PREDICTION OF FINANCIAL HEALTH USING FACTOR ANALYSIS AND DATA ENVELOPMENT ANALYSIS

Emil Exenberger – Jozef Bucko

Abstract

To effectively address economic threats, businesses need to identify these threats soon enough. One of the biggest economic threats that can threaten the very existence of a company is the threat of bankruptcy. This study aims to design a process model for the analysis of a company's financial health to predict the possible danger of bankruptcy in time. The proposed process model uses factor analysis and the DEA model, and we tested the success of bankruptcy prediction on companies in the IT sector in the Slovak Republic for the years 2013 to 2017. All the research was done in the R program, which is the free language and environment for statistical computing and graphics. We calculated the index of correct classification and the index of warning reliability for each year examined. We compared the results of the research with similar studies and, given its high success rate, we recommend it for predicting the risk of bankruptcy in companies.

Key words: factor analysis, data envelopment analysis, financial health, prediction

JEL Code: C51, C55, M21

Introduction

Every company has to face different threats during its operation. A prerequisite for successfully dealing with problems is to register these problems well in advance so that the company can adequately prepare for their solution. One of the most serious problems a company can face is the threat of bankruptcy. Financial distress that is not always clearly visible in their beginnings can be a harbinger of a company's threat of bankruptcy, so potential bankruptcy must be predicted through more complex empirical methods.

Bankruptcy prediction defines Chaudhuri and Ghosh (2017) as a process in which bankruptcy is expected along with financial distress in companies. Currently, the methods used for this prediction include Beaver's model (Kovárník and Hamplová, 2016), Altman's Z-score (Ahmadi et al., 2018), Zmijewski's X-score (Singh and Mishra, 2016), Neural networks (Hosaka, 2019) and other. However, many of these traditional models were developed earlier.

In the meantime, the market has changed and, despite efforts to modify these models to adapt them, newer models have been developed that take into account the current market. One such method for financial analysis of a company's health is the Data Envelopment Analysis (DEA) method, which is currently used by many authors (Mendelová and Bieliková, 2017; Kingyes et al., 2016; Paradi et al., 2004, Horváthová et al., 2019).

A prerequisite for the successful use of the DEA model for the analysis of financial health is the appropriate selection of input data and their pre-processing. Although the process of using DEA models is well known, the process of selecting data and preprocessing it for its successful use in analyzing a company's financial health to predict bankruptcy is insufficient in the current literature. Therefore, in this paper, we will focus on the design of a process model for the selection and pre-processing of data that would lead to the successful use of the DEA model for financial analysis to predict possible bankruptcy. We will compare the results of the research with the research carried out by Mendelová and Bieliková (2017), where they proposed a method of selecting input data to the DEA method to analyze the financial health of companies.

1 Research method

In examining the current literature, we met with only one source (Mendelová and Bieliková, 2017), which dealt with a detailed description of the process of selecting input parameters to the DEA method to analyze the financial health of the company. To fill this gap in the literature, we will design and describe an algorithm for data selection and processing for the mentioned analysis in this paper. We will perform the whole analysis in program R, which is the free language and environment for statistical computing and graphics. Finally, we compare the success of the prediction of the DEA method, in which the inputs were selected based on our proposed process model; the process model we developed in our previous research; and a process model proposed by Mendel and Bieliková (2017).

A prerequisite for the correct application of the DEA method is the need to analyze companies from one sector. We chose the IT industry for the analysis because it is currently one of the most widespread industries in Slovakia. As part of the analysis, the input data will represent the calculated values of financial indicators. For these financial indicators to represent the financial health of the company as reliably as possible, they must be in sufficient quantity. The input data thus represent the values of 27 calculated financial indicators of companies in

the IT industry for the years 2013 to 2017 obtained from the FinStat internet portal, which is listed in Tab. 1.

Tab. 1: Data - 27 calculated financial indicators from the FinStat internet portal

revenues, including revenues from the sale of fixed assets and securities	gross margin	return on long-term capital (EBIT)
profit before tax and interest	EBITDA margin	receivables turnover time
EBITDA (earnings before interest, taxes, depreciation, and amortization)	operating margin	time of collection of short-term receivables
sales adjusted = for inventories and capitalization	profit margin	repayment period of liabilities
accounting cash flow	accounting cash flow margin	time of repayment of liabilities concerning sales
costs of goods and services sold	liabilities / EBITDA	time of repayment of trade payables
gross generation of resources from operating activities	total insolvency	effective tax rate
net operating profit after tax	2nd level liquidity	coverage of personnel costs and charges
net debt	3rd level liquidity	surcharge

Source: Authors´ calculations

Another important variable in the input data is the value of **equity of next year (ENY)**, which represents the value of capital that the analyzed companies achieved in the following year from the observed period. If we analyze the financial health of company X in year n, we will monitor the value of ENY in year n + 1. If this value is greater than 0, then it is a financially healthy company and the DEA method in the analysis of company X in year n should also predict its financial health in year n + 1 and vice versa in the case of financial distress company.

