FINANCIAL DISTRESS CRITERIA DEFINED BY MODEL BASED CLUSTERING

Mária Stachová – Lukáš Sobíšek – Michal Gerthofer – Karel Helman

Abstract

One of the important steps in financial distress analyses is to correctly and reasonably mark a company whether is, or it is not in financial distress risk. There are many definitions used in the past. Most of them are based on time static point of view and thus use only one year data. In this paper, we continue with our previous work that examined possibilities of the companies clustering in order to identify homogeneous clusters regarding to their financial distress by using micropanel data. Financial distress can be described as a situation when a company cannot pay or has a difficulty to pay off its financial obligations. In our analysis we consider three criteria to define this situation: the equity, the earnings after taxes and the current ratio value. These financial indicators data were collected over a few consecutive years and thus create a longitudinal data set. We compare a model based partitioning and *k*-means partitioning to cluster the time trajectories of these three criteria. We use packages “mixAK” and “kml” of the statistical system R.

Key words: financial distress, longitudinal data clustering, model based partitioning

JEL Code: C38, G33

Introduction

In the field of companies decision making process various data mining tools and statistical methods play their role. From one point of view it is important (and useful) to be able to correctly and with high accuracy predict the risk of financial distress or bankruptcy by employing proper prediction models. On the other hand, a decision what to consider a financial distress state is also a non-negligible question.

As it is mentioned in one of our previous papers (Stachová & Sobíšek, 2016), in general, financial distress of a company is understood as its inability to pay off its liabilities or a difficulty to meet its financial obligations. Authors in (Baixauli & Modica-Milo, 2010) use four financial indicators of profitability and an expert opinion to define financial distress. Li & Liu (2009) state that company is in financial distress if its economic results in two consecutive years are negative. In the paper (Stachova & al., 2015) the financial distress is defined as a situation when a company went bankrupt or has some ongoing liquidation or has some overdue obligations (on an obligation).

Our approach is based on that one mentioned in (Boďa & Úradníček, 2016): "An enterprise was considered financially distressed if

a) its Equity was negative,
b) its EAT (Earnings After Taxes) were negative,
c) its Current ration attained a value lower than 1.

All three conditions have to be satisfied for the purpose of an enterprise to be considered financially distressed." The reasons of this decision are described in the paper (Stachová & Sobíšek, 2016).

The issue of estimating (constructing) a "good" model for classifying and predicting financial distress of companies has become a subject of many studies. The well-known Altman's Z-score (Altman, 1968) has to be mentioned at the beginning and its revision (Altman, 1983) as well. This approach continues to be popular till nowadays. Many of following studies and approaches are based on static classification models constructed by employment of various statistical methods as discriminant analysis, logistic regression, decision trees (Boďa & Úradníček, 2016; Brezigar-Masten, 2012). We assume that incorporating a time dynamic into these static models has the potential to improve their predictive accuracy. This idea is supported by studies presented in (Kráľ & al., 2014; Stachová & al. 2015).

We consider the idea of finding proper way how to recognize a company as a one, which is being financially distressed, by using the information about the negative dynamics of the financial indicators to their static cut-off values is an important initial step in financial distress prediction. Thus, the goal of this work is to use the model based clustering (Komárek & Komárekova, 2014) to verify whether this algorithm is able to identify homogeneous clusters with respect to the companies' financial distress by using the financial indicators collected over four consecutive years. This algorithm will be subsequently compared to the method used in paper (Stachová & Sobíšek, 2016) that is based on K-means clustering.

1 Data and Methodology

Our dataset is the same one used in paper (Stachová & Sobíšek, 2016). It consists of 3 numeric financial distress indicators for 2,900 companies. These companies represent the sector of Manufacturing, Construction and Wholesale and retail trade, repairs of motor vehicles and motorcycles, in accordance with SK NACE classification. The dataset was purchased from

Slovak corporate analytical agency CRIF – Slovak Credit Bureau, s.r.o. (http://www.crif.sk) and covers the time period from 2010 to 2013. The selected companies belong to the riskier field of economic activities according to the number of bankruptcy declarations. Descriptive statistics of inter-yearly percentage change of indicators can be found in Table 1.

Tab. 1: Descriptive Statistics of Mean Annualized Percentage Change of Indicators

Financial Indicator	Mean	Standard deviation	Median
Equity	1.24	34.5	0.75
EAT	-17.86	35.55	-24.5
Current ratio value	1.47	22.45	0

Source: the authors.

