CAN LONGITUDINAL CLUSTERING HELP TO DEFINE FINANCIAL DISTRESS CRITERIA?

MÁRIA STACHOVÁ
Matej Bel University in Banská Bystrica, Faculty of Economics,
Department of Quantitative Methods and Information Systems,
Tajovského 10, Banská Bystrica, Slovakia
e-mail: maria.stachova@umb.sk

LUKÁŠ SOBÍŠEK
University of Economics in Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill Sq. 4, Prague, Czech Republic
e-mail: lukas.sobisek@vse.cz

Abstract

One of the main task in each analysis of companies' financial status is to correctly define the criteria that can describe the financial health or financial distress of these enterprises. In general, the financial distress is a situation in which company cannot pay or has difficulty to reach its financial obligations. Our data set consists of three financial indicators of Czech enterprises. These longitudinal data are collected over a few consecutive years. We applied the model-based partitioning and the K-means partitioning to these longitudinal data to cluster the time trajectories of these criteria and subsequently we compare the accuracy of these algorithms. We use packages "mixAK" and "kml" of the statistical system R in our analysis.

***Keywords**: financial distress, longitudinal data clustering, K-means partitioning, model-based partitioning*

***JEL Codes**: C38, G33*

1. Introduction

Not only, but especially financial lenders (banks, investment companies, etc.) employ the prediction statistical models (in their risk management processes) to be able to correctly and with high accuracy predict the risk of debtors' financial distress or bankruptcy. The issue of estimating 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 and Úradníček, 2016; Balcaen and Ooghe, 2006; Brezigar-Masten, 2012).

We assume that the power of static financial distress predictive models might be enhanced by adding information about negative dynamics of the financial indicators to their static cut-off values. This idea is supported by the studies presented in (Kráľ *et al*., 2014; Stachová *et al*. 2015). We follow previous studies about Slovak companies (Stachová and Sobíšek, 2016; Stachová *et al*. 2017), where the aim of those two papers was to investigate whether it is possible to identify homogeneous clusters regarding the companies' financial distress by using the financial longitudinal data collected over four consecutive years.

In order to meet this goal, we used the *K*-means partitioning modified for the longitudinal data (Genolini *et al.*, 2015) and the model-based clustering (Komárek and Komárkova, 2014) to verify whether these algorithms are able to identify homogeneous clusters with respect to the companies' financial distress by using the financial indicators collected over four consecutive years. These two methods have been chosen because they represent different approaches to clustering that we want to compare. The results showed that there is an evidence of companies that should be recorded as being at the risk of financial distress according to their decreasing values of selected financial criteria. These companies would not be found using the static expert method that takes into account only positivity/negativity of one-year indicator values. However, we also found, that the both statistical methods of clustering of trajectories are very rough and they are not able to create more evenly distributed clusters in size.

Not many favorable results (described in the previous paragraph) make a question, whether the applied methods are not proper for the particular Slovak data set or they are not usable for this type of financial data in general. The main goal of the current contribution is to run the identical clustering analysis as in (Stachová and Sobíšek, 2016; Stachová *et al.* 2017) on the Czech real data companies. The reason to do this re-analysis on a different real, independent data set is to evaluate (the robustness and) the reproducibility of our previous results.

The paper is organized as follows. In section 2 we shortly present the data set and the methodology; Section 3 consists of the results achieved in our analysis and in Section 4 we conclude and discuss our achievements.

2. Data and Methodology

We use data set that consists of 3 numeric financial distress indicators – Return on Assets (ROA), Current Ratio and Return on Equity (ROE) from 4338 companies. These data set was extracted from the Czech data repository Albertina, covering processing industry area denoted according to NACE classification as CZ-NACE C category and includes years from 2012 to 2015.

Our hypothesis is that negative changes in the values of the indicators mentioned above over time can indicate the negative changes in financial health of enterprise and thus we should recognize these changes and monitor the susceptible company even more if these changes decrease rapidly. ROA is a ratio computed as a net income divided average total assets, it measures how efficiently an enterprise can manage assets to produce profits during a time period. Current ratio is a comparison of current assets to current liabilities. ROE indicator is ratio of the net income and the average equity. We suspect to find companies that are not labeled as financially distressed from a static expert's point of view (negative ROA and negative ROE and Current ratio is less than 1), but that could be labeled as companies at risk according to their past trends in values of ROA, current ratio and ROE. As an expert's point of view we took a work of (Boďa and Úradníček, 2016).

