Faculty of Management
University of Economics, Prague
&
Faculty of Economics and
Administration, Masaryk University

Identifying Corporate Performance Factors Based on Feature Selection in Statistical Pattern Recognition

METHODS, APPLICATION, INTERPRETATION

Pavel Pudil, Ladislav Blažek,
Ondřej Částek, Petr Somol, Jana Pokorná,
Maria Králová

2014

Identifying Corporate Performance Factors Based on Feature Selection in Statistical Pattern Recognition

METHODS, APPLICATION, INTERPRETATION

VŠE muni PRESS
UNIVERSITAS MASARYKIANA BRUNENSIS • FACULTAS RERUM OECONOMICARUM •

Identifying Corporate Performance Factors Based on Feature Selection in Statistical Pattern Recognition
METHODS, APPLICATION, INTERPRETATION

Pavel Pudil, Ladislav Blažek, Ondřej Částek,
Petr Somol, Jana Pokorná, Maria Králová

FACULTY OF ECONOMICS AND ADMINISTRATION
MASARYK UNIVERSITY
BRNO
2014

Team of authors:
prof. Ing. Pavel Pudil, DrSc. (3, 4)
prof. Ing. Ladislav Blažek, CSc. (1, 6)
Ing. Ondřej Částek, Ph.D. (1, 2, 5)
RNDr. Petr Somol, Ph.D. (3, 4)
Ing. Jana Pokorná (2)
Mgr. Maria Králová, Ph.D. (5)

Translation by Mgr. et Mgr. Bc. Milan Boháček, M.A.
Corrections by Steven Del Riley, BA
Cover designed by Petr Somol

Reviewers:
prof. Ing. Michal Haindl, DrSc.
doc. Ing. Robert Zich, Ph.D.

This work has been supported by the Grant Agency of the Czech Republic – project No. P403/12/1557

ISBN 978-80-210-7672-3 (online : pdf)
ISBN 978-80-210-7557-3 (print)
DOI: 10.5817/CZ.MUNI.M210-7557-2014

Contents

Introduction

The issue of finding competitiveness factors of companies represents a very topical and interesting subject in terms of both practice and theory, with many publications dealing with this issue. However, these works do not present a uniform way of thinking, as they vary in their approach to the issue, terminology used, applications of the methods, solution justifications, and even in the credibility of the results achieved and their interpretation. This raises such questions as (1) how to express corporate competitiveness, (2) how to measure the effect of potential factors of competitiveness in a multidimensional space under the conditions typical for business research, and finally, (3) what are the competitiveness factors and how can we interpret their effects?

This publication presents the results of research focusing precisely on this area. It deals with the issue of applying selected statistical methods to identify the competitiveness factors of companies so as to respect the synergistic effect of their influence.

The theoretical and methodological part of this monograph describes the approach to the way of expressing the financial performance of companies, and especially approaches to the identification of factors that affect the financial performance. Experiments with an application of these approaches are described and analysed using empirical data of companies from manufacturing and construction industries, including interpretation of the results achieved.

The research was conducted in the period of 2012-2014 under a project of the Grant Agency of the Czech Republic entitled, "Developing Methods for Identifying and Evaluating Factors That Critically Affect Corporate Performance"[1] as a joint research project of the University of Economics in Prague, Faculty of Management in Jindřichův Hradec, and Masaryk University, Faculty of Economics and Administration in Brno.

---

[1] This work has been supported by the Grant Agency of the Czech Republic – project No. P403/12/1557.

1 Formulation of objectives and methodological approach

The presented work deals with analysing and formulating methodology on the search for competitiveness factors of companies, including its verification and interpretation on an extensive sample of empirical data.

This work builds on several years of research in the given field, conducted by the research team of the Centre for the Competitiveness of the Czech Economy established at the Faculty of Economics and Administration at Masaryk University. The results of this work are summarized in the monographs by Blažek et al. (2007), Blažek et al. (2008) and Blažek et al. (2009). In order to better understand the focus and purpose of the presented work, we consider it necessary to briefly summarize these research activities.

1.1 Summary of previous research activities

The main focus of the research was on the competitiveness of companies, namely the search for the factors that affect it. In other words: finding the reasons why some businesses are competitive and others are not. In addition to this, it found certain types of businesses that were competitive, and conversely other types of businesses that in turn are not.

The researchers pursued the idea that these factors can be identified on the basis of a sophisticated analysis of empirical data and that it is also possible to identify typical configurations of values of these factors which impact the level of competitiveness that individual businesses achieve. In other words: the same or similar level of competitiveness can be achieved in different ways. Two companies reaching the same level of competitiveness may differ from each other very significantly.

When elaborating this issue, it was of course necessary to proceed to the operationalization of the key concepts. This concerned mainly the concept of "competitiveness", which proved to be too vague, elusive and difficult to measure for the sake of further analysis. Therefore, it was replaced with the term "economic success".

The population was defined as follows:

- companies based in the Czech Republic (territorial aspect),
- companies in all industries (industry aspect),
- companies with 50 or more employees (size aspect),
- joint-stock companies and private limited companies (aspect of a legal form).

The empirical survey was conducted in two stages. The subject of the first one included those industries that represent key areas of the business sector of the Czech economy, i.e. the manufacturing and construction industries. In the second stage, other sectors were analysed. Since the number of companies that fit the definition above is too low for a statistical analysis in some of these industries, the size limit was reduced to 10 workers in the second stage.

Population specified this way included about 8,000 companies at the time of the survey.

There were two primary sources of information for solving the task in question:

- data from the questionnaires, capturing the selected characteristics of companies,
- data from public databases (Albertina Data), capturing the selected financial indicators.

Those companies that were in bankruptcy or liquidation at the time of the survey, as well as those businesses whose data were missing in the database, were excluded from the population.

The population that was subsequently used included 4,915 companies. This set was considered the population for the purpose of the research.

However, it was obviously impossible to subject such a large set to a questionnaire survey. There were two main reasons:

- the over-whelming effort,
- the unwillingness of most businesses to participate in the questionnaire survey.

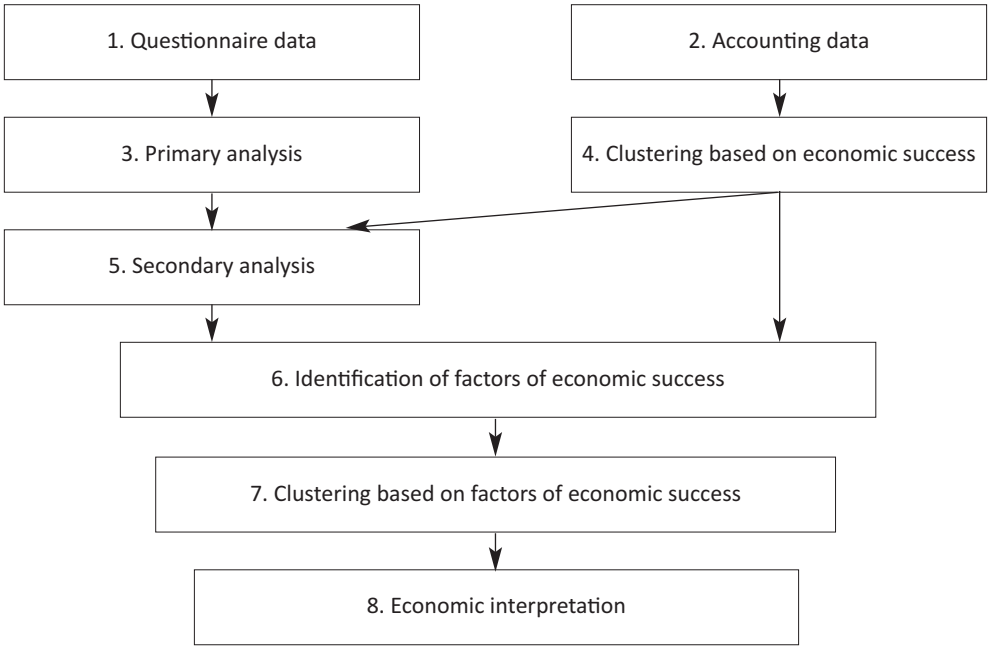
Companies from the population were asked to participate in the questionnaire survey in such a way that self-selection determined by the willingness or unwillingness of the surveyed companies could be restricted or eliminated by a quota selection. In accordance with the definition of the population, the quota variables included territory (regions), industry, size, and legal form of business. The percentage representation in the population and the sample showed very close values on average. This indicates that the sample can be considered representative. The questionnaire survey was conducted in collaboration with Augur Consulting.

The methodology of the task in question is shown in Figure 1-1.

1. Questionnaire data

As previously mentioned, the questionnaire survey was conducted in two stages. In the first sub-stage (2007), which applied to companies from the manufacturing and construction industries, 432 companies participated in the empirical survey. In the second sub-stage (2009), which applied to companies from other industries, the survey included 267 companies.

Figure 1-1: The methodology of previous research activities



2. Accounting data

In the case of the first sub-stage, data from the financial statements for the years 2002 to 2006 were used; in the second sub-stage, data from the financial statements for the years 2002 to 2007 were used.

3. Primary analysis

The primary analysis involved evaluating frequencies of responses to the questions of the questionnaire; the questionnaires produced almost 700 variables. The results were presented in the form of tables and graphs, including comments briefly interpreting the presented numerical and graphical information. The evaluation was performed for the sample as a whole as well as for the sub-samples, categorized by industry, company size, and legal form of business. A number of important pieces of information concerning the businesses were obtained in this way (Blažek et al., 2007, p. 41–287; Blažek et al., 2009, p. 29–262). However, the main output of the primary analysis was the processing of the variable set, entering the secondary analysis.

4. Clustering based on economic success

Economic success was evaluated on the basis of two indicators, i.e. return on assets and asset growth indicators. In the first sub-stage, the sample of 432 companies

was structured based on the application of a cluster analysis into three groups (A, B and C), from the most efficient to the least efficient (Blažek et al 2008, p. 53–66). In the second sub-stage, this methodology was largely modified. The sample of 267 companies was also structured into three groups (A, B and C); however, in this case the basis was not a cluster analysis but the so-called "coefficient of economic success" (Blažek et al., 2009, p. 53–66).

5. Secondary analysis

The purpose of this analysis was to transform the set of variables from the primary analysis into a set of variables as potential factors of economic success, suitable as an input into the methods of multidimensional statistical pattern recognition. Using the application of common statistical methods, the relationship between different variables and economic success of companies was analysed. There was a significant reduction in the number of variables; variables with a low variation and variables with a high proportion of missing values were excluded. In many cases, it was necessary to do a recoding into new variables, suitable for the subsequent mathematical and statistical procedures and interpretation of the obtained results (Blažek et al., 2008, p. 31–39 and p. 67–88, Blažek et al., 2009, p. 275–277).

6. Identification of factors of economic success

In order to ensure the exact mastering of this statistically demanding task, selected methods of statistical pattern recognition were used for the multidimensional analysis. Specifically, these methods included Individual Best Search (IBS), Sequential Floating Search (SFS) and Sequential Floating Forward Search (SFFS). The outcome of the application was to identify which variables tested by these methods are the factors that influence decisively the economic success of the companies analysed; moreover, the outcome did not seek to identify the factors that influence the economic activity separately, i.e. each factor individually, but integrally, i.e. interrelated (Blažek et al., 2008, p. 89–93, Blažek et al., p. 275–280).

7. Clustering based on factors of economic success

After identifying the factors of economic success, it was necessary to find such typical configurations of values of these factors that are produced by certain types of economic success achieved by companies. Clusters based on factors of economic success were created using the application of the cluster analysis method within each of the groups A, B and C, created on the basis of the economic success of the businesses. (Blažek et al., 2008, p. 93 and 94).

8. Economic interpretation

Groups of companies, established on the basis of such clustering, were characterized with values of economic-success factors, and supplemented with several other variables with regard to increasing the rate of illustrativeness. First, the

interpretation of characteristic features at the level of the entire sample was made, using the comparison of factor values and selected variables between the groups of companies A, B, and C, grouped according to their economic success. Subsequently, a similar interpretation was conducted at a more detailed level – within the group of the economically most successful companies (group A); it included the comparison of values of the same factors and variables as in the previous case, but between groups of companies created by secondary clustering, i.e. by factors of economic success (Blažek et al., 2008, p. 95–120).

After some time it can be stated that the research described above represents a challenging and somewhat unique event, both in terms of the methodological approach and the extent of empirical investigation. Despite the obvious benefits, however, many of the problems that occurred during the solution and the complexity of some of the results achieved remained unresolved.

This primarily concerns the methodological level. When searching for the best way to apply approaches based on pattern recognition when identifying the factors of economic success, a number of experiments were performed. Although some useful experience was obtained thanks to these experiments, the results obtained did not lead to unambiguous conclusions. The researchers failed to overcome the negative phenomenon of excessive sensitivity of outputs to relatively small changes in input variables as well as a relatively strong influence of the choice of a particular evaluation algorithm and its parameters on the obtained results.

The quality of the results, both in the primary and secondary analyses and in identifying the factors of economic success of businesses, was also undoubtedly limited by the character of the industries in question. In the first stage, the investigation focused on manufacturing and construction industries, i.e. industries producing material goods, industries with relatively strong internal homogeneity, exclusive entrepreneurial orientation, and consequently with many companies whose legal form was private limited company or joint-stock company. However, in the case of the second stage, the situation was quite different: it was basically a conglomerate of nine industries, many of which showed diametric differences with respect to one another. The main factor that fundamentally undermined the homogeneity of the sample was the fact that only some of these industries can be considered entrepreneurial, while in others entrepreneurship is practically non-existent. This is closely connected with the representation of private limited companies and joint-stock companies. The representation of the respondents in each industry was therefore very uneven. Three of the nine industries analysed included fewer than 10 respondents. The effort for the widest possible coverage unfortunately resulted in comparing, or evaluating jointly, industries that are difficult to compare by nature, or even incomparable.

1.2 Methodology of current research

These facts became a challenge for further research, which was conducted within the project of the Grant Agency of the Czech Republic entitled "Approaches to identification of corporate performance factors with emphasis on methods of feature selection in statistical pattern recognition." The outcomes of this research are presented in this publication.

The objective of the research conducted within this project is formulated as follows:

To develop and verify a methodology for identifying and evaluating factors which have a significant effect on the performance of companies.

The solution involved these successive steps:

a) Search of relevant literature, summary of theoretical knowledge.
   - Approaches to expressing competitiveness, financial performance measurement methods.
   - Methods of feature selection in statistical pattern recognition.

b) Testing on the task in question
   - Testing methods for factor identification. Selecting the most appropriate method, improving and adjusting it. Factor identification.
   - Regression analysis: explaining financial performance on the basis of identified factors.

c) Economic interpretation and overall evaluation
   - Identification and economic interpretation of typical combinations of factor values leading to certain types of financial performance.

Step a) formulates theoretical grounds and creates a methodological background for the solution of the given task. It is contained in chapter 2. Competitiveness and its measurement, and in chapter 3. Feature Selection Methods in Statistical Pattern Recognition.

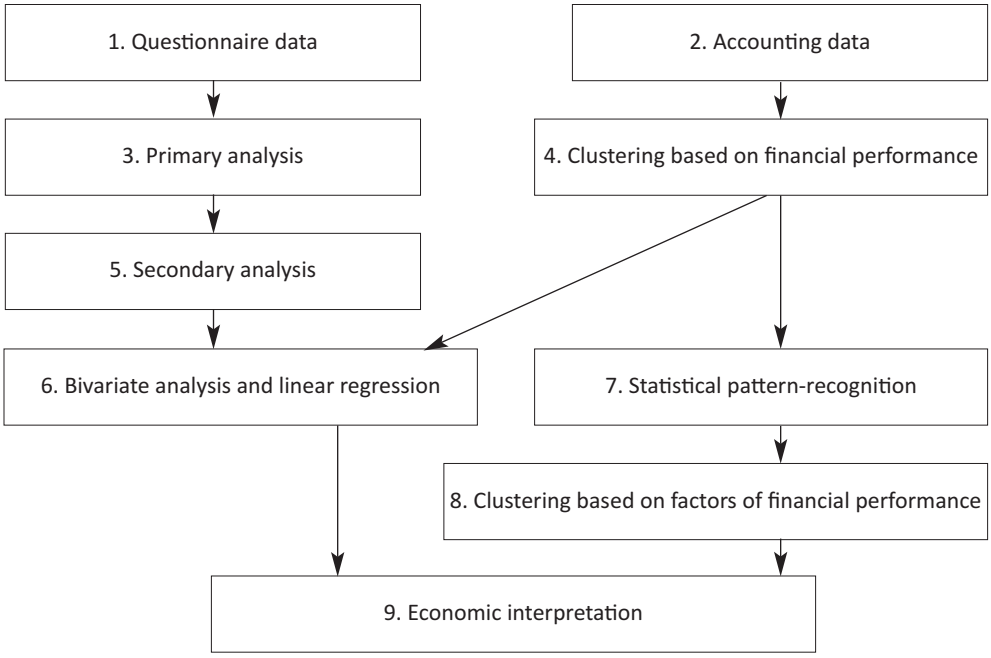
Steps b) and c) concerning testing and interpretation were conducted following step a). These research activities and their results are contained in chapters 4. Testing approaches and methods based on learning methods for identifying factors of competitiveness, 5. Identifying factors of competitiveness using bivariate analyses and linear regression analyses, and 6. Interpretation of the results achieved.

We will use the comments to Figure 1-2. as the first introduction to the solution of the task in question including its relation to previous research activities.

1. Questionnaire data

With respect to the above-mentioned facts, the research used data from the questionnaire survey conducted in 2007, which applied to companies from the manufacturing and construction industries. Within these two major industries, 432 companies participated in the empirical survey. Businesses from other industries were not included.

Figure 1-2: Methodology of current research activities



2. Accounting data

The data used came from the financial statements for the years 2004 to 2010. The relationship between cause and effect was substantially respected in this way, where the cause refers to the company characteristics identified through a questionnaire survey, and the result is the development of a company's financial indicators. Another important reason for choosing this period was the effort to capture the development of these indicators during the crisis when the competitiveness of companies was tested in very difficult conditions.

3. Primary analysis

The starting point was the primary analysis from the previous research. With regard to the application of subsequent procedures and methods, the available data were further modified.

4. Clustering based on financial performance

The term "economic success" was replaced with the term "financial performance" with regard to better operationalization. The indicators of return on assets and asset growth were used again, but the method of their calculation was changed substantially (see Chapter 2).

5. Secondary analysis

Similarly to the primary analysis, the secondary analysis was also based on the previous research. However, the reduction of the number of variables entering the experiments of statistical pattern recognition did not have to be so dramatic because the newly applied DAF algorithm is not so sensitive to the ratio of the number of input variables to the number of available observations (i.e. companies). Therefore, the number of variables from the questionnaire was reduced on the basis of value variability and the number of missing values to 74 variables instead of 37 variables used in the SFFS algorithm.

6. Bivariate analysis and linear regression

Usual methods of bivariate analysis, i.e. t-tests, ANOVA and correlation and association analyses were used in parallel with the experiments of statistical pattern recognition. In addition, partial linear regression models were also formulated. These analyses provided a number of findings on the relationships between the aforementioned 74 variables and financial performance (see Chapter 5). When comparing the power of the relationship found by these procedures and the error of estimation of a nonlinear model based on the variables selected by the DAF method, the advantage of using the techniques of statistical pattern recognition is obvious. This section thus serves as a direct comparison of the results of both approaches as well as assistance for economic interpretation (see the final step).

7. Statistical pattern-recognition

Unlike in the previous research, a different approach of multidimensional analysis was applied in this fundamental research activity. Specifically, it was the DAF algorithm, which shows a much higher stability of results for the given task than the SFFS algorithm which was originally used. Moreover, the non-linear regression model was employed as a better classifier than the original k-NN classifier; the DAF algorithm was also improved in several aspects, which resulted in an increase in the accuracy of the result.

8. Clustering based on factors of financial performance

The procedure was similar to the previous research; however, the difference consisted in clustering that was conducted using different factors of financial performance and within different groups of companies, created based on their financial performance (see Chapter 6).

9. Economic interpretation

The researchers again proceeded similarly as in the previous research, but with different variables and a different grouping of companies according to their financial performance. This time, the companies were divided into quartiles. They were labelled as groups A, B, C and D, where group A included businesses with the highest financial

performance and group D businesses with the lowest financial performance. First, an interpretation was performed at the level of the entire selected sample, using the comparison of factor values and selected variables between the A, B, C and D company groups. Next, a similar interpretation was performed at a more detailed level – within the group financially most powerful companies, in order to interpret the mainly quantitative representation into the representation of the mainly qualitative nature (see Chapter 6).

This book also summarizes and builds on some already published outputs of this project, particularly the following ones. In the work-in-progress paper, Pudil et al. (2012) compared the performance and stability of algorithms SFFS and DAF combined with the k-NN classifier. The result was that DAF algorithm achieves a slightly higher accuracy but significantly higher stability. Částek et al. (2013) proved the necessity of using advanced multivariate methods by comparing different approaches on the same data set and paper. Pokorná and Částek (2013) reviewed methods of corporate performance measurement and experimentally verified a chosen way, longitudinal combination of ROA and Growth of Assets. Pudil et al. (2013) finished the comparison of SFFS and DAF; this time these algorithms were combined with a non-linear regression instead of the k-NN classifier. Again, this approach improved both the accuracy and stability of the results and again the DAF algorithm was found higher performing in both criterions. The final step was taken in the paper by Somol et al. (2014). Here, the distance function and kernel width used were optimized and accuracy has risen once again with no loss of stability.

The use of DAF algorithm together with multivariate regression model for predicting financial performance measured by adding up ROA and Growth of Assets became the basis for selecting the factors of corporate performance. Besides that, the reader will also find a procedure for interpreting such factors in this book, including the actual interpretation of the effect of these factors in Czech construction and manufacturing companies.

2 Competitiveness and its measurement

The term competitiveness is widely used but not a uniformly understood one. Authors of various studies often create their own definitions, which imply different approaches to measuring competitiveness. Therefore, in this chapter we focus on the term competitiveness, explain the possibilities of its measurement, and finally, justify the choice of the methodology used to measure competitiveness in this research.

2.1 The term competitiveness

Competitiveness can be viewed from a macroeconomic or microeconomic perspective. The macroeconomic view deals with competitiveness of countries and their ability to produce products and services (through businesses) that are successful in international markets. The European Union understands competitiveness as the ability of an economy to increase productivity, as the only way to achieve sustainable growth in per capita income, which subsequently leads to the growth of living standards (European Commission, 2011). The World Economic Forum puts greater emphasis on the economy as a set of institutions and policies in understanding competitiveness (Schwab, 2012), through which the level of productivity of enterprises and thus the entire country is influenced.

