Prediction of a financial crisis in Latin American companies using the mixed logistic regression model

Viviana Giampaoli, Karin A. Tamura, Norma P. Caro, Luiz J. Simões de Araujo

Artículo publicado en Chilean Journal of Statistics
Volumen 7, Número 1, 2016 – ISSN 0718-7912 / e-ISSN 0718-7920

Chilean Journal of Statistics
Month 200x

Prediction of a financial crisis in Latin American companies using the mixed logistic regression model

Viviana Giampaoli[1,*], Karin A. Tamura[1], Norma P. Caro[2], and Luiz J. Simões de Araujo[3]

[1]Department of Statistics, University of São Paulo, São Paulo, Brazil,
[2]Faculty of Economic Sciences, National University of Córdoba, Córdoba, Argentina,
[3]Faculty of Economics, Administration and Accounting, University of São Paulo, São Paulo, Brazil

(Received: 00 Month 200x · Accepted in final form: 00 Month 200x)

Abstract

The development of statistical methods for predicting the financial crisis of a company is a real contribution to scientific research. These methods identify possible adverse financial situations of the companies, through the behavior of their financial indicators. The contribution of this work is to compare the binary classification by different prediction methods of mixed logistic models to predict a future financial crisis in new companies. The results based on an application involving companies from the Argentina, Peru and Chile Stock Exchange showed that all prediction methods were able to predict with high accuracy the financial crisis of the next year.

Keywords: Prediction· Logistic Mixed Model · Financial Crisis

Mathematics Subject Classification: Primary 62M20 · Secondary 65C60, 62P05.

1. Introduction

Faced with the negative impact of the financial crisis in companies, caused by social, economic and employment factors, different economic entities (financial institutions, investors, suppliers, etc.) have been shown the necessity to anticipate or predict crises. The information provided in a companys financial statement and the ability to analyze the evolution over time of financial ratios allows building of models for predicting the risk of a crisis. A financial crisis is defined as the inability to meet payment obligations, resulting in huge losses and even extreme situations such as bankruptcy.
The analysis and interpretation of accounting statements (balance sheets) is the information system directive that investigates what the situation of the company is, in order to determine the causes and suggest more appropriate courses of action which depend on the intended purpose.

Caro (2014) provides a review of the literature and shows that many papers have been published with the aim of proposing models that predict insolvency, based on the information contained on balance sheets. In developed economies, we can mention Altman (1968), Olson (1980), Jones and Hensher (2004), among others. Importantly, these studies have been replicated in emerging economies (Altman et al. (1979), Sandin and Porporato

*Corresponding author. Email: vivig@ime.usp.br

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

(2007), Caro et al. (2013), among others). Most of these, based on the models of Altman (1968), use cross-sectional methodology. The article by Jones and Hensher (2004) was one of the first to use mixed logistic models in Australian companies and Caro et al. (2013) does it in Argentine companies.

This paper proposes the use of a mixed logistic regression model to predict the probability of a companys financial crisis (FC) based on the information obtained from the financial statements of companies from three Latin American countries (Argentina, Chile and Peru), which are available at the Buenos Aires, Santiago de Chile and Lima Stock Exchange, respectively, for the decade of 2000. The innovative contribution of this work is that it enables the prediction of crises in new companies or firms that were not part of the database for the fit of the model.

The paper is organized as follows. Section 2 illustrates the data set about FC, providing a brief explanation of the construction of data sets and covariates considered, while Section 3 describes the mixed logistic model. Section 4 presents the different prediction methods. The proposal results in the application are presented in Section 5. Finally, Section 6 discusses the proposed analysis. All the analysis has been performed by R version 2.10.1; see http://www.r-project.org.