The first step in data processing is to reduce it so that maximum information is maintained. There may be a correlation between some of the 27 financial indicators. The multidimensional statistical method called **factor analysis** allows this correlation to identify and then create a new variable (factor) that will represent these financial indicators. This reduces the number of monitored variables by creating several factors while maintaining the maximum information from the original input data (Kráľ et al., 2009).

The data must first be tested for suitability for factor analysis. Whether there is a correlation between the variables is tested by the **Bartlett test** (Bartlett, 1937), which analyzes the hypothesis: "All population variances are equal". According to the Bartlett test, data are appropriate when we reject the null hypothesis in favor of the alternative: "At least two variants of the population are different." In this case, there is a correlation between the variables in the

dataset, so the assumption of using factor analysis is fulfilled. In the R environment, perform a Bartlett test with the command $bartlett.test\ (data\_list)$.

We test the **Kaiser-Meyer-Olkin (KMO) test** to determine whether the identified correlation between variables is sufficient to perform factor analysis. The result of the test is a KMO-Criterion that takes values from 0 to 1. The data are suitable for the use of factor analysis if the KMO-Criterion is greater than 0.5 according to Kaiser and Rice (1974).

If the data are suitable for factor analysis, it is necessary to identify how many resulting factors (new variables) will be after its use. We identify the resulting number of factors by analyzing the main components using the command $fa = printcomp(data, cor = TRUE, rotate = FALSE)$, while the $cor = TRUE$ argument ensures that the correlation matrix is used instead of the variance matrix. In the output, we get the calculated values of Cumulative proportion, which represent what number of components represents what percent of the variability. According to Meloun et al. (2005), based on this value, we should select such a number of result factors that the value of Cumulative proportion is at least 70%, ie that the result factors represent at least 70% of the variability of the input data.

Next, we will use the command $fa\$communality$ to monitor how many percent of the variability of individual input variables is explained by the resulting factors. If some variables have this value lower than 0.5, then we should remove these variables from the dataset and perform factor analysis again until the resulting factors represent at least 50% of the variability of each input variable.

After performing the factor analysis, we divide the resulting factors into inputs and outputs to the DEA method, as Mendelová and Bieliková (2017) did in their research. Based on the value of equity in the following year, we will divide the companies into financial healthy and distressed companies so that if the value of $ENY < 0$, we will consider the company as financial distress; if the value of $ENY > 0$, then we will consider the company as financially healthy. We then calculate the mean values of the factors for both groups and compare these values. Factors that will have a mean value in the group of financially healthy companies higher than in the group of financial distress companies will thus represent inputs to the DEA method; otherwise, the factors will be outputs to the DEA method.

After identifying the inputs and outputs to the DEA method, we proceed to use the DEA method itself. In the R environment, we used the $deaR$ package for the DEA model. Because it is not necessary to select a specific orientation of the DEA model, we decided to use the SBM model for the VRS condition (Tone, 2001), which will be used to quantify the distances of companies from the curve of production possibilities.

The output of the use of the DEA model will be the assignment (prediction) of whether the company will be in the financial healthy zone, in the gray zone, or the financial distress zone next year. For a company to be able to identify and respond to an existing threat of bankruptcy soon enough based on a forecast, it needs to be classified as a gray zone or financial distress zone based on a forecast. Therefore, it is necessary to calculate the reliability with which the proposed model can warn the company of the possible danger of bankruptcy. For this quantification, we calculate the **Index of warning reliability (I_{WR})**, the calculation of which is expressed in Formula 1.

$$I_{WR}\ (Index\ of\ warning\ reliability) = \frac{n_A + n_B}{n_A + n_B + n_C} = \frac{n_A + n_B}{n_{distress}} \quad (1)$$

where:

- n_A is the number of companies in financial distress classified to the financial distress zone;
- n_B is the number of companies in financial distress classified to the gray zone;
- n_C is the number of companies in financial distress classified in the financial healthy zone;
- $n_{distress}$ is the total number of companies in financial distress.

The next result will be calculated values of **Index of Correct Classification (I_{CC})** (Mendelová and Bieliková, 2017) for each of the analyzed years, which represent the success of the prediction of financial health of companies of the DEA model. The calculation of the Index of Correct Classification is shown by Formula 2.

$$I_{CC} = \frac{n_A + n_F}{n} \quad (2)$$

where:

- n_A is the number of companies in financial distress included in the financial distress zone,
- n_F is the number of companies in financial health included in the financial health zone,
- n is the total number of companies.

