ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume LIX 6 Number 7, 2011

THE FACTORS AFFECTING COMPETITIVENESS OF COMPANIES: CONTRIBUTION AND LIMITS OF THE STATISTICAL PATTERN RECOGNITION METHODS

L. Blažek, P. Pudil, J. Špalek

Received: May 11, 2011

Abstract

BLAŽEK, L., PUDIL, P., ŠPALEK, J.: *The factors affecting competitiveness of companies: contribution and limits of the statistical pattern recognition methods.* Acta univ. agric. et silvic. Mendel. Brun., 2011, LIX, No. 7, pp. 69–80

The paper elaborates the methodical side of empirical research of factors influencing the economic success of companies. The analysis is based on the selective sample of more than 400 stock listed (share holding) companies and limited partnerships located in the Czech Republic. The main goal of the research is to verify, methodically and theoretically, the hypothesis that there is significant mutual dependency between certain types of economic success of companies and a certain typical configuration of values of selected characteristics which describe these companies. The paper concentrates on an analysis of applying the statistical pattern recognition methodology in the course of verifying this hypothesis. Our analysis confirms the potential gains connected with the method. Within the sample we identified group of potential factors of competitiveness which can characterize the interdependence between competitiveness and economic performance.

statistical pattern recognition, business economics, economic success, financial performance, empirical research

1 INTRODUCTION

Searching for factors of companies' economic success represents a current and attractive topic for the business economy theory as well as for the management theory. One of the teams of the Center for Research of Competitiveness of the Czech Republic's Economy has concentrated on the analysis of factors of the economic success of a representative sample of selected companies located in the Czech Republic.

The submitted paper concentrates on the methodical side of the carried out research.[1] It provides brief information about data gathering methods and their preliminary processing. The main focus is, however, concentrated on evaluating the application process of self-learning methods of statistical patterns recognition for the purposes of this specific topic.

2 DATA AND METHODS

This research task concentrates mainly on verifying this hypothesis: whether there is a meaningful dependency between a certain type of economic success of companies and a certain typical configuration of values of selected characteristics that describe these companies. The key factor for the evaluation of the hypothesis lies in good data.

1 More detailed information about this research is available mainly in Blažek (2008) and Blažek, Částek (2009).

Taking into consideration the overall focus of the carried out research, and considering the availability of needed information, the primary set of researched companies was defined by:

a) territorial aspect – companies located in the Czech Republic, further divided by regions,
b) business sectors (industry segment) aspect – sector D Manufacturing industry and sector F Construction business (according to sector classification of economic activities),
c) size aspect – companies with more than 50 employees,
d) legal form of business – limited partnership companies, and share holding companies (corporations).

The set of companies that would satisfy the given criteria, after excluding companies in liquidation proceedings, bankruptcy, or under court execution totaled, at the time of empirical survey, 4483 subjects.

Considering the fact that information from companies' balance statements needed for subsequent analysis were acquired from the Albertina Data database, it was necessary in selecting the companies targeted for questionnaire survey to analyze the extent and quality of information contained in this database available for our use. Based on this analysis we selected 2187 companies that satisfied, from the standpoint of complexity and quality of accounting information, the desired data quality. This selection became the primary data set for our research.

2.1 Data sources

In order to acquire the values describing the primary data set of companies two primary sources of information were selected through empirical inquiry:

a) public information sources,
b) information from questionnaire surveys.

Public information sources consisting mainly of:

- companies' web site information,
- web published analyses on ipoint.financninoviny.cz,
- information from business index published on web site portal.justice.cz,
- information from CreditInfo database which is part of Albertina Data database.

The first three of the above-mentioned sources were used to create a brief characteristic of the companies.

The information from Albertina Data database provides economic data of individual companies compiled from yearend balance statements. This information was used to calculate financial indicators used to evaluate economic success of companies.

The second of the primary information sources was information from the questionnaires filled out during the empirical search by the researchers in cooperation with their respondents.

The structure of the questionnaire and the focus of individual questions were based on the stakeholder approach that considers the success or failure of a company as a result of conflict of interest between individual groups of stakeholders – mainly the owners, employees, customers, suppliers, state and the community[2]. Considering this approach, the individual parts of the questionnaire pertained to:

- owners and assets (the type of ownership, the effect of owners on management, the amount of tangible assets, software applications),
- employees (employee structure, employee turnover, wage variability, benefits, employee education, and etc.),
- consumers and customers (business strategy, type of customers and their stability, proportion of exports, product specifics, and etc.),
- suppliers (type of suppliers and their stability, proportion of foreign imports, specifics of supplies, and etc),
- corporate responsibility to codices and certificates.

In the effort to minimize the extent of research, the questionnaire almost entirely focused on information that cannot be easily acquired from publicly available sources or by other means. The purpose of the questionnaire was mainly to get opinions and qualified estimates of strategically thinking company representatives.

Representativeness of the selective sample of companies

The questionnaire was filled out during respondent's meeting with the interviewer. Interviewers were specially trained how to fill out the questionnaire and get a deeper understanding of the survey purpose.

The 432 companies that participated in the empirical survey represent 15.33% of the primary data set. A quota system was used to choose which companies will be included in the selected sample. Quota variables consisted of – territory, where the company is located (region); industry segment in which the company operates (manufacturing industry and construction business, within the manufacturing segment 20 more detailed industry specific sectors); size aspect – companies with: 50–99 employees, 100–249 employees, 250 and more employees; and legal form of business – limited partnerships, and share holding companies.