The microeconomic perspective of competitiveness, which we follow in this research, concerns the competitiveness of smaller units, namely businesses. The Organization for Economic Co-operation and Development (OECD) defines competitiveness of a business as its ability to compete, increase profits and growth. Approaches of other authors (Jirásek, 2000; Blažek, 2009) can be summarized in a widely accepted definition of competitiveness of a business as the ability of a business to operate on the market in the long term and sustainably in competition with other businesses. This view of competitiveness was used by the Research Centre for the Competitiveness of the Czech Economy (Blažek et al., 2009), which this research partly follows. Factors leading to a higher ability of a business to stay in the market are the subject of this research.

2.2 Approaches to measuring competitiveness

Higher competitiveness of a company is reflected in the long-term growth of its profits or increasing its market share and the consequent growth of the company as such. However, there is no single correct approach to measuring competitiveness, as it is always necessary to particularly consider the purpose and possibilities of this measurement. The research is therefore based on the assumption that in the long run, a company's competitiveness affects the success or failure of a company in competition (Blažek et al., 2009). From this perspective, the success of a company can be identified with its performance for the purpose of its measuring. Thus, business performance is conditioned by its competitiveness and it should be true that if a company is competitive, it is also efficient (Suchánek, 2005). As a result, we transformed the problem of identifying factors of competitiveness of companies into the form of corporate performance factor identification.

Business performance in general represents the degree of achieving the set objectives. Based on an organization's objectives, we identify financial performance, operational performance and overall effectiveness (Hult et al., 2008).

The widest approach is undoubtedly measuring a company's overall effectiveness. This includes an analysis of achieving the long-term strategy of a company in all areas of its activity. However, measuring it is very problematic and often made more difficult by considerable subjectivity, as it includes variables such as reputation, perceived performance, achievement of goals, or survival. Analysing the overall effectiveness of a company requires extensive knowledge about the organization and its field, and it can be applied in qualitative research.

Operational performance focuses on the non-financial dimension of performance, such as measuring the quality of products and processes in a company, the cycle time and productivity. The results are especially useful for business management in supporting the improvement of business processes or even benchmarking.

The prevailing approach to performance measurement in quantitative research is the design of financial performance indicators (Hult et al., 2008), measuring the economic situation of a company. Financial analysis indicators based on accounting data and information from financial statements are widely used. Indicators of a company's market position offer an additional view. The advantage of this approach is especially the easy accessibility of financial statements that are publicly accessible in the Register of Companies or other databases. Indicators use "hard" data from financial statements; therefore, they are not affected by the evaluator's subjective view. Based on the purpose of the research, a historical view of a company, as well as possible oversimplification or disregarding industry specificities can be seen as a disadvantage. A possible disadvantage of long-term studies is the risk of errors arising from changes in the accounting procedures of the company.

The quantitative nature of this research with a large number of monitored companies allows us to measure performance solely from the financial perspective. Understanding the specific circumstances of each company would not be practically feasible, whereas data from the financial statements are available from databases such as Albertina CZ. Due to the possibility of comparing the results of this research with other studies and also for practical reasons, we decided to measure business performance using financial indicators.

Financial performance

The most commonly used indicators are the indicators of financial analysis, which can be obtained from the financial statements of a company. This includes indicators of profitability, frequently used by researchers, as well as indicators of debt, liquidity and activity. Their popularity is due to the good availability of data, simplicity of their design, easy interpretability and wide applicability. Ratio indicators allow us to compare companies of different sizes and from different industries, or to compare the results of a company with the recommended values. Another advantage is that the data for calculating these indicators are subject to internal and external checks and they are not influenced by "market sentiment" (Krivogorsky & Grudnitski, 2010, p. 182). On the other hand, they may be negatively affected by differences in accounting procedures (concerning the comparison between countries) or even manipulations such as undervaluing assets (Sanchez-Ballesta & García-Meca, 2007).

Another group is represented by market measures. Shareholder-value measures are based on the perception of a company as a tool whose purpose is to increase the value of the owners' investment. These indicators are based on the value of a company's shares on the market. This, however, may be affected even by external factors besides business performance. These indicators are not useful for quantitative research in the Czech Republic since there are only a few dozen listed companies on the Stock Exchange. The category of market measures also includes the indicators of market share, which, however, are not commonly accessible.

Hybrid indicators try to capture business performance from many perspectives, combining multiple indicators from both the previous groups. They include solvency and bankruptcy models that predict the financial health of a company in the future. Another well-known indicator of this field is Economic Value Added. Hybrid indicators often use estimates and work with interest rates of government bonds. Their design is often complex, making them more difficult to interpret.

The most common indicators used in the conducted economic studies include variously designed profitability (most frequently Return on Assets) and indicators based on revenues (Hult et al., 2008).

2.3 Financial performance indicators used

We used the assumption that higher business performance is in the long run reflected in its profitability or growth in market share and the consequent growth of the company as such as the investigation by the Research Centre for the Competitiveness of the Czech Economy used when choosing the appropriate indicator to measure financial performance. The experienced authors of the most successful performance measurement system worldwide, the Balanced Scorecard method, comment: "*Financial objectives typically relate to profitability – measured, for example, by operating income and return on investment. Basically, financial strategies are simple; companies can make more money by (1) selling more, and (2) spending less. Everything else is background music. ...Thus, the company's financial performance gets improved through two basic approaches – revenue growth and productivity.*" (Kaplan and Norton, 2004, p. 36).

The growth strategy usually consists of entering new markets and developing new products that require the introduction of additional production and service processes with which the company has not had substantial experience yet. An effort to standardize and streamline existing processes is the essence of the strategy to increase productivity, which is applied to activities usually performed for a longer period, in a familiar environment, and for well-known customers. Only excellent companies are able to implement both strategies at the same time, and thus achieve exceptionally high financial performance (Blažek et al., 2011).

If we want to capture both possible strategies for achieving business performance, i.e. the growth strategy and the strategy of increasing productivity, it is necessary to use at least two indicators. However, a larger number of indicators used increases the time and cost of obtaining and processing data, and the amount of missing data increases as well. In addition, the combination of all indicators becomes more complex and more difficult to interpret. Therefore, it was decided to use two indicators, each of which will represent one of the two mentioned strategies.

We decided to measure the strategy of increasing productivity, or profitability, using the indicator Return on Assets (ROA), which expresses the profitability of all resources in the company, regardless of their origin. This approach is in line with the stakeholder view of the company, used in the investigation of potential factors of performance; moreover, its result is not influenced by the type of capital structure of the company. ROA is one of the essential and most frequently used indicators of profitability (Richard et al., 2009). It was designed as follows:

$$ROA_t = \frac{NOPBT_t}{TA_t} \times 100$$

where:
NOPBT – net operating profit before taxes
TA – total assets
t – year

The calculation uses operating profit, which is not affected by supplementary activities of the company and expresses its success on the market. Relating this value to the amount of assets gives the indicator an essential relative form, which allows the comparison of different-sized businesses.

To capture the degree of implementing the growth strategy, we used the Assets Growth (AG) indicator. The rising value of the total assets and including a greater amount of resources at the same time is caused by the higher production of the company. During long-term growth in demand for its products, the company has to invest in expanding its production capacity, which increases the overall amount of assets. The second option is to select directly the indicator of sales growth, but it was not possible to calculate this indicator clearly from the database used (due to different procedures used in the database to determine the indicator of total sales, consisting in confusing total sales with revenues from the sale of goods, or products and services, total revenues or the performance indicator). The Assets Growth indicator was designed as follows:

$$Assets\_Growth_t = \left(\frac{TA_t}{TA_{t-1}} - 1 \right) \times 100$$

TA – total assets
t – year

2.4 Period of performance measurement

We understand competitiveness as the ability of a company to operate on the market in the long term, sustainably in competition with other businesses. For this reason, it was necessary to measure business performance for a longer period of time. Using a time series also helps to reduce random fluctuations as well as intentional optimizations (such as creating reserves, etc.) in the performance of individual companies. From the perspective of corporate economy, a long period is understood to cover a period longer than 3 or 5 years. Even Krištof (2006) recommends following a five-year and longer series of indicators, with Kirby (2005) requiring an even longer period of about ten years. This requirement is based on the environment of North America, where a decade covers two terms of an executive director of a company on average. However, such research time may be too long in the current turbulent times. Factors applicable before this period may no longer be valid with such a considerable time interval. This is probably the reason why there are not many studies using a longer period than five years (e.g. Artiach (2010) – a five-year period, Abor (2007) – a six-year period, Hansen (1989) – a five-year period). On the other hand, many studies follow performance on the basis of only one-year results, e.g. Andrews (2010), Bottazzi (2008), Kessler (2007) and many others.

Another fundamental fact of our research aimed at finding competitiveness factors of a company is the existence of a time lag between the potential causes leading to increased performance and its actual rise. For this reason, it is first necessary to determine the possible factors of high performance and measure this performance subsequently. This ideally means determining the values of competitiveness factors in year 0 and then measure performance in years 1 to 10; moreover, we should allow for a delay in the publication of financial statements of the company. Nevertheless, it is not possible to postpone the analysis for so long due to the above-mentioned utilization of research findings in the current turbulent environment.

Thus, it was necessary to examine a sufficiently long period of time, but at the same time not to delay the analysis too long after data collection. The questionnaire survey, which investigated the potential performance factors, was conducted in 2007. To analyse the performance, compromise was made: the time series of seven consecutive years for the period 2004–2010 was examined. The data for calculating financial performance for later years were not available for a sufficient number of companies at the time of the analysis, as entering financial data into the Albertina database and the Register of Companies is conducted gradually and with delay. Seven years is a sufficiently long period of time for inferring the long-term business performance. Even older data obtained from the extension of the time series backwards were no longer desirable due to a too long interval from the questionnaire survey.

2.5 The development of performance measurement methodology

Selecting two different indicators of financial performance led to the question of whether to analyse a company's performance separately by ROA and separately by the Assets Growth, or whether to try to aggregate both the results into a single summary benchmark. The first approach is more common as it is used by approximately two thirds of economic studies, while aggregation of multiple indicators into one is used by approximately one third of studies (Richard et al., 2009). Given the purpose of the research, it was desirable for further analysis to group or sort the businesses within a single variable based on their financial performance. Only one performance indicator was required by the intended use of some special statistical methods. For this reason, we decided to merge the two indicators into a single one.

Great attention was paid to finding the most appropriate way of analysing financial performance from the beginning of the previous research conducted by the Research Centre for the Competitiveness of the Czech Economy because it significantly determines the results of subsequent experiments. During the research, more possible solutions to the problem outlined were found and the model was revised and adjusted several times so that the selected indicator would reflect the real financial

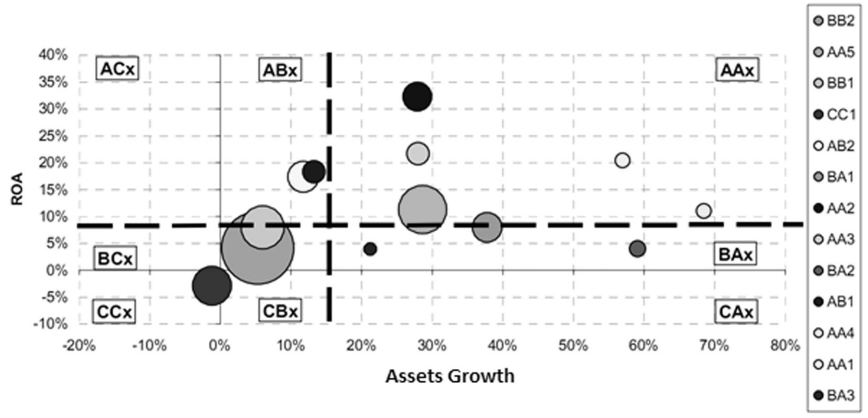
performance of the company as well as possible. Previous methods of ROA and Assets Growth aggregation into a single indicator are described in detail in the individual monographs summarizing the research conducted by the Research Centre for the Competitiveness of the Czech Economy (Blažek et al., 2008, 2009, 2011), and their comparisons are included in Pokorná and Částek (2013). Therefore, they will be only briefly described here.

2.5.1 Cluster analysis

In the first phase of finding methods for analysing business performance, Šiška (2008) first had to deal with the fact that finding the causes for higher business performance in the questionnaire survey, which was conducted in 2007, followed the time period for which it was possible to analyse the financial performance, i.e. 2002-2006. Although we can assume certain sluggishness in established business processes, it was appropriate to take into account the temporal distance of financial data in analysing business performance. Therefore, in order to reinforce the importance of financial data from the years closer to the empirical survey, he applied the weight of 1-2-3-4-5 to both analysed indicators from 2002–2006. Thus, the weight was greatest for the latest data from 2006 and smallest for indicators from 2002. Besides weighing the data by years, they were also standardized using the z-score to eliminate the effects of different scales of both indicators.

The first selected tool for analysing the financial performance was cluster analysis. Using 27 iterations of non-hierarchical k-means clustering, Šiška revealed 13 relatively homogeneous clusters associating a different number of businesses with similar values and development of ROA and Assets Growth over the five-year period analysed (graph 2-1).

Graph 2-1: Clusters of cluster analysis



Source: Šiška, 2008

Distribution of companies in this number of groups made it considerably difficult to interpret the results and represented a problematic use in the subsequent statistical analysis. The clusters formed a categorical variable, which prevented a further statistical analysis using procedures requiring at least ordinal or cardinal variables. Therefore, this method was modified to cluster grouping based on the quadrant in which the clusters were located. This led to the creation of AA, AB, BA, BB and C groups (containing all clusters in areas marked with the letter C). Five categories of this variable proved more useful, but they still were not ordinal. With the need of at least an ordinal expression of financial performance, we used only businesses located in areas AA, BB and C. This made the dependent variable become a higher level variable (from categorical to ordinal); but it meant that the experiments could not include all businesses – businesses in the areas AB and BA were omitted. For these reasons, this method was not used further.

2.5.2 Hyperbola

In the next stage we were looking for a method to evaluate business performance, which would make it possible to put businesses into a small number of cardinal groups for the sake of interpretation – ideally only into two groups: companies economically successful and economically unsuccessful.

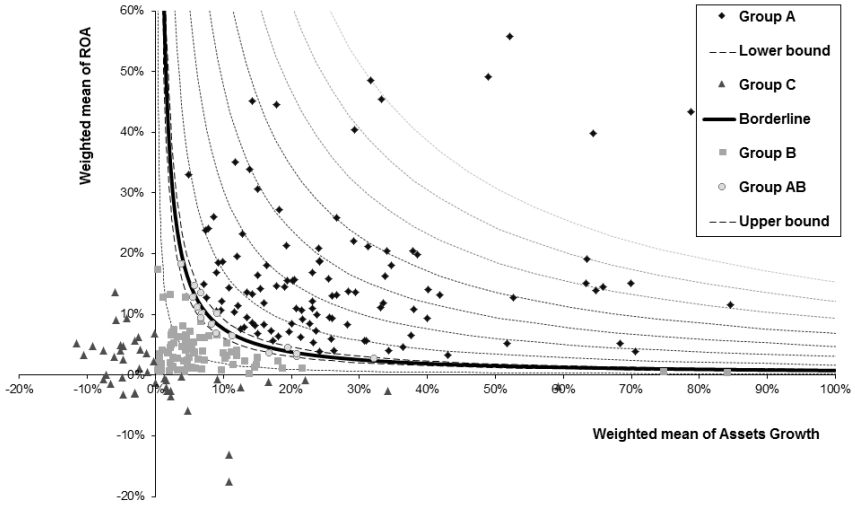
When designing new methodology, Šiška assumed that the growth strategy and strategy of increasing productivity were mutually exclusive (Kaplan, Norton, 2004). This assumption was expressed in this methodology as an inverse proportion between the indicators used of ROA and Assets Growth. The product of the values of the two indicators represented a single-digit indicator of financial performance of a company, referred to as the coefficient of financial performance. When the value of one or two input variables was negative, i.e. a company was becoming smaller and/or depreciating its assets, the company was rated as economically unsuccessful regardless of the outcome of the mathematical product of both indicators.

Ordering companies based on their performance was done using the coefficient of financial performance for the remaining businesses. A company can achieve the same financial performance for different strategies, e.g. if it weakens its focus on productivity and tries to get more customers instead, and expects that an increase in customers will lead to assets growth. We obtained the financial performance curve (graph 2-2) by connecting the points with the same values of the coefficient of financial performance but different combinations of ROA and Assets Growth values. Companies with a higher coefficient of financial performance are located on the higher financial performance curve.

Companies located above the curve, which connected businesses with the financial performance coefficient of a median value, were identified as financially efficient; companies under the curve were marked as less financially efficient. Companies with a negative value of one or both indicators of financial performance,

i.e. businesses from the second, third and fourth quadrant of graph 2-2, were marked as clearly inefficient. The name of the method – **Hyperbola without weights** – was derived from the geometric shape of the financial performance curves.

Graph 2-2: Hyperbola



Source: Blažek et al., 2009

Similarly to the previous method working with clusters, the indicator values used for calculating the financial performance coefficient were assigned weights reflecting the level of variable significance uniformly decreasing toward the past. In further research the authors stopped using the weights because their application caused large distortion (see experiments in the next chapter); thus, indicators for each year were assigned the same weight. This method was called **Hyperbola with weights**.

Based on previous experience, Šiška (Blažek et al., 2009) tried to take into account the variability vs. stability of the results of companies in the next stage during the period analysed. He assumed that 1) a company with long-term stable values of financial indicators can probably face competition better than a company with the same average performance, which, however, strongly varies over time, and 2) all stakeholders consider a stable company, i.e. a low-risk company with balanced indicators, more attractive. Therefore, he purged risk from the average values of both financial indicators using a standard deviation. The calculation based on the method called **Risk-purged hyperbola** for a six-year ROA average was as follows (and similarly for Assets Growth):

$$ROA_{2003-2008} = \frac{\overline{ROA_i}}{1 + \delta_i}$$

When searching for ways to refine the results obtained, the authors also considered and tested the possibility of rating business performance separately in each industry. The results of a method called **Sector-hyperbola**, however, failed to be more accurate than in previous methods (see section 2.6).

The method variants with a hyperbola meet the criteria of a cardinal variable and a small number of the final business groups based on their performance; however, the explanation of the correlation between the two indicators, which was expressed as their product in the calculation of the financial performance coefficient, appeared problematic. This method of calculation provided a substantial undesirable distortion of business analysis, favouring those companies that pursued both strategies simultaneously (e.g. an average company with ROA 5% times Assets Growth 5% = financial performance coefficient of 25) compared to businesses that focused only on one of them although they achieved high success in this one strategy (e.g. ROA 15% times Assets Growth 1% = financial performance coefficient of 15). A fundamental problem was also assessing companies with zero Assets Growth and high profitability, and vice versa, as inefficient due to multiplication of a large number by zero. For these reasons, the method was subsequently used only as a dichotomous variable that separated businesses to those that were found above the hyperbola (successful) and those that were found below the hyperbola (unsuccessful). To highlight the differences between these two categories, borderline businesses, located in close proximity to the hyperbola, were omitted for some experiments. Typically, these were 5% of companies above and 5% of companies below the hyperbola. In graph 2-2 these borderline businesses are identified as Group AB. This made it possible to overcome the aforementioned drawbacks; however, the variable could not be used as a cardinal variable.

2.5.3 Summation

In further search for an appropriate methodology, we took into account the shortcomings of the previous methods. The output was still expected to be a continuous variable or a discrete ordinal variable with fine resolution, which could subsequently be divided into a small number of company groups based on financial performance.

When working with two indicators it was again necessary to convert the values of ROA and Assets Growth to the same scale. We used the method of standardization, which preserves the information value of the resulting z-score compared to normalization.

In the design of a single-digit indicator, we put emphasis on the equivalence of the growth strategy and the strategy of increasing productivity, including any possible combination of these two strategies in achieving performance. Therefore, the standardized indicators of ROA and Assets Growth are of the same weight and they can be simply added up without the risk of distortion. This method eliminates the

aforesaid problems associated with the hyperbola method. The resulting financial performance coefficient is a one-dimensional variable, which can be graphically represented with a curve. In the following text, this method is referred to as **Summation.**

A variant of the summation method is the **Exponent** method, which we considered and in which we also tried to highlight the differences between companies using a power relationship between the variables:

$$X = \frac{a^{ROA} + a^{Assets\_Growth}}{2}$$

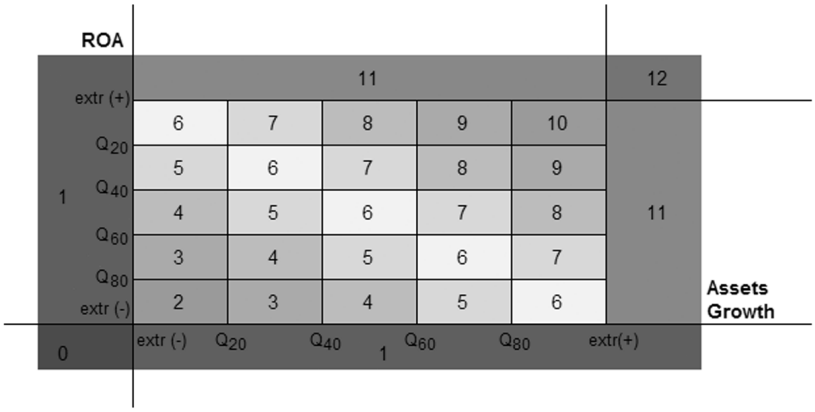
where a>1.

Variable a that influences the degree of variance of the financial performance coefficient was arbitrarily set as a=2 in further experiments.

2.5.4 Quintiles

Another method was designed by J. Šiška and M. Králová (2012). They operate with a graphical representation of indicators on the two main axes, similarly to the hyperbola or cluster method (Fig. 2-3). They divided the resulting space of the upper right quadrant in particular into several segments and assigned each segment points expressing a company's performance.

Figure 2-3: Quintiles method



Source: Pokorná, Částek, 2013, modified according to J. Šiška and Králová, 2012

We adjusted the proposal to our needs by changing the boundaries between the segments. The lowest score was assigned to companies that represent up to 5% of the sample for both indicators; on the other hand, the highest score was assigned to companies that belong in the top 5% in both indicators. We separated the remaining space between the two segments by quintiles. This method satisfies the ordinal variable requirement although its significant disadvantage is a subjective determination of the boundaries between the individual segments. Due to a large number of categories we can consider the variable quasi-interval.

