amburgey et al., 1993). Other studies have also shown that the risk of failure for companies that experienced a blueprint change is much larger than that of firms that did not (Hannan, 2005). Further studies have found that the impact of organizational change on the death rate of firms is higher at the beginning of their life cycle than at the end (Singh et al., 1986). We hypothesize that these changes, if they result in firm failure, may necessarily distort firm financial accounts, since early signals of bankruptcy can be found in those accounts. Given that such changes can be considered symptoms of failure that may appear long before bankruptcy occurs, one can imagine they can be used to design models that would be able to detect the occurrence of a bankruptcy several years in advance. We also hypothesize that all firms do not respond identically to organizational changes, that there must exist groups of companies, each of which is characterized by a specific degree of variation of its financial health in response to such changes, and that within a given group some firms are better able to withstand the impacts of these changes than others. Therefore, we propose a method that is rooted on the estimation of these impacts: within a sample of firms, we estimate the degree of stability of firm financial health over several years, then we group firms sharing the same level of stability and we build as many models as there are groups, using traditional classification techniques (discriminant analysis...) and ensemble techniques (bagging...). The forecasts of these models are then compared to those that are estimated with the same modeling methods, but without taking into account any group of firms.

The current study relies on those by du Jardin & Séverin (2011) and du Jardin (2015), whose models were solely based on the estimation of the general evolution of firms' financial health. Here, we also estimate such an evolution, but we complement this estimation with a measure of the magnitude of annual variations that can be observed within financial accounts, that may affect this evolution, and that reflect organizational changes to some extent.

3. Data, samples and variables

3.1. Samples

Data were collected using the Diane database (bureau Van Dijk), which contains financial statements of French firms that file their annual reports with the French commercial courts, over three periods of time so as to control for the economic environment effect on model accuracy. Within each period, we collected two samples, one to estimate model parameters and one to test model accuracy. Firms were selected so as to estimate an out-of-sample and out-of-time error (Stein, 2007). Thus, none of the firms that belong to a learning sample is part of a test sample, and data from these two samples were gathered with a lag of one year. The proportion of failed and non-failed firms within each sample is the same as that which characterizes the periods we studied, where the average failure rate is slightly lower than 2%. Firms were then selected if at least six consecutive years of financial data were available in the database. We choose six years so as to estimate the changes of their financial health over a period that is sufficiently large and to assess model accuracy at 5 different horizons. Once these conditions were well fulfilled, firms were randomly drawn from the database. Table 6 presents the composition of each sample for each period. Thus, during the first period, the learning sample is made up of 95,910 non-failed firms and 1,920 failed firms, and the test sample is made up of 91,050 non-failed firms and 1,820 failed firms. Within the learning sample, firm status (non-failed vs. failed) was determined in 2003 and financial accounts were selected between 1997 and 2002. Within the test sample, firm status was defined with a lag of 1 year, in 2002, and firm financial accounts were also selected with a lag of 1 year, between 1996 and 2001.

3.2. Variables

Firm financial accounts were used to calculate ratios and these ratios were used to design models. But before choosing these ratios, we determined the financial dimensions that are generally considered in the literature to embody the key factors of bankruptcy. We selected six dimensions: liquidity, turnover, profitability, activity, solvency and financial structure. Then, for each of them, we chose a set of ratios that are known in the literature to present good discrimination ability, both at a 1 and a 3-year horizon, so as to make models as

insensitive as possible to any short-term changes that may occur within the firm economic environment. The ratios are listed in Table 7.

4. Modeling methods

Models were estimated using seven classification techniques that are traditionally used in the literature (discriminant analysis, logistic regression, C4.5 as a decision tree, Cox's model as a survival analysis technique, a feedforward neural network, an extreme learning machine and a support vector machine) and five ensemble techniques (bagging, boosting with AdaBoost, random subspace, Decorate and rotation forest). All of them are briefly presented below.

4.1. Classification methods

4.1.1. Discriminant analysis

This method is used to classify observations into groups. To classify n firms that are characterized by p variables (x_1, ..., x_p) into two groups, discriminant analysis estimates a set of coefficients (α_1, ..., α_p) that are used to calculate a z score associated to each firm. The estimation is achieved using a method that attempts to maximize the variance between groups while minimizing the variance within groups. Once computed, the score is compared to a threshold that represents the boundary between groups. Depending on whether the score lies below or above the threshold, the firm is classified within one of the two groups. The z score of firm i is defined as: $z_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$

The discrimination ability of this method is optimal when variance-covariance matrices of all groups are equal and when the joint distribution of explanatory variables is multivariate normal (Wald, 1944). These conditions are rarely fulfilled when explanatory variables are financial ratios, but the method is sufficiently robust to account for departure from these conditions.

4.1.2. Logistic regression

Logistic regression can also be used to classify observations. This method estimates a probability that a given observation belongs to a given group. If one considers a binary variable y that represents the probability that a company belongs to the group of firms that

are likely to go bankrupt (if $y = 1$), the logistic regression function can be expressed as follows:

$P(y_i = 1|x_{1i}, x_{2i}, \ldots, x_{pi}) = \frac{1}{1+e^{-(\alpha_0+\alpha_1 x_{1i}+\alpha_2 x_{2i}+\cdots+\alpha_p x_{pi})}}$

where $P(y_i = 1|x_{1i}, x_{2i}, \ldots, x_{pi})$ is the probability that firm i fails given its characteristics measured using variables $x_{1i}, x_{2i}, \ldots, x_{pi}$. The α coefficients are calculated using maximum likelihood estimation. This method is often used to overcome the constraints that discriminant analysis imposes on data distribution and especially when explanatory variables are financial ratios.

4.1.3. C4.5

C4.5 belongs to the family of decision trees. A decision tree is a classification technique that has been often used to forecast bankruptcy (Frydman et al., 1985). A tree can be depicted as a set of branches, each of them being created by a variable that splits the decision space. If the variable to be explained is binary, each split is also binary. A tree is then made up of a set of nodes where each node splits the decision space into two branches. The C4.5 algorithm (Quinlan, 1993) uses two processes to create a tree. First, the tree is grown using a measure of entropy to perform each split and the growth is stopped depending on a stopping criterion. Second, to reduce its size, the tree is pruned by removing branches that provide little discrimination ability. The pruning process is performed with data from a test sample and it is stopped when the accuracy of a given tree has reached a desired minimum. Then, nodes that cannot be split are assigned to a given group (failed or non-failed) and are used to make forecasts.

4.1.4. Cox's model

Cox's model is a survival analysis method used to estimate the time period that will elapse before a particular event occurs, such as a bankruptcy. The method relies on the estimation of a survival function $s(t)$, that represents the probability that a firm manages to survive beyond time t, and a hazard function $h(t)$, which represents the failure rate at time t. With Cox's method, the estimation is achieved as follows:

$h(t|x_{1i}, x_{2i}, \ldots, x_{pi}) = h_0(t)e^{(\alpha_1 x_{1i}+\alpha_2 x_{2i}+\cdots+\alpha_p x_{pi})}$

where $h(t|x_{1i}, x_{2i}, \ldots, x_{pi})$ is the hazard function, measured at time t, of firm i which is

characterized by variables $x_{1i}, x_{2i}, \ldots, x_p$, and $h_0(t)$ the baseline hazards that represent the way the hazard function changes over time. The survival function of a given company is defined as:

$s(t) = e^{-h(t)}$

The α coefficients are estimated using a method similar to that used for logistic regression and the classification of firms is performed using their survival function (i.e. their probability of failure).

4.1.5. Neural network

A neural network makes it possible to calculate the probability of an observation belonging to a given class and hence, the probability of a firm belonging to a group of failed companies. The neural network used in this study is a feedforward network that can be depicted as a directed, loopless graph where each vertex corresponds to a neuron. It is usually made up of three layers of neurons when one wants to perform a classification task: an input layer, a hidden layer and an output layer used to classify observations. Neurons from each layer are connected to all neurons from the previous and/or the following layer and a connection is represented with a weight. Each neuron from the hidden and output layers computes its output using a weighted sum of its inputs and an activation function, most often a hyperbolic tangent or a logistic function. The output o_j of a neuron is given by:

$o_j = f(w_{0j} + \sum_{i=1}^{n_j} w_{ij} x_i)$

where w_{0j} is the bias connected to neuron j, n_j the number of neurons connected to neuron j, w_{ij} the weight between neuron j and neurons from the previous layer, x_i the outputs of neurons from the previous layer and f the activation function of neurons.

This method relies on a learning process that is used to progressively change the values of all weights so as to elicit a desired response from the network. The changes are performed using a cost descent in the weight space, step by step. Here as well, once the model is designed, the classification of companies is performed using their probability of failure.

4.1.6. Extreme learning machine

Extreme learning machines (ELM) were first proposed by Huang et al. (2006). An ELM is an optimization technique used during the learning process of a neural network such as

the one presented above, an SLFN. With this technique, the weights between the input layer and the hidden layer are randomly drawn but the weights between the hidden layer and the output layer are estimated analytically. This estimation is solely possible if the neuron activation function is infinitely differentiable. This method is extremely faster than usual optimization techniques used to estimate SLFN weights and makes it possible to design models that present a better generalization ability than those estimated with an SLFN model that is estimated with traditional methods.

4.1.7. Support vector machine

A support vector machine (SVM) is a relatively new classification technique grounded in the research by Boser et al. (1992). To perform a classification into two classes, an SVM looks for the hyperplane that maximizes the distance between the nearest observations of these two classes. These observations are called "support vectors" and the distance between them and the hyperplane is called "margin".

With non-linear problems, an SVM seeks for the hyperplane that, in a space whose dimension is greater than that of data (transformed feature space), may lead to a linear separation. But looking for a hyperplane in a high dimensional space can be extremely time consuming. To simplify the search, one uses a kernel function (polynomial, gaussian radial basis...) that makes it possible to reduce the search space. However, finding a linear separation in a transformed feature space is often impossible. This is why the concept of "flexible" margin was created so as to define a hyperplane which tolerates a classification error to be set up *a priori*. The problem is then finding a hyperplane that minimizes this error using a parameter that is intended to affect the trade-off between the search for a fairly wide margin and the search for an acceptable error rate. SVMs are able to handle non-linear relationships between a forecast and a set of explanatory variables, and are as accurate as (and sometimes more than) neural networks.

4.2. Ensemble methods

Ensemble techniques rely on the idea that several models provide an informational gain that is higher than that obtained using a single model. This gain is a function of model diversity (Kuncheva & Whitaker, 2003) and is due to the fact that each model has a particular

expertise in a part of the decision space. Several ensemble techniques exist. Indeed, models can be designed in parallel, without depending on each other, as demonstrated by methods such as bagging (Breiman, 1996) or random subspace (Ho, 1998), and at the same time they can rely on a single modeling method (neural network...) or on different methods, such as stacking (Wolpert, 1992). They can also be built iteratively and depend on each other or not. They can finally be designed using the entire decision space, and in this case diversity will *a posteriori* lead each model to fit a particular region of the decision space, or by fitting *a priori* each rule to a particular region of the decision space. Among all existing methods, we chose those that are commonly used to forecast bankruptcy: bagging, boosting, random subspace, Decorate and rotation forest. These methods are presented below.

4.2.1. Bagging

Bagging was proposed by Breiman (1996). With this technique, each model is designed using a bootstrap sample that is randomly drawn with replacement from an original sample. Once the ensemble is set up, all individual forecasts are combined using a majority vote to form the final forecast. Bagging depends on the number of models that are estimated. This number is problem dependent and must be chosen carefully depending on the problem at hand.

There are many variants of bagging that are based on the use of a single learning sample (nice bagging, subbagging, trimmed bagging...), or on that of both a learning and a test sample (lazy bagging), or on a combination of the principles governing these two techniques. According to Breiman, bagging may decrease the variance of the estimation achieved with the base model. But according to some works carried out later on (Grandvallet, 2001), bagging especially decreases the influence of some particular observations (outliers, extreme values...), and when it decreases this influence, it also decreases the variance of the estimation.

4.2.2. Boosting

Boosting is rooted in the works by Schapire (1990) and Freund (1990). Unlike bagging, boosting designs models iteratively so as to allow them, during a given step, to correctly classify observations that were wrongly classified during the previous steps. The algorithm imposes a constraint on the classifiers using a weight assigned to each observation; the more

an observation is wrongly classified, the larger its weight, so as to increase its influence on the estimation process. The final forecast is performed by all models using a weighted majority voting scheme. Here as well, this method depends on the number of models that are calculated which must be experimentally assessed.

Many versions of boosting have been developed (AdaBoost, TotalBoost, GentleBoost, Madaboost, LPBoost...) in order to deal with some of its weaknesses (sensitivity to noise, computing time...). In our study, we used AdaBoost.

4.2.3. Random subspace

Random subspace (Ho, 1998) designs models by combining rules that are made up of variables chosen at random among an initial set of variables. Once the models are designed, the final forecast is performed using a majority voting scheme. Random subspace also depends on the number of models which must be chosen depending on the issue to be tackled.

Unlike bagging and boosting, very few works have been carried out to create variants of the original technique, but rather to studies that analyzed variable selection issues or different ways of combining random subspace with other ensemble tecnhiques.

4.2.4. Rotation forest

Rodriguez & Kuncheva (2006) have proposed a technique that makes it possible to create an ensemble of models whose estimation is not performed using the original data but a set of features that are extracted from these data. The method consists in splitting the set of variables of a learning sample into several subsets. With each subset, data are bootstrapped, a principal component analysis is run on the bootstrapped data and all principal components are retained. Then, a rotation matrix is created and the principal components are arranged in this matrix in such a way that the components match the position of the feature in the original training set. Finally, the training set is projected on the rotation matrix and a model is estimated using the projected data.