We can say that, with the exception of some of the regions and segment sections, a high rate of accordance was achieved in the proportions of the

2 Compare for example Berman, Wicks, Kotha (1999), resp. Mitchell, Agle, Wood (1997).

primary and selected samples, which had a positive effect on the representativeness of the selected sample.

2.2 Methods

2.2.1 Primary statistical analysis

The starting dataset of information for every company was composed of 683 variables acquired from the questionnaires and 123 variables representing financial data from the CreditInfo database. Data refining and preparation of data for transformation for further use was an integral part of the primary analysis.

We analyzed the frequency of answers to individual questions of the questionnaire. The results were presented in a uniform way using standardized tables and standardized graphs, including commentary (notes) that briefly interpret presented numerical and graphic information. The evaluation of the data was conducted for both the entire data set as well as the selected data set, and divided into partial sets by industry segments, company size, and legal form of business. This way we acquired a great deal of information about "how the companies are". We created base points for interpreting the acquired data and have also gained many additional inputs for further analysis.

2.2.2 Creating groups of companies based on their economic success

The next procedural step to validate the above mentioned hypothesis was grouping the companies based on their economic success. After analyzing various approaches to this task, and taking into consideration available data, we selected the description of a company's economic success as being represented by two indicators – profitability of assets and growth of assets.

The selective sample, i.e. 432 above mentioned companies, was structured based on cluster analysis. This was carried out using the K-means cluster analysis method that divided the companies into relatively homogenous groups (clusters) based on minimal inter-cluster distance between individual components/members expressed by Euclidian metrics. Considering the fact that the two selected indicators take on variably high values, it was necessary to standardize the given values to ensure comparability. Z-scores were used for standardization.

Using five-year time series for these indicators,[3] there were thirteen groups of companies created based on the cluster analysis and these were, based on the economic analysis, aggregated into 5 and subsequently 3 typical groups A, B, and C.

Group A was classified as companies that showed above average values of both indicators, and companies that showed value of one of the indicators above the average, with the second value below the average, however not in the negative. This group included 250 companies. These companies can be considered economically successful.

Group B was categorized by companies that showed below average values of both indicators for the selective group, however not negative values. This group included 185 companies which can be placed in the middle of the economic success scale.

Group C consists of companies that showed negative values of both mentioned indicators. This group consisted of 42 companies. These companies, with negative financial performance, can be considered to be economically unsuccessful. [4]

2.2.3 Statistical pattern recognition

However, within the framework of the preceding text, from the methodical stand point the most challenging step of verifying the hypothesis of the existence of the dependency between acertain type of economic success of a company and certain typical configuration of values of selected characteristics describing these companies it was necessary to analyze which of the characteristics, acquired through empirical search, describing individual companies of the primary data set influence the economic success of these companies, and thus affect their inclusion into the above mentioned groups.

Taking into consideration that the hypothesis rightfully assumed that it is not a particular effect of individual characteristics, but always an integral effect of certain groupings of selected characteristics, it was not possible to consider using simple statistical methods. Instead, methods implementing multidimensional statistical analysis needed to be used.

Considering the nature of the analyzed problem an approach was chosen based on the methods of statistical recognition and classification of patterns together with the method of dimensionality reduction, specifically the method of selecting of the most informative features[5]. This approach was successfully applied for example in healthcare services to classify mammographs (see Somol, Pudil, Kittler, 2004; Somol, Novovičová, Pudil, 2008; Somol, Novovičová, Grim, Pudil, 2008). Within the framework of tasks with economic focus the

3 i.e. the vector for each company had 10 coordinates

4 Substantiation of the selection of indicators, as well as the whole cluster methodology, including the economic interpretation is presented in Šiška's publication (2007).

5 These approaches together with corresponding algorithms are developed on a long-term basis in a Joint laboratory of Faculty of Management VŠE and ÚTIA AV ČR. The results have been published in tens of publications, that have in SCI, SSCI or Scopus citation index/databases more than 1000 citations.

approach was concerned with finding significant factors of the acquisition process (Pudil, Pirožek, Somol, 2000 or Pudil, Pirožek, Somol, 2002) or analysis of creditworthiness of insurance companies' clients (Somol, Baesens, Pudil, Vantihienen, 2005).

The application of the given approach for analyzing the task at hand represents broadening the scope of its use by the sphere of business economics and management.

The methodology of reducing the dimensionality of decision-making problem of classification type (or the method of selecting the most informative features) is very extensive and its more detailed description is beyond the scope of this paper. For this reason, we are concentrating mainly on the basic concepts of the "self-learning" methods of recognizing and selecting the most informative features from the viewpoint of their application in the framework of given task.

Characteristics of learning approach to the recognition problem

The principal goal of methods categorized as *statistical pattern recognition* methods can be, in a simple way, characterized as a goal to classify *patterns* (real world objects representation) into a finite, usually not large number of groups[6]. In the case of two groups, which are most common, we are talking about so-called dichotomic classification. This method is based on the assumption that every pattern falls exactly into one of the groups, at this is his *classification feature*.

The underlying matter of the problem is the fact that at the time when a decision needs to be made – i.e. a classification into one of the groups, the given classification feature is not known, nor is it directly measurable. It is thus possible to carry out the classification, all things considered, only by using other measurable features of the pattern, whereas the decision making rule is derived by learning (or training) from past occurrences.

However, the fundamental hypothesis is that this other measurable data (in the statistical pattern recognition terminology *features*) is at least statistically related to the actual classification of the pattern into a group. The task, therefore, requires instead of the analytical approach the use of the so called self-learning approach that is based on the idea that sufficient information needed to classify or recognize a pattern is contained in the data describing past experiences.