2. Data Set Description

The study of company crisis considers three Latin American countries (Argentina, Chile and Peru) and was developed within 239 companies observed from 2003 to 2011, where 7% of them went into crisis the following year. For analysis purposes, the companies were considered as present (state: 1) or not present in financial crisis (state: 0), which is the binary response variable. In this work, we defined the companies in crisis as those that were classified as insolvent or bankrupt. The date on which the company enters this state was considered the year of the crisis. For each of the companies that comprise the sample, we consider the four previous financial statements of the year when there were signs of crisis. For the "financially healthy" companies, the financial statements of the same periods of the business problems were considered. The aim of the study is to predict whether a company will present a state of crisis in the next year, given its financial indicators in the previous periods. We considered for the analysis a mixed logistic model. The database was divided into two parts: construction data set (a balanced sample of companies from 2003-2008) and future prediction data set (companies from 2009-2011). In order to make predictions for future years, in the construction data set, we considered 64 companies, balancing the companys state (0: financially healthy or 1: crisis), that is, with 50% of financially healthy companies and 50% of companies in crisis. Thus, the cutoff for classifying a company in crisis was 50%. The other companies, from 2009 on, were considered as a future prediction database, i.e., based on the estimates provided by the model from construction database, we predicted the crisis of the companies for the next year.

The ratios defined by Jones and Hensher (2004) and Altman (1993) were considered as covariates. These are calculated based on the information contained in financial reports issued by the Stock Exchange, after the presentation on the balance sheet by companies and are listed as follows:

- Working capital (current assetscurrent liabilities) by total assets (CT-AT)
- Cash on total assets(E-AT)
- Net operating cash flow by total assets (FF-AT)
- Total sales revenue by total assets (V-AT)
- Returns by total assets (GE-AT)

The mixed logistic model was built considering indicators analyzed over time, i.e., taking into account the historical data available for each company. Note that companies may have different historical data. Therefore, we consider the company as a group and its set of indicators over time as observations within the group (company).

3. Model Specification: Mixed Logistic Model

Mixed effects logistic regression is used to model the binary outcome, in which it is possible to model the probability of response as a function of predictor variables. This model incorporates a potential clustering structure of the data through the inclusion of random effects. Let y_{ij} be independent of Bernoulli, in which i indexes the group, $i = 1, \ldots, q$, and j indexes the observation within the i-th group, $j = 1, \ldots, n_i$ conditional on α_i, we consider the random intercept model given by

$$\text{logit}[P(y_{ij} = 1|\alpha_i)] = \log[\frac{p_{ij}}{1 - p_{ij}}] = \mathbf{x}_{ij}^t\boldsymbol{\beta} + \alpha_i, \tag{1}$$

in which $\boldsymbol{\beta}$ is an unknown vector of fixed effects $(p \times 1)$, $\mathbf{x}_{ij}^t$ of known covariates $(1 \times p)$ is associated with $\boldsymbol{\beta}$, defined by $\mathbf{x}_{ij}^t$=(1, x_{1ij}, x_{2ij},..., $x_{(p-1)ij}$). The parameter α_i is an unknown random intercept with α_i i.i.d. with $\alpha_i \sim \mathcal{N}(0,\sigma^2)$, in which σ^2 is the unknown variance. The usual way to estimate the parameters of the model (β, σ) is to integrate the conditional likelihood $L(\boldsymbol{\beta}|u)$ on the random effect α_i,

$$L(\boldsymbol{\beta}|D(\sigma)) = \int_{\alpha_i} L(\boldsymbol{\beta}|u)\phi(\alpha_i)d\alpha_i \tag{2}$$

in which $\phi(.)$ represents the normal density. In this article, we consider the Laplace Approximation to calculate (2), although other techniques may be used in a straightforward manner.

4. Prediction methods for new groups

In this section, we introduce some prediction methods developed for mixed logistic regression to predict the outcome of a future period, such as naive, empirical best prediction, linear regression and nearest neighbor. These methods were adopted for the mixed logistic model, and are described in the next subsections. These techniques have been developed for k random effects, but in this article, we considered $k = 1$, i.e., one random effect that is the random intercept of the mixed logistic model (1).

4.1 Naive

The simplest method is the method called naive. This method is based on the assumption of the model (1), that $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q$ are i.i.d. with $\boldsymbol{\alpha}_i \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$.