2.6 Assessing the appropriateness of methods to measure financial performance

As seen in the previous section, there are several options in dealing with the selected indicators so that the result would be a one-dimensional variable. Unfortunately, from a theoretical point of view it is not possible to justify one of these options as better than others. Therefore, we decided to use experiments based on the following assumptions:

1. there is a statistically measurable correlation between the factors analysed of competitiveness and financial performance,
2. if the factors and the method used to measure correlation (association, tightness, determination) remain unchanged, then the value of the correlation coefficient for different representations of financial performance (association, tightness, determination) will represent the degree determining the appropriateness of the selected expression of financial performance.

Should the aforementioned be true, it is necessary to choose a suitable method of measuring correlations. It is important that such a method in particular:

1. is not affected by the shape of tightness (e.g. by linearity as correlation), because this shape can change if the expression of financial performance is changed,
2. does not assume an a priori model (e.g. structural modelling), which would need to be changed when expressing financial performance,
3. is able to measure the impact of several independent variables with possible interdependencies.

For the above reasons, it is therefore possible to exclude the majority of the commonly used methods from application: correlation, linear regression, structural modelling, decision trees, partially methods of multiple regression and multiple correlation. On the contrary, these shortcomings are not present in the selected algorithm of sequential floating forward selection (SFFS) in combination with the k-Nearest Neighbours classification method, presented in next chapters. It should be pointed out that the SFFS algorithm reduces the set of so-called symptoms; in our case, these are the variables describing businesses so as to preserve the most

informative symptoms, i.e. variables contributing most to separating companies into classes. We understand classes as categories of financial performance. This process therefore requires that the variable analysed is categorical, which was convenient in our situation as all types of financial performance evaluation proposed in the previous section could be converted to a categorical variable, while variables with a lower level of quantification could not be converted for example to a cardinal variable if required. For methods that produced cardinal variables, we separated the companies into two classes, i.e. above average and below average, while omitting about 10% of borderline businesses in order to distinguish better more competitive ones from less competitive ones.

2.6.1 Experiment settings

To evaluate the informativeness of the tested variable sets, we used the classification method of *k*-Nearest Neighbours, which groups samples (companies) into individual classes (categories of importance) by *k* of the nearest neighbours. The number of nearest neighbours *k* is therefore an optional parameter of the experiments. Since it is not possible to clearly decide what size of *k* should be used, the experiments were carried out separately with *k* set to 1, 3 and 5. Even values are not usually used in order to avoid undecided situations (Blažek et al., 2008, pp. 48–49). On the other hand, the number of companies in the individual categories of importance is too low for a value greater than 5. After each experiment, the authors chose the size of *k* that helped reach the most informative variable set as the most suitable. The same procedure was used in the search for competitiveness factors (Blažek et al, 2008, p. 89). In the following text, the individual *k*'s used will be referred to as methods 1NN, 3NN and 5NN.

Each company was described with 37 variables representing potential competitiveness factors, referred to as "symptoms" in the terminology of statistical pattern recognition. More information about working with these variables before the application of statistical methods and about the SFFS algorithm can be found in the article on the benefits of this algorithm in the search for competitiveness factors (Špalek, Částek, 2010).

2.6.2 Experiment output

The table below shows the descriptive statistics of informativeness values obtained using different methods of analysing financial performance. The methods are ranked on the basis of average informativeness values achieved. It is clear that the performances produced by the first three methods differ within three per thousand, while the fourth method lags behind the best one by almost two percentage points. Hence, the most accurate method seems to be summation, followed by hyperbola without weights and the exponent method closely trailing.

Figure 2-4: Descriptive statistics of informativeness values produced by different methods

Fin. perf. measurement method	Minimum	Maximum	Mean	Std. Deviation
Summation	0.739	0.782	0.765	0.022
Hyperbola without weights	0.747	0.787	0.762	0.022
Exponent	0.751	0.773	0.761	0.011
Quintiles	0.745	0.751	0.749	0.003
Hyperbola with weights	0.730	0.755	0.744	0.012
Sector hyperbola	0.709	0.725	0.717	0.008
Risk-purged hyperbola	0.701	0.715	0.706	0.008

However, besides informativeness, we can use other criteria to assess the efficiency of methods. One of them can be the rate of fluctuation for selected variables. In the next table, we can see how the SFFS algorithm output can be summarized. It lists the values of informativeness, the number of variables in the set of variables with the highest informativeness, and variable codes contained in this most informative set for classifiers 1NN, 3NN and 5NN.

Figure 2-5: Summary of the SFFS algorithm output for the exponent method

Classifier	Properties of the most informative variable subset		
	Informativeness	Number of variables	Variables
1NN	0.750685	9	**3** 7 *9* 14 **18** *31* **32 33** *34*
3NN	0.772603	13	**4 5 12** 14 **19 24 25 26 27** *31 34* **35 36**
5NN	0.758904	4	**8** *9* 14 **17**

Note: **bold** – variables selected once
italics – variables selected twice
underlined – variables selected always

Ideally, not only should the value of informativeness be high, but different classifiers should also agree on the same variables. However, in the case of the exponent method we see that all three methods selected only one variable, which is number 14, three variables were selected twice (9, 31, 34), but seventeen variables were selected only once. To compare it with the other methods, we can design the following analysis of variable fluctuations in the most informative set:

$$1 - \frac{Number\ of\ different\ variables}{(\frac{Number\ of\ items}{3})}$$

where number of items is the sum of the selected variables; so for the above mentioned case the number of different variables is 1 (number 14) + 3 (numbers 9, 31 and 34) + 17 (other variables) divided by the Number of items, i.e. 9 + 13 + 4, and divided by three (because every time the experiment was conducted for 1NN, 3NN and 5NN). The formula would look as follows:

$$1 - \frac{21}{(\frac{25}{3})}$$

This means that the result equals 1.52; in fact, the result can range within the interval from zero to two, with zero being the lowest fluctuation (every kNN selects the same variables) and 2 meaning the maximum fluctuation (each kNN selects a completely different variable). Fluctuation in this case is actually the equivalent of generalizability: if it is low, we can expect the same result with higher probability in multiple repetitions with a random variable. The table below shows the average data on the most informative variable sets.

Figure 2-6: Fluctuation analysis

Fin. perf. measurement method	**Average values**		
	Informativeness	Fluctuation	Number of variables in the most informative subset
Summation	0.765	1.25	4.00
Hyperbola without weights	0.762	1.54	8.67
Exponent	0.761	1.52	8.67
Quintiles	0.749	1.18	16.00
Hyperbola with weights	0.744	1.02	15.33
Sector hyperbola	0.717	1.36	4.67
Risk-purged hyperbola	0.706	0.80	3.33

Let us assume, like in the case of informativeness, that if the experiment setting is the same and differs only in the method used to divide the companies, lower fluctuation of selected variables means that the applied method divides the businesses better. Then we can better identify the success of the best three methods assessed for analysing financial performance, whose informativeness value is very similar. The method of summation, which had the highest informativeness value, is foremost among these three methods even from the perspective of the fluctuation criterion. The difference against the hyperbola without weights and the exponent method is now much stronger than it was for informativeness. The choice of the summation method can be further confirmed by the average number of selected variables in the most informative set, which is less than a half for the summation method compared to the hyperbola without weights and exponent methods. For summation, the kNN classifiers identified a set of six variables one time and a set of three variables two times (i.e. 9 different variables) as most informative, while for the hyperbola without weights it was six, sixteen and four variables (i.e. 22 different variables), and for the exponent method it was nine, thirteen and four variables (i.e. 21 different variables). In the latter two cases, more than half of the input variables were selected as most informative.

Given the results of the experiments and other advantages of the summation method (e.g. problem-free work with negative values compared to methods based on a hyperbola), the summation method was selected for further utilization.

2.7 Description of the methodology used to measure performance

To analyse the performance, the ROA and Assets Growth indicators were used for a 7-year period from 2004 to 2010. Both the indicators were calculated from data obtained from the Albertina CZ database, which contains data from the financial statements of companies registered in the Czech Republic. Yet, there were missing data in the database that had to be found in other publicly available data sources (such as the Register of Companies, digitized volumes of the Commercial Bulletin, or the Albertina database). Even after this manual completion of the database, some data were still missing. As for the ROA indicator, we accepted a maximum of two missing figures from the seven-year time series; due to their volatile development in time they were not replaced with the average values for the company or for all companies in a given year, but the indicator value was not calculated for the year with the missing data. As a result, the overall ROA value was averaged for the number of years for which it was possible to calculate the indicator (i.e. for a minimum of 5 years in the case of 2 missing figures). As for the Assets Growth indicator, we evaluated those companies whose Assets Growth indicator could be calculated for at least 5 years from the period

analysed of 2004–2010. In practice, however, this criterion did not change the number of un-/analysed businesses: it was possible to calculate Assets Growth for all businesses for which it was possible to calculate the ROA. The analysis of financial performance was conducted for 411 companies out of 432; in the case of the remaining 21 companies, too many figures were missing.

Some companies from the basic sample showed an increase in Assets Growth by as much as tens of thousands of a per cent for the observed seven-year period. This increase, however, cannot be considered a long-term success of a company in its growth strategy, since according to the data from the Register of Companies these were mostly newly established businesses, or companies which were involved in a merger in the given year, etc. To prevent such extreme values from distorting the results of the financial analysis, the data on Assets Growth were adjusted so that values above 1000% were disregarded and replaced with an empty value. Such companies did not occur in the selected sample. As for the businesses from the sample that lacked (a maximum of 2) figures representing the amount of total assets at the end of the period analysed, these values were replaced with a value that equalled the value from the previous year increased by the average Assets Growth of the basic sample in a given year. Analogously, the missing values for the beginning of the period were replaced with the total assets value in the following year, reduced by the average Assets Growth in a given year. This adjustment affected 47 companies from the selected sample. In the case of the ROA indicator, the annual values above 200% and below -200% were replaced with an empty value.

For the purpose of further data processing, it was necessary to use the same scale for the values of both financial indicators. Standardization was conducted using the following formula:

$$Z_i = \frac{X_i - \overline{X_s}}{\sigma(X_s)}$$

where i=(1, 2, 3, 4, 5, 6, 7) for ROA and the mean value in the formula is the median of the whole population calculated for each year separately from the Albertina CZ database, and the standard deviation is a deviation of the whole population for each year separately. We eliminated the influence of different economic developments in individual years by using the standardizations for every year. The resulting value of the ROA indicator for one company is then the arithmetic average of seven standardized values.

The Assets Growth indicator for 7 years represents the standardized geometric mean for the seven-year period of 2004[2]–2010. The geometric mean is usually used to calculate the average growth rate.

---

[2] The seven-year average of Assets Growth is based on the values of total assets for 2003 and 2010, as the initial value of total assets for 2004 equals the final value of total assets for 2003.

The geometric average of Assets Growth:

$$Geomean\_AG = \sqrt[7]{\frac{TA_{2010}}{TA_{2003}}} - 1$$

TA – Total Assets

Standardization was conducted similarly to ROA; the difference was that only one value per company (geometric Assets Growth 2004-2010) was standardized.

Using standardization retains the explanatory power of z-score: z-score equals zero for an average company. In the case of the normal distribution, the standardized value of ROA (or Assets Growth) for 68.26% of companies is in the range of <-1; 1>. There are only 13.59% of companies with the standardized ROA value above 1.00, while companies whose value of the standardized ROA is above 2.00 belong to the exceptionally performing businesses; there are only 2.14% of equally good or better companies.

Converting values of both indicators to a single scale using standardization makes it possible to compare the extent of implementing the growth strategy and the strategy of increasing productivity, which are regarded as equivalent. Using the above-mentioned method of summation, i.e. a simple summing up of both the standardized indicators, we obtained the desired one-dimensional coefficient of financial performance, which indicates the performance of a company regardless of the strategy chosen. This methodology marks companies whose financial performance coefficient is greater than 0 as above average; this means that the higher the coefficient is, the more efficient a company is (and vice versa). Businesses that notably increase in size (Assets Growth > 0) while showing average profitability (z-score for ROA = 0) will be marked as above-average as well; this is also true for companies whose productivity is substantially increasing (ROA > 0) while showing an average growth (z-score for Assets Growth = 0), or they are above-average in both strategies at the same time (z-score for both indicators > 0). A company that is becoming smaller in the long term may not yet be included among inefficient businesses if the standardized rate of its profitability is higher than the standardized rate of its (negative) growth (i.e. a situation where the absolute value of the negative z-score for Assets Growth is smaller than the absolute value of the positive z-score for ROA).

3 Feature Selection Methods in Statistical Pattern Recognition

3.1 Introduction

With the ever-increasing specialisation and diversification of scientific disciplines, it is common that similar problems are being tackled in other branches of science, usually without an awareness of the respective research and application communities. It is just this case where the methods developed in a relatively distant field of statistical pattern recognition (SPR) can be used to solve the earlier defined problem of identifying factors of corporate competitiveness, a problem from the field of business and management. Statistical pattern recognition is a discipline comprised of analytical and adaptive methods for processing large datasets, selecting useful information aimed at reducing the data dimensionality and finally classifying these data. Pattern recognition is actually closely connected with machine learning and therefore it is considered to belong to the field of artificial intelligence. One of the fundamental problems of statistical pattern recognition is representing patterns in a reduced number of dimensions, which means a dimensionality reduction. The methods of feature selection are used in statistical pattern recognition to solve this task.

As the methods of statistical pattern recognition represent the essential part of our methodology used for solving the research problems specified in the project, we consider devoting a special chapter to presenting their fundamentals important.

A broad class of decision-making problems can be solved by the *learning approach*. This can be a feasible alternative when neither an analytical solution exists nor the mathematical model can be constructed. In these cases the required knowledge can be gained from the past data, which form the so-called "learning" or "training set". Then, the formal apparatus of statistical pattern recognition can be used to learn the decision-making. The first and essential step of statistical pattern recognition is to solve the problem of variable (feature) selection or more generally of dimensionality reduction, which can be accomplished either by a linear or nonlinear mapping from the measurement space to a lower dimensional feature space.

We will examine some of the most popular tools under various settings to point out several pitfalls often omitted in current literature. In the following we will prefer the term *feature selection* to variable selection in accordance with statistical pattern *recognition conventions*.

3.1.1 Common Research Issues in Machine Learning and Management

Though managers, economists and researchers or practitioners from other fields have different priorities in research issues, issues common to both the fields exist. Such an issue is the problem of selecting only that information which is necessary (and if possible also sufficient) for decision-making. We strongly believe that statistical pattern recognition is the discipline capable of providing a common methodology.

A typical problem which managers often encounter is the problem of too many potential inputs into their respective decision-making problems. This phenomenon has been extensively studied in mathematics and in artificial intelligence. The "curse of dimensionality" problem, as coined by the famous American mathematician Richard Bellman, can perhaps be found in all the fields of science and application areas including economics and management. Without going into the details of this phenomenon, it can be stated that in order to make reliable decisions (or more exactly, to learn to make them based on past experience and the available data) the need for the amount of data dramatically grows with the number of inputs. Mathematically it means that the sample size required grows exponentially with the data dimensionality. This problem is very relevant particularly to the field of management and economics, as the process of managerial or economic data acquisition is usually both time-consuming and costly. Consequently, the data sets acquired are usually too small with respect to their dimensionality.

Though managers consider their respective problems to be of a somewhat different nature (perhaps understandably from their professional point of view), from the point of view of mathematics, the same problem exists formally. It is just a question of different terminology and abstraction needed to find a unified look at the problem. Let us just give an example what we mean by considering which sets of inputs managers use for their decision-making: economic indices, financial data, time series, prediction estimates, etc.

For a mathematician, however, these sets can be looked upon as a set of variables, forming the input vector (in pattern recognition which deals with this problem the term *feature vector* is used). In the majority of practical cases, the dimensionality (the number of inputs) of the original input space can be rather high. It is just a natural consequence of the well-known fact that in the design phase of any system for supporting decision-making, it is extremely difficult or practically impossible to evaluate directly the "usefulness" of particular input variables.

Managers certainly face this problem many times, when having to make the decision. For instance, a manager could have a large number of economic variables

to potentially consider for performing a multiple regression analysis (too many potential regressors). Yet, owing to the often very complex relations and dependencies (sometimes rather strong) among all the respective inputs, economic variables exist (let us speak generally just about variables) which can be left out from decision-making without a great loss of information content. The theory of information, a special mathematical discipline, defines this as the existence of "redundancy" in the set of variables. However, even if managers were aware of this phenomenon, the problem of solving the task of finding the redundant variables for complex problems with many potential inputs would be beyond human capabilities.

The reasons for trying to reduce the set of our inputs into the decision-making process by eliminating redundancies have both practical and theoretical foundations. The practical ones perhaps do not need to be discussed – the reduced cost of the data acquisition is sufficient to substantiate the reduction. On the theoretical front we should like to mention a famous theorem by another American mathematician, S. Watanabe. He founded the mathematical theory of cognition by paraphrasing a world-known fairy tale by the Norwegian author Hans Christian Andersen, where he formulated "the ugly duckling theorem". Roughly stated, it says that no cognition is possible unless our perceptions (input variables) are weighted and, consequently, many of them are given a null weight and thus eliminated from the cognition process.

Each of the fields considered (managerial and clinical decision-making) has its own specificity, and accordingly, different ways of treating the problem. On the other hand, with the ever-increasing specialization and diversification of scientific disciplines, it is not an uncommon fact that similar problems are being tackled in other branches of science, usually without the awareness of respective research and application communities. Yet the results and methods from one scientific discipline can be applied not only to solve problems in another quite different discipline, but they can also often enrich its methodology.

It is our belief that novel methods developed recently in the field of statistical pattern recognition to solve the problem of feature selection can enrich the methodology of selecting the most useful information for decision-making problems in management. Conversely, solving these problems can enrich the methodology of statistical pattern recognition and feature selection.

3.2 Dimensionality Reduction

We shall use the term "pattern" to denote the D-dimensional data vector $x \in X \subseteq \mathbf{R}^D$ of measurements, the components of which are the measurements of the features of the entity or object. We also refer to $\mathbf{x}$ as the feature vector. Let $Y = \{f_1, \bullet\bullet\bullet, f_{|Y|}\}$ be the set of $D = |Y|$ features, where $|\bullet|$ denotes the size (cardinality). The features are the variables specified by the investigator. Following the statistical approach to pattern recognition, we assume that a pattern x is to be classified into one of a finite set of C different classes

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$. A pattern x belonging to class ωi is viewed as an observation of a random vector **X** drawn randomly according to the known class-conditional probability density function $p(\mathbf{x}|\omega_i)$ and the respective *a priori* probability $P(\omega_i)$.

One of the fundamental problems in statistical pattern recognition is representing patterns in a reduced number of dimensions. In most practical cases, the pattern descriptor space dimensionality is rather high. It follows from the fact that it is too difficult or impossible to evaluate directly the "usefulness" of particular input in the design phase. Thus, it is important to initially include all the "reasonable" descriptors the designer can think of and reduce the set later on. Obviously, information missing in the original measurement set cannot be substituted later. Dimensionality reduction (DR in the following) is an important step in data pre-processing in pattern recognition and machine learning applications. In general, it can be shown that such tasks as classification or approximation of the data represented by so-called "feature vectors", can be carried out in the reduced space, more accurately than in the original space, as demonstrated by the so-called "Peaking Phenomenon" (see Fig. 3-7).

The aim of dimensionality reduction is to find a set of new d features based on the input set of D features (if possible d<<D), so as to maximize (or minimize) an adopted criterion.

DR Categorization According to Nature of the Resulting Features

There are two main distinct ways of viewing DR according to the nature of the resulting features:

- DR *by feature selection* (FS)
- DR *by feature extraction* (FE).

The FS approach does not attempt to generate new features, but to select the "best" ones from the original set of features. Depending on the outcome of a FS procedure, the result can be a set of weighting-scoring, a ranking or a subset of features. The FE approach defines a new feature vector space in which each new feature is obtained by combinations or transformations of the original features (see Figs. 3-8a; 3-8b).

FS leads to savings in measurements cost since some of the features are discarded and the selected features retain their original physical or economic interpretation. On the other hand, transformed features generated by feature extraction may provide a better discriminative ability than the best subset of given features (gained by FS), but these new features may not have a clear physical or economic meaning and interpretation.

DR Categorization According to the Aim

DR can be alternatively divided according to the aim of the reduction:

- DR *for optimal data representation*
- DR *for classification.*

The first aims to preserve the topological structure of data in a lower-dimensional space as much as possible, while the second one aims to enhance the subset of

Figure 3-7: Peaking Phenomenon – correct classification rate is a nonmonotonous function of the number of features

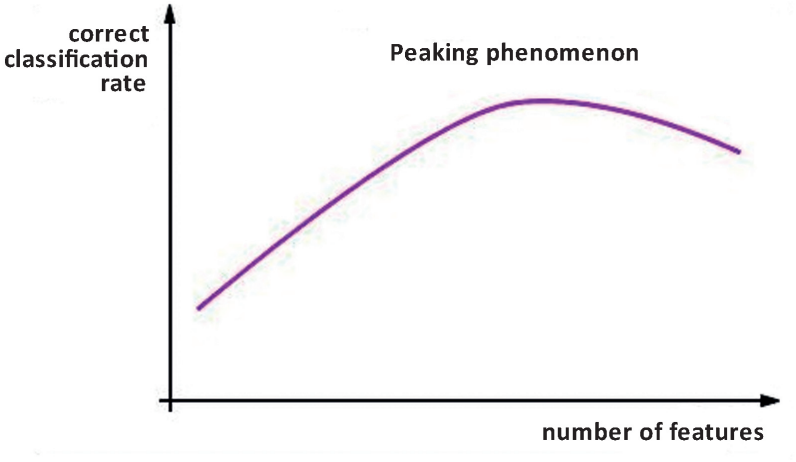


Figure 3-8a: Scheme of FS

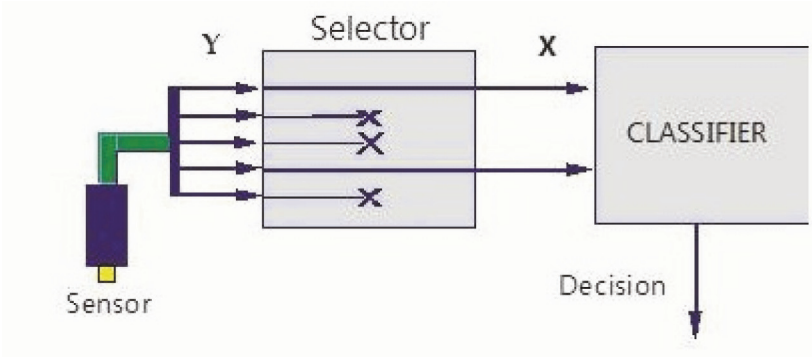
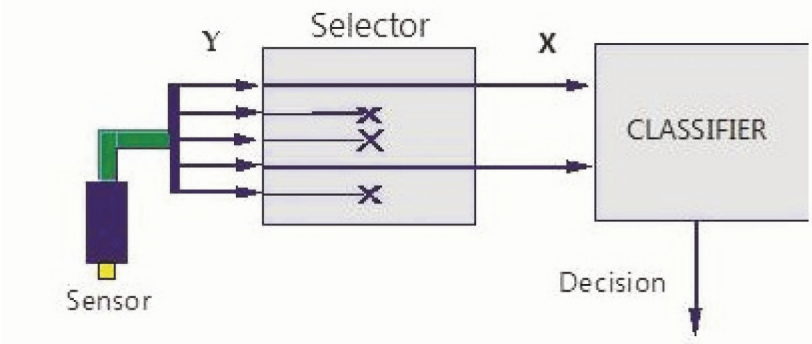


Figure 3-8b: Scheme of FE



discriminatory power. Although the same tools may be often used for both purposes, caution is needed. An example is Principle Component Analysis, one of the primary tools for representing data in lower-dimensional space, which may easily discard important information if used for DR for classification. In the following we shall concentrate on the feature subset selection problem only, with classification being the primary aim. For a broader overview of the subject, see Duda et al. (2000), McLachlan (2004), Ripley (2005), Theodoridis et al. (2006), or Webb (2002), for example.