The aim of this method is both to improve the individual accuracy of each model and to enhance model diversity.

4.2.5. Decorate

Decorate (Diverse ensemble creation by oppositional relabeling of artificial training examples) was proposed by Melville & Mooney (2004). This method does not solely rely on original data belonging to an existing dataset to create diversity, but also on artificial data added to the original dataset. These artificial data are generated randomly and are added to a learning sample used to design a set of models. Then, these data are given a group label that differs totally from the label of the group they would have been assigned by the ensemble of models. Ensemble methods that rely on re-sampling or re-weighting schemes may produce limited diversity if the size of the learning sample is small. In such circumstances, Decorate is able to increase the diversity of an ensemble by adding a large amount of artificial data, consequently leading to higher model accuracy than that of other ensemble techniques when the training set is small.

4.3. Self-organizing map

A Self-organizing map, also called Kohonen map, is a clustering method that makes it possible to quantize a set of observations that are measured in a high dimensional space by projecting them onto a low dimensional space, which is most often a two-dimensional map. The resulting space can be considered a simple, condensed and ordered representation of the input space, which can then be used to design groups. More precisely, a map is made up of a set of neurons and each of them is represented by an n-dimensional weight vector. The number of weights is identical to that of the variables used to characterize observations one wishes to quantize with the map. A self-organizing map requires training before being used (Kohonen, 2001). When the learning process is done, if this method is applied to data that characterize sound and unsound firms, it leads to a set of neurons that represent all possible states of financial health and that preserve the topology of the input space. Such a map can then be used to analyze the evolution of the financial health of a sample of companies over time, find those whose evolution is stable or erratic, and then typify particular forms of evolution.

5. Experimental settings

5.1. Variable selection

Before designing models, we selected some variables among those presented in Table 7. We chose those that exhibited a good discrimination ability and whose correlations were as low as possible. The discrimination ability was estimated in two ways. Firstly, we assessed the discrepancies between failed and non-failed firms at a 1, 2 and 3-year horizon, using a Mann-Whitney test since none of the variables were normally distributed. Secondly, using a resampling scheme, we estimated the confidence interval of the quartiles of each variable, at a 1-year horizon with data from non-failed firms, and at a 1, 2 and 3-year horizon with data from failed firms so as to avoid variables that were too sensitive to short-term variations of the economic environment.

For each period, performing calculations with the Mann-Whitney test, we chose the ratios that presented a significant difference at the threshold of 1%, and that at the same time presented a majority of quartile intervals that did not overlap. Then, we analyzed the correlations between variables with the highest discrimination ability, and we finally selected those with correlations that did not exceed 0.6. Table 8 presents, for the sake of economy of space, the quartiles of each variable used to design models with data from 2003; these statistics were calculated using variables that were winsorized and transformed, so as to use data with 0 mean and unit variance. All variables present significant differences between failed and non-failed firms, at the threshold of 1%, and at a 1, 2 and 3-year horizon. Once this selection process was completed for each period, we chose from among the resulting sets, the variables that were finally used to design each type of model. The final selection process is presented below within the sections that explain the way we designed all models.

5.2. Design of traditional models

5.2.1. Single models

With each modeling method, single models were estimated using a specific variable selection technique. We tested several techniques belonging either to the filter or the wrapper category, and using different combinations of the three key-components of any selection procedure: a search method to explore the variable space, an evaluation criterion of the

solutions and a stopping criterion. We chose filter techniques with discriminant analysis, logistic regression and Cox's model, and wrapper techniques with C4.5, the neural network, the extreme learning machine and the support vector machine and, with each technique, we selected a stepwise or a backward search procedure and a stopping criterion based on a statistical test.

We made these choices for two reasons. First, filter techniques used in conjunction with discriminant analysis, logistic regression and Cox's model can easily rely on robust statistics and statistical tests to evaluate different combinations of variables. However, using statistical tests with techniques such as neural networks, support vector machines, extreme learning machines or decision trees is much more difficult since they often require the estimation of a relevance measure distribution using time-consuming bootstrap or cross-validation procedures (Leray & Gallinari, 1998). Second, we ran a set of experiments and, on average, our choices led to the best results.

With discriminant analysis, we used a stepwise search, a Wilk's Lambda as an evaluation criterion and a Fisher F test as a stopping criterion. With logistic regression and the Cox's model, we also used a stepwise search, a likelihood statistic to compare the different subsets and chose the "best" one and a Chi2 to interrupt the search.

With C4.5, variable selection was performed during the pruning process. Trees were grown using 50% of each sample and an entropy measure was used to estimate the heterogeneity of all nodes. Then the pruning process was performed with the remaining 50% of each sample. During this process, the performance of each sub tree was assessed after the removal of a node, and the process was stopped after the removal of all nodes. We then chose the sub tree whose accuracy was statistically close to the highest accuracy and that was made up of the smallest number of variables.

The procedure used with the neural network was twofold. We first determined network architecture, as recommended by Leray & Gallinari (1998). We chose the Levenberg-Marquardt algorithm as an optimization technique and a 3-layer structure for the network with one hidden layer, one output node, one bias per layer and the hyperbolic tangent as an activation function. Then, we ran a set of experiments to assess the number of hidden nodes and network weights. We drew, at random, 100 sets of variables among the final set chosen for

each period. With each set, we tested different sizes of the hidden layer (between 2 and 35) and different learning rates (between 0.05 and 0.50 with a step of 0.01). Model parameters were estimated using 50% of a learning sample and model accuracy was assessed using the remaining 50%. Then, the performance of each network was averaged over the 100 sets of variables. We selected the network that led to the lowest error, then we looked for the networks with error statistically close to the lowest one, and we finally chose among the latter, the network with the smallest number of hidden nodes. Second, once the network architecture was estimated, variables were selected. We used a backward search and an evaluation criterion of the different solutions inspired by weight-pruning methods, as presented in Leray & Gallinari (1998). The selection process was performed until no selected variable remained. When all variables were removed, we chose the subset with error statistically close to the lowest error and made up of the smallest number of variables.

With the extreme learning machine, the size of the hidden layer of each network was assessed using the same procedure as that used to estimate neural networks models. Then variables were selected using a backward search; 50% of each sample were used to estimate a model, and the remaining 50% to assess its accuracy. Variables were removed one at a time and the ELM was retrained after every removal. The variable selection process was performed until no selected variable remained and we chose, among all sets of variables that led to an error statistically close to the lowest error, the subset with the smallest number of variables.

Finally, with the support vector machine, we chose a radial basis function as the kernel function, and we ran a set of experiments to set up model parameters: C (term used to penalize the error) and γ. We tested different values of C and γ ($C = 2^{-5}, 2^{-4}, \ldots, 2^{-10}; \gamma = 2^{3}, 2^{2}, \ldots, 2^{10}$). For this purpose, we drew 100 samples made up of a random number of variables. 50% of each sub sample was used to estimate SVM model parameters and the remaining 50% to assess their performance. All results were averaged and we finally chose the parameter that led to the lowest error. Then, with each learning sample, variables were selected using the same procedure as that used with the extreme learning machine.

5.2.2. Ensemble-based models

We designed as many ensemble models as there are modeling methods and ensemble techniques. Hence, with bagging, we estimated seven sets of models (one with discriminant analysis, one with logistic regression, one with the Cox's method, one with the neural network, one with the decision tree, one with the extreme learning machine and one with the support vector machine) and we did the same with boosting, random subspace, Decorate and rotation forest. So as to determine the number of classifiers to be used with each ensemble technique, we estimated a set of ensemble models where the number of classification rules ranged from 10 and 500. Rule parameters were assessed using 50% of a learning sample and their error was estimated using the remaining 50%. The best results were achieved with ensembles where the number of models ranged from 70 to 145 and we chose 100 models with each technique. With each ensemble method, models were estimated using 150 iterations. Finally, with each model that was designed using a given modeling method, and that was used with bagging, boosting and Decorate, variables were selected in the same way as for single models. With random subspace and rotation forest, variables were chosen by the ensemble technique itself. The parameters for neural networks, extreme learning machines and support vector machines were assessed as they were estimated with single models. The final forecasts were performed using a majority vote with bagging, boosting, random forest and Decorate, and a weighted majority vote with random subspace[4].

5.3. Design of new models

5.3.1. Modeling principles

The models we designed rely *a priori* on a partitioning of each learning sample in as many categories of firms as there are different types (or prototypes) of variation of their financial situation over time. The estimation of theses prototypes was performed with a measure of financial stability. We consider that a firm is financially stable if its exposure to a bankruptcy

[4]Most calculations were performed using Weka (single-model and ensemble-based model estimations and validations) and algorithms that are still embedded or that can be embedded into Weka, and that were developed by third-parties. We also developed a certain number of routines using Eclipse, routines that were then embedded into Weka. All calculations related to the estimation of the Kohonen maps were developed using Visual Basic. And finally, computations of ratios and their analysis were conducted using SPSS.

risk also remains stable. To estimate this measure, we designed a self-organizing map with each learning sample, and with data characterizing firms over the last year for which we collected their financial accounts. This map makes it possible to quantize a set of firms using a space at risk that represents all financial situations these firms may experience. Then, we divided the space at risk into several regions using a clustering technique and we ranked these regions depending on the average of all variables used to design a map. Each region was labeled with an index k ($k = 1, 2 \ldots, n$). Once this ranking was done, we estimated the way firms moved from one region to another over a six-year period. Since the different regions at risk were ranked using an index that represents a sort of scale of financial health, we considered this index a quantitative variable that we could use to assess how the different positions of firms change over time. Each variation of the financial situation of a company over two consecutive years was then estimated using the difference between the index of the region that corresponded to an estimated risk of bankruptcy at time t, and that of the region that corresponded to the same risk at time $t + 1$. Once these differences were calculated, they were quantized using another self-organizing map so as to determine a few prototypes of financial evolution. Each firm was then assigned the prototype that best represented the variation of its own financial health, and finally with each subset of firms that shared the same prototype, we designed a model using the techniques mentioned above. This process is presented below in details.

5.3.2. Design of financial evolution prototypes

Prototypes were estimated with a subset of variables chosen among those that are presented in Table 7. This subset is made up of one variable by financial dimension. We made this choice because we wanted to categorize firms using all dimensions with equal weight. Therefore, we chose, for each dimension and each learning sample, the variable with the larger discrimination ability assessed with the criteria presented in Section 5.1, and whose correlation with other selected variables was lower than 0.6. These variables were used to design one map per sample. Since all samples are made up of a very disproportionate number of failed and non-failed firms, we chose to balance each class before designing the maps so as to properly quantize the minority class. Different techniques are available (He & Gar-

cia, 2009) and can be used to balance data. We chose a technique that relies on the under sampling of the majority class, but that makes it possible to limit the loss of information. For this purpose, we quantized data of non-failed firms using a map that was made up of as many neurons as there were failed firms within a learning sample. Once the majority class was quantized, data of failed firms and those that result from the latter quantization were then quantized together so as to design the desired space at risk. So, with data from 2003, some 95,910 non-failed firms were quantized with a map that was made up of 1,920 neurons (40 per line and 48 per column). Then data that characterize the 1,920 failed firms from 2003 and the 1,920 neurons that have just been estimated were finally quantized so as to assess the space at risk of the first period we studied.

The size of the different maps was estimated as follows. With each learning sample, we first designed a set of maps whose sizes ranged from 50 to 200 neurons. Then, those maps were compared using an index proposed by Kaski & Lagus (1996) that measures both their ability to preserve the continuity of the mapping and to minimize the quantization error, and which is insensitive to their size. We finally chose the map that led to the smallest value of this index.

Once the maps were designed, we divided them into regions at risk. For this purpose, neurons were first labeled with the label of the group (failed vs. non-failed) they were considered prototypes; the label was assigned based on the number of failed and non-failed firms that were the closest to a given neuron. Then, neurons of each group were grouped using a hierarchical ascending classification and a Ward criterion, and the homogeneity of different partitions of each group was then assessed. As we were seeking a small number of regions at risk, we tested several partitions that were made up of two to six meta-classes per group of firms. Partitions were evaluated with the three best internal indices that were studied by Milligan (1981) and we chose the most homogeneous one. Then, we calculated the mean of each variable used to design a map within each meta-class and we used those means to rank the meta-classes. The map that was designed with data from 2003 is depicted on Figure 1. It is made up of 11 meta-classes: 6 of them are made up of neurons that mainly embody non-failed firms, and 5 of neurons that embody failed firms. Figure 1 especially shows that the 6 meta-classes that represent non-failed firms are the first 6 of the hierarchy calculated

using variable means. The corresponding means are shown in Table 9.