Our analysis is based on an idea that (for an individual company) changes in the values of each of these three criteria can signalize changes in financial health of the monitored enterprise. The stronger the change, the stronger the signal.

To achieve the aim of our work, i.e. to find the proper algorithm that is able to identify homogeneous clusters regarding the companies' financial distress by using the financial longitudinal data collected over four consecutive years, we use the Multivariate mixture generalized mixed model (MMGLMM) based clustering. It is algorithm included in package "mixAK" of statistical system R (R Core Team, 2013; Komárek 2009).

1.1 MMGLMM based clustering

Initially the model based clustering will be introduced and applied on described multivariate generalized linear mixed model (MGLMM) theory resulting into MMGLMM. Further, due to the calculation complexity of the model the Bayesian inference (especially MCMC) will be used for the parameters estimation and clustering procedure.

Model based clustering

Before the description of the clustering procedure several assumptions must be stated. We assume that number of clusters is known and equals to K. Further, we introduce the unobservable component allocations $U_1, \dots, U_N \epsilon \{1, \dots, K\}$,

$$P(U_i = k; \mathbf{w}) = w_k, \; i = 1,\dots,N, \; k = 1,\dots,K \tag{1}$$

where $\boldsymbol{w} = (w_1, \dots, w_K)^{\boldsymbol{T}}$ is a vector of unknown probabilities. Additionally, $U_i = k$ symbolizes the fact that i-th subject $\boldsymbol{Y}_i$ was generated by the k-th model conditional density $f_{i,k}(\boldsymbol{y}_i; \boldsymbol{\xi}, \boldsymbol{\xi}_k)$, where $\boldsymbol{\xi}$ is a vector of common parameters and $\boldsymbol{\xi}_k$ is a vector of cluster specific unknown parameters. Further, marginal density of $\boldsymbol{Y}_i$ is defined as the mixture density as follows

$$f_i(\boldsymbol{y}_i;\boldsymbol{\theta}) = \sum_{i=1}^{K} w_k\, f_{i,k}(\boldsymbol{y}_i;\xi,\xi_k), \tag{2}$$

where overall parameter $\boldsymbol{\theta} = (\boldsymbol{w}^T, \boldsymbol{\xi}_1^T, \dots, \boldsymbol{\xi}_K^T, \boldsymbol{\xi}^T)^{\boldsymbol{T}}$ represents a vector of all unknown model parameters. Finally, clustering procedure is based on estimated value $\hat{p}_{i,k}$ of individual component probabilities

$$p_{i,k} = p_{i,k}(\boldsymbol{\theta}) = P(U_i = k \,|\, \boldsymbol{Y}_i = y_i;\boldsymbol{\theta}) = \frac{w_k\, f_{i,k}(\boldsymbol{y}_i;\xi,\xi_k)}{f_i(\boldsymbol{y}_i;\boldsymbol{\theta})}, \tag{3}$$

and the classification of a subject i into a cluster $g(i)$ is performed using the criterion $\hat{p}_{i,g(i)} = max_{k=1,\dots,K}\hat{p}_{i,k}$. There are also other options how to set the criterion.

Multivariate mixture generalized linear mixed model

The first step is to derive $f_i(\boldsymbol{y}_i;\boldsymbol{\theta})$ for aforementioned random vector $\boldsymbol{Y}_i = (Y_{i,1,1}, \dots, Y_{i,R,n_i})^T$ within MMGLMM. At the beginning, we start with MGLMM. Under well-known assumptions of MGLMM, we have cluster specific density $f_{i,k}$ given as follows

$$f_{i,k}(\boldsymbol{y}_i;\boldsymbol{\xi},\boldsymbol{\xi}_k) = \int_{\mathbb{R}^q} \left\{ \prod_{r=1}^{R} \prod_{j=1}^{n_i} f_{D_r}\big(y_{i,r,j};\boldsymbol{\alpha}_r,\phi_r,\boldsymbol{b}_{i,r}\big) \right\} \varphi(\boldsymbol{b}_i;\boldsymbol{\mu}_k,\mathbb{D}_k)d\boldsymbol{b}_i, \tag{4}$$