3.3 Feature Subset Selection

Given a set Y of |Y| features, let us denote X_d the set of all possible subsets of size d, where d represents the desired number of features. Let $J(\mathrm{X})$ be a criterion function that evaluates the feature subset $\mathrm{X} \in \mathrm{X_d}$. Without any loss of generality, let us consider a higher value of J to indicate a better feature subset. Then the feature selection problem can be formulated as follows:

Find the subset $\tilde{X}_d$ for which $J(\tilde{\mathrm{X}}\, d\,) = \max_{\mathrm{X} \in X_d} J(\mathrm{X})$. (1)

Assuming that a suitable criterion function has been chosen to evaluate the effectiveness of feature subsets, feature selection is reduced to a search problem that detects an optimal feature subset based on the selected measure. Note that the choice of d may be a complex issue depending on problem characteristics, unless the d value can be optimized as part of the search process.

One particular property of feature selection criterion, the monotonicity property, is required specifically in certain optimal FS methods. Assuming we have two subsets $\mathrm{S_1}$ and $\mathrm{S_2}$ of feature set Y and a criterion J that evaluates each subset $\mathrm{S_i}$, the *monotonicity condition* requires the following:

$$\mathrm{S_1} \subset \mathrm{S_2} \Rightarrow \mathrm{J(S_1)} \leq \mathrm{J(S_2)}. \quad (2)$$

That is, evaluating the feature selection criterion on a subset of features of a given set yields a smaller value of the feature selection criterion.

3.3.1 FS Categorization With Respect to Optimality

Feature subset selection methods can be split into basic families:

- *Optimal methods*: These include exhaustive search methods, for instance, which are feasible only for small sized problems and accelerated methods, mostly based on the Branch & Bound principle (Somol et al., 2004). All optimal methods can be expected to be considerably slow for problems of high dimensionality.
- *Sub-optimal methods*: They essentially trade the optimality of the selected subset for computational efficiency. They include, e.g., Best Individual Features, Random (Las Vegas) methods, Sequential Forward and Backward Selection, Plus-l-Take Away-r, their generalized versions, genetic algorithms, and the Floating and

Oscillating algorithms in particular (Devijver et al., 1982; Pudil et al., 1994; Somol et al., 2000; Somol et al., 2008b).

Although an exhaustive search guarantees the optimality of a solution, in many realistic problems it is computationally prohibitive. The well-known Branch and Bound (B&B) algorithm guarantees selecting an optimal feature subset of size d without involving explicit evaluation of all the possible combinations of d measurements. However, the algorithm is applicable only under the assumption that the feature selection criterion used satisfies the monotonicity condition (2). This assumption precludes the use of the classifier error rate as the criterion (cf. wrappers, Kohavi et al. (1997)). This is an important drawback as the error rate can be considered superior to other criteria, Siedlecki et al. (1993), Kohavi et al. (1997), Tsamardinos et al. (2003). Moreover, all optimal algorithms become computationally prohibitive due to their problems of high dimensionality. In practice, therefore, one has to rely on computationally feasible procedures which perform the search quickly but may yield sub-optimal results. A comprehensive list of sub-optimal procedures can be found in Devijver et al. (1982), Fukunaga (1990), Webb (2002) and Theodoridis et al. (2006) among others. A comparative taxonomy can be found in Blum et al. (1997), Ferri et al. (1994), Guyon et al. (2003), Jain et al. (1997), Jain et al. (2000), Yusta (2009), Kudo et al. (2000), Liu et al. (2005), Salappa et al. (2007), Vafaie et al. (1994) or Yang et al. (1998). Our own research and experience with FS has led us to the conclusion that *no unique generally applicable approach* to the problem exists. Some approaches are more suitable under certain conditions; others are more appropriate under other conditions, depending on our *knowledge of the problem.* Hence, continuing effort is being invested in developing new methods to cover the majority of situations which can be encountered in practice.

3.3.2 FS Categorization With Respect to Selection Criteria

Based on the *selection criterion* choice, feature selection methods may roughly be divided into:

- *Filter methods* (Yu et al., 2003; Dash et al., 2002) are based on performance evaluation functions calculated directly from the training data such as distance, information, dependency, and consistency, and select feature subsets without involving any learning algorithm.
- *Wrapper methods* (Kohavi et al., 1997) require one predetermined learning algorithm and use its estimated performance as the evaluation criterion. They attempt to find features better suited to the learning algorithm aiming to improve performance. Generally, the wrapper method achieves better performance than the filter method, but tends to be more computationally expensive than the filter approach. Also, wrappers yield feature subsets optimized for the given learning algorithm only – the same subset may thus be unsuitable in another context.

- *Embedded methods* (Guyon et al., 2003, but also Kononenko, 1994 or Pudil et al., 1995; Novovičová et al., 1996) integrate the feature selection process into the model estimation process. Thus, devising a model and selecting its features is one inseparable learning process that may be looked upon as a special form of wrappers. Therefore, embedded methods offer performance competitive to wrappers, enable a faster learning process, but produce results tightly coupled with the particular model.
- The *Hybrid approach* (Das, 2001; Sebban et al., 2002; Somol et al., 2006) combines the advantages of more than one of the approaches listed. Hybrid algorithms have recently been proposed for dealing with high dimensional data. These algorithms mainly focus on combining filter and wrapper algorithms to achieve the best possible performance with a particular learning algorithm, with the time complexity comparable to that of the filter algorithms.

3.3.3 FS Categorization With Respect to Problem Knowledge

From another point of view there are perhaps two basic classes of situations with respect to a priori knowledge of the underlying probability structures:

- *Some a priori knowledge is available*: it is at least known that probability density functions are unimodal. In these cases, one of probabilistic distance measures (Mahalanobis, Bhattacharyya, etc., see Devijver et al. (1982)) may be appropriate as the evaluation criterion. For these types of situations we recommend either the recent prediction-based B&B algorithms for optimal search Somol et al. (2004), or sub-optimal search methods in an appropriate filter or wrapper setting (Sect. 3.4).
- *No a priori knowledge is available*: we cannot even assume that probability density functions are unimodal. For these situations, either a wrapper-based solution using sub-optimal search methods (Sect. 3.4) can be found suitable, or, provided the size of training data is sufficient, it is possible to apply one of the embedded mixture-based methods that are based on approximating unknown class-conditional probability density functions by finite mixtures of a special type (Pudil et al., 1995; Novovičová et al., 1996).

3.4 Sub-optimal Search Methods

Provided a suitable FS criterion function has been chosen, feature selection is reduced to a search problem that detects an optimal feature subset based on the selected measure. Then the only tool needed is the search algorithm that generates a sequence of feature subsets to be evaluated by the respective criterion (see Fig. 3-9).

Despite advances in optimal search (Somol et al., 2004; Nakariyakul et al., 2007), we have to resort to sub-optimal methods for larger than moderate-sized problems, of which a very large number of various methods exist.

Figure 3-9: Feature selection algorithms can be viewed as black box procedures generating a sequence of candidate subsets with respective criterion values, among which intermediate solutions are chosen.

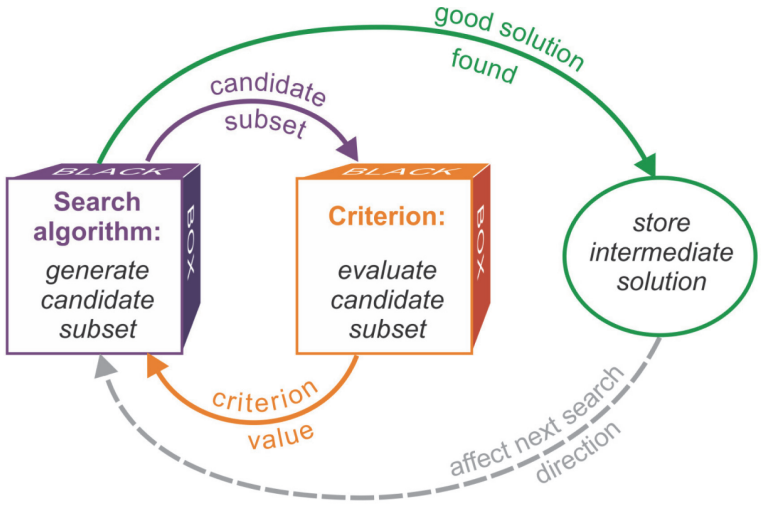
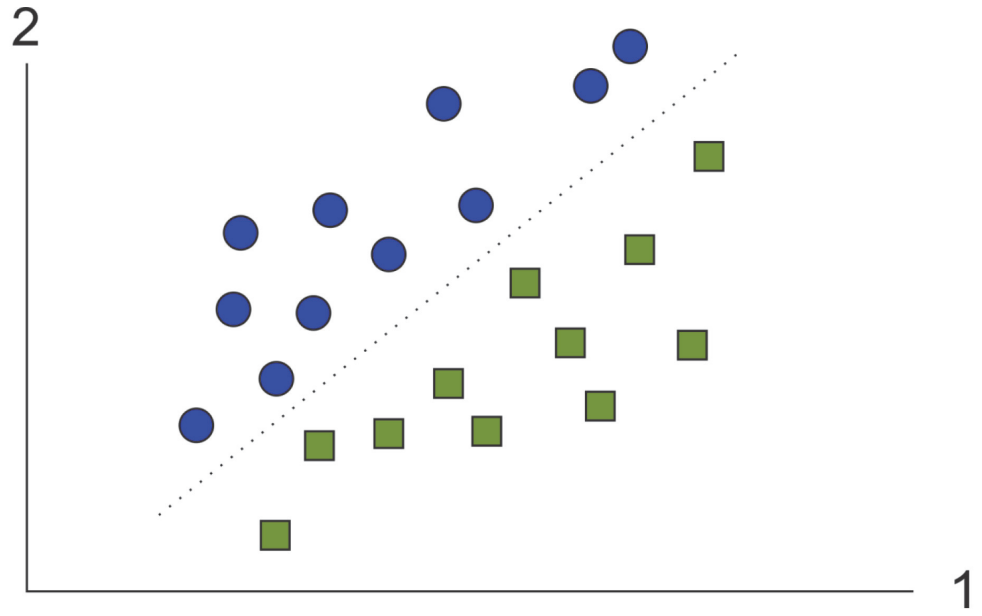


Figure 3-10: In this 2D case nor feature 1 nor 2 is sufficient to distinguish patterns from classes of rectangles and circles. Only when information from both features is combined, classes can be separated (dotted line)



In the following, we present a basic overview over several tools that are useful for problems of varying complexity, based mostly on the idea of sequential search (Section 4.2). An integral part of any FS process is the decision about the number of features to be selected. Determining the correct subspace dimensionality is a difficult problem beyond the scope of this chapter. Nevertheless, in the following we will distinguish two types of FS methods: *d*-parametrized and *d*-optimizing. Most of the available methods are *d*-parametrized, i.e., they require the user to decide what cardinality the resulting feature subset should have.

3.4.1 Best Individual Features

The Best Individual Features (BIF) approach is the simplest approach to FS. Each feature is first evaluated individually using the chosen criterion. Subsets are then selected simply by choosing the best individual features. This approach is the fastest but weakest option. It is often the only applicable approach to FS in problems of very high dimensionality. BIF is standard in text categorization (Yang et al., 1997; Sebastiani, 2002), and genetics (Xing, 2003; Saeys et al., 2007), etc. BIF may be preferable in other types of problems to overcome FS stability problems (see Sect. 6.1). However, it completely ignores inter-feature relations and as such can not reveal solutions, where combinations of features are needed (see Fig. 3-10).

It should be noted that BIF is often considered the method of choice when the problem to be solved is difficult due to unfavourable properties of training data. The fact that BIF has limited optimization power becomes an advantage, e.g., with very high-dimensional data and/or with low-sample size data. See Sect. 6 for more discussion of the problem of overfitting, which in fact is not uncommon in economics where high data acquisition cost and privacy issues often prevent practitioners from obtaining large enough training sets.

3.4.2 Sequential Search Methods and their Evolution

More advanced methods that take into account relations among features are likely to produce better results. Several such methods are discussed in the following.

When feature relations are taken into account, the basic feature selection approach is to build up a subset of required number of features incrementally starting with the empty set (*bottom-up approach*) or to start with the complete set of features and remove redundant features until d features are retained (*top-down approach*). The simplest widely used choice, the *Sequential Forward* (Whitney A. W., 1971) or *Backward* (Marill T., Green D., 1963), *Selection methods* – SFS (SBS) – iteratively add (remove) one feature at a time so as to maximize the intermediate criterion value until the required dimensionality is achieved. Earlier sequential methods suffered from the so-called "nesting" of feature subsets, which significantly deteriorated the performance. The first attempt to overcome this problem was to employ either the

Plus-l-Minus-r, also known as "(l,r)" or "+L-R" (Stearns S. D., 1976) which involves successive augmentation and depletion process, or generalized algorithms (Devijver P. A., Kittler J., 1982). Among later approaches the following two families of methods can be pointed out for general applicability and performance reasons: *Sequential Forward* (or Backward) *Floating Search* methods SFFS, SBFS (Pudil P., Novovičová J., Kittler J., 1994), and *Oscillating Search* (OS) methods (Somol P., Pudil P., 2000). An overview of the evolution of sequential search methods is given in Table 3-1.

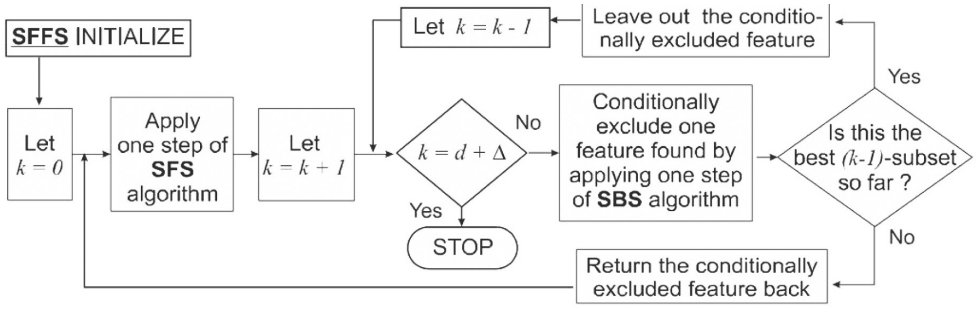
Table 3-1: Evolution of sequential search methods

Method (Simplest first)	**Properties / Improvement over previous method**
Best Individual Features (BIF)	Evaluate each variable separately, completely ignore variable relations
SFS / SBS (Sequential Selection)	Sequentially build subset, in each step with respect to the currently included features
GSFS / GSBS (Generalized Seq. Sel.)	As SFS/SBS, but in each step evaluate groups of features instead of single features to reveal more complicated dependencies
Plus-l-Minus-r	Prevent "nesting": alternate the adding and removing of one feature based of parameters L and R
GPlus-l-Minus-r (Generalized P-l-M-r)	Same as Plus-l-Minus-r, but in each step evaluate groups of features instead of single features to reveal more complicated dependencies
SFFS / SBFS (Floating Search)	Automatically determine the sequence of additions and removals – to avoid user parameters and improve search effectiveness
GSFFS / GSBFS (Generalized Float. S.)	As SFFS/SBSF, but in each step evaluate groups of features instead of single features to reveal more complicated dependencies
ASFFS / ASBFS (Adaptive Float. Search)	Automatically adjust the size of feature groups evaluated in each step to better focus on desired dimensionality
OS (Oscillating Search)	Focus straight on the desired dimensionality + enable greater flexibility: optional randomized search, result tuning, time-constrained search etc.

Floating search methods

The Sequential Forward Floating Selection (SFFS) procedure consists of applying a number of backward steps (removing the feature, that causes the least criterion decrease) after each forward step (adding the feature that maximizes the criterion the most) as long as the resulting subsets are better than previously evaluated ones at that level (see Fig. 3-11). Consequently, there are no backward steps at all if the intermediate result at the actual level (of corresponding dimensionality) cannot be improved. The same applies for the backward version of the procedure. Both algorithms allow a "self-controlled backtracking" so they can eventually find good solutions by adjusting the trade-off between forward and backward steps dynamically. In a certain way, they compute only what they need without any parameter setting (unlike Plus-l-Minus-r).

Figure 3-11: Sequential Forward Floating Selection Algorithm



A formal description of this now classical procedure can be found in Pudil P., Novovičová J., Kittler J. (1994). Floating search algorithms have been critically acclaimed as universal tools not only outperforming all predecessors, but also keeping advantages not met by more sophisticated algorithms (e.g. Kudo M., Sklansky J., 2000). They find good solutions in all problem dimensions in one run and the overall search speed is high enough for most practical problems.

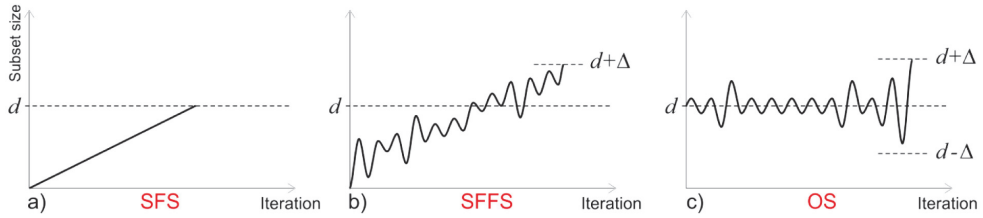
Recent experiments show that the floating search principle overcomes the optimization performance vs. generalization trade-off problem exceptionally well (see the experiments below). The idea of the floating search was later extended in the *adaptive floating* search algorithms (Somol P., Pudil P., Novovičová J., Paclík P., 1999).

Oscillating search method

The *Oscillating Search* (OS) (Somol P., Pudil P., 2000), can be considered a "higher-level" procedure, that takes use of other feature selection methods as sub-procedures within the main course of search. The concept is highly flexible and enables

modifications for different purposes. Unlike other methods, the OS is based on repeated modification of the current subset X_d of d features. In this sense the OS is independent from the pre-dominant search direction. This is achieved by alternating the so-called "*down*- and *up*-swings". Both swings attempt to improve the current set X_d by replacing some of the features by better ones. The *down*-swing first removes the worst feature(s), then adds the best ones back, while *up*-swing first adds, then removes them. Two successive opposite swings form an *oscillation cycle.* The OS can thus be looked upon as a controlled sequence of oscillation cycles of specified depth (number of features to be replaced in one swing). The course of the OS search is compared to SFFS and SFS in Fig. 3-12.

Figure 3-12: Comparing the course of search (current subset size depending on time) in standard sequential search methods



Every OS algorithm requires some initial set of d features. The initial set may be obtained randomly or in any other way, e.g., using some of the traditional sequential selection procedures. Furthermore, almost any feature selection procedure can be used to accomplish the *up*- and *down*-swings. This makes the OS more a framework than a single procedure. The OS can thus be adjusted for an extremely fast search (low cycle depth limit, random initialization) in problems of high dimensionality or a very intensive search aimed at achieving the highest possible criterion values (high cycle depth limit, repeated runs from different random starting points to avoid local extremes, or other complex initialization). The OS can be used to tune solutions obtained elsewhere.

As opposed to all sequential search procedures, OS does not waste time evaluating subsets of cardinalities too different from the target one. This "focus" improves the OS ability to find good solutions for subsets of given cardinality. The fastest improvement of the target subset may be expected in initial phases of the algorithm run. This behaviour is advantageous as it gives the option of stopping the search after a while without serious result-degrading consequences (OS is thus usable in real-time systems). Moreover, because the OS processes subsets of target cardinality from the very beginning, it may find solutions even in cases where standard sequential procedures fail due to numerical problems.

3.4.3 Non-sequential and alternative methods

A broad range of alternative approaches to feature selection is available in addition to sequential search methods, often having properties targeted at particular problems. Other alternatives aim at making the best of existing methods (not necessarily FS methods only) by means of combinations. Many methods inherit both the search procedure and subset evaluation criteria in one indivisible unit.

Randomized methods. Sub-optimal sequential methods are prone to getting stuck in local extremes. Randomizing may help overcome this problem. It may also help find solutions in a significantly shorter time, although this is not guaranteed. The optimization power of purely randomized procedures like *genetic algorithms* (Hussein F., Ward R., Kharma N., 2001; Mayer H. A., Somol P., Huber R., Pudil P., 2000), has been found slightly inferior to sequential methods. Extending sequential methods to include limited randomization may be a good compromise, as is the case with a repeatedly randomly initialized *oscillating search* (Somol P., Pudil P., 2000). A well-known procedure performing the search semi-randomly with an inherited evaluation criterion is the *relief* algorithm Kononenko I. (1994).

Hybrid methods. The motivation to take the best from various approaches led to development of the so-called "hybrid methods". These usually attempt to make use of the better properties of several existing methods while suppressing their drawbacks; the search then often consists of steps performed by means of various sub-methods. Attempts have been made to achieve Wrapper-like performance in Filter-like search time, etc. The idea of hybridization is studied in detail in Liu H., Yu L. (2005).

Mixture-modelling based methods. Mixture-modelling approaches are suitable especially when the data is large and suspected to be multi-modal or otherwise complex in structure. Mixture modelling methods enable simultaneous construction of decision rules and feature selection (Novovičová J., Pudil P., Kittler J., 1996; Novovičová J., Pudil P., 1997; Pudil P., Novovičová J., (1998).