Then, we assessed the evolution of firm financial health with the maps that were estimated. To do so, we projected on the maps the data belonging to each sample and that were collected over 6 years. Hence, for each company, we calculated a set of 6 positions that corresponded to its own individual financial evolution. Then, we calculated the magnitude of the changes in position using the difference between the position estimated over year t and that estimated over year $t-1$. We then got 5 measures of changes per company. These individual measures were finally quantized using another map so as to determine prototype sequences of firm financial evolution. We analyzed different partitions of sequences (between 2 and 15 sets of sequences) using the same clustering technique and the same indices as those used to group neurons into meta-classes, and we chose the most homogeneous partition. Figure 2 represents the partition that was calculated with data from 2003 and the map depicted on Figure 1. It is made up of 8 sequences. On each graph that corresponds to a given sequence, the X-axis represents the 5 intervals during which the difference between two positions estimated in t and $t-1$ was calculated, and the Y-axis represents the magnitude of the differences. As there are 11 meta-classes in 2003, the magnitude of the maximum difference is 10. On Figure 1, the first two sequences mainly correspond to that of non-failed firms, the last four to that of failed companies, and the third and fourth to sequences that correspond both to certain subgroups of failed and non-failed firms. These graphs clearly indicate that sequences that embody few changes between meta-classes, that is to say rather stable firm financial evolution, mostly typify the evolution of firms that are likely to stay alive, whereas those that embody large variations mostly typify the evolution of failed companies. Besides, these graphs can be interpreted in light of Table 10, which is also calculated with data from 2003. Table 10 shows the percentages of firms that have spent time within the failed zone by number of years and firm status (failed vs. non-failled). We can notice that non-failed companies share a rather stable financial health, since most of them stayed within the non-failed zone nearly over the whole 6-year period. Roughly 46.6% stayed within this zone throughout the period and 18.3% stayed within this zone for 5 out of 6 years. Conversely, failed firms experienced rather erratic evolutions since 18.9% stayed within the non-failed zone for 2 years, 28.1% for 3 years and 19.7% for 4 years.

5.3.3. Design of models

Once the sequences were estimated, we designed, with each modeling method and each learning sample, as many models as sequences. Hence, with data from 2003, we estimated 8 models.

5.3.4. Result estimations

To assess new model accuracy, we first had to assign a prototype sequence to each firm belonging to a test sample. For this purpose, we estimated, for each firm within each test sample, its position over 6 years on the corresponding self-organizing map. Then, we calculated its changes of position on the map over the 6 years: these changes form the individual sequence of a given firm. Finally, we looked for the prototype sequence that was the closest to an individual sequence, and we used the model that fitted the latter prototype sequence to make forecasts. Results were summed up to assess the accuracy of single or ensemble-based models. With traditional and new models, model accuracy was estimated based on the comparison between the current status of firms belonging to test samples and their predicted status.

We used three statistics to evaluate model performance. First, we estimated type-I and type-II errors, which are very useful measures when the costs of misclassification are equal and when the sizes of each group to be discriminated are identical. To analyze type-I and type-II errors, we used the decomposition of the error suggested by Kohavi & Wolpert (1996). Kohavi & Wolpert (1996) show how to split the error into three components: a bias term that corresponds to the error due to the distance that exists between the accuracy of a given classifier and that of the best classifier that may exist within the chosen family of classifiers; a variance term that represents the error due to the training set used; and a noise term which represents the statistical uncertainty. Since the noise and the bias error cannot be estimated separately, we compared the bias and variance of the different models we designed. The estimation of the bias and the variance requires at least the use of different learning samples. Bouckaert (2008) presents different ways of building learning and test samples that have been used for the same purpose in the literature. Based on these experiments, we performed our estimations using the following procedure. We drew 100 bootstrap samples from each

original learning sample with the same number of companies. Models were designed using these samples and were then tested using the original test samples. All results were averaged over the different data sets.

In the field of bankruptcy prediction, the costs and the sizes of the groups are very uneven. This is why we used a second statistic, the area under the receiver operating characteristic curve (AUC), which makes it possible to estimate a performance that is independent from the costs of misclassification and from the sizes of each group.

Finally, we used a third measure proposed by Hand (2009), the H-measure. Hand has demonstrated that using the AUC to compare two classifiers may lead to a biased conclusion as the AUC relies on a metric that depends on the classifier whose performance is to be evaluated. The H-statistics overcome this limitation.

6. Results and discussion

6.1. Error of models

We first calculated the error of each model over the three periods studied by 5 different forecasting horizons. The estimations were made using a cut-off value that maximizes the overall correct classification rate. Table 11 presents the results achieved with traditional models. We grouped the results by type of model: single models, bagging-, boosting-, random subspace-, Decorate- and random forest-based models. Table 11 indicates that model accuracy slightly changes from one period to another. At a 1-year horizon, the average error ranges from 16.0% to 17.3%, from 17.7% to 19.0% at a 2-year horizon, from 19.8% to 20.6%, 22.5% to 23.3% and 24.7% to 25.6% at a 3, 4 and 5-year horizon respectively. The changes that occur within the economic environment certainly explain a part of these variations. Table 11 also shows that the error increases regularly as the horizon of the predictions recedes, the average difference between a 1-year and a 5-year forecast being equal to 8.4 percentage point, regardless of the period and the modeling method. When one analyzes the results in detail by type of model, one can notice that ensemble-based models are, on the whole, more accurate than single models at a 1-year horizon; the difference ranges from 0.65 percentage point to 2.89 percentage points. But this difference is considerably reduced at a 5-year horizon, since it only ranges from 0.01 percentage point to 0.84 percentage point. This

clearly shows that ensemble-based models, if they manage to improve model performance when the horizon of the prediction is short, do not solve the question that is related to the decrease in performance when this horizon recedes.

Then, we compared the error achieved with traditional models to that achieved with models that fitted different prototypes of firm financial evolution, we called "new models". The discrepancies between these two errors are presented in Table 12. Here as well, we can notice that the error slightly fluctuates from one period to another, but also that new models tend to be more accurate than the others since more than 81% of all differences that were calculated are in favor of new models, and when this is the case, 89% are statistically significant at the threshold of 5%. We can finally notice that the average difference between a forecast at a 1-year horizon and that at a 5-year is smaller than before and reaches 4.96 percentage points instead of 8.4 percentage point, regardless of the period and the modeling method, that is to say a decrease of 3.44 percentage point.

6.2. Error by type of model

To understand these differences, we estimated, for each type of model, the gain brought by new models. The results are indicated in Table 13. Panel A shows the differences between the error achieved using new models and that achieved using traditional models, and Panel B shows the p-values of a test for differences between proportions that make it possible to assess those that are statistically significant. On the whole, nearly all differences are significant. However, they are rather low at a 1-year horizon, with an average rate of 0.44 percentage point in favor of new models, but still tend to increase in favor of new models, as the horizon increases to finally reach 3.88 percentage point. On the whole, our method significantly improves forecasts when the horizon of a prediction exceeds 2 years, which corresponds to a characteristic that is rather desired by financial institutions.

If one analyzes these differences by firm status more closely, one may notice that the discrepancy between forecasts at a 1-year horizon and at a 5-year is on average 8.19 percentage points for non-failed firms, and 8.66 percentage points for failed firms, with traditional models, and is 5.15 percentage points for non-failed firms, and 0.87 percentage point for failed firms, with new models. New models are thus able to significantly reduce type-I error and

at the same time they provide better forecasts than those of traditional models.

Then, we studied the differences between ensemble-based models and traditional single models used by banks. Table 14 presents these results and shows, in Panel A, the differences between traditional single models and traditional ensemble-based models, and in Panel C, the same differences but between traditional single models and new ensemble-based models. Panel A shows, as mentioned earlier, that traditional ensemble-based models are of no use when one seeks to decrease the mid-term error achieved with single-models. If the average difference between the former and the latter is 2.05 percentage points at a 1-year horizon, in favor of the former, it is 0.21 percentage point at a 5-year horizon. However, the difference is 2.47 percentage points at a 1-year horizon in favor of new ensemble-based models, and 4.27 percentage points at a 5-year horizon, still in favor of new models, and all differences are significant (Panel D). Therefore, the gain brought by new models is higher than 4 percentage points compared to models used by banking institutions, and this gain is significant.

We then calculated the bias and the variance components of type-I and type-II errors by type of model. Table 15 shows the percentages of the bias and the variance components of each error. Models were estimated using bootstrap samples drawn from the original learning samples, and the errors were assessed using the original test samples. On the whole, boosting – used in conjunction with both traditional and new models – leads to the best results when it comes to reducing the bias component of the error, followed by random subspace, rotation forest, single models, Decorate and bagging. On the other hand, bagging leads to the best results when it comes to decreasing the variance component, followed by rotation forest, Decorate, random subspace, boosting and single models. These results are rather consistent with what the literature has shown (Zhang et al., 2012). Hence, our method does not fundamentally change the way ensemble-based models can influence the bias or the variance of a classification error, especially when using samples that are made up of a huge number of observations. Moreover, if one analyzes the ratio between the variance and the bias of the error, for each type of model and each type of error, one may notice that new models lead to ratios that are, on the whole, higher than those of traditional models. Incidentally, the variance component of the error achieved with traditional models is indeed lower than that of new models both in relative and absolute terms. Now, if one deepens the

differences between the two components, one also notices that, with traditional models, the difference between the average bias of type-II error and the average variance of the same error is 39.8 percentage points, whereas with type-I error, the difference is 29.2 percentage points. However, with new models, such a difference is significantly reduced with type-II error, as it is 21.5 percentage points, and slightly reduced with type-I error, as it is 20.9 percentage points. These figures show that our method makes it possible to better balance the bias and the variance of type-II error.

6.3. Model general performance

The previous results do not take into account misclassification costs, which are extremely asymmetric. Indeed, the cost of type-I error, which can lead to a bank granting a loan based on a wrong decision, is much larger than that of type-II error, which can lead to the refusal of a grant, also based on a wrong decision. In the first case, the error involves the loss of credit that will not be reimbursed, while in the second, it involves the loss of a potential bargain. This is why we estimated model performance using two measures: the AUC and the H-measure. Table 16 presents the AUC estimated using results achieved with traditional models (Panel A) and with new models (Panel B). Table 17 presents the H-measures that correspond to traditional (Panel A) and new models (Panel B).

All these results lead to several conclusions. First, they confirm the previous ones. The differences between AUCs are extremely low when forecasts are made at a 1 or 2-year horizon; the performances of traditional and new models are quite similar. Conversely, the differences tend to increase in favor of new models when the horizon reaches 3 years, and keep on increasing up to 5 years, still in favor of new models. Moreover, a part of these results is due to the ability of new models to significantly decrease type-II error compared to traditional ones, when the horizon increases. Second, H-measures slightly show the same things: at a very short horizon, traditional and new models perform equally, but beyond, new models are rather more accurate. In four out of five cases, and with both statistics, new models provide better results than the others. Third, they show what Table 14 (Panel C) already showed: boosting used in conjunction with new models presents, at a 5-year horizon, the most significant difference compared to traditional single models (4.92 percentage points),

followed by random forest and bagging (4.73 percentage points and 4.22 percentage points respectively). The same prevalence can be found within the results presented in Tables 16 and 17.

These results soundly confirm the added value of our method since it significantly improves model performance, regardless of its measure, and particularly of model ability to correctly forecast the fate of failed firms. It thus provides a reliable solution to a very current problem that financial institutions face. We know that model diversity makes it possible to better identify the boundary between classes than single models do because each rule has a specific expertise in a given region of the decision space (Kuncheva & Whitaker, 2003). But most of the time this diversity relies on the sole hazard that governs model estimation. Our study shows that this hazard can usefully be complemented *a priori* by knowledge, which confirms that incorporating domain knowledge into data mining techniques may lead to substantial performance improvement (Sinha & Zhao, 2008).

6.4. Contributions to the literature and implications

Our study outlines how another way of incorporating time into a model (other than that traditionally used in the literature) makes it possible to significantly improve results achieved with failure models. Its contribution to the literature is three-fold.

The first one is conceptual since the study we have conducted shows the true interest of designing models based on sound theoretical considerations, whereas most previous studies (cf. Table 2) do not rely on any theoretical basis. Indeed, the models we propose derive from the concept of dynamics of organizational change, grounded in the Hannan & Freeman (1984)'s structural inertia theory, a concept that has been empirically validated several times. One knows that the sole, pure empiricism is often subject to data variability and experimental conditions.

The second contribution is methodological, since our study highlights the limits of models that are solely based on an automatic (and sometimes random) extraction of knowledge and suggests that a successful complementarity exists between automatic procedures and knowledge that is forged *a priori*. Moreover, it shows once more how the time dimension is an important variable that models should take into account, but in a way that radically breaks

with traditional practice; as already pointed out, if a consensus exists around the usefulness of the time dimension as an explanatory factor, however, traditional measures of this dimension do not work well. Our work reinforces the idea that other modeling approaches make it possible to provide an answer to this issue.

The third and final contribution is empirical. Our results strengthen the very few previous studies that have shown that when failure classification models are based on a prior segmentation of data, they are more likely to achieve better results than those of models that are not. It is an important further step since such models are quite rare in the bankruptcy literature (Tsai, 2014). They also demonstrate that our modeling framework significantly improves the accuracy of traditional models used by financial institutions at a 5-year horizon and also, when used in conjunction with traditional ensemble techniques, it significantly improves both short- (1 year) and mid-term (5 years) forecasts. Moreover, these improvements appear to be robust because models were designed with big, random samples, that do not lead to overestimations, especially of type-I error (Zmijewski, 1984; Platt & Platt, 2002). They also appear to be robust because models were validated with different criteria and using real operating conditions, that is to say conditions that perfectly reflect the way models are operationally used by financial institutions, and that are embodied by the estimation of out-of sample and out-of time errors (Stein, 2007). This last point is very important because most studies cited in Table 1 present highly biased estimations: samples are often small and not drawn randomly, and most of the time models are not validated using the same experimental conditions as those they would be likely to face if they were used by banks (cf. Table 5).