To implement the self-learning method we need to employ the so-called training dataset. This dataset must consist of patterns with known classification. The solution is finding above described features and finding a rule for how to assign individual patterns according to these features into individual groups.

Using the above characterized approach to solve given tasks we made following application and interpretation:

Individual companies of the selective sample were considered as patterns for the purpose of the given application. Individual groups of companies were created based on cluster analysis according to the selected financial indicators – profitability and growth of assets (see paragraph 2.1) were considered as classes. As features we considered the variables from previous set of characteristics (variables), acquired from the questionnaire or database respectively, that have, in their mutual context, a significant effect on classifying the companyto one of the mentioned groups. These are the factors influencing economic success.

Reduction of dimensionality

The general goal of the presented approach is the classification of a given pattern (i.e. a company) – deciding into which class or group (cluster based on the economic success) it belongs. Since the grouping was, by definition, made based on the financial indicators (see paragraph 2.1) the classification was solved prior to the application of the presented approach. The focus of the application was in essence narrowed down to finding causes of why companies achieve such values of financial indicators that assign them into individual classes. It is then a matter of finding the most descriptive features that characterize, in the best way, different classes – i.e. factors of economic success that are generally valid for all companies of the primary data set.

For the purpose of statistical pattern recognition based on a selected set of features (factors) it was necessary to appropriately prepare the original, fairly extensive and heterogeneous set of acquired information (see paragraph 2.1). It is possible to, generally speaking, describe a given pattern by a number of characteristics that we can denominate as a set of variables D_0. It is however not suitable to use all of these variables for the actual selection of the most informative symptoms for the primary input. The reason can be, for example, their mutual correlation, as well as the fact that a number of them can be redundant or irrelevant for the given task. It is therefore necessary to reduce the number of these variables. Within the framework of the given analysis, we have used traditional statistical methods to carry out the preliminary reduction based on two successive steps:[7]

1. **primary data analysis** – that lead to the elimination of those variables that showed

6 As we will describe later, in this task we have divided the companies into different numbers of groups based on the value of two financial indicators. The number of groups varied from relatively high (13), to classification dichotomy.

7 For more detailed description of applying these methods see Špalek (2008: chapter 4).

a high level of missing values, or in some cases that presented very low variability between individual classes (groups of companies created based on economic success),

2. **bivariate data analysis** – that lead in particular to the elimination of those variables that showed mutual correlation, or lead to creating new (dummy) variables.

Subject analysis based on economic aspects of analyzed relationships was carried out along with these statistical methods to accomplish this reduction.

Despite a substantial reduction carried out this way, the acquired set D_1 remained extensive. For the given task (see paragraph 4) it was a matter of 37 variables. Knowing that the relationship between variables (between features) are within this task very complex and heterogeneous and that it is impossible to determine in advance which variables to eliminate and which to keep in this process of preliminary reduction of dimensionality, we carried out the reduction very carefully.

Types of search strategies

The second phase of dimensionality reduction, carried out using the statistical pattern recognition method, was based on searching for sets of features d (factors of economic success) within the set of variables D_1 so that the chosen criterion was maximized. It was a matter of determining:

- efficient strategy of searching for an optimal set of features,
- criterion to rate the quality of this set of features.

Individual techniques of dimensionality reduction, focused on preserving the differentiation of the classes, can be divided according to two approaches – feature selection and feature extraction, based on transforming the feature space. We will be further concerned only with feature selection methods that are adequate for the task within the given research framework.

By selecting appropriate criterion (criterion function) to evaluate the quality of feature subsets, the actual process of selecting features (factors) is transferred into a search problem – detection of optimal subset of features in terms of the selected criterion. It has been proven that in order to find a guaranteed optimal subset of d features from given D of observations (properties, indexes, attributes, etc.) it is necessary to search through all possible ds. The exhaustive search procedure will find the optimum thanks to the fact that it is searching through all subsets of value d.

As we know, those can be $\binom{D}{d}$. The above-mentioned process is, however, not very practical since the number of examined subsets can quickly reach a disproportionately large size and thus unreasonably prolong the computing time.[8]

For this reason, current trends concentrate on constructing a more sophisticated and efficient search method that has less time requirements than the demands for an exhaustive search would be.

The main trend in the feature selection field concentrates on sub-optimal search strategies. A whole array of sub-optimal search methods has been published, the most known and nowadays the most quoted ones are methods of sequential floating search of subsets (see Pudil, Novovičová, Kittler, 1994). The floating search encompasses in reality two methods. Even though they both alternate adding and subtracting features into and out of the working dataset, they can be distinguished as two separate algorithms based on the prevailing direction of the search. It can be:

- forward algorithm, also known as sequential forward floating selection – SFFS,
- backward algorithm, known as sequential backward floating selection – SBFS.

Both algorithms together are known as floating methods, because the resulting dimensionality that follows the individual steps of the algorithm does not change monotonously, but in reality "floats" up and down. By combining the original (after optimal steps) SFS (sequential forward selection) and SBS (sequential backward selection) methods, the floating methods are approaching the optimal solution. Their main characteristic is that, for example, during the backward floating selection (SFBS) the once eliminated features can be during the selection process added back to the selected features, if that improves the criterion value.

In some cases (for example – a large number of primary features, as it is the case of given analysis) it is possible to "fine tune" the solution (the most informative feature subset of dimensionality d) using the oscillatation algorithm (Somol, Pudil, 2000). Oscillation search can be considered as a "higher level" procedure that uses a different feature selection method than a sub-procedure of the main course of the search. The whole concept is characterized by high flexibility and allows modification for different purposes.