Problem-specific methods. In many fields standard methods can be used only with difficulties or not at all, often due to extreme dimensionality and a small number of samples in the input data. This is the case in genetics (Alexe G., Alexe S., Hammer P.L., Vizvari B., 2006), or text categorization (Forman G., 2003), where the *individually best* feature selection is often the only applicable procedure. Defining highly specialized criteria suitable for the particular task compensates for the deficiency of the BIF search.

Many other methods exist (in all senses of the term "FS Method"), among others the generalized versions of the ones listed above, various randomized methods, methods related to use of specific tools (FS for Support Vector Machines, FS for Neural Networks) etc. For an overview see, e.g., A. K. Jain, R. P. W. Duin, and J. Mao (2000), H. Liu, L. Yu, (2005).

3.4.4 Pitfalls of feature subset evaluation – experimental comparison of criterion functions

As stated before, in certain types of tasks it is important to judge the importance of individual features. Although the importance of every feature may be evaluated in decision theory, in practice 1) we usually lack enough information about the real underlying probabilistic structures and 2) analytical evaluation may become computationally too expensive. Therefore, many alternative evaluation approaches were introduced. It is generally accepted that in order to obtain reasonable results, the particular feature evaluation criterion should relate to a particular classifier. From this point of view, we may expect at least slightly different behaviour of the same features with different classifiers. In fact, even more differences can be observed between feature evaluation made using *wrappers* and *filters*.

Table 3-2: Single features in descending order, first best 7 then last worst 7, according to individual criterion values (i.e., "individual discriminative power"), 4-class, 38-dimensional Australian credit scoring data

Bhattacharyya	37 24	29 23	31 13	19 32	6 0	28 12	20	...	14
Divergence	37 24	31 23	29 13	6 32	19 0	28 12	20	...	14
G.Mahalanobis	29 25	30 17	31 13	28 24	37 35	2 12	26	...	0
Patrick-Fisher	29 13	6 32	19 0	10 1	25 37	20 36	18	...	12
Gauss. cl. (10-f. CV)	6 8	10 9	35 20	25 33	29 31	19 37	2	...	34
1-NN (10-fold CV)	29 37	30 12	31 14	28 36	19 1	16 2	26	...	35
SVM lin (10-f. CV)	29 27	3 22	16 15	18 6	14 7	31 24	4	...	20

In the example in Table 3-2 we demonstrate the differences between some standard criterion functions – both the probabilistic measures (*filter* setting: Bhattacharyya, Divergence, generalized Mahalanobis, Patrick-Fisher distances) and the classification accuracy (*wrapper* setting: *Gaussian classifier, 1-Nearest Neighbour, Support Vector Machine* with linear kernel, classification accuracy evaluated by means

of *10-fold cross-validation*). We evaluated single features of the *Australian credit scoring* data using each of the criteria and ordered them descending according to the respective criterion values. In this way the more distinctive features should appear in the left part of the table, while the noisy and less important should appear in the right. The differences in feature ordering illustrate the importance and also the possible pitfalls of the choice of suitable criterion. Although some features are evaluated as good by most of the criteria (Salappa A., Doumpos M., Zopounidis C., 2007; Somol P., Baesens B., Pudil P.,Vanthienen, J., 2005) and some as bad (Jain A. K., Duin R. P. W., Mao J., 2000), with many others the results vary considerably and may show conflicting evidence Ferri F. J., Pudil P., Hatef M., Kittler J. (1994), Kononenko I. (1994). This is an undesired effect illustrating how difficult it may be to draw general conclusions about which features are generally best to select – it can also be taken as an argument in favour of using *wrappers* instead of *filters*, to identify features with more certainty with respect to the given decision rule.

Following the examples above, it can be concluded that by employing classifier-independent criteria one accepts certain simplification and possibly misleading assumption about data (note that most of probabilistic criteria are defined for unimodal normal distributions only). Nevertheless, classifier-independent criteria may prove advantageous to prevent over-fitting in cases when *wrapper* based feature selection fails to identify feature subsets that generalize well.

3.4.5 Summary of recent sub-optimal feature selection methods

Our own research and experience has led us to the conclusion that *no unique generally applicable approach exists* to the feature selection problem. Some feature selection approaches and methods are more suitable under certain conditions; others are more appropriate under other conditions, depending on the properties and our knowledge of the given problem. Hence, continuing effort is invested in developing new methods to cover the majority of situations which can be encountered in practice.

Recent developments in algorithms for optimal search have led to considerable improvements in the speed of search. Nevertheless, the exponential nature of optimal search remains and will remain one of the key factors motivating the development of principally faster sub-optimal strategies. *Floating search* and *oscillating search* methods deserve particular attention among the family of sequential search algorithms as a practically useful compromise between speed and optimization performance.

We can give the following recommendations based on our current experience– the *floating search* can be considered the first tool to try, as it is reasonably fast and generally yields very good results in all dimensions at once, often succeeding in finding global optimum with respect to the chosen criterion. The *floating search* also shows to be a good compromise to deal with the *optimization efficiency* versus *generalization* (impact on classifier performance on unseen data) trade-off. The oscillating search may become a better choice when: 1) the highest possible criterion value must be

achieved but optimal methods are not applicable, 2) a reasonable solution needs to be found as quickly as possible, 3) numerical problems hinder the use of standard sequential methods, 4) extreme problem dimensionality prevents any use of standard sequential methods, or 5) the search is to be performed in real-time systems. The *oscillating search* shows an outstanding ability to avoid local extremes in favour of finding the global optimum, especially when repeated with different random initial feature subsets.

It should be stressed that, as opposed to the optimal *branch & bound* algorithm, the sub-optimal sequential methods are tolerant to deviations from monotonic behaviour of feature selection criteria. It makes them particularly useful in conjunction with non-monotonic FS criteria like the error rate of a classifier (cf. *wrappers* Kohavi R., John G.H., 1997), which according to a number of researchers seem to be the only legitimate criterion for feature subset evaluation. The superior performance of *wrappers* over *filters* has been verified experimentally (e.g. Pudil, Somol, 2008).

3.4.6 Dependency-Aware Feature Selection (DAF)

A relatively new member of the family of FS methods is the so-called "Dependency-Aware Feature Selection" (Somol P., Grim J., Pudil, P., 2011). Though it is not a "classical search method" but rather a ranking method, it can be used for feature selection too. Since the DAF method has been used in our research (as described in the following chapters) we shall describe it briefly in the following.

Problems stemming from an insufficient sample size with respect to problem dimensionality are not typical only for a high-dimensional FS; such situations can appear even in low-to-mid-dimensional problems as is common in economics or medicine (Liu H., Yu L., 2005) for instance, where the number of observed cases is often too limited. In cases of an insufficient data sample size it may be questioned what information about the features can be reliably obtainable from the data at all (Pudil P., Novovičová J., Kittler J., 1994; Pudil P., Novovičová J., Kittler J., 1994; Raudys Š., 2006). The commonly suggested work-around is to refrain from complex analysis of feature subsets in favour of simpler FS methods or even trivial feature ranking, also known as the Best Individual Features (BIF) method (Ripley B., 1996; Salappa A., Doumpos M., Zopounidis C., 2007). It is commonly assumed that ignoring inter-feature dependencies is less harmful than obtaining misleading information through serious estimation errors due to over-fitting.

Assume a general pattern recognition problem (typically a classification or clustering problem) in an N-dimensional feature space. In the particular case of classification, some objects described by means of features $f_1, f_2, \ldots f_N$ (real valued or discrete) are to be classified into one of a finite number of mutually exclusive classes. The common initial step in classifier design is to choose a reasonably small subset of informative features by using a feature selection method.

Denoting F the set of all features

$$F=\{f_1, f_2, \ldots f_N\}$$

we assume that for each subset of features $S \subset F$ a feature selection criterion $J(\bullet)$ can be used as a measure of quality of S (typically but not necessarily from the classification point of view). According to the standard FS paradigm, the resulting feature subset is obtained by maximizing $J(S)$ over the class of all subsets $S \subset F$.

Here we do not impose any restrictions on the function $J(\bullet)$ except that we expect it to be capable of reflecting feature behaviour in context, i.e., it should provide more than just combined information on individual feature merit.

A specific FS problem arises in case of very high dimensionality, e.g., $N \approx 10^3 \div 10^6$ or even more. Even the simplest sub-optimal optimization techniques are exceedingly time-consuming in such cases and, consequently, only very basic tools can be used to optimize features. The common approach is a simple ranking of features based on the individual feature quality (cf. BIF). By ordering the features according to the inequality

$$J(\{f_{i_{n-1}}\}) \le J(\{f_{i_n}\}); \; n=2, 3, \ldots, N$$

we can easily identify a subset of d individually best features $f_{i_{N-d+1}}, f_{i_{N-d+2}}, \ldots, f_{i_N}$ but, in this way, we completely ignore the potentially crucial dependence among features and the resulting subset thus may be far from optimal.

The DAF method attempts to generalize the idea of individually best ranking by evaluating the quality of each feature repeatedly in the context of randomly chosen feature subsets. In other words, we evaluate the quality $J(\{f_n\} \cup S)$ for a sufficient number of random subsets $S \subset F$, $(f_n \notin S)$ and compare the corresponding mean value S with the analogous mean of $J(S)$ for subsets $S \subset F$ not containing the feature f_n, (i.e. $f_n \notin S$).

This idea is based on the intuitive assumption that "good" features exhibit reasonably consistent behaviour in context with other features, that this information is obtainable easily enough and that it can improve upon the information about individual feature quality. An analogous mechanism has been shown to perform well in Fast Branch & Bound algorithm (Somol P., Pudil P., Kittler J., 2004), where feature behaviour is studied in variable context and the averaged information is utilized to predict J $(\bullet)$ values, enabling considerable acceleration of the search process.

The starting point of the proposed dependency-aware feature ranking is a randomly generated sequence of feature subsets, to be denoted probe subsets

$$S = \{S_1, S_2, ..., S_K\}, S_j \subset F, j = 1, 2, ..., K,$$

where each subset is evaluated by a criterion function $J(\bullet)$. The cardinality of the subsets $S \in S$ should vary and the resulting sequence S should be long enough to "approximate" the class of all possible subsets of F in a reasonably uniform way. For each feature $f \in F$ there should be enough subsets in S that do contain it as well as enough subsets that do not.

To generate a random probe subset we use the following simple procedure: first the subset size d is randomly chosen so that $d \in [1, \min\{N, \tau\}]$ where $\tau \in [1,N]$ is an optional user-specified upper limit. Next, indexes of features to be selected are randomly generated from [1, *N*] as long as the number of unique feature indexes is lower than *d*.

Given a sufficiently large sequence of feature subsets *S*, we are interested in utilizing the information contained in the criterion values $J(S_1)$, $J(S_2)$, ..., $J(S_K)$ in depth. Instead of measuring the classification "power" of individual features $\boldsymbol{f} \in F$, we compare the quality of probe subsets containing ***f*** with the quality of probe subsets not including ***f***.

A straightforward idea is to compute the mean quality μ_f of subsets $S \in \mathrm{S}$ containing the considered feature $\boldsymbol{f} \in F$

$$\mu_f = \frac{1}{|\mathbb{S}|_f} \sum_{S \in \mathbb{S}_f} J(S), \qquad \mathbb{S}_f = \{S \in \mathbb{S}: f \in S\}$$

and the mean quality $\bar{\mu}_f$ of subsets $S \in \mathrm{S}$ not containing the considered feature ***f***:

$$\bar{\mu}_f = \frac{1}{|\bar{\mathbb{S}}|_f} \sum_{S \in \bar{\mathbb{S}}_f} J(S), \qquad \bar{\mathbb{S}}_f = \{S \in \mathbb{S}: f \notin S\}$$

with the aim of using the difference of both values as a criterion for ranking the features:

$$DAF_0(f) = \mu_f - \bar{\mu}_f, \qquad f \in F$$

Note that the "dependency aware" ranking criterion DAF_0 does not measure the individual quality of a feature $f \in F$ separately by means of the criterion value $J(f)$, but takes into account the quality of feature *f* in the context of other features occurring in the sets $S \in \mathrm{S}$. The value $DAF_0(f)$ can be viewed as the average benefit of including the feature f into the feature subsets $S \in \mathrm{S}$.

To conclude, the DAF method ranks the features according to the average benefit of including a feature into a number of randomly generated feature subsets. The benefit is expressed as the difference of mean criterion values computed for subsets that do and do not contain the feature, based on arbitrarily chosen feature selection criterion. This simple idea has been shown quite suitable for high and very-high-dimensional feature selection problems where it is capable of considerably over-performing the commonly used individual feature ranking approaches due to its favourable mix of properties: the ability to reveal contextual information, reasonable speed, and generalization ability. Moreover, it has also been proven to have good properties with respect to the stability of solution, discussed in the following.

3.5 Performance Estimation Problem

A comparison of FS methods needs to be performed with caution, paying attention to the further mentioned aspects of the problem. It is very different whether we compare concrete method properties or the final classifier performance determined by use of particular methods under particular settings. Researchers frequently break the basic requirement of using different data sets for training and testing. In such a case (denoted as *resubstitution*), the results are biased and too optimistic. The performance on an independent data set (and thus the required ability of learning methods to generalize) is generally worse.

Certainly, final classifier performance (on independent test data) is the ultimate quality measure. However, as stated in more detail in Section 6, misleading conclusions about FS may be easily drawn when evaluating nothing else, as classifier performance depends on many more different aspects then just the actual FS method used. Nevertheless, in the following we will adapt classifier accuracy as the main means of FS method assessment.

There seems to be a general agreement in the literature that wrapper-based FS enables the creation of more accurate classifiers than filter-based FS. Nevertheless, this claim is to be taken with caution, while using actual classifier accuracy as the FS criterion in wrapper-based FS may lead to the very negative effects mentioned above (over-training). At the same time, the weaker relation of filter-based FS criterion functions to particular classifier accuracy may help better generalization. However, these effects can be hardly judged before the building of the classification system has actually been accomplished.

We will focus only on wrapper-based FS in the following. Wrapper-based FS can be accomplished (and accordingly its effect can be evaluated) using one of the following methods:

- *Re-substitution* – In each step of the FS algorithm all data is used both for classifier training and testing. This has been shown to produce strongly optimistically biased results.
- *Data split* – In each step of the FS algorithm the same part of the data is used for classifier training and the other part for testing. This is the correct way of classifier performance estimation, yet it is often not feasible due to the insufficient size of available data or due to the inability to prevent bias caused by unevenly distributed data in the dataset (e.g., it may be difficult to ensure that with two-modal data distribution the training set will not represent one mode by coincidence and the testing set the other mode)
- *Cross-Validation (CV)* – Training data is split into several parts. Then in each FS step a series of tests is performed, with all but one data part used for classifier training and the remaining part used for testing. The average classifier performance is then considered to be the result of FS criterion evaluation. Because

a different part of data in each test is used for testing, all data is eventually utilized, without actually testing the classifier on the same data on which it had been trained. This is significantly better than re-substitution.

- *Leave-one-out* – A special case of CV with just one sample left for testing in each data split. This is computationally more expensive, but better utilizes the data.
- *Hold-Out (HO)* – Training data is randomly sampled. A series of tests is performed in each FS step, with a part of the training data randomly sampled for classifier training and another part randomly sampled for testing. The average classifier performance is then considered to be the result of a FS criterion evaluation. Unlike CV, this may avoid possible bias caused by deterministically and evenly split data, but possibly requires a higher number of trials than CV.

3.6 Problem of Feature Selection Overfitting and Stability

The prevailing approach to FS method performance assessment in older literature was to evaluate the ability to find the optimum, or to get as close to it as possible, with respect to some criterion function defined to distinguish classes in classification tasks or to fit data in approximation tasks. Recently, emphasis has been put on assessing the impact of FS on generalization performance, i.e., the ability of the devised decision rule to perform well on independent data. It has been shown that similarly to classifier over-training, the effect of feature over-selection can hinder the performance of a pattern recognition system (Raudys, 2006), especially with small-sample or high-dimensional problems. Compare Figures 3-13 and 3-14 to see an example of the effect.

Figure 3-13 shows the maximal criterion value obtained by each method for each subset size. It can be seen that the strongest optimizer in most of the cases is OS, although SFFS falls behind just negligibly. SFS's optimization ability is shown to be markedly lower, but still higher than that of BIF's.

Figure 3-14 shows how the optimized feature subsets perform on independent test data. From this perspective the differences between methods is largely diminished. The effects of feature over-selection (over-fitting) affect the strongest optimizer – OS – the most. SFFS seems to be the most reliable method in this respect. SFS yields the best independent performance in this example. Note that although the highest optimized criterion values have been achieved for subsets of roughly 6 features, the best independent performance can be observed for subsets of roughly 7 to 13 features. The example thus is effective in illustrating one of the key problems in FS – the difficulty to find subsets that generalize well, related to the problem of feature over-selection (Raudys, 2006).

The speed of each method tested decreases with its complexity. BIF runs in linear time. Other methods run in polynomial time. SFFS runs roughly 10× slower than SFS. OS in the slow test setting runs roughly 10 to 100× slower than SFFS.

Figure 3-13: Sub-optimal FS methods' optimization performance on 3-NN wrapper

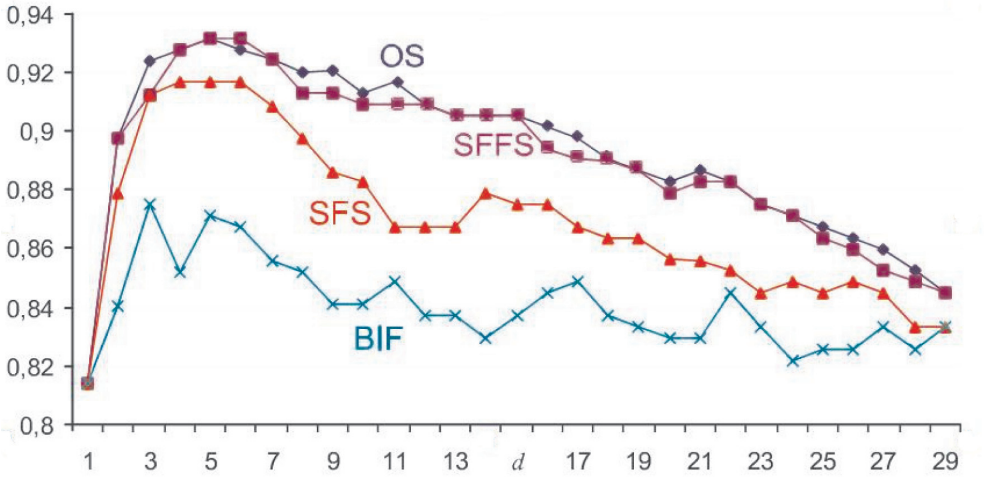
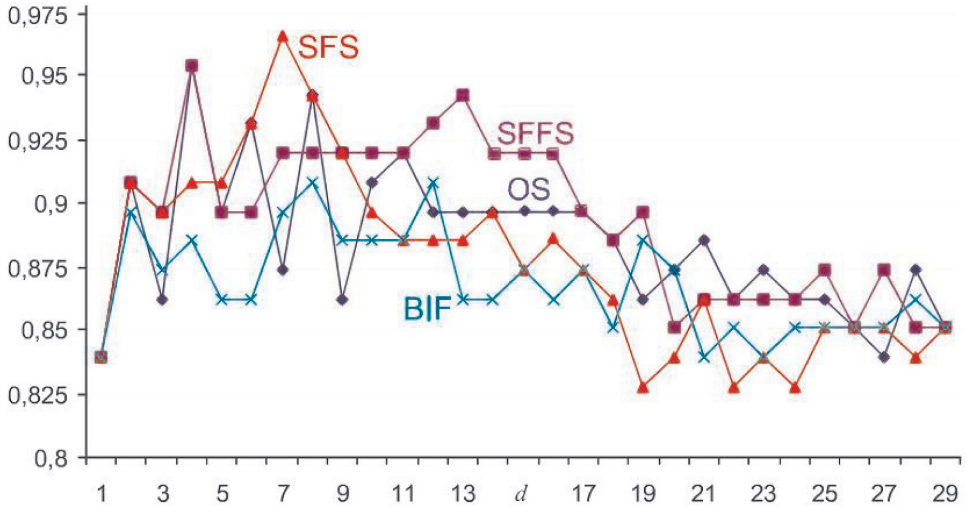


Figure 3-14: Sub-optimal FS methods' performance verified using 3-NN on independent data



It has also been pointed out that independent test data performance should not be neglected when comparing FS methods (Reunanen, 2003). There seems to be a general agreement in literature that wrapper-based FS enables the creation of more accurate classifiers than filter-based FS. This claim is nevertheless to be taken with caution, while using actual classifier accuracy as FS criterion in wrapper-based FS may lead to the very negative effects mentioned above (overtraining). At the same time, the weaker relation of filter-based FS criterion functions to particular classifier

accuracy may help better generalization. Yet these effects can be hardly judged before building the classification system has actually been accomplished. The problem of classifier performance estimation is by no means simple. Many estimation strategies are available, the suitability of which is problem dependent (re-substitution, data split, hold-out, cross-validation, leave-one-out, etc. – see the Section 5). For a detailed study on classifier training related problems and work-around methods, e.g., stabilizing weak classifiers, see Skurichina (2001).

3.6.1 Problem of Feature Selection Stability

As already stated before, it is common that classifier performance is considered the ultimate quality measure, even when assessing the FS process. However, misleading conclusions may be easily drawn when ignoring stability issues. Unstable FS performance may seriously deteriorate the properties of the final classifier by selecting the wrong features. Following Kalousis et al. (2007) we define the *stability* of the FS algorithm as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution. FS algorithms express the feature preferences in the form of a selected feature subset $S \subseteq Y$. Stability quantifies how different training sets drawn from the same generating distribution affect the feature preferences. Recent works in the area of FS methods' stability mainly focus on various stability indices, introducing measures based on Hamming distance (Dunne et al., 2002), correlation coefficients and Tanimoto distance (Kalousis et al., 2007), consistency index (Kuncheva, 2007) and Shannon entropy (Křížek et al., 2007). Stability of FS procedures depends on the sample size, the criteria utilized to perform FS, and the complexity of FS procedure, Raudys (2006).

Two new measures called *consistency* and *weighted consistency*, which express the stability or robustness of FS method with respect to various data samplings, have been introduced in Somol et al. (2008a).