The method assumes that the value of the random effects for the new groups is simply the mean of the distribution to be considered. Thus, the methodology ignores the existence of the random part of the model. Therefore, making a prediction for observations belonging to new groups, the naive method considers only the fixed part estimated by the mixed logistic model. Thus, the predicted probability of j-th observation of the i-th new cluster is given by

$$p_{ij} = P(y_{ij} = 1|\boldsymbol{\alpha}_i = 0) = \frac{\exp\{\mathbf{x}_{ij}^t\hat{\boldsymbol{\beta}}\}}{1+\exp\{\mathbf{x}_{ij}^t\hat{\boldsymbol{\beta}}\}}. \quad (3)$$

In a traditional logistic model, which considers only the fixed part of the linear predictor, the prediction would be performed the same way. The difference in this case is given by the fact that to make a prediction for new groups, the estimates used are those obtained from the fixed parameters of the random effects model (1).

4.2 Empirical Best Prediction (EBP)

Since the purpose of predicting the outcome of the i-th new group in the observation level of model (1) based on the conditional expectation of the random effect, Tamura and Giampaoli (2010) propose the use of the Empirical Best Prediction (EBP) method based on approach of Jiang and Lahiri (2001). Empirical Bayes predictors in the terms of interest are the means of the empirical posterior distribution (with the parameters substituted by their estimates), i.e., $\hat{\varsigma} = E(\varsigma|y)$. It has the property that minimizes the mean-squared error of prediction (MSEP) $E(\varsigma' - \varsigma)^2$ for predictor ς' of ς over the joint distribution of (ς) and responses. It can be calculated as follows

$$\varsigma = \frac{E(\varsigma(\boldsymbol{\beta},\alpha_i)\exp(S_i(\boldsymbol{\beta},\alpha_i)))}{E(\exp(S_i(\boldsymbol{\beta},\alpha_i)))} = \frac{\int \varsigma(\boldsymbol{\beta},\alpha_i)\exp(S_i(\boldsymbol{\beta},\alpha_i))f_v(\alpha_i)d\alpha_i}{\int \exp(S_i(\boldsymbol{\beta},\alpha_i))f_v(\alpha_i)d\alpha_i}, \quad (4)$$

with $S_i(\boldsymbol{\beta},\alpha_i) = \sum_{j=1}^{n_i}[y_{ij}h(\mathbf{x}_{ij}^t\boldsymbol{\beta}+\alpha_i) - \log(1+\exp(\mathbf{x}_{ij}^t\boldsymbol{\beta}+\alpha_i))]$. Jiang and Lahiri (2001) suggested to use $S_i(\boldsymbol{\beta},\alpha_i)$ as $S_i(\boldsymbol{\beta},\sigma\xi) = \sum_{j=1}^{n_i}[y_{ij}\sigma\xi - \log(1+\exp(\mathbf{x}_{ij}^t\boldsymbol{\beta}+\sigma\xi))]$, in which $\alpha_i = \sigma\xi$ with $\xi \sim \mathcal{N}(0,1)$.

The objective is to particularize equation (4) to predict a new value of response variable. Thus, we call function ς as the outcome defines the logistic regression model which we are interested in predicting the function

$$\varsigma(\boldsymbol{\beta},\alpha_i) = p_{ij} = \frac{\exp(\mathbf{x}_{ij}^t\boldsymbol{\beta}+\alpha_i)}{1+\exp(\mathbf{x}_{ij}^t\boldsymbol{\beta}+\alpha_i)}.$$

With this aim, we assumed that $y_{i.} = n_{i.}/2$ and $\sum_{j=1}^{n_i} a_{ij} = n_{i.}$, i.e, the prior assumption of equal success and failure probabilities. Then, the mixed logistic model is given by (5) in which the expectations are taken with respect to ξ, the estimated probability for the j-th observation of the i-th new cluster presenting the event 1 is given by:

$$\hat{p}_{ij}(\hat{\boldsymbol{\beta}},\hat{\sigma}\xi) = \frac{E\left(\frac{\exp(\mathbf{x}_{ik}^t\hat{\boldsymbol{\beta}}+\hat{\sigma}\xi)}{1+\exp(\mathbf{x}_{ik}^t\hat{\boldsymbol{\beta}}+\hat{\sigma}\xi)}\cdot\exp(y_{i.}\hat{\sigma}\xi - \sum_{k=1}^{n_i}\log(1+\exp(\mathbf{x}_{ik}^t\hat{\boldsymbol{\beta}}+\hat{\sigma}\xi)))\right)}{E(\exp(y_{i.}\hat{\sigma}\xi - \sum_{k=1}^{n_i}\log(1+\exp(\mathbf{x}_{ik}^t\hat{\boldsymbol{\beta}}+\hat{\sigma}\xi))))}. \quad (5)$$