For the analysis of economic success we chose, in consideration with the above mentioned matters, the SFS (sequential forward selection) method.

Selecting the most informative features using the classification method as a selection criterion

An appropriate criterion function, capable of describing how informative is any feature subset

8 Let us present for illustration, that if we want to select 10 symptoms out of 60 available measurements we will need to evaluate more than 7×10^{10} data subsets of symptoms. The number of possible combinations for selection of 30 symptoms from the original 60 measurements is even bigger than the number of molecules in the universe.

under consideration, is needed in order to apply the search method. A number of functions that can be used for a common data recognition task is well known, however the specifics of the given analysis requires a custom/made definition of the function. Before we are able to define this function, we need to summarize some important observations from the pattern recognition field.

It is nowadays common to divide the feature selection methods into *filters* and *wrappers*, where the difference is in the principle of the use of the criterion function[9]. Wrappers are computationally more demanding, however more common in practice. Instead of criterion functions, they use one of the concrete decision rules (classifier). Evaluating feature subsets in the wrapper type selection then looks as follows: part of the data is used to train the classifier, this part is afterwards with the help of the classifier repeatedly classified – the percentage of the samples that have been correctly classified this way is subsequently used in the feature selection method in place of the criterion function value. It is clear from this observation, that evaluating every analyzed feature subset is very time consuming because of the need to train the classifier. On the other hand, it is possible to select features that with use of a given classifier will lead to more accurate classification than it would be possible using the features selected by the *filter* method.

The number of known classifiers is enormous. For our purposes we selected the *K-Nearest Neighbors* classifier, known under the abbreviation kNN, that by using its properties deals well with compromises that need to be considered: kNN is parameter free (we do not input additional assumptions, which is convenient considering the type of given analysis); kNN is easily implemented and quite demanding as far as the computing power and memory is concerned (which is not a problem in our case as we do not require interactive response); kNN's decision capability is one of the strongest among available classifiers considering the nature of the task of generalization, or classification based on unknown data it is not a serious problem.

Briefly summarized, the kNN principle is as follows: starting from the situation where the goal is to classify a new pattern, while we have available a training data set with known affiliation to classes. We will analyze the distance (using for example – Euclidian) of the analyzed pattern to all components of the training data set and find this way k of its nearest neighbors. The sample will then be assigned to the class where the majority of these k neighbors belong.

Value of k is a user parameter and, in general, it can be any nonnegative number. In practice, it is however necessary to account for the way this value affects the classification abilities of the kNN classifier. For $k = 1$ the hypothetical boundary separating in multidimensional space individual classes is composed of straight areas (the boundary then has edges and vertices). With increasing k the border starts to round out and starting with a certain value of k we start losing the ability to distinguish details. On the other hand, a higher value of k can be a safeguard against the influence of isolated non/typical patterns – outliers. In a number of applications it is common to use k values equal to 1, 3, and 5 (odd numbers are preferable to suppress indecisive situations).

For further information about classification and decision problems in general we refer you to books providing an overview – Duda, Hart, Stork, (2000) and Theodoridis, Koutroumbas, (2006), resp. paper Jain, Duin, Mao, (2000).

For our purposes we use the kNN classifier in the role of the criterion function that evaluates how informative is a data subset as follows: For a given feature subset, we will attempt to assign each of the patterns (a company) into a class based on its *k's* closest neighbors and establish a relative success of correct classification. A proportion of correctly (re) classified patterns always analyzed for a currently studied feature subset will be used within the framework of the search algorithm for further direction of the search. The result is the subset of d features that provides, in this sense, the most accurate classification.

Technical difficulties and question of selecting the most informative variables

The above-mentioned methodology of selecting the most informative variables is based on the assumption that it is possible to interpret the distance (similarity, mutual dependency) between random patterns (companies). The standard form of the kNN classifier can be used for numeric data without missing values.

The analyzed dataset of variables that characterizes companies, however, has two disadvantageous characteristics. Not all of the features are known for all patterns (companies did not provide some data and it was not possible to find it), and not all of the features are numerical (for example – the business segment, or legal form of business are not by their nature indexed). The distance of two patterns cannot, in this case, be evaluated using simple Euclidian metrics, however it is necessary to define our own combined metrics that will allow the implementation of kNN as a criterion of a feature selection.

9 Division implemented by Kohavi, John, (1997)

These combined metrics are generally expressed as:

$$Dist(A,B) = \sqrt{\sum_{i=1}^{D} \|a_i - b_i\|},$$

where the meaning of $\|a_i - b_i\|$ depends on the type of values that the i-th feature takes. For all features with numeric, arrangeable values (real values, whole number values that quantify a certain property of the pattern – for example the number of employees, amount of resources allocated for education, and etc.) we maintain the standard Euclidian meaning – $\|a_i - b_i\| = (a_i - b_i)^2$. For categorical features (for example the above mentioned legal form of business – limited partnership, or share holding company) we define $\|a_i - b_i\| = 1$ in cases where $a_i = b_i$ and $\|a_i - b_i\| = 0$ in all other cases.

The majority of variables characterizing companies are captured on a scale – most frequently on a five point scale. Quality is then quantified using a point assessment. From the definition of the above mentioned metrics we accept the simplified assumption that the distance between individual levels on the points scale is the same and thus the assigned numbers can be treated as regular numbers. The analysis of returned questionnaires suggests, however, that this need not to be the case in all instances. A possibility presents itself to differentiate between the individual values of the point scale to take into consideration the nature of each question and the way the answer was interpreted.