The average of the calculated I_{WR} and I_{CC} values of all analyzed years will represent the success of the prediction of our proposed process model. We will then compare the success of the prediction with the average I_{WR} and I_{CC} values from our previous research and the research of Mendelová and Bieliková (2017). We will also compare other average values of results representing the accuracy of the DEA model, which represent forms of an error rate of results.

2 Results

The read data consisted of 28 columns - 27 values of calculated financial indicators and the value of ENY (equity of next year) for each of the analyzed companies.

Tab. 2: Bartlett test and Kaiser-Meyer-Olkin test

Year	Bartlett test p-value	Bartlett test p-value < alpha (0.05)	KMO-Criterion	KMO-Criterion > 0.5	Data suitable for factor analysis
2013	< 2.2e-16	YES	0.67	YES	YES
2014	< 2.2e-16	YES	0.71	YES	YES
2015	< 2.2e-16	YES	0.73	YES	YES
2016	< 2.2e-16	YES	0.71	YES	YES
2017	< 2.2e-16	YES	0.73	YES	YES

Source: Authors´ calculations

The suitability of the data for the use of factor analysis was tested by Bartlett test and KMO-Criterion, while the results of these tests are shown in Tab. 2. After performing the Bartlett test for each of the years, the p-value in each year was <2.2e-16. For a significance level of 0.05, we rejected the null hypothesis in favor of the alternative for each year: "At least two variants of the population are different.". This means that a correlation is present in the data in each of the years examined, and thus the first assumption for the use of factor analysis is fulfilled. When analyzing the KMO-Criterion for the data in each year, we found that for all years the KMO-Criterion is > 0.5, which means that the correlations between the variables are sufficient to perform factor analysis. The tests performed show that the data are suitable for performing factor analysis because there are correlations that are sufficient each year.

After performing the tests, we needed to find out the number of factors that will be calculated for each year. After analyzing the main components, we selected the number of factors so that the Cumulative proportion is> 70% (Meloun et al., 2005), i.e. that the resulting factors explain at least 70% of the variability of the original financial indicators.

Subsequently, we found that the monitor how many percent of the variability of individual input variables is explained by the resulting factors, and we removed from the data those financial indicators whose variability was explained by factors less than 50%. We repeated this procedure until the resulting factors explained at least 50% of the variability of each variable. We only performed this procedure once each year.

After removing the selected variables, we sorted the factors into inputs or outputs to the DEA model each year. To do this, we divided the companies into two groups each year, namely financial healthy and financial distress companies, and in each group, we calculated the

arithmetic averages of these factors. If the average factor was larger in the financial healthy group than in the financial distress group, it represented an input to the DEA method, otherwise, it represented an output to the DEA method.

Tab. 3: Results of DEA analysis accuracy

Year	2013	2014	2015	2016	2017	Average
A (distress to distress zone)	3	5	6	6	4	4.8
B (distress to gray zone)	1	2	4	2	3	2.4
C (distress to healthy zone)	8	6	11	4	9	7.6
D (healthy to distress zone)	34	53	48	42	53	46
E (healthy to gray zone)	15	35	21	51	33	31
F (healthy to healthy zone)	157	202	260	218	255	218.4
Total number of companies	218	303	350	323	357	310.2
I_{CC} (Index of correct classification)	73.39%	68.32%	76.00%	69.35%	72.55%	71.92%
I_{WR} (Index of warning reliability)	33.33%	53.85%	47.62%	66.67%	43.75%	49.04%

Source: Authors´ calculations

Tab. 3 represents the processing of the accuracy of the results of the DEA model. The average value of the Index of correct classification is 71.92% and the average value of the Index of warning reliability is 49.04%. We compared these results with the results of our previous research and with the results of Mendelová and Bieliková (2017). For a correct comparison with our previous research, we repeated the whole process with the difference that we selected the same 100 financial healthy and 10 financial distress companies each year as in the previous research. Also, in previous research, instead of factor analysis, we removed multicollinearity from the dataset and those financial indicators in which we did not reject the null hypothesis in the Mann-Whitney U test: companies "There is no difference within the tested financial indicator between financially healthy companies and companies in financial distress".