Note that in order to predict the response variable, we do not know the true value of $y_{i.}$.

Thus, it was assumed $y_{i.} = n_{i.}/2$, with probability 50% that event 1 may happen. To analyze the influence of this assumption over the results, we considered different values for $y_{i.}$. Through simulation studies, we concluded the position among the predicted probabilities of the observations were the same and the order of the predicted values did not depend on $y_{i.}$.

4.3 Linear Regression Predicton Method (LRPM)

The Linear Regression Prediction Method (LRPM) was developed by Tamura and Giampaoli (2013) for the logistic mixed model with k random effects. The methodology considers the adjustment of an additional regression model to predict the random effects based on a construction data set. This model considers that the response variable is the predicted random effect $\hat{\boldsymbol{\alpha}}_i$ (with $i \in G$). Hereafter, the estimated parameters of these regression models are used to predict the random effects for the l-th new group, now with $l \notin G$.

Thus, after obtaining the estimate parameters of model (1), it is necessary that all of the covariates available at the observation level should be aggregated at the group level, i.e.,

$$\mathbf{w}_i^t = (\mathbf{x}_{i.}^t, \mathbf{z}_{i.}^t), \tag{6}$$

with $i \in G$, because the linear regression modelo is ajusted in the group level.

We consider for each m-th random effect estimate, with $m = 1, \ldots, k$, a model of the form

$$\hat{\alpha}_{mi} = f(\mathbf{w}_{mi}^t \boldsymbol{\lambda_m}) \tag{7}$$

that was able to explain the relationship between the covariates and the random effects, where $\boldsymbol{\lambda_m} = (\lambda_{m1}, \ldots, \lambda_{mp})^t$ is the vector of unknown regression coefficients and $\mathbf{w}_{mi}$ is the vector of known covariates $(p \times 1)$ of the i-group and the m-th random effect, aggregated at the group level as described in (6). The general model (7) may be particularized to a linear regression model, given by

$$\hat{\alpha}_{mi} = \mathbf{w}_{mi}^t \boldsymbol{\lambda_m} + \varepsilon_{mi}, \tag{8}$$

with $\varepsilon_{mi} \sim \mathcal{N}(0, \sigma_m^2)$, independent. The parameter λ_{m1} is the fixed intercept and $(\lambda_{m2}, \ldots, \lambda_{mp})^t$ are the fixed slopes of the vector $\boldsymbol{\lambda_m}$. In the assumption of model (1), as $\boldsymbol{\alpha_i}$ has a multivariate normal distribution, then each marginal m of random effects (α_{mi}) has univariate normal distribution, thus the parameters of each m-linear independent models can be performed by the usual estimation methods, such as Least Squares or Maximum Likelihood.

In the application data set for predicting the random effects for the l-th new group with $l \notin G$, we use the following regression equation:

$$\hat{\alpha}_{ml}^* = \mathbf{w}_{ml}^t \hat{\boldsymbol{\lambda}}_{\boldsymbol{m}}. \tag{9}$$

Thus, it is possible to predict the outcome probability for j-th observation within the m-th new group by using the logistic function of the mixed logistic model considering (9).

4.4 Nearest Neighbor Prediction Method (NNPM)

Tamura, et al. (2013) developed the Nearest Neighbor Prediction Method (NNPM) considering a logistic mixed model with k random effects. In this proposal for a new group, the prediction of the random effect is based on a vector of covariates or feature vector by using a distance (e.g. Euclidian, Mahalanobis, City Block, etc) and a centrality measurement (e.g. mean, median, medoid, etc) of the known random effects of the l nearest neighbors. The advantage of this technique does not require any distribution of the empirical random effect, for more details, see Tamura, et al. (2013).