It is also necessary to keep in mind the two following circumstances that can influence the selection of the most informative features (factors). They are:

- **Number of classes (groups of companies) into which the pattern are divided**

  Classification tends to be more accurate with lower numbers of classes into which it is necessary to classify the samples. We can view the task in a hierarchical way – other feature subsets will most likely be the most informative in order to differentiate all of the classes together, while others will be the most informative in order to differentiate certain actual pairs (or other subgroups) of classes. The hierarchical approach to analysis can lead to a more refined differentiation of the importance of competitiveness factors.
- **Separability of classes (groups of companies)**

  Potential difficulty can occur in the separability of classes, which tends to be a common problem stemming from the internal character of analyzed data. In some types of the problem it might be more appropriate to consider multiple classification of a sample into multiple classes especially when the classes strongly overlay each other in the multidimensional space and are, from the theoretical viewpoint, distinguishable only with a certain minimal error. The methods of statistical pattern recognition applied in our analysis presume however an unequivocal classification of the patterns to classes, which leads to better defined and interpretable results. The results of economic success analysis are, in this sense, dependent in principle on the definition of the classes (groups of companies) with the different nature of economic success.

The stated circumstances lead, among other things, to the assignment of experiments described in following paragraphs.

3 RESULTS

A number of experiments were carried out to verify whether the methodology of statistical pattern recognition is applicable for a given task – i.e. finding factors, that in their mutual interaction influence the economic success of companies and whether the acquired results are credible. The task of individual experiments was based on combining a selection of available variations of this methodology and selecting inputs.

As we stated above, the Sequential Forward Floating Selection (SFFS) method of statistical pattern recognition was selected. This method generally provides the best results[10] and therefore it was used in all of the experiments presented below.

To evaluate the information value of tested datasets the k-Nearest Neighbors (kNN) classification method was used. The experiments were carried out with the variable pre-set of k values to 1, 3, and 5. After each experiment, the most appropriate value of k, that provided us with the most informative dataset of variables, was selected. In the following text the three variations of selected k are denoted as 1NN, 3NN and 5NN methods respectively.

The selection of variations of inputs was concerned with the selection of the number and character of classes (groups), into which the companies were assigned based on their economic success.

As a criterion of mutual comparison and evaluation of results of individual variants we chose the probability with which the patterns (companies) are assigned into the appropriate classes based on selected features.

3.1 Basic experiment

Data set D_1 containing 37 selected variables was entered into the primary experiment. The relationship between informativeness (in our case

10 This was, to a certain extent, verified in our other experiments that are not presented here because of the limited extent of this article.

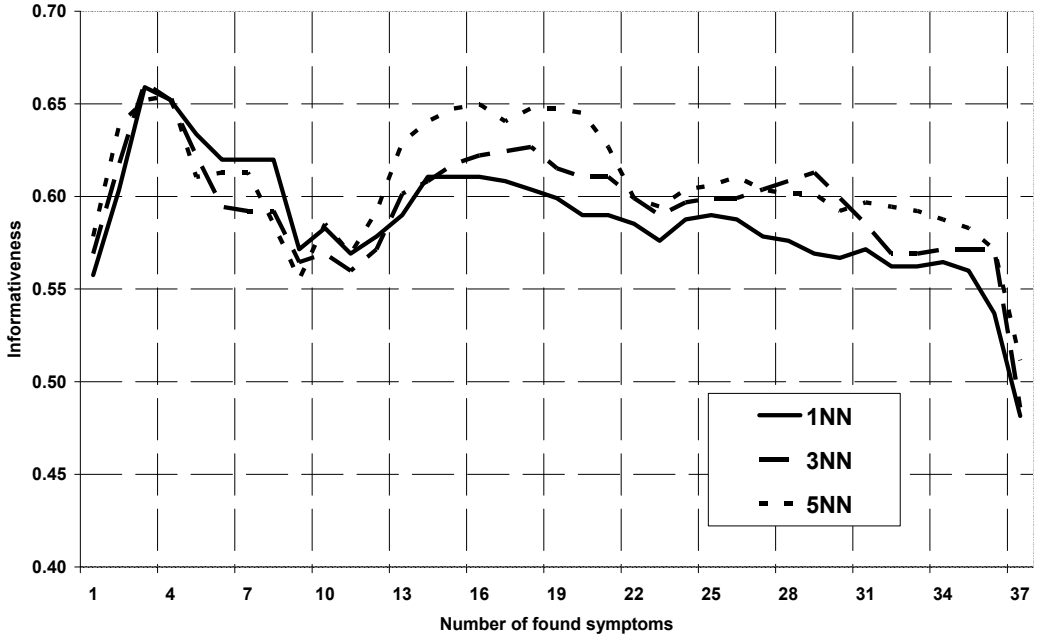
evaluated with the help of the probability of correct classification) and the number of features (variables) of the given method selected in individual steps as factors of economic success is presented in Figure 1.

The results comparing the use of methods 1NN, 3NN a 5NN are shown here. It is obvious that a different number of "the closest neighbors" clearly affects achieved maximum values of informativeness. Using the 1NN the highest value of informativeness is 0.6590, for 3NN it is 0.6613 and using 5NN it is 0.6544.

The typical evolution of the value of informativeness is also obvious from this regardless to the number of the "the closest neighbors". It starts with rapid growth and reaching the global maximum for a dataset of three to four variables, and then decline. The decline of the value of informativeness after reaching the optimum does not, however, decline with the increase of dataset variables continuously nor without exception. As we can see, the decline of informativeness does not happen monotonously, but there are certain "swings" connected to the existence of local maximums. Looking for and evaluating these local maximums were the next steps of the experiment.