There are two main implications of all these contributions. The first concerns the academic community. Very little research has been undertaken to deepen our understanding of the influence of the time dimension on the failure process and, by ricochet, to offer a reliable modeling framework of this dimension. Past studies have shown the limits of most statistical measures that tried to embody the history of companies. One may then think that the method we suggest might inspire other researchers and lead them to study other types of techniques that would make it possible to better account for the dynamics of firm financial evolution. The second implication concerns the business community. Most failure models

that are currently used by banks and financial companies mainly rely on discriminant analysis, and sometimes on logistic regression. We have long known that many other methods (especially the non-linear one) are far more accurate, but banks seem to be reluctant to use non-parametric techniques. Our study shows that they could improve their own models without sacrificing the choice of their preferred modeling method.

6.5. Limits and disadvantages of our method

First of all, the limits of our method are conceptual. Indeed, the time dimension is solely estimated using a single variable: the magnitude of variations that may occur within financial accounts over time. Since bankruptcy is a complex phenomenon, if one may think that it is not reducible to a single model, one may also think that it is not reducible to a single measure of its dynamics.

Second, they are methodological. One of the key factors of model robustness lies in the quality of data used to estimate model parameters. This is the reason why many banks design their models with variables whose characteristics are stable over time. Therefore, our measure of the dynamics of firm financial evolution needs to be studied so as to confirm its characteristics and stability. The same reasoning can be applied to the classes that are estimated with the segmentation process, for which we know little.

Third, they are practical. Our models require data that are measured over 6 years, and cannot be applied to firms for which one has solely been able to collect data over a short period of time, either because data are difficult to gather or because data do not exist, as is the case with very young firms. In the latter case, the drawback is not that severe as the real risk incurred by financial institutions mainly concerns firms that are precisely not young. Indeed, because of the liability of newness, very young companies experience a much higher bankruptcy rate than older firms do. Moreover, our models are not as meaningful as a discriminant function can be. A financial institution using our models would not be able to justify a decision that would be taken based on one of their predictions. Finally, they do not estimate a probability of failure. A good output of a model should be a measure of a risk, and more precisely of the intensity of a risk, rather than a simple measure indicating a distance from a threshold that separates two groups. As it happens, banking models are

not solely used to estimate a probability of default; they are also useful for designing classes of risk to analyze how firms move within these classes over time.

7. Conclusion

The results of our study corroborate the hypothesis that we mentioned at the very beginning of this article where we assumed that a model embodying the changes that may affect firm financial evolution over time might achieve better forecasts than those of traditional models. When one compares the results achieved with any modeling techniques to those achieved with the same techniques but using the proposed segmentation, one may notice that the latter results are better than the former, but the improvement solely occurs when the horizon of the forecast is at least 3 years. However, when our method of segmentation is used in combination with ensemble-based models, model accuracy is always improved whatever the forecasting horizon, compared to that of traditional models used by financial institutions. This result is rather interesting because the improvement both concerns results achieved with single and ensemble-based models. It shows that the intrinsic performance of ensemble-based models has a limit that can be exceeded by introducing an additional step during the modeling process: the decision space is no longer divided into subspaces where each rule is designed to fit a particular region of this space, but into subspaces that are themselves re-divided into smaller subspaces where each rule is estimated. It is certainly this additional subdivision, which brought a sort of informed diversity, which explains the performance of our method.

All these findings may give rise to several new research avenues. First, the output of our models could be improved. Instead of estimating a score value, the models might calculate a probability of bankruptcy. In such a case, the models might be used to assess classes of risk. This potentiality is very important for banks because they need to follow how the risk associated to a firm is likely to change over time. One could then study how to build these classes and especially how to assess their robustness. Second, our method relies on single-criterion segmentation. It would therefore be necessary to study the usefulness of other types of criteria, or the relevance of the combination of different criteria and, more generally, to assess the conditions that are required to perform an efficient segmentation.

Finally, we propose in this paper a sort of time-modeling technique that appears to be more efficient than those studied in the past. One may then wonder whether more relevant ones might exist.

References

Alfaro, E., Garcia, N., Games, M., & Elizondo, D. (2008). Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks. *Decision Support Systems*, *45*, 110–122.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*, 589–609.

Altman, E. I., Haldeman, R., & Narayanan, P. (1977). Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking Finance*, *1*, 29–51.

Altman, E. I., Kim, D. W., & Eom, Y. H. (1995). Failure prediction: Evidence from Korea. *Journal of International Financial Management and Accounting*, *6*, 230–249.

Amburgey, T. L., Kelly, D., & Barnett, W. P. (1993). Resetting the clock: The dynamics of organizational change and failure. *Administrative Science Quarterly*, *38*, 51–73.

Aziz, A., Emanuel, D. C., & Lawson, G. C. (1988). Bankruptcy prediction: An investigation of cash flow based models. *Journal of Management Studies*, *25*, 419–437.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, *38*, 63–93.

Bardos, M. (2007). What is at stake in the construction and use of credit scores? *Computational Economics*, *29*, 159–172.

Bardos, M. (2008). Scoring sur donnees d'entreprises: Instrument de diagnostic individuel et outil d'analyse de portefeuille d'une clientele. *Modulad*, *29*, 159–177.

Basel Committee on Banking Supervision (2009). *Guiding principles for the replacement of IAS 39*. Technical Report.

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, *23*, 129–143.

Betts, J., & Belhoul, D. (1987). The effectiveness of incorporating stability measures in company failure models. *Journal of Business Finance and Accounting*, *14*, 323–334.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, *12*, 1–25.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). ACM Press.

Bouckaert, R. R. (2008). Practical bias variance decomposition. In W. Wobcke, & M. Zhang (Eds.), *21st Australian Joint Conference on Artificial Intelligence Auckland* (pp. 247–257). Springer, Berlin, Heidelberg.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Castagna, A. D., & Matolcsy, Z. P. (1981). The prediction of corporate failure: Testing the Australian experience. *Australian Journal of Management*, *6*, 23–50.

Dambolena, I., & Khoury, S. (1980). Ratio stability and corporate failure. *Journal of Finance*, *33*, 1017–1026.

D'Aveni, R. A. (1989). The aftermath of organizational decline: A longitudinal study of the strategic and managerial characteristics of declining firms. *Academy of Management Journal*, *32*, 577–605.

Deakin, E. B. (1972). A discriminant analysis of predictors of business failures. *Journal of Accounting Research*, *10*, 167–179.

Dimitras, A. I., Zanakis, S., & Zopounidis, C. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, *90*, 487–513.

Emery, G. W., & Cogger, K. O. (1982). The measurement of liquidity. *Journal of Accounting Research*, *20*, 290–303.

European Central Bank (2014). *SME access to finance in the Euro area: Barriers and potential policy remedies*. Technical Report.

Fanning, K. M., & Cogger, K. O. (1994). A comparative analysis of artificial neural networks using financial distress prediction. *International Journal of Intelligent Systems in Accounting, Finance and Management*, *3*, 241–252.

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *The third annual workshop on computational learning theory* (pp. 202–216).

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The British Accounting Review*, *40*, 269–291.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*, 236–247.

Gepp, A., & Kumar, K. (2008). The role of survival analysis in financial distress prediction. *International Research Journal of Finance and Economics*, *16*, 13–34.

Gombola, M. J., Haskins, M. E., Ketz, J. E., & Williams, D. D. (1987). Cash flow in bankruptcy prediction. *Financial Management*, *16*, 55–65.

Grandvallet, Y. (2001). Bagging can stabilize without reducing variance. In *International Conference on Artificial Neural Networks* (pp. 49–56). Vienna, Austria.

Gupta, Y. P., Rao, R. P., & Bagchi, P. K. (1990). Linear goal programming as an alternative to multivariate discriminant analysis: A note. *Journal of Business Finance and Accounting*, *17*, 593–598.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Hannan, M. T. (2005). Ecologies of organizations: Diversity and identity. *Journal of Economic Perspectives*, *19*, 51–70.

Hannan, M. T., & Freeman, J. (1984). Structural inertia and organizational change. *American Sociological Review*, *49*, 49–164.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions On Knowledge And Data Engineering*, *21*, 1263–1284.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 832–844.

Hu, Y.-C., & Ansell, J. (2007). Measuring retail company performance using credit scoring techniques. *European Journal of Operational Research*, *183*, 1595–1606.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*, 489–501.

Huang, S.-C., Tang, Y.-C., Lee, C.-W., & Chang, M.-J. (2012). Kernel local Fisher discriminant analysis based manifold-regularized SVM model for financial distress predictions. *Expert Systems with Applications*, *39*, 3855–3861.

du Jardin, P. (2015). Bankruptcy prediction using terminal failure processes. *European Journal of Operational Research*, *242*, 286–303.

du Jardin, P., & Séverin, E. (2011). Predicting corporate bankruptcy using a self-organizing map: An empirical study to improve the forecasting horizon of a financial failure model. *Decision Support Systems*, *51*, 701–711.

Kaski, S., & Lagus, K. (1996). Comparing self-organizing maps. In J. C. V. C. von der Malsburg, W. von Seelen, & B. Sendhoff (Eds.), *International Conference on Artificial Neural Networks* (pp. 809–814). Springer, Berlin, Heidelberg volume 1112 of *Lecture Notes in Computer Science*.

Keasey, K., McGuinness, P., & Short, H. (1990). Multilogit approach to predicting corporate failure: Further analysis and the issue of signal consistency. *Omega*, *18*, 85–94.

Kinney, W. R. (1973). Discussion of a prediction of business failure using accounting data. *Journal of Accounting Research*, *11*, 183–187.

Kohavi, R., & Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. In *Machine Learning: Proceedings of the Thirteenth International Conference*.

Kohonen, T. (2001). *Self-organizing maps* volume 30 of *Springer Series in Information Sciences*. (3rd ed.). Springer.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*, 181–207.

Laitinen, T. (1991). Financial ratios and different failure processes. *Journal of Business Finance and Accounting*, *18*, 649–673.

Leray, P., & Gallinari, P. (1998). Feature selection with neural networks. *Behaviormetrika*, *26*, 145–166.

Marques, A. I., Garcia, V., & Sanchez, J. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, *39*, 10244–10250.

Melville, P., & Mooney, R. J. (2004). Creating diversity in ensembles using artificial data. *Journal of Information Fusion: Special Issue on Diversity in Multiclassifier Systems*, *6*, 99–111.

Miller, D., & Friesen, P. H. (1977). Strategy-making in context: Ten empirical archetypes. *Journal of Management Studies*, *14*, 253–280.

Milligan, G. W. (1981). A Monte-Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, *46*, 187–199.

Mossman, C. E., Bell, G. G., Swartz, L. M., & Turtle, H. (1998). An empirical comparison of bankruptcy models. *Financial Review*, *33*, 35–53.

Nam, J. H., & Jinn, T. (2000). Bankruptcy prediction: Evidence from Korean listed companies during the imf crisis. *Journal of International Financial Management and Accounting*, *11*, 178–197.

Platt, H. D., & Platt, M. B. (2002). Predicting corporate financial distress: Reflections on choice-based sample bias. *Journal of Economics and Finance*, *26*, 184–199.

Pompe, P. P. M., & Bilderbeek, J. (2005). Bankruptcy prediction: The influence of the year prior to failure selected for model building and the effects in a period of economic decline. *Intelligent Systems in Accounting, Finance and Management*, *13*, 95–112.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufman Publishers.

Rodriguez, J. J., & Kuncheva, L. I. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*, 1619–1630.

Rose, P. S., & Kolari, J. W. (1985). Early warning systems as a monitoring device for bank conditions. *Quarterly Journal of Business & Economics*, *24*, 43–60.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*, 197–227.

Sheppard, J. P. (1994). Strategy and bankruptcy: An exploration into organizational death. *Journal of Management*, *20*, 795–833.

Singh, J. V., House, R. J., & Tucker, D. J. (1986). Organizational change and organizational mortality. *Administrative Science Quarterly*, *31*, 587–611.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems*, *46*, 287–299.

Skogsvik, K. (1990). Current cost accounting ratios as predictors of business failure: The Swedish case. *Journal of Business Finance and Accounting*, *17*, 127–160.

von Stein, J. H., & Ziegler, W. (1984). The prognosis and surveillance of risks from commercial credit borrowers. *Journal of Banking and Finance*, *8*, 249–268.

Stein, R. M. (2007). Benchmarking default prediction models: Pitfalls and remedies in model validation. *Journal of Risk Model Validation*, *1*, 77–113.

Sun, J., Jia, M.-Y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert Systems with Applications*, *38*, 9305–9312.

Tian, S., Yu, Y., & Zhou, M. (2015). Data sample selection issues for bankruptcy prediction. *Risk, Hazards & Crisis in Public Policy*, *6*, 91–116.

Tsai, C. F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, *16*, 46–58.

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Stat.*, *15*, 145–162.

Wilcox, J. W. (1973). A predicton of business failure using accounting data. *Journal of Accounting Research*, *11*, 163–179.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.

Zhang, C. X., Wang, G. W., & Zhang, J. S. (2012). An empirical bias-variance analysis of decorate ensemble method at different training sample sizes. *Journal of Applied Statistics*, *39*, 829–850.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59–82.