Tab. 4: Comparison of results of similar research

	I_{CC} (Index of correct classification)	I_{WR} (Index of warning reliability)
Research of this paper	71.92%	48.65%
Adjusted research of this paper	53.09%	76.00%
Previous research	78.73%	48.00%
Research of Mendelová and Bieliková (2017)	78.50%	60.00%

Source: Authors´ calculations

Tab. 4 shows the values of the I_{CC} and I_{WR} indices using the proposed process model in this study; using the same process model on data from previous research; results from previous research and results from research Mendelová and Bieliková (2017). A comparison of the results shows that the highest I_{CC} value was obtained from our previous research and the highest I_{WR} value was obtained from the process model described in this paper, which was applied to data from previous research. It follows that it is more appropriate to use factor analysis and the DEA model to predict bankruptcy than to analyze multicollinearity and perform the Mann-Whitney U test in conjunction with the DEA model.

Conclusion

The motivation of this paper was to design an effective and inexpensive method for assessing the financial health of companies to predict the risk of bankruptcy. We tested the proposed process model in the R program at companies in the IT sector in the Slovak Republic.

We compared the research results with our previous research and with the research by Mendelová and Bieliková (2017). The result of the comparison was the finding that the procedure described in this paper makes it possible to predict the threat of bankruptcy with the greatest success among the comparative studies. For this reason, we recommend the proposed process model in this paper to companies that want to predict the threat of bankruptcy next year.

The disadvantage of comparing the results is that the proposed procedure was performed on other data as it was in the research Mendelová and Bieliková (2017), therefore the comparison of these specific research may not be sufficient, which we consider a lack of research.

The conditions for the replication of the proposed analysis and its testing are the need to have a large amount of input data in the form of financial indicators calculated for a large number of companies. In future research, we plan to test the proposed process model described in this

paper on similar data as Mendelová and Bieliková (2017) to adequately compare the success of these different process models.

Acknowledgment

The research was realized within the national project "Decision Support Systems and Business Intelligence within Network Economy" (Contract No. 1/0201/19) funded by Grant Agency for Science; Ministry of Education, Science, Research and Sport of the Slovak Republic.

References

Ahmadi, Z., Martens, P., Koch, C., Gottron, T., & Kramer, S. (2018, October). Towards bankruptcy prediction: deep sentiment mining to detect financial distress from business management reports. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 293-302). IEEE.

Chaudhuri, A., & Ghosh, S. K. (2017). Bankruptcy prediction through soft computing based deep learning technique. Springer.

Horváthová, J., Mokrišová, M., & Vrábliková, M. (2019). Integration of balanced scorecard and data envelopment analysis to measure and improve business performance. *Management Science Letters*, *9*(9), 1321-1340.

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. Expert systems with applications, 117, 287-299.

Kingyens, A. T., Paradi, J. C., & Tam, F. (2016). Bankruptcy prediction of companies in the retail-apparel industry using data envelopment analysis. In *Advances in Efficiency and Productivity* (pp. 299-329). Springer, Cham.

Kovárník, J., Hamplová, E. (2016). The comparison of prediction ability of selected bankruptcy models in the glassmaking industry in the Czech Republic. The 10th International Days of Statistics and Economics.

Kráľ, P., Kanderova, M., Kaščáková, A., Nedelova, G., & Valenčáková, V. (2009). Viacrozmerné štatistické metódy so zameraním na riešenie problémov ekonomickej praxe. *Banská Bystrica: Ekonomická fakulta UMB*.

Mendelová, V., & Bieliková, T. (2017). Diagnostikovanie finančného zdravia podnikov pomocou metódy DEA: Aplikácia na podniky v Slovenskej republike [Diagnosing of the Corporate Financial Health Using DEA: an Application to Companies in the Slovak Republic]. *Politická ekonomie*, *2017*(1), 26-44.

Paradi, J. C., Asmild, M., & Simak, P. C. (2004). Using DEA and worst practice DEA in credit risk evaluation. *Journal of productivity analysis*, *21*(2), 153-165.

Singh, B. P., & Mishra, A. K. (2016). Re-estimation and comparisons of alternative accounting based bankruptcy prediction models for Indian companies. Financial Innovation, 2(1), 6.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and psychological measurement*, *34*(1), 111-117.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *160*(901), 268-282.

Meloun, M., Militký, J., & Hill, M. (2005). Počítačová analýza vícerozměrných dat v příkladech.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European journal of operational research*, *130*(3), 498-509.

Contact

Emil Exenberger

Technical University of Košice, Faculty of Economics, Department of Applied Mathematics and Business Informatics

Němcovej 32, 040 01 Košice, Slovak Republic

emil.exenberger@tuke.sk

Jozef Bucko

Technical University of Košice, Faculty of Economics, Department of Applied Mathematics and Business Informatics

Němcovej 32, 040 01 Košice, Slovak Republic

jozef.bucko@tuke.sk