3.7 Summary

We have shown in this chapter that a battery of dimensionality reduction methods exists within the context of statistical pattern recognition. We claim that many of these methods are directly (or with just minor adjustments) applicable to solving managerial problems. In particular the problem of identifying factors of competitiveness in companies can be considered a special case of the general feature selection problem known within statistical pattern recognition. Managerial problems of this type can be approached using feature selection tools to great advantage, following from the fact that such dimensionality reduction techniques are well established and understood.

4 Testing approaches and methods based on learning methods for identifying factors of competitiveness

4.1 Introduction

The issue of searching for corporate competitiveness factors represents an attractive and very topical matter for both practice and theory and is discussed in more detail in the introductory chapter of this monograph. Summarising studies quote dozens of works (e.g. 82 studies in the meta-analysis by Allouche and Laroche, 2005) and bibliographic databases index thousands of articles annually on this topic. However, these works do not present a unified school of thought, as they differ in their approaches to the issue, terminology used, application of methods and reasons for dealing with the issue as well as the credibility of the results and the ways they are applied (Ambastha and Momaya, 2004).

Within the context of corporate competitiveness factors, i.e. the causes that influence competitiveness substantially, we focus on one of the key issues – the synergy problem. This refers to the reality that a firm's competitiveness is not the result of a partial effect of individual factors, but of their synergistic effect. This is often mentioned in the literature focused on mergers or inter-company and intra-company cooperation (Carter, 1977; Williamson and Verdin, 1992), while papers focused on other potential sources of competitive advantage view competitiveness as a multidimensional concept (Fraj-Andrés et al., 2008).

The literature analysed implies that for the purposes of the search and evaluation of corporate competitiveness factors, there is quite a wide variety of different mathematical and statistical methods and techniques. First, we should mention bivariate techniques. From earlier studies, we can mention White (1986) and his study of the influence of generic strategies on return on investment and sales growth, where he used correlation, frequencies and averages. Further, we can cite Hansen and Wernerfelt (1989), who researched the influence of economic and organisational factors on ROA using correlations. As far as more recent studies are concerned, we should stress Artiach et al. (2010), who used correlations and t-tests to verify the influence of

selected factors on the position of companies in the Dow Jones Sustainability Index, or Liu et al. (2004), who also used correlations. However, bivariate techniques cannot capture the complexity of reality nor the above-mentioned synergistic effect because they always analyse only the relationship between two variables.

The authors believe that corporate competitiveness, even if narrowed down to Corporate Financial Performance (CFP in further), depends on many factors and that the influence of these factors needs to be examined as mutual associations between a number of interconnected variables and CFP. Our view is in accordance with empirical studies of corporate competitiveness using advanced statistical methods which range from multiple regression (Homburg et al., 1999; Cagwin, D., 2006), multiple logistic regression (Kessler, 2007) through structural modelling (Yilmaz et al., 2005), to decision trees (Molina et al., 2004). Unlike bivariate techniques, these procedures test the dependence of diversely measured performance or corporate competitiveness on multiple independent variables. This usually provides a better explanation of the variability in the performance of companies. However, their use requires certain restrictions, such as the normality of the input data or a robust a priori model. Thus, each of these methods has its limitations, requirements and drawbacks. Similarly, though the first results (cf. Blazek et al., 2008) of the Centre the Research Centre for Competitiveness of the Czech Economy (RCCCE) were based mainly on bivariate analysis of these variables, the fact that they are generally not mutually independent resulted in using methods of multivariate analysis. The authors aim to show that the methodology of feature selection in statistical pattern reduction, rarely used in this context in the past (cf. Pudil et al., 2002), is a well-developed methodology fulfilling this task.

The fundamental problem to be solved is to analyse the characteristics of companies and to identify the factors on which their competitiveness depends. A large dataset is available for this purpose. As described in more detail in the following paragraphs, it consists of approximately 400 companies where each company is characterized by a relatively high-dimensional vector (it means in our case by some 70 variables). Furthermore, each company is evaluated by one or several values that express its successfulness, assessed by means of expert knowledge. However, the data set is not complete. Some values are missing; moreover, different values are missing for different companies. Another very important characteristic of the data set is in the fact that it consists of different types of variables (features). Some of them are quantitative (numerical); some are ordinal or even categorical.

There is no single and unique approach to the solution of this very complex problem since it depends on several options. The first choice is whether the problem should be looked upon as a classification or as a regression problem. These two types of problems can be solved by different sets of tools and the respective solutions have different properties.

The input data contain one specific continuous value for each company, denoting a measure of its successfulness. Consequently, it offers the possibility of considering

the problem at hand as a regression problem, while at the same time it is possible to transform the problem into a classification problem. This can be done simply by discretizing the value of successfulness (e.g. ROA in our case) into several values, which effectively means assigning the companies according to their value of successfulness into several classes (e.g. successful, neutral, unsuccessful). By transforming this into a classification problem, we lose some information and impose a potentially inexact division on the data, which increases the danger of achieving misleading results. However, on the other hand, the classification task may be solved more easily and thus the resulting classification rule may better describe the less precisely formulated problem. Therefore, the classification approach can perform a more successful "data-mining" when considering the whole problem. Obviously, it cannot be guaranteed in the case of very unfavourable properties of data (a small sample size, high dimensionality, a very complex distribution which cannot be assessed with a small sample in a sufficiently robust manner). In our case, we face these problems to a large extent as a preliminary analysis of the data set shown. The data we have at disposal are multimodal, strongly non-linear, incomplete and non-homogeneous.

Apart from the issues of regression or classification there is another question – how to evaluate the dependence of the company quality (or successfulness) on the variables (features) characterizing companies. In principle, we could try to attain this information directly from the data without any modelling. However, this approach gives very limited options. As an example, it is possible to evaluate individual correlations of features with the successfulness of a company. The introductory analyses have shown that such an analysis did not lead to the goal. No single features that could describe the successfulness of companies sufficiently well were found. As expected, these attempts have shown that it is necessary to utilise multi-dimensional feature subspaces in order to reveal at least some reasonable link between the features and successfulness. To make it more illustrative, we can use as an example the two following figures.

It can be seen that although we can have two features where each of them has a very low discriminative power (therefore very low informativeness I (F1), I (F2)) when considered separately (or isolated), by considering them both as a pair, the informativeness can substantially increase and even be several times higher than the sum of individual measures of informativeness.

Based on all these reasons, it is instead a more reasonable "intrinsic analysis" to approach the data analysis by means of modelling as described in the following. The usage of model presents the possibility of gaining some knowledge on data distribution, to a certain extent including the subspaces with few or no samples. It also makes it possible to generalize, e.g. in prediction context. By modelling we understand creating a simplified formal description of data which substantially simplifies further analyses and makes it possible to find out generalizing information.

Figure 4-15a: Informativeness of isolated features

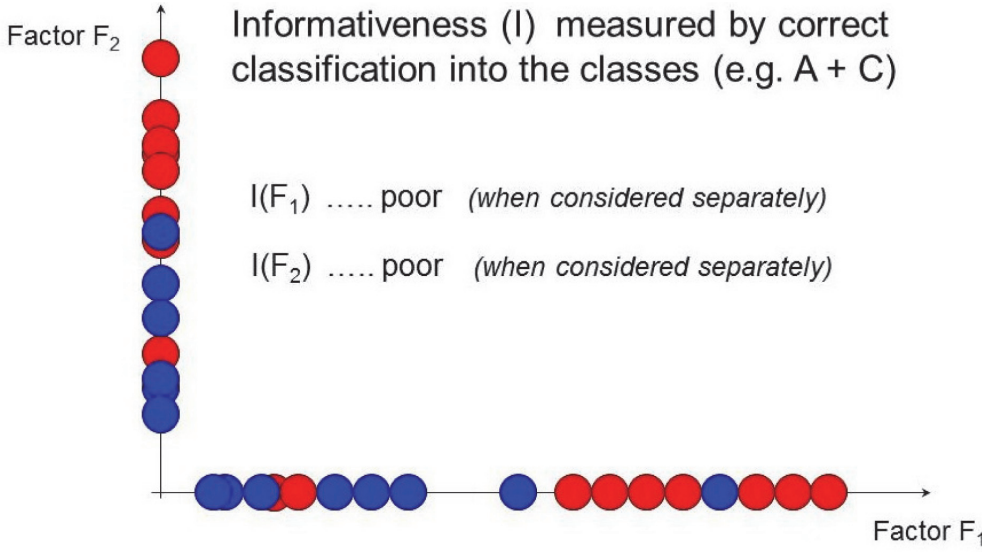
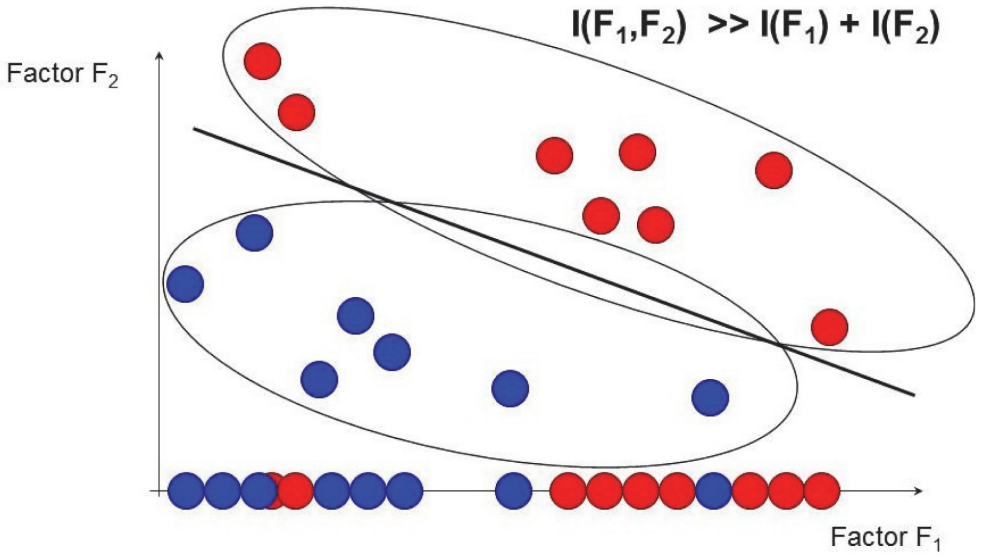


Figure 4-15b: Synergic effect of more features



In order to verify the above-specified hypothesis, two approaches have been investigated, differing in the way the factors group is selected and how the CFP is defined:

1. "Pattern classification approach" – by means of the informativeness for classification into a pre-defined set of mutually exclusive classes of companies;

2. "Regression approach" - by means of the non-linear regression model accuracy.

The first approach was the primary focus of our paper at the 8th European Conference on Management Leadership and Governance (cf. Pudil et al., 2012). The second approach was presented (for different datasets) in our papers at the 1st and the 2nd International Conference on Management Leadership and Governance (cf. Pudil et al., 2013; Somol et al., 2014).

For the purpose of classification, (cf. Pudil et al., 2012) companies were grouped into three classes on the two-indicator space. The companies with an above average value of the CFP coefficient were denoted as having a good CFP, thus "successful" (group A, included 165 companies). On the other hand, the companies having a negative assets growth or/and with a negative ROA were denoted as "unsuccessful" (group C, included 82 companies). The remaining companies became the "intermediate" group B.

For the regression analysis, we do not need the class information but a single continuous value expressing the overall CFP. For this purpose, we used the sum of Assets Growth and Return on Assets (more in chapter 2).

4.2 Feature selection based evaluation of competitiveness factors

As stated above, the multivariate analysis has to be used for the investigated task. In the context of machine learning, however, multivariate analysis is common and the available analysis frameworks have been considered vital in various recognition tasks (credibility scoring, image analysis, automated medical diagnostics /cf. Theodoridis and Koutroumbas, 2006/). One of the key approaches to multi-variate analysis is feature selection (FS), described in more detail in the previous chapter.

The problem we face here is selecting a smaller subset of the most informative characteristics from the set of all the characteristics measured (variables, features). The informativeness is measured here by the ability of a subset to correctly discriminate classes according to their financial performance (groups A, B and C). Selecting a smaller subset of characteristics means effectively performing a dimensionality reduction on the original data. As stated in the previous chapter, the principal goal of FS is to select a small subset of variables with the aim of discriminating among classes of observations. We used FS methodology to search for a subset of characteristics which discriminate between the group of A companies and the group of C companies as much as possible. The idea about using the statistical pattern recognition approach for specifying competitiveness factors is based on the analogy of both the problems:

Table 4-3: Analogy of problems of statistical pattern recognition and the problem of determining competitiveness factors

Statistical pattern recognition	Determining competitiveness factors
– "pattern" is described by a real *D*-dimensional vector y = (y_1, y_2, ...y_D) – pattern is to be classified into one from a final number of different classes – it is a classification task combined with the reduction of dimensionality of feature vectors on which the classification is based – feature selection methods enable selecting a reduced subset of features, optimal with respect to the discriminative ability (and the following classification of other patterns) – complete patter recognition task consists of two main stages: – 1/ feature selection – 2/ pattern classification	– all companies are represented by D indices (characteristics, variables, features) – they are classified into one from several classes (in our case two or three classes) – the classification task is not the goal here; however, the dimensionality reduction helps to identify "key factors" of competitiveness – feature selection methods enable selecting a reduced subset of characteristics (features, optimal with respect to the discriminative ability related to their competitiveness) – only the first stage is used in our problem – there is no need to determine a classification rule – we use it only as an FS criterion

4.2.1 Feature Selection Methodology

The methodology of feature selection, or more generally dimensionality reduction in machine learning, is very extensive and has been briefly described in the previous chapter. We utilize FS methodology in two ways:

1/ to search for a subset of characteristics which best discriminate between companies of groups A, B and C as much as possible;

2/ to search for a subset of characteristics that minimizes the regression error.

It should be noted that the resulting optimized subsets of these two different tasks might be different.

The main advantage of non-trivial FS methods is their ability to evaluate characteristics in context, possibly extracting more information than is customary with commonly used ranking methods. In machine learning it is well known that "the two individually best features may not be the best pair". Very often either the two individually best might prove redundant (each of them provides almost exactly the same information despite their seemingly different nature), or none of the features proves sufficient to reveal the true structure in the data (see Figure 4-16 for an illustration of a two-dimensional case; note that the same effects can take place in multiple-dimensional subspaces).

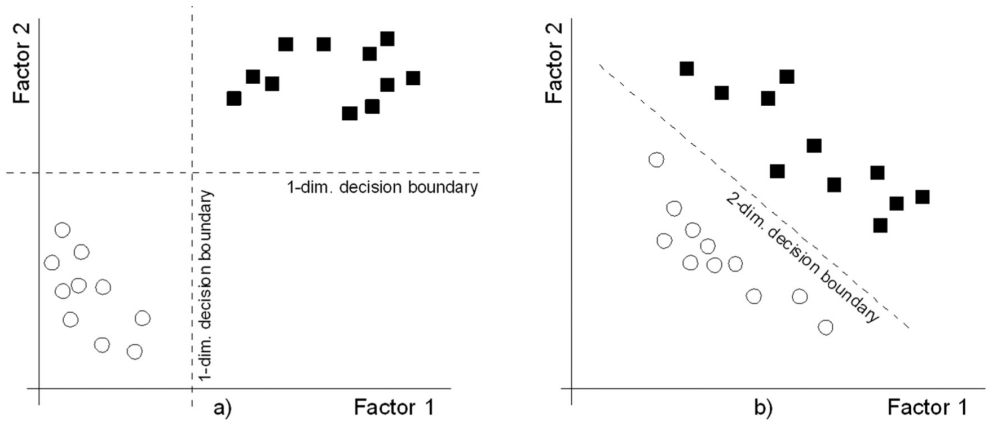
As already stated in the previous chapter, in order to find the most informative subset of features, we need:

– an efficient strategy for traversing the space of possible subsets towards the optimum

– a criterion of evaluating the quality of arbitrary subsets of features.

A wealth of search strategies and criteria exists. There is no single strategy, nor a single criterion function known to be best in all cases. Moreover, our data had the limitation of a relatively high dimensionality and a low sample size which is rather typical in this type of analyses – the more characteristics we try to collect, the less companies we have available. With respect to this fact, we decided to first apply two conceptually different FS methods and to evaluate their performance not only in terms of results achieved, but also in terms of their stability (cf. Somol and Novovovicova, 2010). Further analysis (Section 4.3.) will then be performed using the more stable (robust) method identified in the following Section 4.2.2.

Figure 4-16: Cases of possible univariate analysis failure (illustrative example). Let the dots represent the companies of poor CFP and the rectangles represent the well performing companies. Univariate analysis is not capable of revealing a) competitiveness factor redundancy, b) multi-variate factor dependency leading to crucial model accuracy improvement in higher than one-dimensional subspace.



4.2.2 Evaluating Stability of Feature Selection Methods

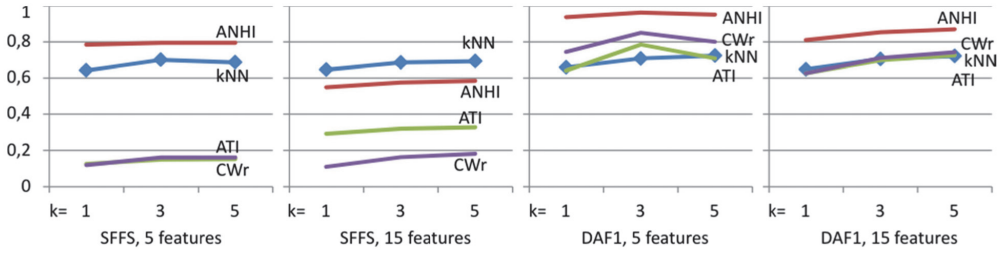
To evaluate candidate feature subsets in the stability test, we estimated the accuracy of the well-known *k*-Nearest Neighbour classifier (cf. Theodoridis and Koutroumbas, 2006)). The advantage here is the simplicity, non-linearity (high descriptive power), adjustability of sensitivity to outliers through the *k* value, and most importantly, the possibility to redefine *k*-NN to support a mix of feature types: numerical and nominal. This was achieved by normalizing all numerical feature values to <0,1>, and defining the distance between nominal values as 0 in case of equality and 1 otherwise.

As a search strategy, we chose a) Sequential Forward Floating Search (SFFS) (cf. Pudil et al., 1994), generally known for its good optimization performance but also

good search speed, and b) Dependency-Aware Feature Ranking (DAF) (cf. Somol et al., 2011) known to be very robust against over-fitting.

To compare the stability of results yielded by the two chosen FS methods, we repeated each experiment for each of the methods 20x, each time using a different randomly chosen 80% part of the available data. In each run, a subset of features was selected, possibly different from those of the other runs. The similarity of the various results is evaluated using various *stability measures*. Here we evaluate ANHI, ATI and CWr, (for an explanation and details cf. Somol and Novovovicova, 2010) in order to get a broader picture (no single measure is able to capture all aspects of the result observed). The goal is to possibly reveal the principal differences between the two FS methods employed with respect to solving our particular FS problem. The graphs in Figure 4-17 summarize the findings. It can be clearly seen that the stability of SFFS is considerably worse than that of DAF, regardless of the size of subset found and the k parameter of the employed FS criterion. Especially the low CWr values in case of SFFS suggest that SFFS tends to over-fit too much.

Figure 4-17: Comparing SFFS and DAF stability and performance when selecting 5 or 15 features out of 37, with k-NN for k=1, 2, 3. Diamonds show k-NN accuracy achieved; other lines show stability



Based on the aforementioned observation, we decided to select features using the DAF method. The goal of this second experiment is to identify which features are better than the others, possibly yielding information about how much better they are. The criterion is again the estimated accuracy of *k*-NN classifier, now for *k*=5 as higher k values proved better to further prevent over-fitting (cf. Theodoridis, Koutroumbas, 2006).

4.3 Introducing the modified feature selection methodology

In order to perform a feature selection process (analysis of the dependency between particular competitiveness factors and overall CFP over the available training data) using the chosen feature selection method in our particular case, we first need to impose a suitable model on the data that would describe the underlying data structure as accurately as possible. A good model can then be used to evaluate the quality of candidate subsets of characteristics in the process of feature selection. In the following, we consider two principally different approaches – classification and regression. The difference is in the measure of accuracy to be optimized. In the case of classification, we aim to identify such a subset of competitiveness factors, for which the model proves most accurate in distinguishing among the company groups A, B, and C (cf. Section 4.4). In the case of regression, we avoid the grouping and aim to minimize the prediction error of a dependent variable expressing the overall CFP.

Non-Parametric Model

From the vast battery of existing models (cf. Theodoridis and Koutroumbas, 2006) our choices are limited due to two specifics of our problem: the data is incomplete (roughly 5% of values are missing for various companies and various characteristics) and some of the available characteristics are non-numeric, i.e., their values are impossible to order. Another concern is the sample-size vs. dimensionality ratio, which in this case prevents application of models requiring large numbers of samples (mixture models or other multi-dimensional models, cf. Theodoridis and Koutroumbas, 2006, where the curse of dimensionality would quickly lead to over-training).

As reported in (Pudil et al., 2012), a suitable approach is the application of k-Nearest Neighbor idea (Cover, Hart, 1967). k-NN is the non-parametric classifier that imposes a non-linear model taking direct use of existing samples in the training data set. Its known disadvantage – the necessity to permanently store the complete data set – is not a disadvantage in our case due to the limited data size and off-line nature of our search process. On the other hand its advantage – a good model fit and accuracy in the case of a limited data size – makes it a good choice for our purpose. The only additional tool needed is a suitably defined distance function capable of expressing the distance between any two samples (companies) in the 36-dimensional space of our data set. The commonly used distance is the Euclidean distance. A modification of standard Euclidean distance is the basis of how we handle missing values and non-numeric values.

Handling Missing Values and Non-Numeric Values

The accuracy of the model is affected not only by its fundamental principle but also by its possible parameters or other setup details. In our case, the handling of missing

values and non-numeric values proves important. Both of these concerns are reflected in the definition of distance function used within the applied models.

We considered two approaches of missing value handling, as expressed by the definition of pseudo-Euclidean distance:

– "standard" one-time substitution of each unknown value by the mean value over all known values for the respective feature in case of numeric features, or substitution by the most frequent value in case of non-numeric values
– "pessimistic" handling of unknown values when computing distance; missing values are always interpreted as if the distance was maximum with respect to the respective feature.

4.4 Pattern classification approach

In the classification approach, the DAF feature selection method (cf. Chapter 2) is applied to maximize the estimated prediction accuracy of the k-NN classifier. In the course of the search, the DAF method generates candidate feature subsets, evaluates each of them and collects the results to obtain a final feature ranking. Each candidate subset is evaluated as follows. The k-NN classifier is used to classify each single known company to one of the groups A, B, C purely based on the characteristics in the current subset. The result of classification is in each case compared to the known assignment of the respective company to one of the classes. Eventually, the percentage of companies classified correctly (predicted to belong to the same group where they actually belong) is used as the measure of subset quality.