The proposed approach is based on the nearest neighbors technique, which is commonly used for supervised pattern classification (see for example, Cover and Hart (1967), Nigsch et al (2006)).

The goal is to assign random effects to new groups based on the covariates in the observation level or aggregated at the group level. Since the random effects are continuous outcomes, we applied the NN technique, in which the assignment is performed by considering some centrality measurement (e.g. mean, median, medoid, etc) of the known random effects of the l-NN.

The NNPM is described as follows, applied to mixed logistic regression (see, Tamura et al. (2014)). For each $i \in G = \{1, \ldots, q\}$, there is a feature vector g_i, and a known random effect vector $\hat{\boldsymbol{\alpha}}_i=(\hat{\alpha}_{1i}, \ldots, \hat{\alpha}_{ki})$ estimated by model (1). The objective is to predict the values of the random effects for the i'-th new group ($i' \notin G$) represented by $\boldsymbol{\alpha}_{i'}$ from g'_i. The algorithm is described as follows:

1:For i' in 1 to q' {
2: For i in 1 to q {
3: Compute the distance $d_{(i',i)}$ between $g_{i'}$ and g_i;
4: }
5: Sort the elements of $d_{(i',.)} = (d_{(i',1)}, d_{(i',2)}, \ldots, d_{(i',q)})$ in increasing order;
6:}
7:For l in 1 to q {
8: For i' in 1 to q' {
9: Compute a centrality measurement of the known random effects, $\boldsymbol{\alpha}_{i'}=(\bar{\alpha}_{1.}, \ldots, \bar{\alpha}_{k.})$, corresponding to the l first elements of the sorted $d_{(i',.)}$;
10: The random effects $\boldsymbol{\alpha}_{i'}$ are inserted in the linear predictor of the mixed logistic regression, providing the outcome probability of the i'-th new group in the observation level;
11: }
12:}
13: Select l which maximizes the performance prediction of the mixed logistic model.

Lines 1-6 compute the distances, which can be stored in a matrix with elements $d_{(i',i)}$. In the following lines, the approach computes l that maximizes the performance prediction of mixed model with the predicted random effects for new groups.

5. Results

The application of the proposed methodology to the problem of prediction of the financial crisis on companies are presented in this section. In both bases construction data set and prediction data set was observed a large dispersion between the values of variables, so it was decided to present the interquatile range (IQR), minimum (Min), median, and maximum (Max) (Table 1). For each variable, the number of observations was 271 and 506 in the construction base and the prediction base, respectively. Considering that the dispersion is large, the medians of all investigated variables are relatively close.

Table 1. Descriptive measures.

Variable	data set	IQR	Min	Median	Max
CT-AT	construction	22.29	-263.86	5.70	99.95
	prediction	26.08	-263.86	10.83	68.51
E-AT	construction	1.44	0.00	0.47	33.61
	prediction	4.63	0.00	2.04	64.12
FF-AT	construction	9.61	-266.70	3.66	126.01
	prediction	10.30	-64.51	8.70	64.74
GE-AT	construction	10.51	-271.55	0.79	45.39
	prediction	12.44	-52.83	4.53	90.57
V-AT	construction	62.08	0.00	38.82	286.41
	prediction	69.01	-6.42	67.02	374.85

For the construction base, the index plot of the classical and robust (based on the Minimum Covariance Determinant estimator-MCD) Mahalanobis distances are shown in Figure 1 (a) e (b) respectively. It can be seen that there are many points as potential multivariate outliers because they are above the horizontal cut off line (97.5% quantile). However, note that in this case the calculations were performed with the estimator for the mean and covariance of the entire set of observations. But when applying the methodology will be considered NNPM groups of neighboring observations to calculate the distances with which distant observations will not be a problem.