In order 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 status, we use two methods. First one is the multivariate mixture generalized mixed model (MMGLMM) based clustering. It is an algorithm included in the package "mixAK" (Komárek, 2009 of statistical system R (R Core Team, 2013). The second approach is the *K*-means clustering for partitioning the time trajectories of selected outcomes which is also implemented in the R, in the package "kml" (Genolini *et al.*, 2015).

2.1 MMGLMM based clustering

In this subsection we take over description of MMGLMM algorithm as it was described in (Stachová *et al.* 2017).

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 the Markov chain Monte Carlo) will be used for the parameters estimation and clustering procedure.

2.1.1 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 $U_1, \dots, U_N \epsilon \{1, \dots, K\}$,

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

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.

2.1.2 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

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

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

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

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 $\boldsymbol{\beta} = E(\boldsymbol{B}_i; \boldsymbol{\theta}) = \sum_{i=1}^K w_k \boldsymbol{\mu}_k$.

2.1.3 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.,

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

Consequently, classifying of each subject is performed according to aforementioned criterion $\hat{p}_{i,g(i)} = \max_{k=1,\dots,K} \hat{p}_{i,k}$. $\max_{k=1,\dots,K}$ 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.

2.2 K-means clustering for longitudinal data

In this subsection can be found short introduction to *K*-means algorithm for longitudinal data as it is described in (Stachová and Sobíšek, 2016).

The *K*-means clustering applied in the “kml” package is modified for the longitudinal data. This algorithm is based on the original *K*-means clustering (MacQueen, 1967). This method minimizes the utility function iteratively for the time *t*, *N* objects according to an assumption of *C* clusters. The utility function can be expressed as follows:

$$\min \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ict} d_{ict}^2, \quad (7)$$

where u_{ict} is a degree of appropriateness of the *i*-th object into the *K*-th cluster in the time *t* with conditions:

$$\sum_{c=1}^{C} u_{ict} = 1, \forall i, t, \quad (8)$$

$$\forall u_{ict} : u_{ict} = \begin{cases} 1 & \|\mathbf{x}_{it} - \mathbf{h}_{ct}\| = \arg\min_i \|\mathbf{x}_{it} - \mathbf{h}_{ct}\| \\ 0 & \text{elsewhere} \end{cases}.$$

We used the Euclidean distance $d_{ict} = \|x_{it} - h_{ct}\|$ between *i*-th vector of objects $x_{it} = (x_{i1t}, \dots, x_{ijt}, \dots, x_{iJt})'$ and *K*-th centroid $h_{ct} = (h_{c1t}, \dots, h_{cjt}, \dots, h_{cJt})'$ in the time *t*. We applied the algorithm to the standardized values of variables.