4.5 Regression approach and pseudo-kernel regression model

Considering the CFP as the dependent variable, we tried to predict this dependent value by means of a special non-linear regression model with other company characteristics such as independent variables. The error of the regression model with given independent variables (companies characteristics) as regressors, thus the prediction error, serves in our approach as a criterion of how good these independent variables are for approximating the real CFP of companies in the data set. Such a subset of characteristics (features in pattern classification terminology, regressors in mathematical statistics terminology) that provides the lowest regression error can be then regarded as the set of key factors of corporate competitiveness.

The only additional tool needed is a suitably defined distance function capable of expressing the distance between any two samples (companies) in the 37-dimensional (or 74-dimensonal) space of our data set. The commonly used distance

is the Euclidean distance. A modification of standard Euclidean distance is the basis of our handling of missing values and non-numeric values.

Applying regression instead of classification, as used in Pudil et al. (2012), has the potential benefit of preventing the impact of possible inaccuracies introduced when pre-processing data (at the moment of initial assignment of companies to classes). The regression model we have used is based on the "Kernel" regressor, where the dependent value is predicted as a linear combination over dependent values of existing samples, where more distant samples get a lower weight. "Kernel" in this context is the weighing function dependent on the distance among the point in question and existing samples (companies).

More exactly, the proposed *pseudo-kernel regression model* is analogous to Parzen kernel models with Gaussian kernels with the main difference being in the fact that we assume only one-dimensional kernels. The unknown dependent value y_k is predicted as the weighted average of dependent variables of known companies. The one-dimensional Gaussian kernel in our case serves as the weight in the definition of distance between two companies, $w(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The Gaussian kernel thus helps to progressively reduce the influence of more distant companies and emphasizes the importance of close companies when predicting the dependent variable y_k for k-th company. The *pseudo-kernel regression model* predicts the unknown competitiveness value y_k from neighbouring competitiveness values as follows:

$$y_k = \frac{1}{n} \sum_{i=1;\, i \neq k}^{n} w(\boldsymbol{x}_k, \boldsymbol{x}_i) \cdot y_i$$

The weighted distance between two companies is computed as follows:

$$w(\boldsymbol{x}_k, \boldsymbol{x}_i) = \frac{1}{\sigma \cdot m \cdot \sqrt{2\pi}} \cdot e^{-\frac{L_p(\boldsymbol{x}_k, \boldsymbol{x}_i)^2}{2(\sigma \cdot m)^2}}$$

where L_p denotes the L_p-distance defined as:

$$L_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt[p]{\sum_{f=1}^{D} |x_i^f - x_j^f|^p}$$

Unless stated otherwise, we assume p=2, i.e., we primarily use the Euclidean distance. Note that σ is estimated from the training set as the standard deviation of pairwise distances of all known companies. The parameter m is to be user-specified, or optimized in the pre-processing stage. Note that a higher value of m leads to a more smoothed-out model; a too high m would lead to a collapse of the model, which would then always predict the same average value of competitiveness. A lower value of m leads to a more detailed model that generalizes less. A too low m would collapse the

model to an analogy of 1-Nearest Neighbour. Let us mention that both the m and p parameters can be optimized to minimize the model prediction error (see Section 4.5).

Note that the above formulation of *pseudo-kernel regression model* is still based on the assumption of companies being represented by numerical vectors without missing values. The resolution of this problem consists in redefining the L_p distance. The problem of categorical variables can be diminished by assuming all numerical values to be normalized to [0,1]. In case of categorical variables, the contribution of a variable to the overall distance would then easily be 0 in case of equality or 1 in case of inequality of variable's values. The problem of missing values can then be resolved by restricting the computation of L_p to only such a subspace on which all values are known (provided the result is properly weighted). The other option is to interpret the missing value in one of two ways: either to indicate pessimistically that the distance is 1 (maximal), or to use the average distance computed over all cases where the variable is not missing.

To explain it less mathematically – we gradually explore increasing subsets of features (companies' characteristics) and use them as regressors for a special regression model. For a given subset of features, CFP for each company (serving as the dependent variable) is predicted by means of our regression model (see Figure 4-18). Now, let us consider a particular company. When trying to predict its CFP, we calculate it as a weighted sum of all the CFP's of respective companies. The weight reflects the distance of the point representing our particular company from the point representing the respective company.

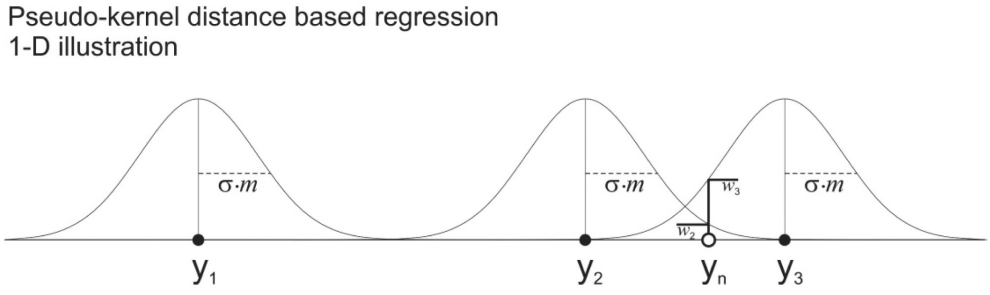
To summarize the explanation presented above, we consider two regression models following an analogous idea as discussed in Sect. 4.3:

1. 1-NN, where the dependent value for any point in the 36-dimensional space is predicted to be equal to that of such an existing sample that is the nearest to the point in question.
2. "Kernel" regressor, where the dependent value is predicted as a linear combination over dependent values of existing samples, where more distant samples get a lower weight. This model enables prediction based solely on the information about the distance between points in space; no coordinate information is used. This facilitates the definition of custom distance functions to accommodate features of various, even non-numeric types. The principle of the model is illustrated in Figure 4-18.

This model has been shown capable of achieving a prediction error lower than 3% in the paper by Pudil et al. (2013), despite the difficult modelled data set.

The idea of "kernel" regression can be viewed as analogous to the idea of k-NN classifier (cf. Sect. 3.3.1). Similar to k-NN, our kernel regressor imposes a non-parametric non-linear model that directly uses information from the available data. When compared to the original kernel regression idea (cf. Nadaraya, 1964; Simonoff,

Figure 4-18: Given existing samples S_1, .., S_3 and distance function d(), the yn value of sample Sn is predicted as sum over w_i * y_i for i=1, .., 3, where wi reflects $d(S_n,S_i)$ distance from 1-D Gaussian kernel centered in S_i



1996) we apply a slight modification. In order to accommodate the solutions to problems described in Sect 3.3.2, we apply only 1-dimensional Gaussian kernels on the space of sample distances. Kernel width is then optimized, starting initially from the estimated average distance between any two samples in the available data.

In the regression approach, the DAF feature selection method (cf. Chapter 3, Sect. 4.6) is applied to minimize the estimated error of predicted dependent values for each known company. The feature selection procedure is analogous to that described in Section 3.4.

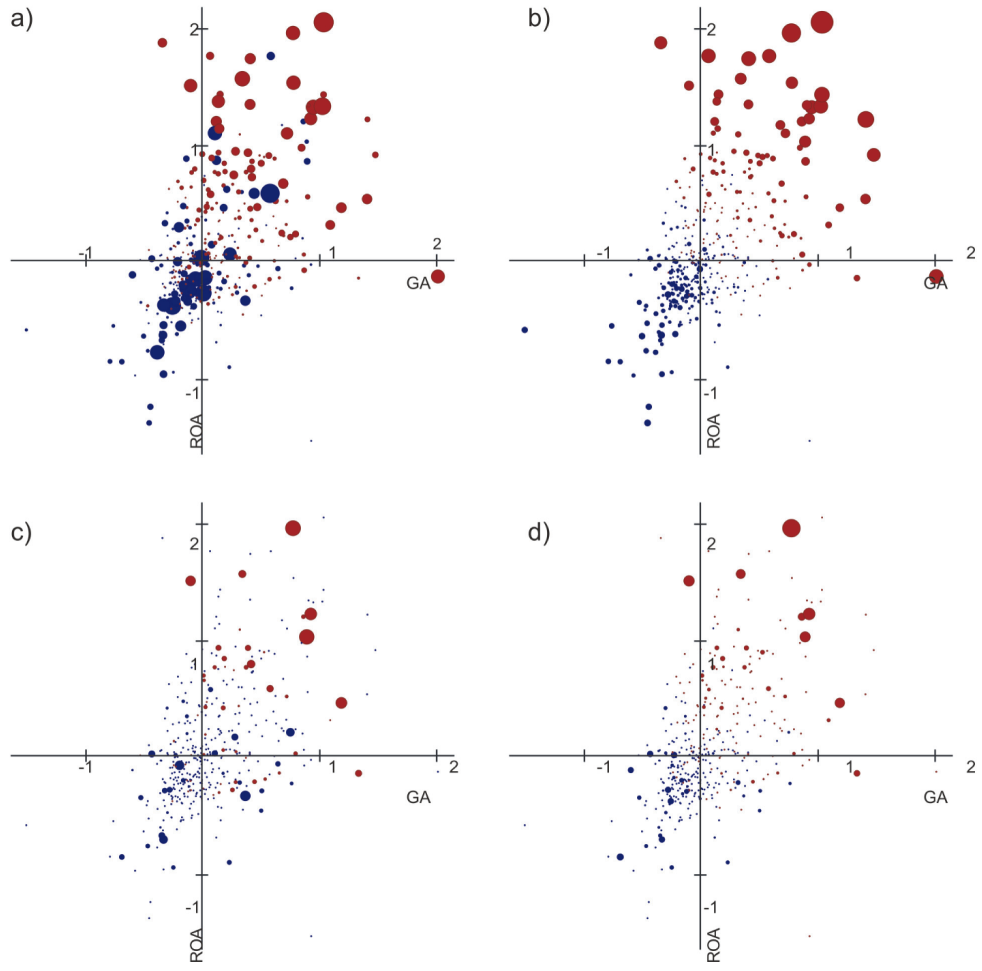
4.6 Experiments and results

We primarily considered four different regression models in our key experiments, obtained by combining two types of regression models (1-Nearest Neighbour and Kernel Regressor) and two types of missing value substitution (substitution by mean value, and pessimistic treatment of each missing value as an indicator of maximum distance). Feature ranking by DAF (Somol et al., 2011) has been computed in each of the four cases so as to minimize the average model error. The DAF method produced a weight for each feature, mirroring roughly the average feature ability to improve the criterion value on addition to a subset, evaluated over a large number and variety of contexts (various subsets). Additionally, we performed classification-based experiments for comparison purposes.

4.6.1 Regression-based analysis results

The best-achieved models of the four types as described in Section 3.5 are illustrated in Figure 4-19.

Figure 4-19: Errors of four regression models – each dot represents a company, positioned according to its Assets Growth (GA) and Return on Assets (ROA) values, a higher dot diameter depicts a higher regression error; black and grey colours depict its positive or negative value, respectively: a) 1-Nearest Neighbour with missing values substituted by mean values, b) Kernel Regressor with missing values substituted by mean values, c) 1-Nearest Neighbour with missing values treated pessimistically, d) (best) Kernel Regressor with missing values treated pessimistically.



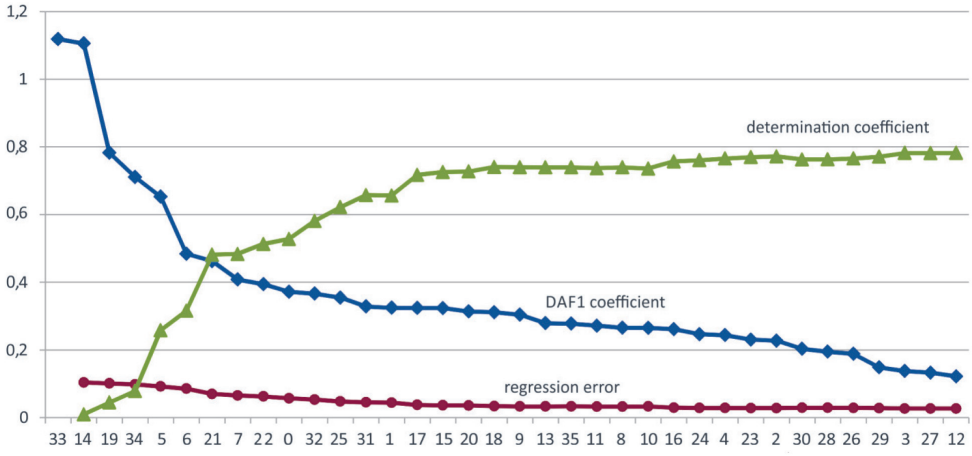
The single best model identified in our survey – Kernel Regressor with pessimistic missing value substitution – achieved the average error of $e = 0.0268903$ and determination coefficient $Dc = 0.781963$. The fact that a 1-dimensional kernel instead of a 36-dimensional kernel is used proved advantageous as it prevented over-fitting issues. Kernel regression thus proved more accurate in our experiments than simple 1-NN regression.

The highest accuracy was achieved with the set of all features. Removal of any feature led to a degradation of results. The result of feature ranking is thus of interest

to compare the importance of various company characteristics though it has not lead to any decrease of dimensionality.

The DAF method (cf. Somol et al., 2011) produced for each feature a weight, mirroring roughly the average feature ability to improve the criterion value on addition to a subset, evaluated over a large number and variety of contexts (various subsets). Figure 4-20 shows the weights obtained in the regression case, ordered in descending order. Note that the DAF weights obtained provide only the information about the relative quality of features when compared to other features. It can be seen that there are roughly up to 8 features that tend to improve criterion value considerably more than the others.

Figure 4-20: Importance of single company characteristics according to the best-achieved regression model. The graph represents growing subsets of features, features added according to the highest DAF1 coefficients. Note that model accuracy markedly improves after adding the first 8 features, and then after adding roughly the next 7 features.



The best 8 factors (regression based) in decreasing order:

Region

– DAF1 coefficient 1.11868, regressor error on single feature not computable due to missing values)

Ratio of Technical and Administrative Staff

– DAF1 coefficient 1.10539, regressor error on 2 best features 0.104015

Strategy

– DAF1 coefficient 0.782508, regressor error on 3 best features 0.101098

Legal Form *(legal form of the company)*

– DAF1 coefficient 0.710418, regressor error on 4 best features 0.0983053

Ownership Type *(5 types of ownership)*

– DAF1 coefficient 0.652801, regressor error on 5 best features 0.0920556

FDI *(domestic owner, foreign, both)*
– DAF1 coefficient 0.483909, regressor error on 6 best features 0.085747
Ration of Exports
– DAF1 coefficient 0.462479, regressor error on 7 best features 0.0698314
Owners in Top Management (yes/no)
– DAF1 coefficient 0.408192, regressor error on 8 best features 0.0653676

Regarded as factors of competitiveness within the framework of this pilot study, 5 out of 8 these features (denoted in italics) correspond with the previous results of Spalek and Castek (2010) based on different analyses of similar data.

4.6.2 Classification-based analysis results

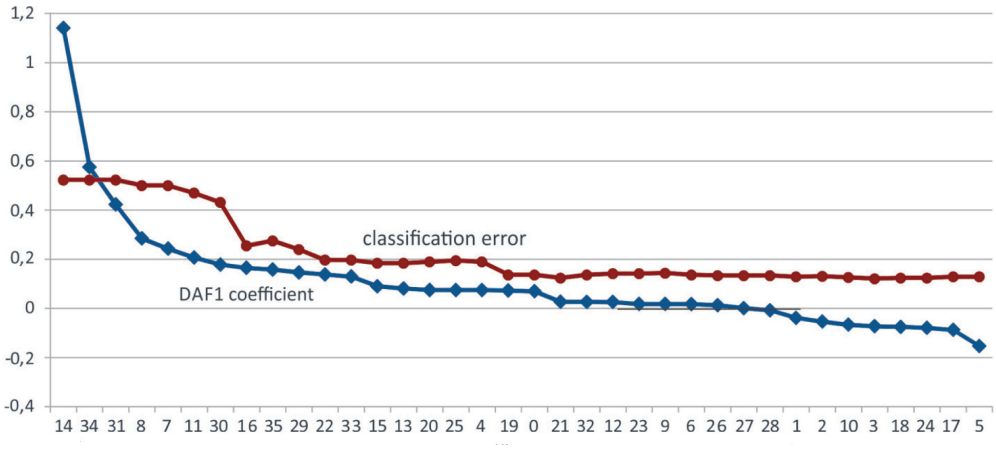
We performed a series of k-NN classifier based experiments for the values of k=1, 2, 3, 5, and two types of missing value substitution (substitution by mean value, and pessimistic treatment of each missing value as an indicator of maximum distance).

The single best classifier identified in our survey – 1-NN classifier with pessimistic missing value substitution – achieved the estimated classification accuracy of 0.873737 for the subset of 32 features out of the complete set of 36. With the complete set, the accuracy is 0.866162. However, as illustrated in Figure 4-21, roughly half of all features are enough to achieve classification accuracy, only negligibly worse than the best achieved. k-NN classifiers for k>1 tend to smooth out the decision boundary, i.e., reduce the influence of outliers at the cost of reducing sensitivity to detail. This effect proved disadvantageous in our case.

The DAF method (cf. Somol et al., 2011) produced for each feature a weight, mirroring roughly the average feature ability to improve criterion value on addition to a subset, evaluated over a large number and variety of contexts (various subsets). Figure 4-21 shows the weights in classification case, ordered in descending order; note that the DAF weights provide only information about the relative quality of features when compared to other features. It can be seen that there are roughly up to 8 features that tend to improve the criterion value considerably more than the others.

The best 8 factors (classification based) in decreasing order:
Ratio of Technical and Administrative Staff
– DAF1 coefficient 1.1404, 1-NN accuracy on the single feature 0.472222
Legal Form *(legal form of the company)*
– DAF1 coefficient 0.574625, 1-NN accuracy on best 2 features 0.472222
ISO 14000-certificate holding *(yes/no)*
– DAF1 coefficient 0.42274, 1-NN accuracy on best 3 features 0.472222
Assets Level
– DAF1 coefficient 0.284363, 1-NN accuracy on best 4 features 0.494949

Figure 4-21: Importance of single company characteristics according to best-achieved classification using 1-NN classifier and 3-class data. The graph represents growing subsets of features, features added according to highest DAF1 coefficients. Note that model accuracy markedly improves after adding the first 8 features, and then after adding roughly the next 10 features.



Owners in Top Management (yes/no)
– DAF1 coefficient 0.242903, 1-NN accuracy on best 5 features 0.494949
Software Applications – SCM module (yes/no)
– DAF1 coefficient 0.206311, 1-NN accuracy on best 6 features 0.525253
Ethical Code Adoption (yes/no)
– DAF1 coefficient 0.177655, 1-NN accuracy on best 7 features 0.563131
Share of Performance-Related Pay
– DAF1 coefficient 0.164487, 1-NN accuracy on best 8 features 0.739899

Regarded as factors of competitiveness within the framework of this pilot study, 4 out of 8 these features (denoted in italics) correspond with the previous results of Spalek and Castek (2010) based on different analyses of similar data.

4.7 Comparing Regression-based and Classification-based analysis results

When comparing the results achieved by both the approaches discussed, we should bear in mind that our goal is not prediction but identification of "informative" factors. Thus, the error in predicting the CFP or in classifying a company is not critically important; what matters more is which factors are assigned to the group of key factors influencing the CFP and thus corporate competitiveness. In this context, we should not be too surprised that the regression approach and the classification one yield somewhat different results. We should keep in mind that the tasks solved by the

respective approaches are not absolutely the same. The way of defining the CFP, implicitly hidden in these tasks, is different. In the regression approach, the CFP is defined by a combination of GA and ROA and attains continuous values. On the other hand, in the classification approach, the CFP is discretized by means of clustering the companies into distinct classes.

Clearly, the classification approach does not utilize "fine" information about the value of CFP as the regression one does. On the other hand, as a generally accepted unique definition of CFP does not exist, clustering its values into distinct classes can also be a reasonable solution. This clustering certainly represents a simplification, representing a certain loss of information but at the same time having the potential benefit of giving more comprehensible information to the observer: the financial performance of a given company (and thus its competitiveness) is a good, average or bad one.

In any case, by comparing these different approaches and analysing the results, we can get a more robust insight into the task.

4.8 Improved Model for Attribute Selection on High-Dimensional Economic Data

This section presents a newly improved procedure for estimating the model parameters, allowing us to achieve a significantly higher precision in feature subspaces. In other words, it improves the results when searching for a small set of factors of competitiveness. The performance of this improved procedure is illustrated in experiments with both the old multidimensional data (D=37) used in previous experiments, and with new data of higher dimensionality (D=74), which enabled us to reveal a not previously considered fact related to the estimates of the multiplication constant of the pseudo-kernel model.

4.8.1 Improvements of the regression model

Theoretical considerations verified in the course of experiments proved that the prediction accuracy of the pseudo-kernel non-linear regression model employed can be improved under certain conditions. These improvements are in possible adjustments to the distance function used and in optimizing the kernel width. They will be discussed in the following.

Varying the distance function

In the paper by Pudil et al. (2013), we assumed the standard L_2 (Euclidean) distance to be used when evaluating distances between samples. Let us recall that for k =1, 2, … the higher order distance in D-dimensional space is defined as

$$L_k = \sqrt[k]{x_1^k + x_2^k + \ldots + x_D^k}$$

In Figure 4-22 we illustrate that for the specific problem of corporate competitiveness, the L_2 distance data is not necessarily the best option.