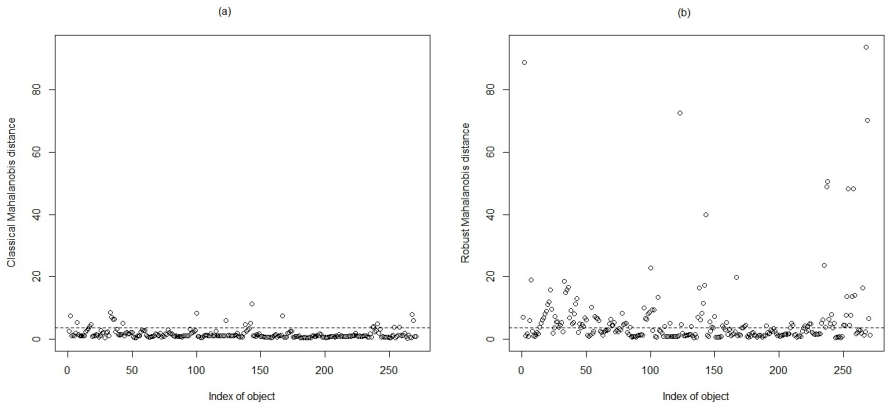

Figure 1. Index plot: (a) classical Mahalanobis distances, (b) robust Mahalanobis distances.

Using the construction database or the construction data set, a mixed logistic model was adjusted to the response, whether or not the company was is in crisis, and the random effect was considered by the company. A process of backward type was applied to the country considered as a fixed effect and all the variables of balance defined in Section 2, the selected variable was Returns (GE-AT); the results are present in Table 2.

Table 2. Output of the mixed logistic model.

	Estimate	Std. Error	z value	p-value
Fixed Effect (Returns)	-0.2320	0.1006	-2.306	0.0211
Random intercept variance	62.442			
Random intercept standard deviation	7.902			

The diagnosis of the selected mixed model, which the estimate parameters are in Table 2, were conducted by using normalized randomized quantile residuals (Dunn and Smyth (1996), Rigby and Stasinopoulos (2005)).

Figure 2 provides the diagnostic plots of mixed model with two random effects (2). Figure 2 (c) provides QQ-plots of the empirical random intercept, indicating some departure of the random effect from a normal distribution. Such a fact may be expected as Huang (2009) indicates in some situations it may be unrealistic to assume the normality of the random effects. However, this is not a cause for concern, because authors Neuhaus et al. (1992), McCulloch and Neuhaus (2011) and Neuhaus et al. (2013) use simulation studies to show that most aspects of statistical inference were robust at the misspecification of the random effects distribution, i.e., the lack of the normality of the random effects. This explains why even in the face of an apparent lack of normality of random effects in this case, the residuals appear satisfactory, as the quantile residuals follow the normal distribution (e.g., Figure 2 (a) and Figure 2 (b)).

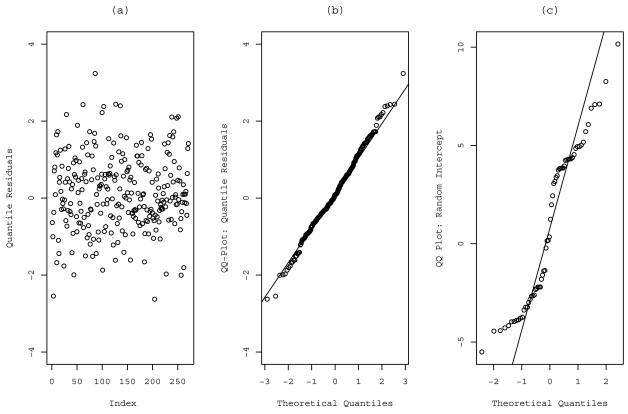

Figure 2. Diagnostic plots for the mixed model with two random effects (2): (a) quantile residuals against index, (b) QQ-plot of the quantile residuals, (c) QQ-plot of the random intercept.

According to the value of the variance of the random intercept, we can conclude that there is a large dispersion between companies. So, it makes sense to consider the mixed model. Furthermore, an estimate of the fixed negative effect indicates that the higher the rate of return, the lower the risk the company runs of going into crisis in the next year.