8. Tables

Table 1: Results of the main studies that have estimated model accuracy up to four and five years ahead

Studies	% of correct classification Year prior to failure				
	1	2	3	4	5
Altman (1968)	95.0	72.0	48.0	29.0	36.0
Altman et al. (1977)	91.0	89.0	83.5	79.8	76.8
Altman et al. (1995)	97.1	88.2	69.7	50.0	68.8
Aziz et al. (1988)	91.8	84.7	78.6	80.2	80.9
Betts & Belhoul (1987)	90.1	72.4	64.7	41.2	37.5
Blum (1974)	95.0	80.0	70.0	80.0	69.0
Castagna & Matolcsy (1981)	92.9	83.3	85.7	78.6	83.3
Dambolena & Khoury (1980)	91.2	84.8	82.6	89.1	
Deakin (1972)	97.0	95.5	95.5	79.0	83.0
Emery & Cogger (1982)	94.0	93.0	93.0	90.0	81.0
Fanning & Cogger (1994)	85.0	78.0	72.0	70.0	70.0
Geng et al. (2015)			78.8	76.0	76.2
			78.4	76.1	70.5
Gepp & Kumar (2008)	95.4	93.0	90.5	88.8	86.7
Gombola et al. (1987)	89.0	86.0	72.0	70.0	
Gupta et al. (1990)	85.0	80.0	65.0	60.0	60.0
Hu & Ansell (2007)	92.7	89.4	88.2	88.2	89.0
Huang et al. (2012)	97.9	96.4	92.3	92.9	90.7
	91.2	89.8	85.4	80.6	79.7
	91.2	89.1	85.4	83.6	78.1
	91.2	89.8	84.7	85.6	82.8
Keasey et al. (1990)	63.0	74.5	64.5	65.0	41.0
Kinney (1973)	88.0	82.0	78.0	75.0	72.0
Laitinen (1991)	88.8	68.8		55.0	
Mossman et al. (1998)	84.0	73.2	72.6	65.9	
Nam & Jinn (2000)	80.4	76.1	76.1	87.0	80.4
Pompe & Bilderbeek (2005)	80.0	70.0	68.0	65.0	63.0
Rose & Kolari (1985)	76.1	77.0	69.0	62.3	65.5
Sheppard (1994)	71.4	69.6	76.8	66.1	64.3
Skogsvik (1990)	90.5	88.8	88.0	87.6	87.3
von Stein & Ziegler (1984)	95.0	89.9	86.6	78.2	71.4
Wilcox (1973)	94.0	90.0	88.0	90.0	76.0
	Accuracy ratio				
Berg (2007)	0.78	0.76	0.73	0.70	0.67
Tian et al. (2015)	0.64	0.39	0.29		0.18

Table 2: Issues and theoretical foundations used in the studies presented in Table 1

Studies	Issues	Theoretical or conceptual basis
Panel A: Single-rule models single-period data		
Altman (1968)	Assess the quality of ratio analysis as an analytical technique	
Altman et al. (1995)	Test a distress classification model for Korean firms	
Castagna & Matolcsy (1981)	Address methodological issues related to bankruptcy models	
Deakin (1972)	Propose an alternative model to that of Altman (1968)	
Emery & Cogger (1982)	Propose liquidity measures to forecast bankruptcy	Theoretical model of firms' liquidity
Fanning & Cogger (1994)	Study the usefulness of a new neural network to forecast bankruptcy	
Gombola et al. (1987)	Study the usefulness of cash flow variables to forecast bankruptcy	
Gupta et al. (1990)	Show the usefulness of linear programming so as to forecast failure	
Hu & Ansell (2007)	Design models for evaluating credit risk in relation to the retailing industry	Theory of competition
Laitinen (1991)	Study the dimensions that influence the financial situation of failed firms	Theoretical model of investment projects
Mossman et al. (1998)	Study the predictive ability of different existing models	
Nam & Jinn (2000)	Design a forecasting model over a period of financial crisis	
Pompe & Bilderbeek (2005)	Study the influence of the period over which models are design	
Rose & Kolari (1985)	Examine the efficiency of a default prediction model	
Sheppard (1994)	Study the effects of strategic factors on firm survival	Theory of organizational decline
Skogsvik (1990)	Test the ability of cost accounting ratios to forecast financial failure	Cost accounting framework
von Stein & Ziegler (1984)	Develop procedures for early warning of latent risks	
Panel B: Single-rule models multi-period data		
Altman et al. (1977)	Design a model using corporate and financial market variables	
Aziz et al. (1988)	Design a prediction model based on a theoretical model of a firm's liquidity	Theory of corporate valuation
Betts & Belhoul (1987)	Test the influence of a measure of earnings stability on model accuracy	
Berg (2007)	Study the efficiency of generalized additive models	
Blum (1974)	Design a model based on a time dimension	Cash flow framework
Dambolena & Khoury (1980)	Design a method that makes it possible to stabilize financial ratios	
Gepp & Kumar (2008)	Use survival analysis techniques to design bankruptcy models	
Keasey et al. (1990)	Study the role of signal consistency when dating corporate failure	
Tian et al. (2015)	Investigate the relative importance of various bankruptcy predictors	
Wilcox (1973)	Design a dynamic bankruptcy prediction model	Theoretical model of failure
Panel C: Multi-rule models single-period data		
Huang et al. (2012)	Propose a novel hybrid classifier based on an SVM	
Panel D: Multi-rule models multi-period data		
Geng et al. (2015)	Study the phenomenon of financial distress in China	

Table 3: Modeling methods, explanatory variables and variable selection methods used in the studies presented in Table 1

Studies	Modeling methods (1)	Models (2)		Variables (3)			Criteria used to select explanatory variables
		SM	EM	SP	MP	Types (4)	
Panel A: Single-rule models single-period data							
Altman (1968)	DA	x		x		FR	Variables used in previous studies - Discrimination ability - Correlations
Altman et al. (1995)	DA	x		x		FR-SV	Variables used in previous studies - Discrimination ability over time
Castagna & Matolcsy (1981)	DA-QDA	x		x		FR	Variables used in previous studies - Stepwise search
Deakin (1972)	DA	x		x		FR	Best variables used in a previous study
Emery & Cogger (1982)	WP	x		x		FR	Theoretical model
Fanning & Cogger (1994)	NN-LR-SA-WP	x		x		SV	Variables used in a previous study
Gombola et al. (1987)	DA-QDA-PR	x		x		FR	Variables used in previous studies - Factor analysis
Gupta et al. (1990)	DA-LGP	x		x		FR	Variables used in a previous study
Hu & Ansell (2007)	NB-LR-DT-NN-SMO	x		x		FR-FV-MEV	Variables that embody firm environment and resources - Forward search
Laitinen (1991)	DA	x		x		FR	Variables that embody the theoretical model used
Mossman et al. (1998)	LR	x		x		FR-FV-SV-MV-VFV	Variables used in previous studies
Nam & Jinn (2000)	LR	x		x		FR	Discrimination ability - Stepwise search
Pompe & Bilderbeek (2005)	DA-NN	x		x		FR	Variables used in a previous study - Stepwise search
Rose & Kolari (1985)	DA-QDA	x		x		FR-SV	Variables that measure firm performance - Discrimination ability - Correlations
Sheppard (1994)	LR	x		x		FV-NFV	Variables that embody different factors of decline
Skogsvik (1990)	PR	x		x		FR-SV	Variables that derive from an accounting framework - Discrimination ability
von Stein & Ziegler (1984)	KNN-DA-QDA	x		x		FR	Variables that can be calculated using balance sheets - Discrimination ability
Panel B: Single-rule models multi-period data							
Altman et al. (1977)	DA-QDA	x		x	x (a)	FR-MV-SV	Variables used in previous studies - Discrimination ability - Stepwise search
Aziz et al. (1988)	DA-LR	x		x	x (b)	FV-MV-VFV	Variables that derive from the conceptual model used
Betts & Belhoul (1987)	DA	x		x	x (c)	FR-SV-VFV	Ratios used in financial analysis - Stepwise search
Berg (2007)	DA-GAM-GLM-NN	x		x		FR-MV-NFV-VFR	Variables used in a previous study
Blum (1974)	DA	x		x	x (d)	FR-SV-MV	Cash flow framework
Dambolena & Khoury (1980)	DA	x		x	x (e)	FR-SV	Variables used in previous studies - Stepwise search
Gepp & Kumar (2008)	DA-LR-SA	x			x (f)	FR	Variables used in a previous study
Keasey et al. (1990)	ML	x			x (g)	FR-MV	Stepwise search
Tian et al. (2015)	HM	x			x (h)	FR-MV-SV	Lasso
Wilcox (1973)	GRM	x		x	x (i)	FR-SV-VFV	Variables that embody the theoretical model used
Panel C: Multi-rule models single-period data							
Huang et al. (2012)	DT-KNN-BN-RBFNN-SVM-KSVM AdaBoost-MultiBoost-Bagging		x	x		FR	Variables found in the database - Stepwise search - Data reduction techniques
Panel D: Multi-rule models multi-period data							
Geng et al. (2015)	DT-NN-SVM-(DT+NN+SVM)		x	x	x (j)	FR-VFR	Accounting standards - Individual discrimination ability

(1) BN: bayesian network; DA: discriminant analysis; DT: decision tree; GAM: generalized additive model; GLM: generalized linear model; GRM: gambler's ruin model; HM: hazard model; KNN: K nearest neighbors; KSVM: kernel support vector machine; LGP: Linear Goal Programming; LR: logistic regression; ML: Multi-logit.
(2) SM: single model; EM: ensemble-based model.

(3) Single-period (SP) variables correspond to variables that are measured once. Multi-period (MP) variables correspond to variables that are measured over several consecutive years or variables that represent a variation of a given quantity over time.
(4) FR: financial ratios; FV: financial variables; NFV: non financial variables; MEV: macro-economic variables; MV: market variables; SV: statistical variables; VFR: variations of financial ratios; VFV: variations of financial variables.
(a) Standard error of estimate of earnings before interest, taxes, depreciation, and amortization/total assets measured over ten year-periods.
(b) Change in current assets, current liabilities, tax liability, cash and marketable securities, common and preferred stock, short-term debt and long/medium term debt over two year-periods.
(c) Standard deviation of all financial variables measured over three year-periods, four measures of trend over three year-periods and four measures of change over the previous year calculated on total assets, total sales, total employees and inventories.
(d) Decrease of net income over two year-periods.
(e) Standard deviation of all financial variables measured over three and four year-periods, standard error of estimate of all variables around a 4-year linear trend, coefficient of variation of all variables over four year-periods.
(f) Financial variables measured over different years and measures of variations of variables over two year-periods.
(g) Financial variables measured over five years.
(h) Financial variables measured over up to five years.
(i) Change in long-term assets and current assets over different years.
(j) Measure of variation of business income, total assets and net profit over two year-periods.

Table 4: Characteristics of learning samples used in the studies presented in Table 1

Studies	Types (1)	Sectors	Countries	Non-failed firms Number	Failed firms Number	Years
Panel A: Single-rule models single-period data						
Altman (1968)	MS	MA	USA	33	33	1946-1965
Altman et al. (1995)		IN-TR	Korea	34	34	1990-1993
Castagna & Matolcsy (1981)	MS	IN	Australia	21	21	1963-1977
Deakin (1972)	MS	IN	USA	32	32	1962-1966
Emery & Cogger (1982)	MS	IN	USA	52	52	1949-1971
Fanning & Cogger (1994)	MS	DS	USA	75	75	1947-1965
Gombola et al. (1987)	MS	MA-RE	USA	122	122	1967-1981
Gupta et al. (1990)	MS	Industry	USA	20	20	1971-1986
Hu & Ansell (2007)		RE	USA	195	51	1994-2002
Laitinen (1991)	MS	DS	Finland	40	40	
Mossman et al. (1998)	MS	DS	USA	$\leq$100	$\leq$100	1980-1991
Nam & Jinn (2000)	MS	DS	Korea	46	46	1997-1998
Pompe & Bilderbeek (2005)	NRS	IN	Belgium	1800	678	1986-1994
Rose & Kolari (1985)	MS	BA	USA	$\leq$71	$\leq$71	1964-1977
Sheppard (1994)	MS	DS	USA	28	28	1980-1987
Skogsvik (1990)	NRS	MA-MI	Sweden	328	51	1966-1980
von Stein & Ziegler (1984)	NRS	DS	Germany	327	117	1971-1978
Panel B: Single-rule models multi-period data						
Altman et al. (1977)	MS	MA-RE	USA	53	58	1969-1975
Aziz et al. (1988)	MS	DS	USA	39	39	1971-1982
Betts & Belhoul (1987)	NRS	IN	USA	93	$\leq$39	1974-1978
Berg (2007)	NRS	DS	Norway	59400	600	1996-2001
Blum (1974)	MS	IN	USA	115	115	1954-1968
Dambolena & Khoury (1980)	MS	MA-RE	USA	23	23	1969-1975
Gepp & Kumar (2008)	NRS	MA-RE	USA	106	65	1974-1991
Keasey et al. (1990)	MS	DS	USA	40	40	1976-1984
Tian et al. (2015)	NRS	DS	USA		1383	1980-2009
Wilcox (1973)	MS	DS	USA	52	52	1949-1971
Panel C: Multi-rule models single-period data						
Huang et al. (2012)	MS	DS	Taiwan	100	50	2000-2007
Panel D: Multi-rule models multi-period data						
Geng et al. (2015)	RS	DS	China	107	107	2001-2008

(1) MS: matched sample; NRS: non-random sample; RS: random sample.
(2) BA: banking; DS: different sectors; IN: industry; MA: manufacturing; MI: mining; RE: retailing; TR: trading.