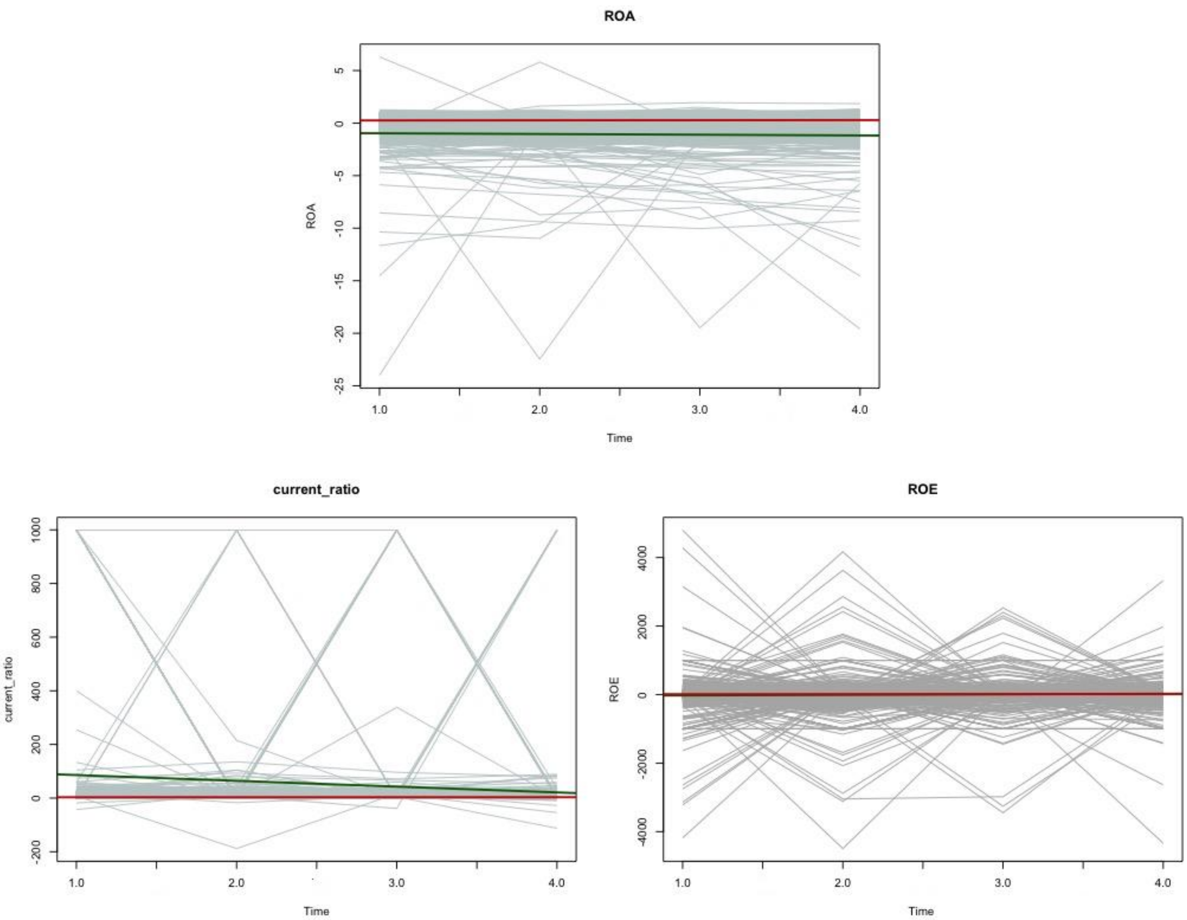
3. Results

If the static expert-based definition of financial distress is applied to our data from the last year, we should find 811 (18.7%) companies financially distressed. In the next subsection we are describing the results obtained by two mentioned clustering methods for longitudinal data.

3.1 Results of MMGLMM clustering

We applied the MMLGLMM clustering on all three financial indicators together and our goal was to recognize companies of two characters- at the risk of financial distress and the financial health and therefore we set the number of desired clusters to number 2. The variable Time was treated as a fixed and a random effect. The Lilliefors normality test was used to assess whether the variable is normally distributed within each timepoint. The normality assumption was met. The number of simulation in MCMC algorithm was set only to 100 because of the computation difficulties. In the Figure 1 we can find the observed data with the results of clustering algorithm, i.e. estimated cluster specific mean longitudinal profiles. The mean longitudinal profile plot depicts the two lines (curves) connecting average values of an indicator at each time, i.e. each line represents by means the companies within the cluster. The profile of the first cluster is drawn in green color and contains 224 (5.2%) companies and the second one is red colored with 4114 (94.8%) companies.

Figure 1: Estimated cluster specific mean longitudinal profiles. Numbers 1 – 4 stand for years 2012 – 2015 respectively.



Source: The author’s work.

It can be seen that in case of the ROA indicator the “red” cluster contains slightly higher values. This fact could indicate that this cluster contains the companies with lower risk of financial distress. However, this could be misleading. As we can see on profile plot drawn for current ratio indicator, it is the “green” cluster that contains the higher values. Still, there is a space for wide discussion because the “green” cluster at the same time contains the values that decrease over time and this could indicate that it contains “susceptible” companies (at the risk of financial distress). In the case of profile plot drawn for the ROE indicator, we can see, that means of two clusters are overlapping and thus we cannot say whether “green” or “red” cluster is the cluster with companies at the risk of financial distress. Finally, after deeper analysis we found, that 123 (only 15.2 %) companies labeled as financially distressed from the expert point view were included into the “green” cluster. Confusion matrix for MMGLMM clustering can be found in Table 1.