The wide dispersion of the intercept is a phenomena present in various contexts and models used for different corporations. Each corporation has quirks and idiosyncrasies typical of its governance, its environment and its business sector, as reflected in the variance. Articulate well-reasoned accounting data, which have been audited and appropriately chosen to apply to the model above described, had proved to be very productive. This articulation can provide relevant support in estimating the probability of financial crisis in the short and medium-terms in these corporations.

To assess model performance in binary classification by model, we consider measures of sensitivity, specificity and KS and accuracy in the cutoff of 50%: The sensitivity measures the proportion of actual positives (company in crisis) which are correctly predicted such as for model (the percentage of companies who correctly identified as having the condition). The specificity measures the proportion of actual negatives ("healthy firms"), and Kolmogorov-Smirnoff (KS) statistic (sensitivity-(1-specificity)) and the accuracy the overall correct classification in relation to total dataset.

From Table 3 it is clear that the mixed logistic model was an excellent predictive tool by the measures of performance obtained in the construction data set.

This model is reduced in terms of the fixed effects that the full model presented by Caro et al. (2013); nevertheless, we obtained a higher accuracy (93.80% vs 91.00%), although this work in addition to the Argentine companies were also considered at Peru and Chile

Table 3. Specificity, sensibility, KS and accuracy of the model in the construction data set.

Specificity	90.63%
Sensibility	96.88%
KS	87.50%
Accuracy	93.80%

companies.

As the estimates are based on the basis of construction, would it be possible to predict the probability of a crisis in the company the following year? To answer this question, we consider 90 companies between the years of 2009 and 2011, and the " financially healthy" companies were 85. If the company were on the basis of construction, the random effect would already be predicted by the model. This happened just for 15% of companies. For other "new" companies, in which the value of the random intercept did not know it was necessary to use predictive methods to predict the probability of the company in or not in crisis.

In the LRPM method, the variables selected for the regression model of random effects were mean and maximum of Cash on total assets ((E-AT). Furthermore, in the NNPM method, the variables in the method selected to provide greater accuracy were mean, median, minimum and maximum of E-AT and FF-AT.

The measures of performance obtained for the different methods present in Section 4 for the prediction data set are presented in Table 4.

Table 4. Specificity, sensibility, KS and accuracy of the model in the prediction data set.

Measure	Naive	EBP	LRPM	NNPM
Specificity	76.5%	68.2%	72.9%	82.4%
Sensibility	100.0%	80.0%	100.0%	100.0%
KS	76.5%	48.2%	72.9%	82.4%
Accuracy	77.8%	68.9%	74.4%	83.3%

The EBP method presented the worst performance presenting the lowest values of the performance measures; possibly due to the fact random intercept does not follow a normal distribution. The LRPM method even considers a normal distribution of the random intercept, allowing adding other variables, which justifies a better classification compared to EBP. The naive method presented a good classification. Although NNPM enables the inclusion of variables in longitudinal level, it was the best method of classification of companies who will be or not be in crisis the next year achieving sensitivity of 82.4% and specificity of 100%.

6. Conclusions

In this paper, we have shown that it is possible not only to obtain estimates of the probability of a company going or not into crisis, but also to make a prediction for the future considering the mixed model and different methodologies for the prediction of random effects for new companies.

The four prediction methodologies were able to predict satisfactorily; however NNPM presented the best performance. This result is understandable in this application due to the fact that empirical random intercept does not follow a normal distribution, and allows inclusion of a new variable longitudinal manner, which improves the prediction of the values of the random effects of new companies, hence better prediction response.

In previous studies, authors did not consider new companies or companies that were not part of database for adjustment of the model which constitutes an important feature of this article. As future work, we intend to extend and compare this proposal with other traditional models presented in the literature (see, Mossman et al. (1998), Wu et al. (2010), Kumar and Kumar (2012)), Hazard model and Bayesian model (see for example, Trabelsi et al. (2014)).
We are in the process of building a database that includes the companies from the São Paulo Stock Exchange for the application of our proposal.

References

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23, (3), 589–609.

Altman, E., Baida, T and Rivero Diaz, L., 1979. Assesing potential financial problems for firms in Brazil. Journal of International business studies: 9 – 24.