where $= 1, \dots, N,\ \ r = 1, \dots, R,\ j = 1, \dots, n_i$. Distribution f_{D_r} symbolizes exponential family distribution with dispersion parameter ϕ_r, mean given by $h_r^{-1}\{E(Y_{i,r,j}|\boldsymbol{B}_{i,r} = \boldsymbol{b}_{i,r};\boldsymbol{\alpha}_r)\} = \boldsymbol{x}^T{}_{i,r,j}\boldsymbol{\alpha}_r + \boldsymbol{z}^T{}_{i,r,j}\boldsymbol{b}_{i,r}$, where h_r^{-1} is link function, $\boldsymbol{\alpha}_r \epsilon \mathbb{R}^{p_r}$ is a vector of unknown parameters (fixed effects), $\boldsymbol{b}_{i,r} \epsilon \mathbb{R}^{q_r}$ is a vector of random effects and $\boldsymbol{x}^T{}_{i,r,j} \epsilon \mathbb{R}^{p_r}$, $\boldsymbol{z}^T{}_{i,r,j} \epsilon \mathbb{R}^{q_r}$ are vectors of known covariates. Further, parameters $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K$ correspond to the means and covariance matrices $(\boldsymbol{\mu}_k, \mathbb{D}_k)$ of the conditional distributions of random effects and $\boldsymbol{\xi}$ corresponds to fixed and dispersion parameters common for all clusters $(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_R, \phi_1, \dots, \phi_R)$. Last but not least, $\varphi(.;.)$ represents density of multivariate normal distribution.

It is worth to note that vector $\boldsymbol{B}_i = (\boldsymbol{B}_{i,1}^T, \dots, \boldsymbol{B}_{i,R}^T)^T \epsilon \mathbb{R}^q, q = \sum_{r=1}^{R} q_r$, given $U_i = k$ follows a multivariate normal distribution with unknown mean $\boldsymbol{\mu}_k \epsilon \mathbb{R}^q$ and unknown $q \times q$ positive definite covariance matrix $\mathbb{D}_k, k = 1, \dots, K$, i.e., $\boldsymbol{B}_i |\ U_i = k \sim \aleph_q(\boldsymbol{\mu}_k, \mathbb{D}_k)$.

Now, using aforementioned formula for marginal density and cluster specific density $f_{i,k}$ from MGLMM, we obtain a likelihood contribution for i-th subject as follows

$$f_i(\boldsymbol{y}_i;\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \left\{ \prod_{r=1}^{R} \prod_{j=1}^{n_i} f_{D_r}\big(y_{i,r,j};\boldsymbol{\alpha}_r,\phi_r,\boldsymbol{b}_{i,r}\big) \right\} \left\{ \sum_{i=1}^{K} w_k \varphi(\boldsymbol{b}_i;\boldsymbol{\mu}_k,\mathbb{D}_k) \right\} d\boldsymbol{b}_i. \tag{5}$$

It is worth to note the dependence among the random vectors $\boldsymbol{Y}_{i,1}, \dots, \boldsymbol{Y}_{i,R}$ representing different markers ($r = \{1, \dots, R\}$) is induced by non-diagonal covariance matrix $\mathbb{D}_k$ of the random effects vector $\boldsymbol{B}_i$ in general.

The last equation indicates that now this model can be interpreted either as a mixture of multivariate generalized linear mixed models (MMGLMM) with normally distributed random effects, or as a multivariate generalized linear mixed model with normal mixtures in the random effects distribution, where the overall mean of the random effects $\boldsymbol{B}_i$ is given by $\boldsymbol{\beta} = E(\boldsymbol{B}_i;\boldsymbol{\theta}) = \sum_{i=1}^{K} w_k \boldsymbol{\mu}_k$.

Clustering procedure

The following step is to estimate the component probabilities $p_{i,k}$. However, we can see that they are functions of unknown vector parameter $\boldsymbol{\theta}$. Therefore, we firstly concentrate on estimation of parameter $\boldsymbol{\theta}$. Nevertheless, we can see that using MLE approach it is not feasible to estimate parameter $\boldsymbol{\theta}$ due to the complexity of the likelihood $L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f_i(\boldsymbol{y}_i;\boldsymbol{\theta})$.

Therefore, the Bayesian approach based on the output from the Markov chain Monte Carlo (MCMC) simulation may be considered as the appropriate way to estimate the unknown parameter vector $\boldsymbol{\theta}$ and consequently the $p_{i,k}$.