Table 5: Criteria used to test model accuracy in the studies presented in Table 1

Studies	Methods (1)	Out-of sample error	Out-of-time error	Performance measures (2)
Panel A: Single-rule models single-period data				
Altman (1968)	TS	Yes	No	t-1 t-II
Altman et al. (1995)	TS	Yes	No	t-1 t-II
Castagna & Matolcsy (1981)	CV	No	No	t-1 t-II
Deakin (1972)	TS	Yes	No	t-1 t-II
Emery & Cogger (1982)	LS	No	No	OCR
Fanning & Cogger (1994)	TS	Yes	Yes	t-1 t-II
Gombola et al. (1987)	CV	No	No	OCR
Gupta et al. (1990)	TS	Yes	No	t-1 t-II
Hu & Ansell (2007)	CV	No	No	OCR-AUC
Laitinen (1991)	LS	No	No	t-1 t-II
Mossman et al. (1998)	CV	No	No	t-1 t-II
Nam & Jinn (2000)	TS	Yes	No	t-1 t-II

Pompe & Bilderbeek (2005)TS	Yes	No	t-1 t-II	
Rose & Kolari (1985)	CV	No	No	t-1 t-II
Sheppard (1994)	CV	No	No	t-1 t-II
Skogsvik (1990)	CV	No	No	t-1 t-II
von Stein & Ziegler (1984)	LS	No	No	t-1 t-II
Panel B: Single-rule models multi-period data				
Altman et al. (1977)	TS-CV	Yes	No	t-1 t-II-MC
Aziz et al. (1988)	TS-CV	Yes	No	t-1 t-II
Betts & Belhoul (1987)	CV	No	No	t-1 t-II
Berg (2007)	TS	Yes	Yes	AR
Blum (1974)	TS	Yes	No	t-1 t-II
Dambolena & Khoury (1980)	CV	No	No	t-1 t-II
Gepp & Kumar (2008)	TS	Yes	No	t-1 t-II-MC
Keasey et al. (1990)	TS	Yes	No	t-1 t-II
Tian et al. (2015)	TS	Yes	Yes	AUC
Wilcox (1973)	LS	No	No	t-1 t-II
Panel C: Multi-rule models single-period data				
Huang et al. (2012)	CV	No	No	OCR
Panel D: Multi-rule models multi-period data				
Geng et al. (2015)	TS	Yes	No	AR

(1) CV: cross validation; LS: learning sample; TS: test sample.
(2) AR: accuracy ratio; AUC: area under the Roc curve; MC: misclassification costs; OCR: overall classification rate; t-1 t-II: type-I type-II errors.

Table 6: Learning and test samples

Period	Sample	Number of firms		Firm status	Data					
		Non-failed	Failed	Year	1 Y	2 Y	3 Y	4 Y	5 Y	6 Y
1	Learning	95,910	1,920	2003	2002	2001	2000	1999	1998	1997
	Test	91,050	1,820	2002	2001	2000	1999	1998	1997	1996
2	Learning	97,940	1,960	2006	2005	2004	2003	2002	2001	2000
	Test	95,420	1,910	2005	2004	2003	2002	2001	2000	1999
3	Learning	94,020	1,880	2010	2009	2008	2007	2006	2005	2004
	Test	96,530	1,930	2009	2008	2007	2006	2005	2004	2003

Table 7: Initial set of variables by financial dimension

Liquidity		Activity	
Cash/Current liabilities	C/CL	Cash flow/Total sales	CF/TS
Cash/Current assets	C/CA	Cash flow/Value added	CF/VA
Cash/Total assets	C/TA	EBIT/Value added	EBIT/VA
Quick assets/Current liabilities	QA/CL	EBITDA/Total sales	EBITDA/TS
Turnover		Net income/Total sales	NI/TS
Accounts payable/Total sales	AP/TS	Net income/Value added	NI/VA
Current assets/Total sales	CA/TS	Value added/Total sales	VA/TS
Current liabilities/Total sales	CL/TS	Solvency	
Inventories/Total sales	I/TS	Financial debt/Cash flow	FD/CF
Net op. work. capital/Total sales	NOWC/TS	Financial expenses/Net income	FE/NI
Receivables/Total sales	R/TS	Financial expenses/Total assets	FE/TA
Total sales/Total assets	TS/TA	Financial expenses/Value added	FE/VA
Profitability		Financial structure	
Cash flow/Shareholder funds	CF/SF	Current liabilities/Total assets	CL/TA
EBIT/Shareholder funds	EBIT/SF	Long term debt/Total assets	LTD/TA
EBITDA/Permanent equity	EBITDA/PE	Shareholder funds/Permanent equity	SF/PE
EBITDA/Total assets	EBITDA/TA	Shareholder funds/Total assets	SF/TA
Profit before tax/Shareholder funds	PBT/SF	Total debt/Total assets	TD/TA

Net op. work. capital: Net operating working capital; EBIT: Earnings before interest and taxes;

EBITDA: Earnings before interest, taxes, depreciation and amortization.

Table 8: Quartiles of variables that were selected to design models using data from 2003

		QA/CL			C/TA			I/TS			R/TS			AP/TS			TS/TA		
		25	50	75	25	50	75	25	50	75	25	50	75	25	50	75	25	50	75
Non-Failed	1 Y	-0.36	0.17	0.81	-0.48	0.13	0.98	-0.73	-0.56	-0.06	-0.98	-0.41	0.26	-0.88	-0.44	0.02	-0.29	0.18	0.82
Failed	1 Y	-0.92	-0.46	0.12	-0.68	-0.44	0.09	-0.69	-0.17	0.83	-0.59	0.22	1.04	-0.51	0.10	1.06	-0.95	-0.61	-0.13
	2 Y	-0.89	-0.39	0.18	-0.76	-0.45	0.11	-0.68	-0.17	0.80	-0.59	0.18	0.92	0.06	0.11	0.17	-0.87	-0.57	-0.15
	3 Y	-0.87	-0.34	0.23	-0.77	-0.39	0.19	-0.68	-0.18	0.79	-0.56	0.18	0.92	0.06	0.11	0.17	-0.88	-0.53	-0.07
		VA/TS			CF/TS			EBIT/VA			EBITDA/TA			CF/SF			EBITDA/PE		
		25	50	75	25	50	75	25	50	75	25	50	75	25	50	75	25	50	75
Non-Failed	1 Y	-0.96	-0.44	0.45	0.13	0.31	0.51	0.10	0.25	0.52	-0.45	0.04	0.61	-0.17	0.01	0.25	-0.13	0.05	0.41
Failed	1 Y	-0.58	-0.01	0.55	-0.82	0.06	0.47	-0.51	0.09	0.31	-0.64	-0.10	0.37	-0.37	-0.06	0.43	-0.59	-0.16	0.11
	2 Y	-0.58	0.02	0.63	-0.68	0.04	0.46	-0.42	0.00	0.29	-0.67	-0.20	0.28	-0.37	-0.03	0.42	-0.46	-0.20	0.03
	3 Y	-0.57	0.03	0.64	-0.50	0.07	0.43	-0.51	-0.02	0.32	-0.68	-0.18	0.39	-0.43	-0.06	0.42	-0.43	-0.18	0.06
		FE/TA			FE/VA			FD/CF			LTD/TA			LTD/TA			CL/TA		
		25	50	75	25	50	75	25	50	75	25	50	75	25	50	75	25	50	75
Non-Failed	1 Y	-0.65	-0.46	-0.06	-0.51	-0.41	-0.13	-0.14	-0.08	0.20	-0.71	-0.42	0.17	-0.22	0.01	0.44	-0.80	-0.40	0.04
Failed	1 Y	-0.56	-0.20	0.42	-0.48	-0.27	0.20	-0.28	-0.14	0.03	-0.65	-0.31	0.48	-0.67	-0.28	-0.01	-0.50	0.07	0.87
	2 Y	-0.56	-0.20	0.42	-0.56	-0.28	0.25	-0.27	-0.15	0.23	-0.69	-0.26	0.54	-0.53	-0.30	0.01	-0.52	0.02	0.70
	3 Y	-0.61	-0.21	0.52	-0.53	-0.27	0.24	-0.30	-0.19	0.18	-0.70	-0.24	0.57	-0.50	-0.30	0.01	-0.56	0.00	0.66

Table 9: Mean of variables in each super-class estimated with data from 2003

Super-classes	1	2	3	4	5	6	7	8	9	10	11
QA/CL	1.30	0.75	-0.14	-0.11	-0.47	0.25	0.74	-0.75	-0.42	-0.69	-0.76
TS/TA	0.12	-0.13	0.01	-0.28	1.54	-0.67	-0.26	-0.56	-0.24	-0.14	-0.52
EBIT/VA	0.52	0.69	0.34	0.14	0.20	0.07	-0.20	0.28	-0.13	-2.17	-0.28
EBITDA/PE	0.24	2.28	0.19	-0.05	-0.15	-0.18	-0.45	0.05	-0.16	-1.20	-0.93
FD/CF	0.13	-0.11	0.16	2.76	0.00	-0.10	-1.00	-0.05	-2.69	-0.30	-0.23
SF/PE	0.55	2.07	-0.07	-0.22	-0.28	0.05	0.34	-0.52	-0.27	-0.66	-0.99

Table 10: Distribution of firms (%) by number of years spent within the non-failed zone of the map presented on Figure 1 and by firm status

Nb. of years	6	5	4	3	2	1	0
Non-Failed	46.6	18.3	11.5	8.6	7.4	4.9	2.7
Failed	8.2	9.6	18.9	28.1	19.7	8.8	6.7

Table 11: Misclassification rates (%) achieved using traditional models

LS	TS	SM							BG							BO						
		DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
2003	1 Y	18.8	18.1	17.5	17.3	18.3	16.2	16.8	17.1	16.0	15.7	15.8	16.8	15.8	15.7	17.7	15.7	13.3	14.3	14.8	13.2	14.6
	2 Y	19.9	19.4	18.7	19.7	19.6	17.4	18.4	19.2	18.7	19.2	16.9	18.3	16.3	17.6	18.2	17.2	17.7	16.7	16.3	14.7	15.6
	3 Y	20.4	21.8	19.9	20.8	21.1	19.2	20.4	20.1	20.5	19.7	19.6	21.9	19.1	19.8	20.5	20.5	19.1	19.9	18.9	17.4	18.1
	4 Y	24.4	22.4	21.7	23.9	24.4	23.7	23.6	23.9	23.3	21.7	22.8	22.3	21.4	21.0	24.3	24.9	22.8	23.8	23.2	20.5	22.6
	5 Y	24.0	23.0	23.0	26.9	25.4	24.1	24.8	24.8	23.6	22.3	25.0	25.0	23.8	25.8	26.8	22.4	21.9	26.6	25.7	26.2	26.5
2006	1 Y	19.0	18.6	18.4	17.8	18.0	17.2	18.5	17.2	16.4	16.1	15.8	16.2	18.0	17.6	15.4	17.0	15.3	14.3	14.9	13.8	16.3
	2 Y	20.0	19.5	18.6	18.8	19.9	18.4	19.4	19.9	18.6	17.6	18.9	19.7	18.7	19.0	17.4	16.8	17.2	16.9	16.7	15.2	16.0
	3 Y	22.4	22.0	20.7	21.6	21.4	21.7	21.2	21.8	19.9	21.6	20.7	21.0	19.3	20.0	20.8	20.9	20.8	19.9	20.8	19.2	20.3
	4 Y	26.6	22.5	21.1	20.7	23.2	23.8	23.4	23.9	22.3	23.4	21.3	22.4	21.9	21.8	24.5	23.3	21.6	22.9	22.8	21.4	22.2
	5 Y	26.5	23.7	23.9	21.4	25.1	25.6	25.5	25.4	23.5	22.8	24.0	25.1	24.4	24.7	23.7	23.6	23.3	25.7	26.3	25.7	24.2
2010	1 Y	19.8	19.4	19.2	18.9	19.9	18.5	18.2	18.2	17.8	17.3	16.4	17.4	16.7	16.7	17.4	16.3	15.5	16.2	16.3	15.3	16.2
	2 Y	20.8	19.5	19.4	18.7	20.3	19.1	18.7	19.6	18.9	19.4	18.9	19.9	18.9	18.4	19.9	18.8	17.6	19.2	18.4	17.7	18.8

	3 Y	20.3	20.7	19.8	19.9	20.7	20.0	19.9	20.7	19.9	20.1	18.7	20.5	20.4	19.3	21.3	18.9	19.4	19.4	19.7	19.3	20.9
	4 Y	25.8	24.8	22.5	23.0	26.8	24.7	24.4	24.8	24.0	22.5	22.7	23.0	21.5	21.9	25.7	24.4	23.4	22.8	23.1	22.5	21.8
	5 Y	28.6	26.3	25.1	24.5	28.6	26.2	26.6	27.0	26.7	26.3	23.8	26.8	25.9	24.5	24.7	25.3	23.8	24.7	24.8	24.5	24.8