Altman, E., 1993. Corporate Financial Distress and Bankruptcy. New York: John Wiley and Sons.

Caro, N., 2014. Modelos de prediccin de crisis financiera en empresas: una revisin de la literatura. Revista Internacional Legis de Contabilidad y Auditora, 58,135 – 183.

Caro, N., Diaz, M. and Porporato, M., 2013, Prediccin de quiebras empresariales en economas emergentes: uso de un modelo logstico mixto. Revista de Mtodos Cuantitativos para Economa y Empresa, 16, 200 – 215.

Cover, T.M. and Hart, P.E.,1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13 (1), 21 – 27.

Dunn, P. K., Smyth, G. K., 1996. Randomised quantile residuals. J. Comput. Graph. Statist. 5, 236–244.

Huang, X., 2009. Diagnosis of random-effect model misspecifcation in generalized linear mixed models for binary response. Biometrics. 65, 361 – 368.

Jiang, J. and Lahiri, P., 2001. Empirical best prediction for small area inference with binary data. Annals of the Institute of Statistical Mathematics, 53, 2, 217 – 243.

Jones, S. and Hensher, D. A. ,2004. Predicting Firm Financial Distress: A Mixed Logit Model. The Accounting Review, v. 79, 4, 1011 – 1038.

Kumar, R. G., and Kumar, K., 2012. A comparison of bankruptcy models. International Journal of Marketing, Financial Services and Management Research, 1, 4, 76 – 86.

McCulloch, C.E., Neuhaus, J.M., 2011. Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. Statist.Science. 26, 3, 388 – 402.

Mossman, C. E., Bell, G. G., Swartz, L. M. and Turtle, H., 1998. An empirical comparison of bankruptcy models. Financial Review, 33, 35 – 54.

Neuhaus, J.M., Hauck, W.W., Kalbfleisch, J.D., 1992. The Effects of Mixture Distribution Misspecification when Fitting Mixed-Effects Logistic Models. Biometrika, 79(4), 755 – 762.

Neuhaus,J.M. , McCulloch, C.E. and Boylan, R., 2013. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. Statistics in Medicine, 32, 14, 2419 – 2429.

Nigsch F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E., Mitchell, J.B., 2006. Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. Journal of Chemical Information and Modeling, 46, 6, 2412 – 2422.

Rigby, R., Stasinopoulos, D, 2005. Generalized additive models for location, scale and shape. Applied Statist. 54(3), 507 – 554.

f. Sandin, A., and Porporato, M., 2007. Corporate bankruptcy prediction models applied

to emerging economies. Evidence from Argentina in the years 1991 - 1998 International Jorunal of Commerce and Management, 17, 4, 295 – 311.

Tamura, K.A. and Giampaoli, V., 2013. New prediction method for the mixed logistic model applied in a marketing problem. Computational Statistics & Data Analysis, 66, 202 – 216.

Tamura, K.A. and Giampaoli, V., 2010. Prediction in Multilevel Logistic Regression. Communications in Statistics-Simulation and Computation, 39, 6, 1083-1096.

Tamura, K.A., Giampaoli, V. and Noma, A., 2013. Nearest Neighbors Prediction Method for mixed logistic regression. In: 28th International Workshop on Statistical Modeling, Palermo, 799 – 802.

Tamura, K.A., Giampaoli, V. and Noma, A., 2014. The impact of the misspecification of the random effects distribution on the prediction of the mixed logistic model. In: 29th International Workshop on Statistical Modeling, Gottinguen, 2, 153 – 156.

Trabelsi, S., He, R., He, L. and Kusy, M., 2014. A comparison of Bayesian, Hazard, and Mixed Logit model of bankruptcy prediction. Computational Management Science , 1 – 17.

Ohlson, J.S.1980.Financial ratios and theprobabilistic prediction of bankruptcy. Journal of Accounting Research, 19, 109 – 31.

Wu, Y. Gaunt, C.,Gray. S., 2010. A comparison of alternative bankruptcy prediction models. Journal of Contemporary Accounting & Economics 6, 1,34 – 45.