Table 1: Confusion matrix for MMGLMM clustering

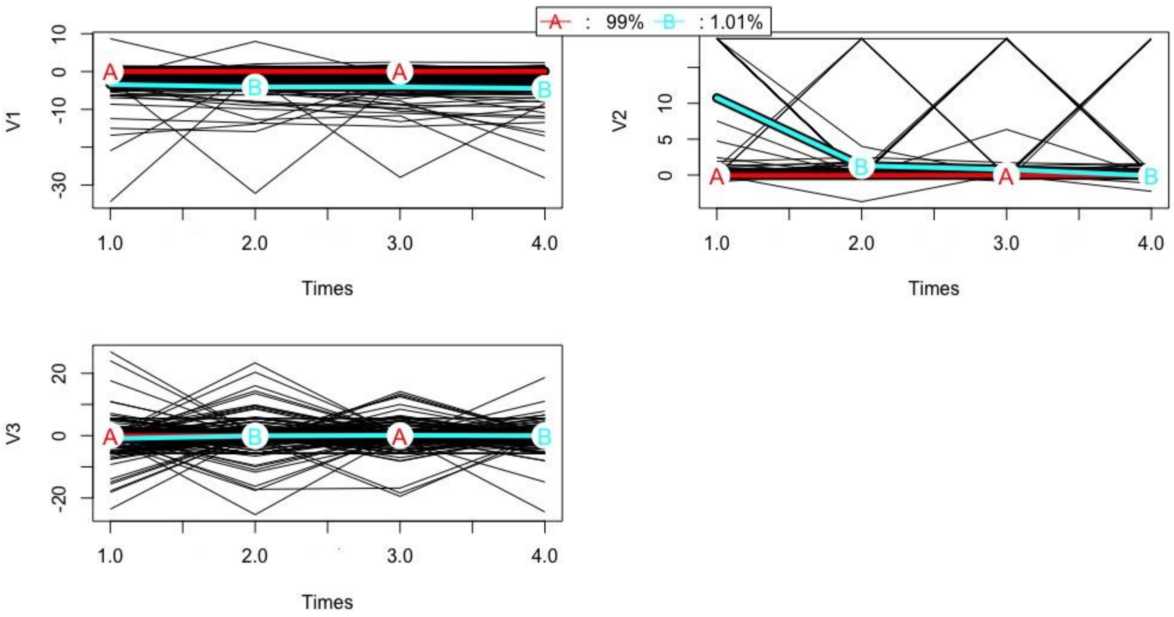
	Cluster “green”	Cluster “red”
Financially healthy	3426	101
Financially distressed	688	123

Source: The author’s work.

3.2 Results of *K*-means clustering for longitudinal data

The *K*-means partitioning algorithm for clustering longitudinal data has been applied on all three financial indicators together. The means of trajectories of created clusters and the percentage of values in each cluster are shown in the Figure 2.

Figure 2: Estimated cluster specific mean longitudinal profiles (V1 stands for ROA, V2 stands for current ratio and V3 stands for ROE). Numbers 1 – 4 stand for years 2012 – 2015 respectively.



Source: The author’s work.

According to the Figure 2 it is obvious that in case of ROA indicator the “A” cluster contains slightly higher values than in MMGLMM algorithm and also, we can see on profile plot drawn for current ratio indicator that this time the “B” cluster contains higher values. For the ROE indicator, we can see, that means of two clusters are overlapping and thus we cannot say whether the “B” or the “A” cluster is the cluster with companies at the risk of financial distress. From this perspective, we could say, that results of these two algorithms are similar. But, if we compare the abundances, the “A” cluster contains 99% of the companies and the “B” cluster only 1%. At the same time, almost half of the companies included in the “B” cluster are labeled as financially healthy from an expert point of view. Thus, we cannot say whether the cluster "A" or cluster "B" is the cluster with financially unhealthy companies. Confusion matrix for *K*-means clustering can be find in Table 2.

Table 2: Confusion matrix for *K*-means clustering

	Cluster "A"	Cluster "B"
Financially healthy	3507	20
Financially distressed	787	24

Source: The author's work.