LS	TS	RS							DE							RF						
		DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
2003	1 Y	17.1	15.2	15.1	14.8	15.3	13.7	15.0	18.1	17.8	17.7	17.1	17.8	16.1	15.4	17.2	15.9	13.8	14.9	15.1	13.9	14.8
	2 Y	17.8	18.4	17.2	17.2	16.8	14.8	15.9	18.9	18.9	17.7	19.0	18.5	17.2	17.8	18.7	17.7	18.3	16.9	17.4	14.6	15.7
	3 Y	19.7	19.2	19.8	19.9	19.1	17.9	19.2	20.4	21.0	19.4	20.9	20.8	18.0	19.1	19.9	20.0	20.0	19.6	19.0	18.8	19.1
	4 Y	24.1	22.8	23.9	21.7	21.2	19.9	23.0	23.8	22.1	21.0	22.7	22.7	22.4	22.0	20.8	23.2	21.5	23.1	21.8	20.9	22.4
	5 Y	24.4	22.7	24.4	21.7	26.1	23.1	24.7	25.9	25.8	25.8	27.0	26.2	24.6	25.4	24.7	24.1	24.5	24.0	26.0	24.4	25.0
2006	1 Y	16.6	16.3	14.9	15.6	15.3	14.6	17.8	17.9	18.4	17.2	17.2	17.7	16.9	17.3	15.7	17.0	15.3	14.7	15.7	13.8	17.1
	2 Y	17.1	17.7	18.7	17.3	16.7	15.7	16.3	19.1	19.4	19.0	18.7	19.5	17.2	18.7	17.7	17.4	18.0	17.2	17.2	15.7	17.9
	3 Y	21.6	20.0	20.9	19.9	20.0	19.9	19.3	21.0	21.0	20.3	20.5	21.1	20.4	19.9	20.8	21.4	20.8	18.9	20.6	18.9	19.8
	4 Y	21.8	23.2	23.7	22.1	22.4	21.3	22.4	23.4	22.1	21.8	22.4	23.6	21.4	21.5	21.9	22.8	21.8	22.6	23.4	21.9	21.2
	5 Y	24.7	23.7	23.3	22.9	26.2	24.7	24.7	25.9	25.2	24.2	26.1	25.7	24.2	24.9	26.9	25.9	24.7	25.2	25.1	25.5	25.1
2010	1 Y	16.2	17.0	16.9	15.9	15.7	15.7	16.7	19.1	19.0	18.3	18.0	19.3	17.2	17.3	17.8	16.7	15.3	16.8	16.7	15.2	16.7
	2 Y	18.8	18.8	17.3	18.4	18.9	18.1	18.4	20.3	19.7	19.2	19.4	20.8	18.0	18.4	18.9	18.5	17.9	19.8	18.9	18.8	19.3
	3 Y	19.4	18.1	20.1	19.0	19.3	19.9	21.4	21.0	20.5	21.4	19.9	21.5	19.3	19.2	22.0	19.1	20.7	20.2	20.4	19.9	19.8
	4 Y	24.8	23.8	22.2	22.8	23.5	22.8	22.2	24.8	24.9	22.5	22.8	23.1	22.4	21.6	22.9	22.3	23.9	22.9	23.5	21.4	22.8
	5 Y	26.3	24.7	25.5	24.1	25.2	24.0	24.3	27.8	26.1	24.9	26.3	26.6	25.6	25.8	26.0	25.6	24.9	26.0	25.6	25.7	24.2

LS: learning sample; TS: test sample.
SM: single models; BG: bagging-based models; BO: boosting-based models; RS: random subspace-based models;
DE: Decorate-based models; RF: rotation forest-based models.
DA: discriminant analysis; LR: logistic regression; NN: neural network; Cox: Cox's model; C4.5: classification tree;
ELM: extreme learning machine; SVM: support vector machine.

Table 12: Differences (percentage point) between misclassification rates achieved using new models and those achieved using traditional models

LS	TS	SM*-SM							BG*-BG							BO*-BO						
		DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
2003	1 Y	0.27	-1.33	-1.11	-0.91	-0.95	-0.45	-0.56	-1.10	1.59	0.62	0.65	0.10	-0.60	0.10	-0.47	0.51	1.88	-0.06	-0.81	0.32	-1.34
	2 Y	-0.04	-0.41	-0.12	-0.38	-0.50	-0.12	-1.02	-0.31	-0.27	-1.17	-0.26	-1.00	-0.22	-0.38	-0.19	0.13	-0.98	-0.97	-0.09	0.60	0.08
	3 Y	-0.03	0.08	-0.32	0.01	-0.17	0.61	-0.11	0.09	-1.48	-0.11	-2.23	-3.44	-1.39	-1.40	-0.57	-1.97	-0.42	-2.58	-0.18	-0.68	-0.30
	4 Y	-1.58	0.08	-0.33	-2.49	-3.14	-3.83	-2.51	-2.53	-2.93	-1.67	-3.28	-1.95	-0.53	-1.29	-4.68	-4.61	-3.84	-5.02	-2.47	-3.09	-3.16
	5 Y	-1.53	0.29	-1.26	-5.83	-2.58	-2.66	-2.79	-3.26	-1.50	-2.19	-4.55	-3.93	-2.88	-5.51	-5.96	-1.46	-1.81	-6.52	-4.45	-6.96	-6.20
2006	1 Y	0.92	-0.69	-0.17	-0.07	0.24	0.06	-1.27	0.64	2.21	1.67	0.98	1.01	-1.69	-2.00	1.80	-0.19	1.59	1.96	0.73	2.19	-2.67
	2 Y	0.88	-0.26	0.70	-0.01	-0.13	-0.67	-0.53	-0.10	0.26	1.63	-0.27	-1.34	-0.36	-1.15	1.17	1.47	0.02	1.55	0.51	2.15	0.87
	3 Y	-2.21	-1.24	-1.11	-1.35	-0.27	-1.26	-0.68	-1.61	-0.55	-2.00	-0.98	-1.14	-0.20	-1.69	-1.38	-1.83	-2.36	-0.84	-1.19	0.35	-2.17
	4 Y	-3.30	-0.51	-0.74	0.17	-1.58	-1.88	-2.76	-2.70	-2.37	-2.52	-0.81	-1.02	-0.14	-1.62	-4.03	-2.91	-2.19	-3.74	-2.44	-0.92	-2.76
	5 Y	-2.63	-1.46	-2.31	0.48	-2.90	-3.42	-4.47	-3.62	-3.02	-2.73	-3.29	-3.37	-2.86	-3.01	-3.06	-2.85	-3.51	-5.61	-4.95	-6.24	-4.72
2010	1 Y	-0.18	-0.68	-1.00	-1.06	-1.18	-0.57	-0.88	-0.94	-0.23	-0.45	-0.72	-0.31	-1.55	-1.90	-1.57	0.11	0.22	-1.50	-1.54	-0.99	-1.98
	2 Y	-0.08	0.42	-1.18	0.19	-0.35	-0.24	0.78	-0.98	-0.08	-0.96	-1.16	-1.96	-3.09	-2.06	-2.11	-2.03	-1.40	-3.53	-1.67	-2.36	-2.87
	3 Y	1.73	-0.01	0.64	0.43	0.64	0.20	1.68	-0.19	-1.00	-1.38	0.40	-0.68	-2.70	-1.27	-2.43	0.07	-1.64	-1.75	-2.56	-3.10	-4.11
	4 Y	-2.42	-3.78	-1.84	-2.75	-4.79	-3.02	-1.79	-3.72	-4.73	-3.09	-3.01	-2.72	-2.14	-2.56	-5.35	-4.64	-5.00	-3.81	-3.14	-3.11	-3.03
	5 Y	-5.34	-4.61	-4.34	-3.94	-4.88	-2.49	-3.04	-4.34	-5.59	-6.44	-3.36	-5.30	-5.96	-3.92	-3.29	-4.94	-4.00	-4.95	-4.46	-4.55	-4.87

LS	TS	RS*-RS							DE*-DE							RF*-RF						
		DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
2003	1 Y	-0.35	-1.62	1.63	-1.14	-0.84	0.02	-1.19	-1.29	-1.63	-1.82	-0.98	-1.70	-0.90	0.63	-0.25	0.72	1.33	-1.64	-0.22	-0.04	-0.80
	2 Y	-0.58	-1.61	-0.35	-1.56	-0.06	0.95	0.59	-1.19	-0.21	0.59	-0.56	-0.70	0.12	-0.69	0.04	-0.45	-1.55	-2.02	-0.02	1.28	0.64
	3 Y	-1.39	0.56	-0.58	-3.34	-1.18	-0.48	-2.14	-0.43	0.17	0.63	-1.07	-1.84	0.64	-0.73	-0.02	-0.80	-0.73	-2.74	-0.21	-1.62	-2.33
	4 Y	-3.87	-1.45	-3.48	-2.62	-1.22	-2.01	-3.07	-1.90	-0.36	0.06	-2.38	-1.96	-2.46	-1.26	-0.87	-2.88	-1.69	-3.74	-1.89	-2.84	-3.44
	5 Y	-2.45	-2.42	-4.19	-1.63	-4.40	-3.18	-3.69	-2.98	-3.33	-4.49	-5.53	-5.16	-3.75	-4.28	-3.90	-2.93	-3.44	-3.64	-4.47	-5.24	-5.33
2006	1 Y	0.28	2.08	0.32	-1.45	0.50	1.73	-3.30	0.31	-2.07	-0.47	-1.00	0.01	0.06	-0.53	1.70	-0.63	1.39	1.19	0.30	1.94	-3.35
	2 Y	1.69	1.17	-3.26	-1.64	1.30	2.02	1.14	0.70	-0.64	0.00	-0.83	-0.56	0.05	-0.27	1.02	0.56	-0.53	1.47	0.89	0.36	-0.72
	3 Y	-3.31	-0.78	-4.16	-2.76	-0.03	0.09	0.18	0.34	-1.60	-0.71	-1.08	-0.16	-0.70	-0.44	-0.91	-1.39	-2.22	0.04	-1.82	0.07	-0.40
	4 Y	-1.51	-2.62	-4.54	-2.19	-1.42	-0.94	-2.38	-0.59	-1.08	-1.02	-2.02	-2.27	-0.30	-0.86	-0.56	-1.50	-1.83	-2.82	-3.94	-2.41	-0.58
	5 Y	-1.21	-2.46	0.38	-1.43	-4.87	-4.33	-4.15	-2.70	-3.85	-3.12	-4.56	-4.33	-2.26	-3.77	-6.29	-4.24	-4.50	-5.13	-3.50	-5.18	-5.76
2010	1 Y	-1.32	-0.90	-2.86	-1.69	-0.58	-1.05	-1.88	-1.39	-1.67	-1.08	-1.30	-2.16	0.27	-0.36	-2.10	0.01	2.62	-2.48	-1.84	-1.39	-2.93
	2 Y	-1.02	-4.00	-1.45	0.08	-3.13	-1.83	-2.44	-0.40	-0.77	-0.60	-1.75	-2.03	0.38	0.01	-1.10	-1.15	-1.22	-4.63	-2.55	-2.88	-2.56
	3 Y	-0.36	-0.66	-2.09	0.24	-0.94	-3.07	-3.70	0.37	-0.34	-1.88	0.02	-0.73	0.42	0.40	-2.93	-0.19	-2.37	-3.54	-1.80	-3.57	-2.47

4 Y	-4.46	-3.86	-2.53	-2.40	-3.21	-2.83	-2.79	-2.85	-4.14	-2.19	-2.34	-0.57	-0.99	-0.04	-1.52	-2.34	-5.29	-4.00	-4.52	-3.03	-3.33
5 Y	-6.17	-1.52	-3.82	-3.95	-4.40	-3.40	-3.60	-6.27	-5.15	-4.31	-4.88	-4.00	-2.50	-2.85	-4.42	-4.88	-5.06	-6.25	-5.97	-4.92	-4.28

New models are denoted by an (*).
Differences presented with normal characters are statistically significant at the threshold of 0.01, those presented with underlined characters are statistically significant at the threshold of 0.05 and those with double-underlined characters and not statistically significant (p-value>0.05).