The reason is given by the nature of the analyzed task. It is not possible to presume that such a complex and challenging phenomenon, that companies' economic success without a doubt is, depends only on the combination of values of three to four variables. In the same manner, it would be difficult, or even impossible, to credibly interpret certain types of economic success using these few variables.

Three datasets of variables actually reach the local maximum by using the 1NN method, the datasets with 14, 15, and 16 variables, while attaining the informativeness value of only 0.6106. Using the 3NN method, a dataset with 18 variables reaches the local maximum with an informativeness value of 0.6267. And finally, using the 5NN method, a dataset with 16 variables reaches the local maximum with informativeness value as much as 0.6498, which is, converted to percentages, not even half a percentage point worse result than the best dataset achieved by this method (foursome variables with informativeness of 0.6544) and roughly one percentage point worse than the best overall informativeness of 0.6613 achieved with



1: *Dependence of informativeness on the number of found symptoms*
Source: Částek (2010), modified by authors

I: *Comparison of global maximums and selected local maximums*

Classification method	global maximum		Selected local maximum	
	Informativeness value	Number of found factors	Informativeness value	Number of found factors
1NN	0.6590	3	0.6106	14
3NN	0.6613	3	0.6267	18
5NN	0.6544	4	0.6498	16

Source: the authors

this 37 member dataset of variables. The results are compared in the following table:

Based on the above noted facts, the factors of economic success of companies were identified as variables contained in the 16 member dataset created using the SFFS 5NN method with the information value of 0.6498.

3.2 Additional experiments

In order to examine the effect of changes in input values on the results of selecting the dataset factors of companies' economic success, additional experiments were carried out.

We focused on monitoring how informativeness is affected by changes in the class structures. We experimented with following structures:[11]

a) Thirteen groups of companies, as they were grouped together based on cluster analyses (more in paragraph 2.6);
b) Three groups of companies, created by aggregation of mentioned thirteen groups (more in paragraph 2.6);
c) Three variations of two dichotomous groups of companies.

In case of c) – variations of dichotomous groups were created in such a way that out of the groups A, B, and C we focused in turns on analyses of one against the merger of the remaining two.

It is clear that in the presented Tab. II, where the highest values of global and selected local maximums of informativeness are captured, in 13 groups the global maximum value of this criterion reaches only value of 0.3086, in 3 groups (A, B, and C) value of 0.6613, and the best option of the dichotomous groups (C versus A + B) even value of 0.9284.

It is worth mentioning the fact that by comparing the three option of dichotomous groups it is obvious how confinement, respectively specifics of given classes (groups) affects informativeness. The highest value of informativeness is shown by selected features (factors) within the "companies with negative profitability and negative growth when compared to other companies" (C versus A + B) dichotomy. Ranked second, the value of informativeness is shown by selected features forming the other view point, meaning the "above average companies compared to the rest of the companies" (A versus B + C) dichotomy, while the worst values of informativeness is then shown by selected features when the companies with the economic success in the middle of the scale are compared against each other group of companies – above the average one and loss making ones (B versus A + C).

Overall, it is clear from the experiments that our method provides the better results the lower the number of classes, meaning the more aggregate the structure of the groups of companies based on economic success is.

A somewhat different conclusion offers itself, when we are reminded, that informativeness is in our case understood as a probability with which the companies are assigned into the appropriate classes (i.e. groups of companies defined by the type of economic success) based on the values of selected dataset of variables (factors). We must take into consideration that without the use of any methodology, solely based on random selection according to statistical probability and the effect of the law of large numbers, a company that is ranked with the existence of 13 groups into the appropriate group with the 1:13 probability (that is 0.0769), compared to with the existence of two groups with the probability of 0.5.

It is clear from the Tab. III that the biggest effect of the methodology application on increasing the probability of appropriate classification of a company against a random selection is in the case of the largest number of companies groups (13), where the increase is more than fourfold (see column p_{ig} / p_n). Compared to that, even the best case in the 2 groups of companies (C versus A + B) shows a increase of probability only in the amount of 1.8568. This interpretation of the experiment results then shows that with the growing number of classes (groups of companies) the effect of the methodology application on informativeness increases.

II: *Relationship between informativeness and change in structure of companies*

Company structure based n economic success	global maximum			Selected local maximum		
	Classification method	Informativeness value	Number of found factors	Classification method	Informativeness value	Number of found factors
13 groups	5NN	0.3086	2	5NN	0.2815	15
3 groups (A, B, C)	3NN	0.6613	3	5NN	0.6498	16
2 groups (A versus B + C)	5NN	0.7413	19	5NN	0.7413	19
2 groups (B versus A + C)	1NN	0.7229	4	1NN	0.6905	19
2 groups (C versus A + B)	3NN	0.9284	4	3NN	0.9215	20

Source: the authors

11 Detailed results of additional experiments with variable petting can be found in Špalek, Částek (2010).

III: *The effect of methodology application on increased informativeness*

Company structure based on economic success	global maximum			Selected local maximum		
	probability of appropriate classification		p_{ig} / p_n	probability of appropriate classification		p_{il} / p_n
	Methodology p_{ig}	Random p_n		Methodology p_{il}	Random p_n	
13 groups	0.3086	0.0769	4.0118	0.2815	0.0769	3.6595
3 groups (A, B, C)	0.6613	0.3333	1.9839	0.6498	0.3333	1.9494
2 groups (A vs. B + C)	0.7413	0.5000	1.4826	0.7413	0.5000	1.4826
2 groups (B vs. A + C)	0.7229	0.5000	1.4458	0.6905	0.5000	1.3810
2 groups (C vs. A + B)	0.9284	0.5000	1.8568	0.9215	0.5000	1.8430

Source: the authors

4 DISCUSSION

Searching for the factors of economic success of companies using the exact approach based on multidimensional statistical analysis presents a demanding task carried out in a so far, fairly unexplored territory. Some of the distinctive facts and related questions are presented here for discussion.