We skip details about MCMC simulations which can be found in Stephens (2000) and move to the usage of output from the MCMC algorithm. As the result of the MCMC simulation we obtain a sample $S_M = \left\{ \left(\boldsymbol{\theta}^{(m)}, \boldsymbol{b}_1^{(m)}, \dots, \boldsymbol{b}_N^{(m)}, u_1^{(m)}, \dots, u_N^{(m)} \right) : m = 1, \dots, M \right\}$ from posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{b}_1, \dots, \boldsymbol{b}_N, u_1, \dots, u_N | \boldsymbol{y})$. This sample is later used within the formula for the estimation of component probabilities $p_{i,k}$.

It is worth to note once more that in Bayesian statistics, the latent quantities, random effects $\boldsymbol{B}_i$ and component allocation U_i, are considered as additional model parameters with the joint prior distribution for all the model parameters.

The last step of the model based clustering is estimation of the individual component probabilities $p_{i,k}$ using sample S_M from MCMC simulation. Within the Bayesian framework, the natural estimates of the components probabilities $p_{i,k}$ are their posterior means. Thus, MCMC estimates are easily obtainable from the generated posterior sample S_M, i.e.,

$$\hat{p}_{i,k} = E\{p_{i,K}(\boldsymbol{\theta})|\boldsymbol{Y} = \boldsymbol{y}\} = P(U_i = k|\boldsymbol{Y} = \boldsymbol{y}) = \int p_{i,k}(\boldsymbol{\theta})p_{i,k}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \approx \frac{1}{M}\sum_{m=1}^{M} p_{i,k}(\boldsymbol{\theta}^{(m)}). \quad (6)$$

Consequently, classifying of each subject is performed according to aforementioned criterion $\hat{p}_{i,g(i)} = max_{k=1,\dots,K}\hat{p}_{i,k}$. Moreover, by using this approach, uncertainty in the classification can be measured by either the full posterior distribution of the component probabilities, or by calculating their credible intervals.

2 Results

If the static expert-based definition of financial distress as in (Boďa & Úradníček, 2016) is applied to our 2013 data, we would find 901 (31 %) companies to be in the financial distress.

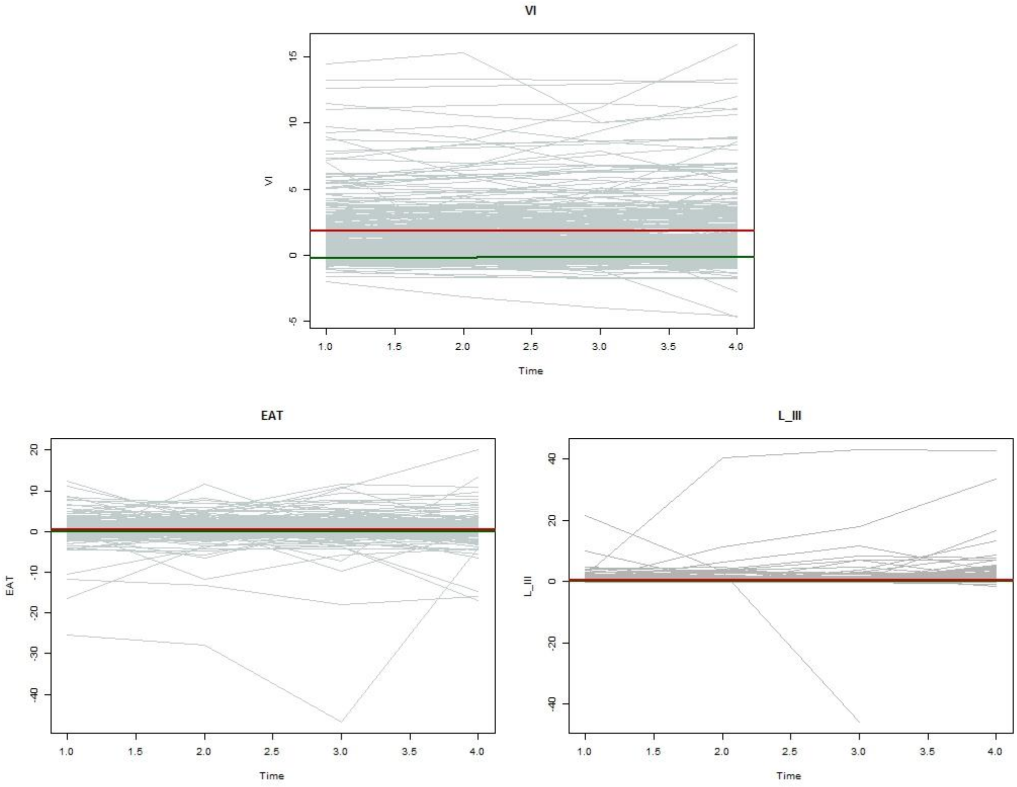
On the contrary with "kml" clustering used in paper (Stachová & Sobíšek, 2016), we do not apply MMGLMM model based clustering on financial indicators separately. This algorithm works on model basis and incorporates all three indicators together (at the same time). Our goal is to classify (recognize) companies of two groups (financially distressed vs. financially healthy) and thus we set number of desired clusters to value 2. In the Figure 1 there are the observed data together with results of the clustering-algorithm, i.e. the estimated cluster specific mean longitudinal profiles. The profile of the first cluster is drawn with green color and contains 2582 (89 %) companies and the second one is red and contains 318 (11%) companies.