Table 13: Differences (percentage points) between misclassification rates achieved using new models and those achieved using traditional models by type of model

Panel A: Differences							
Year	SM*-SM	BG*-BG	BO*-BO	RS*-RS	DE*-DE	RF*-RF	Mean
1 Y	-0.55	-0.10	-0.09	-0.65	-0.91	-0.31	-0.44
2 Y	-0.14	-0.73	-0.47	-0.68	-0.45	-0.73	-0.53
3 Y	-0.13	-1.18	-1.52	-1.43	-0.42	-1.53	-1.03
4 Y	-2.14	-2.26	-3.52	-2.64	-1.50	-2.63	-2.45
5 Y	-2.95	-3.85	-4.54	-3.19	-4.00	-4.74	-3.88

Panel B: P-values of a test for differences							
Year	SM*-SM	BG*-BG	BO*-BO	RS*-RS	DE*-DE	RF*-RF	Mean
1 Y	0.002	0.561	0.579	0.000	0.000	0.057	0.010
2 Y	0.428	0.000	0.006	0.000	0.012	0.000	0.002
3 Y	0.484	0.000	0.000	0.000	0.023	0.000	0.000
4 Y	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5 Y	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 14: Differences (percentage points) between misclassification rates achieved using traditional single models and those achieved using ensemble-base models

Panel A: Differences (percentage point) between traditional SM and ensemble-based models						
Year	BG-SM	BO-SM	RS-SM	DE-SM	RF-SM	Mean
1 Y	-1.60	-2.89	-2.53	-0.65	-2.60	-2.05
2 Y	-0.55	-1.96	-1.82	-0.42	-1.52	-1.25
3 Y	-0.51	-0.92	-1.04	-0.44	-0.77	-0.74
4 Y	-1.14	-0.63	-1.04	-1.07	-1.37	-1.05
5 Y	-0.37	-0.38	-0.84	0.52	0.01	-0.21

Panel B: P-values of a test for differences						
Year	BG-SM	BO-SM	RS-SM	DE-SM	RF-SM	Mean
1 Y	0.000	0.000	0.000	0.000	0.000	0.000
2 Y	0.002	0.000	0.000	0.019	0.000	0.000
3 Y	0.005	0.000	0.000	0.017	0.000	0.000
4 Y	0.000	0.001	0.000	0.000	0.000	0.000
5 Y	0.061	0.055	0.000	0.009	0.944	0.286

Panel C: Differences (percentage point) between traditional SM and new ensemble-based models						
Year	BG*-SM	BO*-SM	RS*-SM	DE*-SM	RF*-SM	Mean
1 Y	-1.70	-2.98	-3.19	-1.55	-2.91	-2.47
2 Y	-1.28	-2.43	-2.50	-0.87	-2.26	-1.87
3 Y	-1.70	-2.44	-2.46	-0.86	-2.29	-1.95
4 Y	-3.40	-4.15	-3.68	-2.57	-3.99	-3.56
5 Y	-4.22	-4.92	-4.03	-3.48	-4.73	-4.27

Panel D: P-values of a test for differences						
Year	BG*-SM	BO*-SM	RS*-SM	DE*-SM	RF*-SM	Mean
1 Y	0.000	0.000	0.000	0.000	0.000	0.000
2 Y	0.000	0.000	0.000	0.000	0.000	0.000
3 Y	0.000	0.000	0.000	0.000	0.000	0.000
4 Y	0.000	0.000	0.000	0.000	0.000	0.000
5 Y	0.000	0.000	0.000	0.000	0.000	0.000

Table 15: Decomposition of type-I and type-II errors into bias and variance components

Panel A: Traditional models					
Type	Year	Type-I		Type-II	
		Bias	Variance	Bias	Variance
SM	1 Y	15.09	10.88	13.04	5.07
	5 Y	23.42	12.49	17.76	6.53
BA	1 Y	19.38	7.13	12.79	3.78
	5 Y	25.43	9.90	19.26	5.00
BO	1 Y	15.30	10.08	9.55	6.13
	5 Y	22.18	11.21	15.21	8.88
RS	1 Y	16.10	10.02	11.02	4.89
	5 Y	22.85	10.40	17.50	5.71
DE	1 Y	17.53	8.39	12.88	4.73
	5 Y	24.89	9.34	18.67	6.67
RF	1 Y	16.05	8.01	10.53	5.25

Panel B: New models					
Type	Year	Type-I		Type-II	
		Bias	Variance	Bias	Variance
SM	1 Y	15.24	11.08	10.16	7.10
	5 Y	17.21	13.17	13.18	8.34
BA	1 Y	18.36	8.73	11.35	5.78
	5 Y	17.92	11.14	13.23	7.20
BO	1 Y	17.68	9.56	8.37	7.48
	5 Y	16.79	11.26	11.21	8.85
RS	1 Y	15.93	12.10	9.35	6.04
	5 Y	15.52	11.25	13.97	7.90
DE	1 Y	17.09	12.12	10.60	5.42
	5 Y	16.31	12.76	13.64	7.59
RF	1 Y	15.38	11.73	9.41	6.37

5 Y	22.05	10.13	16.65	8.48	5 Y	14.79	12.97	11.74	8.29					

Table 16: Area under the ROC curve (AUC) by type of model

Panel A: Traditional models

Year	SM							BG							BO						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.750	0.779	0.800	0.773	0.781	0.804	0.802	0.803	0.804	0.807	0.801	0.806	0.804	0.801	0.805	0.802	0.808	0.804	0.802	0.818	0.810
2 Y	0.760	0.783	0.761	0.746	0.782	0.801	0.803	0.791	0.786	0.777	0.790	0.797	0.799	0.806	0.783	0.768	0.753	0.801	0.789	0.809	0.799
3 Y	0.728	0.731	0.768	0.754	0.736	0.775	0.787	0.782	0.772	0.750	0.781	0.785	0.785	0.789	0.743	0.765	0.765	0.725	0.757	0.762	0.772
4 Y	0.706	0.721	0.722	0.736	0.726	0.750	0.738	0.762	0.746	0.756	0.768	0.723	0.772	0.765	0.740	0.746	0.749	0.760	0.748	0.754	0.749
5 Y	0.684	0.716	0.721	0.708	0.697	0.713	0.709	0.722	0.709	0.733	0.717	0.720	0.728	0.736	0.720	0.732	0.730	0.731	0.721	0.744	0.726
Year	RS							DE							RF						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.807	0.815	0.803	0.804	0.810	0.827	0.793	0.783	0.776	0.801	0.798	0.795	0.804	0.802	0.807	0.802	0.812	0.808	0.796	0.807	0.813
2 Y	0.785	0.793	0.789	0.793	0.791	0.802	0.809	0.766	0.763	0.796	0.775	0.784	0.799	0.801	0.798	0.787	0.796	0.786	0.789	0.797	0.809
3 Y	0.753	0.785	0.747	0.792	0.765	0.776	0.785	0.758	0.756	0.772	0.765	0.756	0.761	0.775	0.770	0.751	0.756	0.744	0.751	0.744	0.785
4 Y	0.711	0.745	0.729	0.716	0.757	0.741	0.746	0.722	0.731	0.745	0.719	0.743	0.739	0.745	0.751	0.741	0.723	0.752	0.734	0.746	0.727
5 Y	0.702	0.729	0.721	0.723	0.723	0.733	0.742	0.705	0.699	0.717	0.710	0.709	0.711	0.710	0.723	0.727	0.712	0.712	0.703	0.727	0.711

Panel B: New models

Year	SM							BG							BO						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.739	0.786	0.777	0.802	0.810	0.802	0.805	0.803	0.821	0.797	0.811	0.810	0.812	0.817	0.827	0.802	0.796	0.816	0.802	0.829	0.828
2 Y	0.756	0.753	0.771	0.778	0.769	0.803	0.800	0.792	0.774	0.770	0.805	0.791	0.812	0.799	0.783	0.777	0.768	0.775	0.787	0.801	0.811
3 Y	0.755	0.745	0.756	0.761	0.766	0.789	0.798	0.791	0.773	0.775	0.799	0.781	0.803	0.793	0.745	0.779	0.767	0.773	0.765	0.786	0.795
4 Y	0.722	0.717	0.723	0.738	0.746	0.741	0.769	0.756	0.751	0.767	0.786	0.783	0.773	0.774	0.779	0.757	0.770	0.777	0.743	0.772	0.796
5 Y	0.728	0.734	0.734	0.721	0.733	0.731	0.744	0.746	0.739	0.763	0.767	0.762	0.764	0.747	0.759	0.760	0.769	0.770	0.743	0.771	0.772
Year	RS							DE							RF						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.814	0.806	0.803	0.825	0.803	0.812	0.823	0.789	0.794	0.812	0.801	0.813	0.807	0.811	0.801	0.802	0.824	0.812	0.810	0.821	0.825
2 Y	0.783	0.792	0.800	0.806	0.798	0.807	0.811	0.766	0.786	0.798	0.787	0.776	0.801	0.798	0.798	0.806	0.813	0.781	0.780	0.815	0.801
3 Y	0.779	0.793	0.795	0.792	0.777	0.796	0.790	0.746	0.762	0.765	0.772	0.788	0.775	0.784	0.785	0.787	0.798	0.772	0.766	0.799	0.795
4 Y	0.778	0.758	0.748	0.791	0.778	0.776	0.762	0.734	0.757	0.755	0.744	0.744	0.756	0.761	0.761	0.762	0.761	0.777	0.758	0.751	0.768
5 Y	0.765	0.751	0.749	0.777	0.762	0.753	0.771	0.713	0.745	0.744	0.740	0.734	0.745	0.746	0.753	0.744	0.743	0.772	0.753	0.757	0.765

Underlined figures correspond to situations where new models achieve lower AUC than those achieved with traditional models.

Table 17: H-measure by type of model

Panel A: Traditional models

Year	SM							BG							BO						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.425	0.396	0.445	0.429	0.427	0.452	0.457	0.453	0.465	0.454	0.438	0.449	0.473	0.466	0.463	0.430	0.479	0.471	0.457	0.519	0.486
2 Y	0.397	0.422	0.428	0.399	0.415	0.444	0.443	0.428	0.428	0.408	0.437	0.433	0.438	0.450	0.433	0.450	0.420	0.438	0.431	0.475	0.454
3 Y	0.391	0.391	0.420	0.401	0.401	0.418	0.439	0.410	0.448	0.402	0.399	0.408	0.423	0.412	0.400	0.411	0.432	0.388	0.415	0.433	0.428
4 Y	0.350	0.376	0.374	0.392	0.389	0.420	0.413	0.429	0.410	0.412	0.431	0.414	0.400	0.421	0.398	0.401	0.418	0.402	0.403	0.415	0.433
5 Y	0.357	0.360	0.386	0.352	0.371	0.377	0.387	0.363	0.343	0.396	0.367	0.388	0.392	0.404	0.348	0.386	0.393	0.403	0.389	0.409	0.418
Year	RS							DE							RF						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.459	0.478	0.455	0.453	0.466	0.515	0.482	0.450	0.438	0.456	0.434	0.447	0.461	0.489	0.500	0.504	0.531	0.518	0.454	0.510	0.517
2 Y	0.429	0.450	0.422	0.432	0.426	0.499	0.479	0.422	0.414	0.434	0.447	0.449	0.439	0.478	0.476	0.468	0.484	0.495	0.462	0.488	0.510
3 Y	0.433	0.429	0.397	0.439	0.411	0.447	0.446	0.413	0.424	0.441	0.423	0.432	0.401	0.454	0.438	0.437	0.476	0.424	0.458	0.434	0.475
4 Y	0.377	0.408	0.398	0.366	0.422	0.411	0.411	0.433	0.399	0.418	0.395	0.404	0.411	0.403	0.445	0.414	0.436	0.442	0.405	0.455	0.432
5 Y	0.347	0.379	0.391	0.394	0.406	0.400	0.423	0.377	0.356	0.388	0.375	0.382	0.385	0.377	0.413	0.423	0.398	0.399	0.412	0.408	0.437

Panel B: New models

Year	SM							BG							BO						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.388	0.413	0.407	0.469	0.469	0.477	0.458	0.450	0.490	0.444	0.489	0.476	0.477	0.484	0.481	0.473	0.447	0.478	0.465	0.510	0.509
2 Y	0.397	0.392	0.429	0.416	0.407	0.437	0.457	0.427	0.398	0.417	0.444	0.423	0.462	0.432	0.436	0.400	0.406	0.396	0.433	0.477	0.475
3 Y	0.395	0.412	0.401	0.422	0.413	0.427	0.445	0.439	0.450	0.388	0.439	0.409	0.424	0.453	0.401	0.419	0.440	0.413	0.416	0.446	0.433

4 Y	0.372	0.391	0.389	0.381	0.401	0.421	0.427	0.438	0.413	0.437	0.440	0.416	0.416	0.428	0.407	0.413	0.428	0.404	0.415	0.424	0.435
5 Y	0.391	0.389	0.398	0.388	0.399	0.409	0.432	0.399	0.392	0.415	0.411	0.402	0.399	0.414	0.409	0.411	0.426	0.430	0.399	0.412	0.433

Year	RS							DE							RF						
	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM	DA	LR	NN	Cox	C4.5	ELM	SVM
1 Y	0.487	0.452	0.457	0.498	0.461	0.490	0.511	0.439	0.445	0.500	0.473	0.489	0.457	0.490	0.486	0.477	0.502	0.493	0.484	0.525	0.520
2 Y	0.412	0.456	0.442	0.450	0.444	0.476	0.496	0.429	0.413	0.461	0.456	0.449	0.455	0.463	0.478	0.469	0.490	0.449	0.467	0.491	0.470
3 Y	0.429	0.432	0.433	0.444	0.421	0.458	0.477	0.418	0.428	0.414	0.424	0.463	0.446	0.458	0.443	0.457	0.488	0.435	0.460	0.485	0.475
4 Y	0.403	0.410	0.410	0.428	0.425	0.444	0.438	0.403	0.404	0.427	0.408	0.424	0.427	0.439	0.447	0.433	0.444	0.452	0.442	0.456	0.443
5 Y	0.428	0.401	0.396	0.412	0.411	0.409	0.427	0.388	0.406	0.410	0.396	0.401	0.396	0.414	0.429	0.426	0.452	0.437	0.421	0.438	0.438

Underlined figures correspond to situations where new models achieve lower H-measure than those achieved with traditional models.

9. Figures

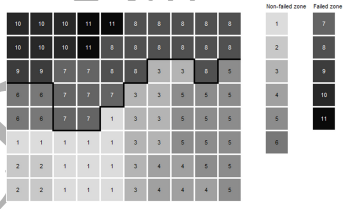

Figure 1: Distribution of meta-classes on the Kohonen map calculated with data from 2003. Neurons are numbered using a scale of financial health: 1 corresponds to neurons that represent firms with the best overall financial situation, and 11 to neurons that represents firms with the weakest situation. The heavy black line corresponds to the boundary between the failed and the non-failed zone.

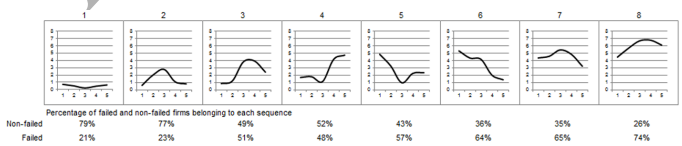

Figure 2: Prototype sequences of change in firm position on the map designed with data from 2003