4.1 Relationship between cause and consequence

The given task fundamentally concerns itself with searching for a relationship between economic success as a consequence and certain characteristics of a company as a cause. There is a logical hypothesis that the cause precedes the consequence. This was not, and could not have been, respected in given analysis. The five year time line of indicators of profitability and growth of assets, from which financial performance was calculated and based on which the economic success was judged, was preceded by periods in which data was collected using a questionnaire. The past economic success was put in relation to the causes of the future economic success. The research team then assumes that information from account statements, from which the financial indicators are derived, will be collected in the future as well and calculations will be made with updated data. This way, the relationship cause – consequence will have the proper time sequence. In this sense, it is however necessary to be reminded, that we can assume, and thus also analyze a reverse causality: to what extent a company did that was in the past economically successful or unsuccessful create ordid not create conditions for future economic success.

4.2 Selection of companies characteristics and their quantification

There is an extensive, through many decades updated theory used to evaluate the financial performance of companies. Many methods, supported by the use of the number of financial indicators stem from this theory. These indicators have a quantifying character and are derived from the accounting data that reflect the value side of described reality.

It cannot be generally stated that accounting valuation of actual events as well as the formation of financial indicators is problem free. In the same manner, the correct selection of concrete financial indicators that would best represent financial performance and economic success of companies was not without difficulties either. Nevertheless, within the framework of the given task, the second side of the analyzed relationship – the characteristics of the companies, is disproportionately analyzed. And specifically in this area we can find problems to which solution the given analysis was trying to contribute the most.

The initial problem is creating a dataset of characteristics of a company (dataset of variables D_0, see paragraph 3.2) that we can reasonably assume contains all of the most informative features – i.e. potential factors of economic success. If this data set does not contain the given feature, it naturally cannot be selected. At the same time, there is no exact methodology to create such data set, nor for testing for risk that one of these symptoms is missing. It was necessary, while analyzing the given task to use heuristic practices, supported by the business economic theory. Specifically, as we mentioned earlier in paragraph 2.3, the application of the stakeholders approach.

Another difficulty is the fact that the sources of companies' characteristics are in most cases subjective opinions by the companies' representatives captured on a point scale. In a substantially lower number of cases are there numerical values available that are based on internal statistics, and in many cases these are only expert estimates. We can assume that within the framework of the selective sample the inaccuracies tied to subjective opinions of individual correspondents (overestimating, underestimating, erroneous or inaccurate opinion, and etc.) will balance each other out, however we cannot dismiss the fact that common difficulties with metrics (see paragraph 3.5) can negatively affect the quality of results.

4.3 Relationship between informativeness and class structure

The result of the theoretical fundamentals of statistical pattern recognition and classification and their previous application is that the informativeness of symptoms of class differentiation grows with a declining number of classes. That is supported by the results of our experiments presented in paragraph 3.2. This fact is not contrary to the interesting finding, that is mentioned at the end of the quoted paragraph – that with the growing number of classes (groups of companies) the effect of applying the methodology on informativeness grows as well. This relationship is however only relative. It is clear for the absolute numbers from the experiment that the methodology assigns patterns (companies) into individual classes more precisely with fewer classes.

It seems that for the factual interpretation within the framework of the given task (and also considering the above noted facts) it is most appropriate to look for the most informative features, meaning factors of economic success, while assigning the companies into two classes – above average and below average companies.

5 CONCLUSION

The presented results allow us, with a considerable level of probability, to conclude that the selected approach based on the statistical pattern recognition method is a suitable tool to look for factors of economic success of companies. Until now, executed analyses confirmed a number of initial hypotheses; however they have also uncovered, respectively specified in more detail, a number of problems representing a challenge for further solution as well.

Currently we are processing the data from the second empirical search that captures (with the exception of earlier analyzed manufacturing industry and construction industry) the majority of business sectors. It is nevertheless carried out principally in the same manner, using more precise methodology that creates positive groundwork for advancing the methodical as well as contextual side of the given task.

SUMMARY

The paper is devoted to the methodical side of empirical research of factors influencing the economic success of companies that was carried out within the Center for Research of Competitiveness of the Czech Republic Economy. The selective sample consisted of more than 400 stock listed (share holding) companies and limited partnerships located in the Czech Republic operating in the manufacturing and construction business segments, with 50 or more employees. The main goal of this research was to methodically and theoretically verify the hypothesis of significant mutual dependency between certain types of economic success of companies by a certain typical configuration of values of selected characteristics describing these companies. The paper concentrates on analysis of the application of the statistic pattern recognition methodology in the course of verifying this hypothesis.

Acknowledgements

This paper was prepared in context to research activities of the Center for Research of Economic Competitiveness of the Czech Republic, that was established and operates with the support of MŠMT 1M0524 project.

REFERENCES

BELLMAN, R. E., 1957: *Dynamic Programming*. Princeton University Press, Princeton, NJ. Republished 2003: Dover, ISBN 0486428095.

BERMAN, S. L., WICKS, A. C., KOTHA, S., JONES, T. M., 1999: Does stakeholder orientation matter? The relationship between stakeholder management models and firm financial performance. *Academy of Management Journal*, 1999, Vol. 42 No. 5, p. 488–506.