It can be seen that probabilities of both clusters are constant over time and the "red" cluster contains slightly higher values. These two clusters are unbalanced as the first one (the green one), with lower values of all three financial indicators, contains almost all companies. Even those that from expert point of view were labeled to be financially healthy. Moreover, twelve "financially distressed" (from expert point of view) companies were classified into the second cluster that contains companies with higher values of financial indicators. It is not what we expected. We expected that all the "financially distressed" companies would belong to the cluster with lower financial indicators' values.

The Figures 2, 3 and 4 display the distribution of annualized percentage change of Equity (VI), EAT and Current ratio (L_III) respectively, calculated from the values measured in years 2010 and 2013. It is obvious, that MMGLMM algorithm should create more balanced clusters with second clusters with higher number of members.

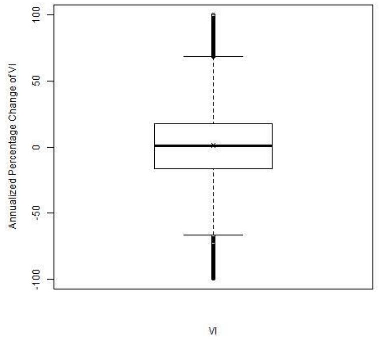
The obtained results are similar as those obtained with "kml" algorithm in (Stachová & Sobíšek, 2016).

Fig. 1: Estimated cluster specific mean longitudinal profiles



Source: the authors.

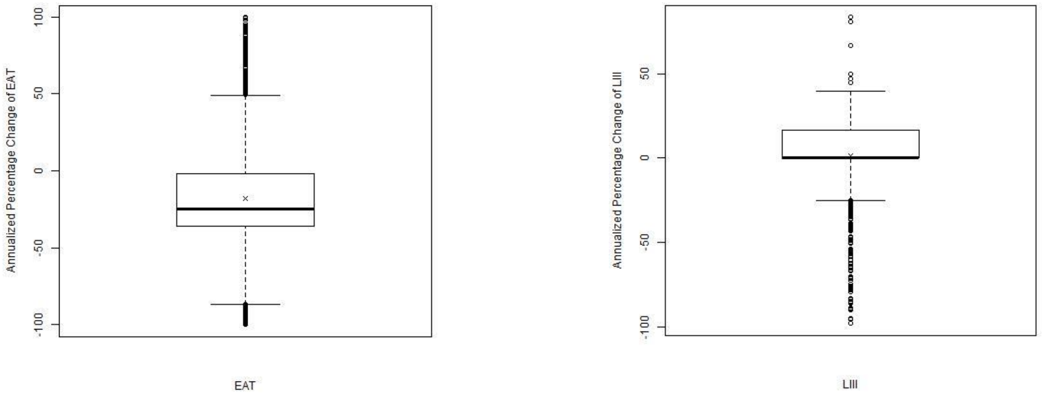
Fig. 2: The box plot of Equity values



Source: the authors.

Fig. 3: The box plot of EAT values

Fig. 4: The box plot of Current ratio values



Source: the authors.

Conclusion

The main goal of this contribution is to investigate whether multivariate MMGLMM model based cluster algorithm included in R package “mixAK” enables identification of the cluster of companies with negative temporal trend better than K-means partitioning (included in R package “kml”). We expected to get better results by employing MMGLMM model based clustering (for our data set) from two reasons: firstly it allows multivariate partitioning of longitudinal data (not possible in “kml” package). The second reason is that the model-based approach enables the estimation of random effects (slope) of time and use these parameter estimates to enhance the partitioning ability. In our analysis we estimated not only random slopes but also random intercepts.