To evaluate the performance of clustering algorithms we compute the average silhouette index. The silhouette index (Rousseeuw, 1987) assesses the performance of clustering based on the properties (within-cluster homogeneity and between-cluster heterogeneity) of the estimated clusters. The formula is

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathrm{b}_i - a_i)}{\max\{a_i; b_i\}} \quad , \tag{9}$$

where n is the number of objects, a_i is an average distance of i-th object from any other objects in the same cluster and b_i is the minimum from average distances between object i and objects in another clusters. The values vary in the interval $\langle -1;1 \rangle$, the higher value, the better result. The MMGLMM algorithm silhouette index is 0.76 and silhouette index for *K*-means clustering was 0.02.

4. Conclusion

The main aim of our contribution is to find the proper algorithm that is able to identify the homogeneous clusters regarding the companies' financial status. For this purpose, we use two clustering algorithms for longitudinal data. First one is included in R package "mixAK" and it is model based clustering, while the second one is *K*-means based clustering included in R package "kml". We have selected two approaches to see whether the results obtained from one algorithm differ from results obtained by another algorithm as it is prevalence. The results can be different even in the different run of one algorithm (for example, caused by random choice of initial point), so the different results obtained by different algorithms based on totally different idea tent to be more than expected. It does not mean, that we suspect that method create the different mean longitudinal profiles, but they can evaluate these profiles in a different way.

The results didn't meet our expectations. The mean profiles trajectories show inconsistent results even the mean longitudinal profiles are very similar. ROE and ROA are constant over time in both algorithms. Moreover, in case of ROA indicators, the means are overlapping and thus is very hard to talk about proper partitioning. On the other hand, the cluster with higher values of ROA indicator, contains decreasing values of current ratio indicators in case of both algorithm and it is again in contrary to our expectation. Finally, the *K*-means algorithm identified only one cluster and achieved the average silhouette index at level 0.02, thus we can say that this algorithm is inapplicable to our data. The MMGLMM algorithm achieved higher average silhouette index (0.76), but still its interpretability is more than debatable.

These results are similar to the results obtained on Slovak companies (Stachová and Sobíšek, 2016; Stachová *et al.* 2017) where the same two algorithms and the same idea of financial distress has been investigated.

This fact leads us to investigate and estimate more appropriate clustering algorithm that would recognize better pattern of the development of repeated measures in time. Such algorithm could be useful not only in the financial distress area but also in other scientific fields where data can be collected over time such as: insurance, pension schemes or medicine.

Acknowledgements

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".

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".

References

[1] Altman, E. I., (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, vol. 23, iss. 4, pp. 583 – 609.

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

[3] Balcaen, S., and 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, vol. 38, iss. 1, pp. 63 – 93.

[4] Boďa, M., and 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, vol. 22, iss. 4, pp. 532 – 553.

[5] Brezigar - Masten, A., and Masten, I. (2012). CART-based selection of bankruptcy predictors for the logit model. Expert Systems with Applications, vol. 39, iss. 11, pp. 10153 – 10159.

[6] Genolini, Ch., Alacoque, X., Sentenac, M, and Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. Journal of Statistical Software, vol. 65, iss. 4, pp. 1 – 34. URL http://www.jstatsoft.org/v65/i04/.

[7] 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 Statistisc and Data Analysis, vol. 53, iss. 12, pp. 3932 – 3947.

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

[9] Kráľ, P., Stachová, M. and 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, pp. 156 – 163.

[10] MacQueen, J., B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 281 – 297.

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

[12] ROUSSEEUW, P.J. (1987). “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis“, Computational and Applied Mathematics, vol. 20, p. 53 – 65. doi:10.1016/0377-0427(87)90125-7

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

[14] Stachová, M., Sobíšek, L. (2016). Financial distress criteria defined by clustering of longitudinal data. In: International Days of Statistics and Economics (MSED) [online]. Praha, 08.09.2016 – 10.09.2016, pp. 1703–1712. URL https://msed.vse.cz/msed_2016/article/248-Stachova-Maria-paper.pdf.

[15] Stachová, M., Sobíšek, L., Gerthofer, M., Helman, K. (2017). Financial distress criteria defined by model based clustering. In: International Days of Statistics and Economics (MSED) [online]. Praha, 14.09.2017 – 16.09.2017, pp. 1511–1520. URL https://msed.vse.cz/msed_2017/article/28-Stachova-Maria-paper.pdf.

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