BLAŽEK, L. a kol., 2008: *Konkurenční schopnost podniků. Analýza faktorů hospodářské úspěšnosti.* Brno: Masarykova univerzita, 2008. ISBN 978-80-210-4734-1.

BLAŽEK, L., ČÁSTEK, O., 2009: Stakeholder approach and the corporate financial performance. *Národohospodářský obzor*, IX., 2, p. 91–106.

ČÁSTEK, O., 2010: *Využití stakeholderského přístupu při strategické analýze podniku*. 1. vyd. Brno: Masarykova univerzita, 2010. 242 s. ISBN 978-80-210-5411-0.

DUDA, R. O., HART, P. E., STORK, D. G., 2000: *Pattern classification.* 2nd Edition. New York: A Wiley Interscience, 2000. 654 p. ISBN 0-471-05669.

JAIN, A. K., CHANDRASEKAR, B., 1982: Dimensionality and sample size considerations in pattern recognition practice. Handbook of Statistics. Amsterdam: P. R. Krishnaiah and L. N. Kanal, 1982, Vol. 2., p. 835–855.

JAIN, A. K., DUIN, R. P. W., MAO, J., 2000: Statistical pattern recognition: A review. *IEEE Transanctions on Pattern Analysis and Machine Intelligence*, 2000, Vol. 22, No. 1, p. 4–37.

KOHAVI, R., JOHN, G. H., 1997: Wrappers for feature subset selection. *Artificial Inteligence,* 1997 No. 1–2, p. 273–324.

MITCHELL, R. K., AGLE, B. R., WOOD, D. J., 1997: Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, October 1997, Vol. 22, No. 4, p. 853–886.

PUDIL, P., NOVOVIČOVÁ, J., KITTLER, J., 1994: Floating Search Methods in Feature Selection. *Pattern Recognition Letters*, 1994, Vol. 15, s. 1119–1125.

PUDIL, P., PIROŽEK, P., SOMOL, P., 2002: Selection of Most Informative Factors in Merger and Acquisition Process by Means of Pattern Recognition. *Signal Processing, Pattern Recognition, and Applicatio, IASTED,* ACTA Press, 2002, s. 224–229. ISBN 0-88986-338-5.

PUDIL, P., PIROŽEK, P., SOMOL, P., 2000: Výběr nejinformativnějších faktorů při akvizici podniků pomocí metod rozpoznávání obrazů. *Acta Oeconomica Pragensia*, 2000, roč. 8(2), s. 143–159.

SOMOL, P., BAESENS, B., PUDIL, P., VANTI-HIENEN, J., 2005: Filter- versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems,* 2005, Vol. 20, No. 10, s. 985–999.

SOMOL, P., PUDIL, P., 2000: Oscillating search algorithms for feature selection. *Proceedings of the 15th International Conference on Pattern Recognition. IEEE Computer Society,* Los Alamitos, 2000, s. 406–409.

SOMOL, P., PUDIL, P., KITTLER, J., 2004: Fast Branch & Bound algorithms for optimal feature selection, *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 26, No. 7 (2004), p. 900–912.

SOMOL, P., NOVOVIČOVÁ, J., PUDIL, P., 2008: Are Better Feature Selection Methods Actually Better? Discussion, Reasoning and Examples, *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies – BIOSTEC 2008*, International Conference on Health Informatics, (Funchal Madeira, Portugal, (2008).

SOMOL, P., NOVOVIČOVÁ, J., GRIM, J., PUDIL, P., 2008: Dynamic Oscillating Search Algorithm for Feature Selection, *ICPR 2008 Proceedings* (Int. Conf. on Pattern Recognition 2008, Tampa, Florida, US, 2008).

ŠIŠKA, L., 2007: Analýza finanční výkonnosti respondentů empirického šetření CVKS. *Working Paper 10/2008*. Brno: Masarykova univerzita, Centrum výzkumu konkurenční schopnosti české ekonomiky, 2007, č. 10, 34 s. ISSN 1801–4496.

ŠPALEK, J., ČÁSTEK, O., 2010: Přínos učících se metod statistického rozpoznávání obrazů při hledání faktorů konkurenceschopnosti českých podniků. *Ekonomický časopis/Journal of Economics*, 58, 9, s. 922–937.

ŠPALEK, J., 2008: Metodický postup definování proměnných. In: BLAŽEK, L. *Konkurenční schopnost podniků (Analýza faktorů hospodářské úspěšnosti)*. 1. ed. Brno: Masaryk University, 2008. p. 31–40.

THEODORIDIS, S., KOUTROUMBAS, K., 2006: *Pattern Recognition*. 3rd edition. USA: Academic Press, 2006. ISBN 0126858756.

Address

prof. Ing. Ladislav Blažek, CSc., Katedra podnikového hospodářství, Ekonomicko-správní fakulta, Masarykova univerzita, Lipová 41a, 602 00 Brno, Česká republika, prof. Ing. Pavel Pudil, DrSc., Společná laboratoř SALOME, Fakulta managementu, Vysoká škola ekonomická, Jarošovská 1117/II, 377 01 Jindřichův Hradec, Česká republika, doc. Ing. Jiří Špalek, Ph.D., Katedra veřejné ekonomie, Ekonomicko-správní fakulta, Masarykova univerzita, Lipová 41a, 602 00 Brno, Česká republika, e-mail: blazek@econ.muni.cz, pudil@fm.vse.cz, spalek@econ.muni.cz