MMGLMM model based clustering identified clusters according to the absolute value of indicators and does not take into the account trends of individual indices, i.e. a cluster with worse (lower) values of indicators containing 89 % of companies. The mean profile trajectories in this cluster are constant (Fig. 1), which is in contrary to our expectations. The boxplots (Fig. 2 - 4) display distribution of individual annualized percentage changes. In the presented charts and in Table 1 it is visible that almost 50 % of companies has the negative annualized change of equity values and current ratio values. The proportion of negative development is even higher for EAT values. This evidence supports the idea that the partitioning algorithm identifies one cluster that gather negative trajectories with mean negative trend in time. Unfortunately, neither algorithm included in “kml: package (Stachová & Sobíšek, 2016) nor algorithm from “mixAK” package was able to identify such a cluster.

Additionally, we also excluded the cluster with higher positive values (11%) and run again the model-based clustering on 89% of companies included in the large cluster in hope to identify the cluster containing “unhealthy” companies with negative trends. Moreover, we tried to partition companies in 3 clusters. Even these additional steps did not identify such a cluster. This fact suggests that in our future work we will try to find a more appropriate clustering algorithm that would recognize better existing patterns of the development of repeated measures in time. We suggest such an algorithm could be useful not only in the field of financial distress prediction but also in other scientific areas where data are collected over time, e. g.: insurance, pension schemes or medicine.

Acknowledgment

Mária Stachová has been supported by the PROJECT VEGA NO. 1/0093/17 Identification of risk factors and their impact on products of the insurance and saving schemes. Lukáš Sobíšek has been supported by the project of the University of Economics, Prague - Internal Grant Agency, project No. 44/2017 “Clustering and regression analysis of micro panel data”.

References

Altman, E. I., (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 583-609.

Altman, E. I. (1983). *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy.* New York: WIley.

Boďa, M., & Úradníček, V. (2016). The portability of Altman’s Z-score model to predicting corporate financial distress of Slovak companies. In Technological and Economic Development of Economy, 22(4), 532-553.

Baixauli, J. S., & Modica-Milo, A. (2010). The bias of unhealthy SMEs in bankruptcy prediction models. Journal of Small Business and Enterprise Development, 17(1), 60-77.

Brezigar - Masten, A., & Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. Expert Systems with Applications, 39(11), 10153–10159.

Komárek, A. (2009). A New R package for Bayesian Estimation of Multivariate Normal Mixtures Allowing for Selection of Number of Components and Interval-Censored Data. Computational Statistics & Data Analysis, 53(12), 3932-3947.

Komárek, A. & Komárková, L. (2014). Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. Journal of Statistical Software, 59(12), 1–38

Kráľ, P., Stachová, M. & Sobíšek, L. (2014). Utilization of repeatedly measured financial ratios in corporate financial distress prediction in Slovakia*:* In the 17th AMSE, international scientific conference, conference proceedings, Poland, 156-163.

Li, D., & Liu, J. (2009). Determinants of Financial Distress of ST and PT Companies: A Panel Analysis of Chinese Listed Companies. Retrieved February 12, 2009, from http://ssrn.com/abstract=1341795

R Core Team 2013: R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2012, http://www.r-project.org/.

Stachová, M., Kráľ, P., Sobíšek, L., & Kakaščík, M. (2015). Analysis of financial distress of Slovak companies using repeated measurement. In 18th AMSE, International Scientific Conference Proceedings, Czech Republic.

Stachová, M., & Sobíšek, L. (2016). Financial distress criteria defined by clustering of longitudinal data. In Conference proceedings : the 10th international days of statistics and economics, September 8–10, 2016, Prague, 1703-1712.

Stephens, M. (2000). Dealing with Label Switching in Mixture Models. Journal of the Royal Statistical Society B, 62(4), 795.

Contact

Mária Stachová
Faculty of Economics, Matej Bel University,
Tajovského 10
975 90 Banska Bystrica, Slovakia
maria.stachova@umb.sk

Lukáš Sobíšek, Michal Gerthofer, Karel Helman
Faculty of Informatics and Statistics, University of Economics Prague
W. Churchill Sq. 1938/4
130 67 Prague 3 – Žižkov, Czech Republic
lukas.sobisek@vse.cz, michal.gerthofer@gmail.com, helmank@vse.cz