Figure 4-22: Comparing Euclidean distance to alternative distances. Higher-order distance functions improve the accuracy of the kernel-based regression model on subspaces roughly of up to D/2 where D is the total number of features

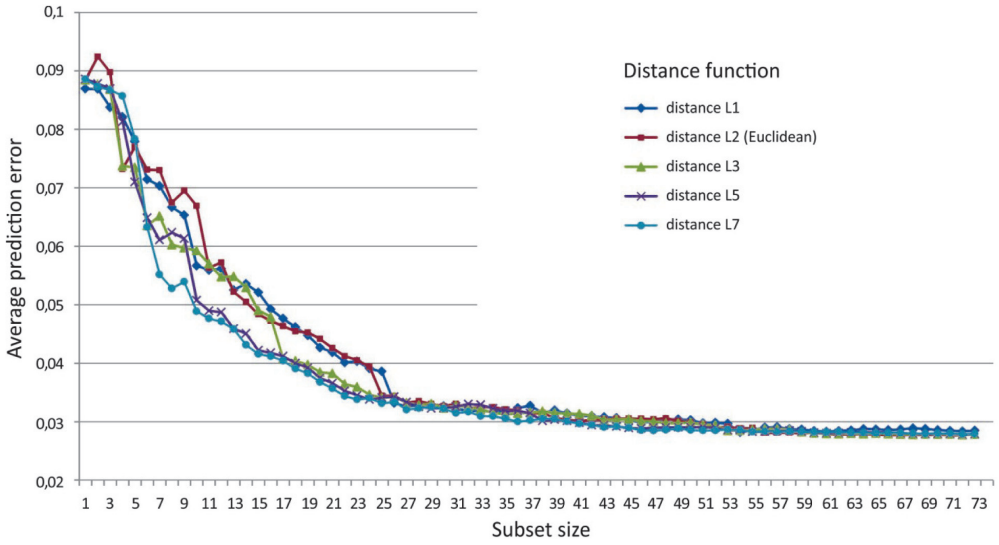
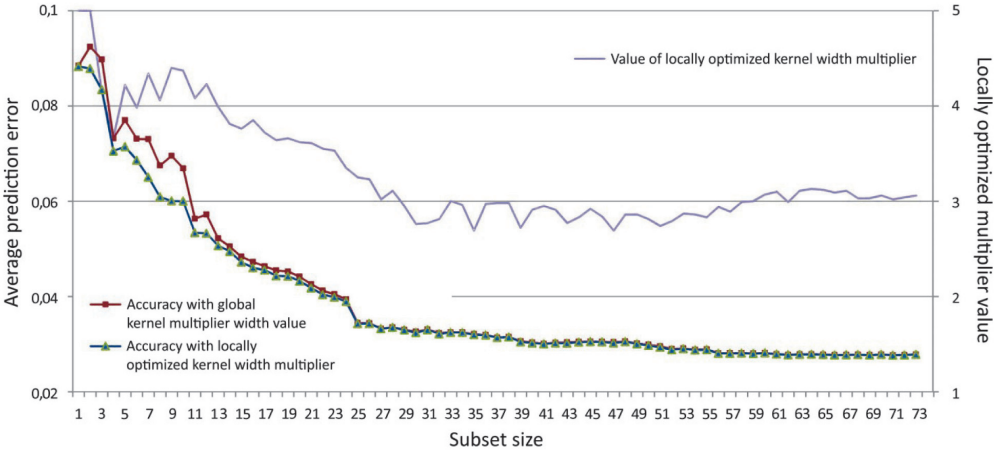


Figure 4-23: Comparing accuracy of a pseudo-kernel regression model with a globally optimized multiplication constant m to a model with a locally optimized m.
Note that a local optimization of m (optimization on subspace instead on full space) improves accuracy when a small number of features is used



Note that higher-order distances of up to L_7 proved capable of improving the model accuracy when a small number of features is used. We could not improve the best-achieved prediction accuracy, which remains roughly the same when all 74 features are used. However, if the number of features considered is restricted to 25, the improvement in accuracy is clearly visible.

The impacts of varying the distance function are presented in the following Table 4-4. The "DAF" columns contain the values of DAF coefficients of corresponding features where these features are ordered in the decreasing order of their DAF

Table 4-4: Comparing feature orderings yielded by model using Euclidean distance and L_7 distance

74-dimensional data, Euclidean distance			74-dimensional data, L_7 distance		
DAF	**Feature**	**Error on subset**	**DAF**	**Feature**	**Error on subset**
0.00378051	Span of Control	0.0884389	0.00869157	Strategy	0.0885776
0.00377241	Ratio of Women	0.0924165	0.00868319	Motivation of Workers	0.0872776
0.00369803	Ratio of Technical and Administrative Staff	0.0897509	0.0085613	Motivation of Top Management	0.0867333
0.00363094	Motivation of Workers	0.073188	0.00740714	FDI	0.0856955
0.00350262	Ratio of Workers	0.0769846	0.00706415	Ratio of Exports	0.0783415
0.00350031	Motivation of Top Management	0.0730957	0.00692042	Span of Control	0.0632577
0.00330279	Span of Control	0.0730181	0.00635151	Ratio of Women	0.0551528
0.00326192	Ratio of Exports	0.0674859	0.00625303	Ratio of Technical and Administrative Staff	0.0527764
0.00324774	Company Size (number of employees)	0.0695187	0.0062338	Ratio of Workers	0.0539102
0.00318828	Ratio of Graduates	0.0668979	0.00591669	Reasons for Staff Turnover	0.0488559
0.00310982	Strategy	0.0563008	0.00566361	Number of management levels per employee	0.0476115
0.00269444	Intensity of Motivation Means Use	0.0572132	0.00551711	Ratio of Graduates	0.0471201
0.00255556	Specificity of Supplies	0.0521837	0.00540293	Foreign Supplies	0.0458394

coefficients. The "Error on subset" means the error of the pseudo-kernel regression model that is based on all the features in the preceding rows including the current feature. The first three columns represent the values for the Euclidean (L_2) distance while the last three columns show the results for the L_7 distance used.

The higher the DAF coefficient (explained in the previous chapter) of a feature is, the more informative (thus more important) the feature is. In the regression approach, the DAF feature selection method is applied to minimize the estimated error of predicted dependent values for each known company.

Kernel width is then optimized, starting initially from the estimated average distance between any two samples in the available data.

Table 4-4 shows the impact of the improved model on the results of feature selection. The ordering of selected features changes notably, but generally those features identified as important using Euclidean distance (L_2) are identified as important also when using L_7 distance, with most of the difference being in relatively limited order shifts. L_2 distance outperforms L_7 in very small subsets (up to 5 features). After that, L_7 is distinctively better than L_2. This can be observed in up to roughly 25 features, after which the error on subsets becomes very similar (these results are not shown due to the paper length limit). The DAF coefficient is, on the other hand, much higher while using L_7 distance from the very beginning.

Kernel width multiplication constant

In the paper by Pudil et al. (2013), the default width of all kernels was set to the estimated standard deviation of the distances between all pairs of samples in the available data set. In Figure 4-18 we suggest that kernel width can be adjusted by adding the multiplication constant m. Note that for values m>1 the model gets smoother, less prone to over-fitting, but also less sensitive to detail. In contrast, values m<1 lead to a more detailed model capable of capturing more detail but at the same time with degraded generalization ability.

In the above-specified paper the constant m was optimized on a full set of features and then fixed globally at the beginning of the training process. In Figure 4-23 (and also Table 4-5) we show that it is possible to improve the accuracy of the model when only a smaller subset of features is used (or searched for), if the constant is optimized specifically for the model on the considered subspace.

The reason for the positive effect of *m* optimization on subspaces is simple; the average distance between points in space increases with increasing dimensionality. Applying a sub-space model with an m constant optimized on a full feature set leads to a model that is smoother than expected, thus reducing its sensitivity to detail.

Note that marked improvement by a local optimization of *m* can be expected only if the difference between the total number of features and the number of features in the considered subspace is large. Only then the inherent difference in distances comes

into play. This effect is observable for subsets of roughly up to 25 features with our 74-dimensional data.

The impacts of optimizing the kernel width are presented in the following Table 4-5. The "DAF" columns and the "Error on subset" have the same meaning as in Table

Table 4-5: Comparing feature orderings yielded by a model using a globally versus locally optimized kernel width multiplier constant. Euclidean distance is used in both cases.

74-dimensional data, globally optimized kernel width multiplier constant *m=3.06*			74-dimensional data, locally optimized kernel width multiplier constant *m*	
DAF	**Feature**	**Error on subset containing all features up to the current**	**Error on subset containing all features up to the current**	**Locally optimized m value**
0.00378051	Span of Control	0.0884389	0.0882787	5
0.00377241	Ratio of Women	0.0924165	0.0878285	5
0.00369803	Ratio of Technical and Administrative Staff	0.0897509	0.0834008	1.14
0.00363094	Motivation of Workers	0.073188	0.0705265	3.67
0.00350262	Ratio of Workers	0.0769846	0.0714226	4.22
0.00350031	Motivation of Top Management	0.0730957	0.0685935	3.98
0.00330279	Number of management levels per employee	0.0730181	0.0650337	4.34
0.00326192	Ratio of Exports	0.0674859	0.0608841	4.06
0.00324774	Company Size (number of employees)	0.0695187	0.0600489	4.4
0.00318828	Ratio of Graduates	0.0668979	0.0600023	4.37
0.00310982	Strategy	0.0563008	0.0533632	4.08
0.00269444	Intensity of Motivation Means Use	0.0572132	0.0532684	4.23
0.00255556	Specificity of Supplies	0.0521837	0.0506469	3.99

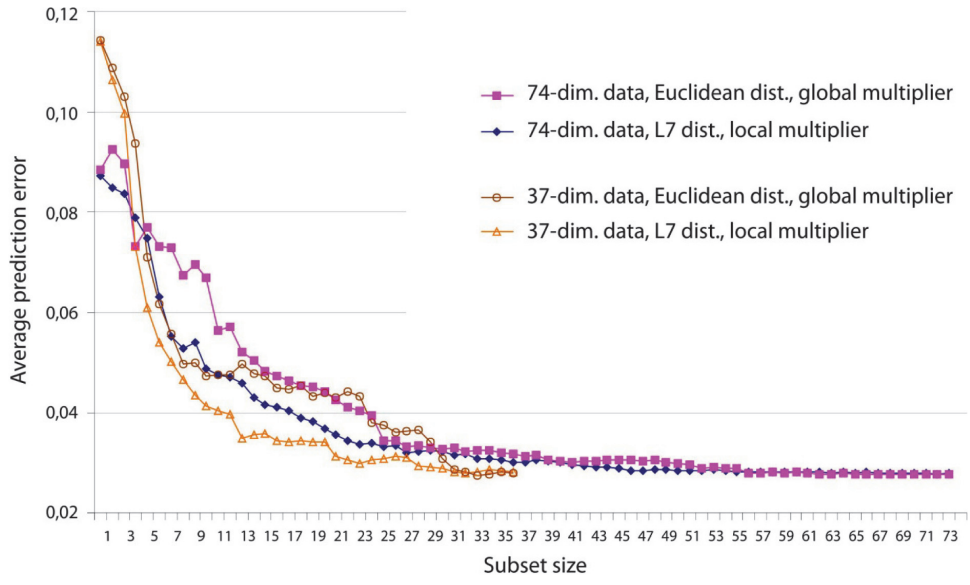
4-4. The first three columns represent the values for the kernel width multiplier constant optimized globally while the last three columns show the results for the same multiplier constant optimized locally.

4.8.2 Optimized model performance on 37- and 74-dim data

We applied the optimization described above to two datasets: the 37-dimensional data used in a paper by Pudil et al. (2013) and to newly acquired 74-dimensional data representing the same domain by a wider set of differently defined features. In Figure 4-24 we provide two comparisons. First, we compare model accuracy on a 37-dimensional dataset to the accuracy on a 74-dimensional dataset. Second, we compare the model accuracy for each of the two datasets in the default setting as described in (Pudil et al., 2013) to the model tuned both with respect to the employed distance measure and with respect to the locally adjusted multiplication constant (see Section 4.8.1).

It is apparent from Figure 4-24 that the proposed improvements to pseudo-kernel regression model did not improve the best achievable prediction error, but it did notably improve model accuracy on subspaces. With 37-dimensional data, all results for subsets of up to 30 features proved to be notably better when the improved model was used. With the 74-dimensional data, the most notable improvement took place on subspaces of up to roughly half of the total number of features. This means the

Figure 4-24: Comparing the accuracy of default model to that of the improved model. Comparison provided separately for 37- and 74-dimensional dataset representing the same domain. Default configuration compared to the best configuration identified in this paper



improvement is achieved when a smaller set of key factors of corporate competitiveness is to be identified.

Also note in Figure 4-24 that the best achievable model setting produces practically equal accuracy if all new 74 features are used or if all old 37 features are used. This can be explained by mutual statistical dependencies of features which become redundant; thus, adding more and more features does not increase the overall information value. The new feature set, however, allows the selection of a small subset that performs better than an equally sized selected subset of the old set. This is apparent in Figure 4-24 for subsets of up to 4 features where the model on 74-dimensional data gives a lower prediction error.

The impacts of extending the original 37-dimensional dataset into a 74-dimensional dataset are presented in the following Table 4-6, with the columns having the same meaning as in the preceding tables.

Table 4-6: Top 25 features selected using the optimized model on 74-dim and 37-dim data

74-dimensional data, L7 distance, localized kernel width multiplication constant			37-dimensional data, L7 distance, localized kernel width multiplication constant		
DAF	**Feature**	**Error on subset**	**DAF**	**Feature**	**Error on subset**
0.00869157	Strategy	0.0885253	0.0225594	Ratio of Technical and Administrative Staff	0.113963
0.00868319	Motivation of Workers	0.0849643	0.0157939	Expenses on Employee`s Benefits	0.106365
0.0085613	Motivation of Management	0.0837816	0.0141935	Expenses on Employee`s Education	0.0997067
0.00740714	FDI	0.0788026	0.0141366	Motivation of Top Management	0.0732506
0.00706415	Ratio of Exports	0.0749111	0.0134088	Strategy	0.0609535
0.00692042	Span of Control	0.0632048	0.0118486	FDI	0.0540098
0.00635151	Ratio of Women	0.0551528	0.0116693	Ratio of Graduates	0.0503014
0.00625303	Ratio of Technical and Administrative Staff	0.0527764	0.0114994	Ratio of Exports	0.0467132
0.0062338	Ratio of Workers	0.0539102	0.0100634	Number of management levels per employee	0.0434888

0.00591669	Reasons for Staff Turnover	0.0488559	0.010026	Specificity of Supplies	0.0414148
0.00566361	Number of management levels per employee	0.0476115	0.00971683	Specificity of Products	0.0405086
0.00551711	Ratio of Graduates	0.0471201	0.00963302	Company Size (number of employees)	0.0397118
0.00540293	Foreign Supplies	0.0458394	0.00932846	Owners in Top Management	0.0349433
0.005298	Specificity of Supplies	0.0431305	0.00908389	Stability of Customers	0.0356843
0.00515788	Access to Funding	0.0415497	0.00904028	Rate of Staff Turnover	0.0359644
0.00509582	Company Size (number of employees)	0.0411758	0.00899967	Supplier Selection Criterion: CSR Compatibility	0.0345245
0.00498899	Specificity of Products	0.0404037	0.00886835	Foreign Supplies	0.0341523
0.00488013	Level of Corruption	0.0390621	0.00855376	Region	0.0344136
0.00477958	Maturity of the Company	0.038239	0.00845204	Legal Form of the Company	0.0342419
0.00475404	CSR	0.0367709	0.00844075	Supplier Selection Criterion: Quality Certificate	0.0342001
0.00450035	Importance of Creditors	0.0356523	0.00828733	OHSAS Certificate Holding	0.0312972
0.00447213	Stability of Customers	0.0343655	0.00822393	CSR	0.030517
0.00442313	Effect of Employee`s Benefits	0.0337949	0.0079642	Assets Level	0.0299106
0.00442278	Intensity of Motivation Means Use	0.0339987	0.00782862	Stability of Suppliers	0.0304842
0.0042628	Other Costs	0.0331099	0.00754582	Ownership Type	0.0308387

4.9 Conclusions

It can be stated that the results presented here confirm the claim of the usefulness of applying learning approaches in searching for factors of corporate competitiveness, which was stated at ICMLG 2013 in Bangkok (Pudil, 2013). Though the methods of machine learning and feature selection (FS) are related to statistical methods in particular, they are less strict with respect to assumptions about the data analysed.

The core idea of how to derive factors of corporate competitiveness is based on selecting those features (variables characterizing companies) which have the highest importance in a special multivariate regression model predicting the CFP and minimizing the error between the predicted and real value of CFP. Therefore, the factors of corporate competitiveness have a straightforward meaning – they are just those features which have the highest impact on corporate financial performance and thus on corporate competitiveness.

We have described several modifications to the pseudo-kernel regression model used in Pudil et al. (2013) for evaluating the importance of corporate competitiveness factors. The modifications proved capable of notably improving model accuracy if applied on subspaces. In other words, if a small subset of factors is to be evaluated, then the improved model allows selecting factors that provide more accurate information than subsets selected in the paper by Pudil et al. (2013).

We have also compared the results on original 37-dimensional data and newly acquired 74-dimensional data obtained on the same domain. Though the effort invested to define more features failed to result in direct improvement of the best achievable accuracy in the full feature space, some newly included features helped to yield better results in very small subspaces.

The implication of our findings for the search for the factors of competitiveness is as follows. One cannot expect competitiveness on the corporate level to be a result of one particular cause. The whole body of literature on this topic works with the notion that competitiveness is a complex phenomenon with multiple causes (Beranova, 2008, Andrews et al., 2010 and others). Our results point to this with the error on subsets decreasing to the very last feature included and evaluated. This can be interpreted in the way that each of the features (company characteristics) has its contribution to the prediction of CFP. However, in practice it would be difficult if not next to impossible to make changes to a business according to e. g. 74 characteristics as in our regression model. A smaller set of characteristics is suitable for this, making the improved model more useful as it selects features with better prediction power. This means that these characteristics are most important for a given set of companies and thus can be regarded as factors of their competitiveness.

To generalize the contribution of our findings, for a researcher it lies in the possibility of processing high-dimensional data in an easier manner and with a better reliability of the result. If applied in the field of competitiveness, then the result points

at characteristics important for the practitioner. The practitioner can therefore adjust his or her own business according to the values of the characteristics found to have the strongest influence on the business performance.

5 Identifying factors of competitiveness using bivariate analyses and linear regression analyses

This chapter explores bivariate associations of the individual variables included in the experiments with the financial performance of companies. In addition, it formulates partial linear regression models. Both are arranged in units corresponding to the division of the original questionnaire survey that followed the logic of the stakeholder view of an enterprise. This means that the corporate characteristics are analysed in groups related to the owners, employees, customers, suppliers, and business environment. However, they are preceded by sections regarding potential internal and external factors of competitiveness and a section on the overall approach to the system of business management. Each of these units represents an independent sub-chapter and a separate partial linear regression model in the following text. These analyses serve as the basis for interpreting the results of the DAF experiments, since these do not provide simple instructions explaining the influence of the predictors examined, as is the case of e.g. linear regression models or t-tests, analyses of variance, or correlation analyses.

Therefore, we will use independent samples t-tests, an analysis of variance, correlation analysis, and linear regression models in this chapter, using Pearson's r, Spearman's r_s and Kendall's τ_c coefficients. While Pearson's r is preferably used when assumptions of this test are met, due to the direct comparability of the results of the correlation analysis with regression models, Spearman's r_s is used when the sample does not follow bivariate normal distribution but still the relationship between two variables is monotonous. Kendall's τ_c is used in the analyses where the independent variable is an ordinal variable with fewer categories, resulting in tied ranks. Kendall's τ_c should be used in such situations (deVaus, 2002; Field, 2009). Coefficient η will be used to report the effect size of associations examined by t-test and by ANOVA and is calculated as follows

for t-test:

$$\eta = \sqrt{\frac{t^2}{t^2 + df}}$$

where
t denotes t-test statistic
df denotes degrees of freedom

for ANOVA:

$$\eta = \sqrt{\frac{SSA}{SST}}$$

where
SSA denotes Sum of squares between groups
SST denotes Total sum of squares)

All associations examined are evaluated at the level of significance of α = 10%, thus reducing the chance of making a Type I error. However, the actual p-values observed are reported for each test together with the number of observations involved in the particular test, together with the effect size found by this test. Thus, our reader can kindly evaluate our results.

5.1 General characteristics

Before we start examining the influence of the variables concerned, it is also necessary to consider the influence of the general characteristics of enterprises that can affect the particular variables or that can modify or disguise their impact on financial performance, or those which give the appearance of such an action. These are particularly the industry in which the company operates, and the size of the company. These two characteristics are usually considered control variables in examining the factors of competitiveness or the financial performance of a company. This is true for a wide range of studies on completely different areas of financial performance factors. We can mention, for example, works by Coles et al. (2012) *Structural Models and Endogeneity in Corporate Finance: The Link between Managerial Ownership and Corporate Performance, Youndt et al. (1996) Human Resource Management, Manufacturing Strategy, and Firm Performance, or Powell and Dent-Micallef (1997) Information Technology as Competitive Advantage: The Role of Human, Business and Technology Resources.*

Effect of company size on the financial performance

Company size is usually measured by the number of employees, the volume of total capital or the volume of total sales. In our case we consider total capital or sales to be rather variables describing the response variable, i.e. business performance. Therefore, a good representative of company size is the number of employees. However, in our sample (as well as in the population), the number of employees is significantly skewed

with a predominance of values on the left, i.e. in smaller businesses and with significant outliers towards large enterprises. Thus, extreme values (up to 5,600 employees) were Winsorized. In some cases it may be preferable to use categorized company sizes, dividing enterprises into categories of 50 to 99 employees, 100 to 249 employees, and 250 or more employees. As the following table shows, such a division is more uniform.

Table 5-7: Frequency distribution of a categorized company size

Valid	Frequency	Percent	Valid Percent	Cumulative Percent
50–99	128	29.6	29.6	29.6
100–249	175	40.5	10.5	70.1
>250	129	29.9	29.9	100.0
Total	432	100.0	100.0	

Since the vector of financial performance variable and size of a company variable measured by the number of employees strongly violates the assumption of two-dimensional normality, we use Kendall's coefficient τ_c to measure the association between the size and the financial performance. The influence of categorized company size is measured by an analysis of variance and its effect by coefficient η.

Table 5-8: The association between the company size and the financial performance

Variable	Coefficient	Association	p (two-tailed)
Categorized size	η	0.122	0.046
Number of employees	τ_c	-0.068	0.032

Both ways of expressing company size were significantly related to financial performance. In both cases, the measure of association is low (one might say trivial in the case of the number of employees) and in both cases it is also true that larger companies show worse financial performance. The average performance of small businesses in the sample is m = 0.416, m = 0.247 for medium-sized businesses, and m = 0.117 for large enterprises.

Effect of industry on the financial performance

The industry in which a company operates can be expressed in our sample at the level of the first two digits of CZ NACE, or at a less detailed level as manufacturing versus construction. The first division is of course finer, but even for the sample size of 408 enterprises there are industries in which the number of observations is very low – in up to seven categories it is lower than 10. Therefore, the bigger part of our analyses uses the less detailed division to manufacturing and construction. However, the table below shows associations and p-values for both independent samples t-test (manufacturing/construction categories) and analysis of variance (CZ NACE categories).

Table 5-9: The association between the industry and the financial performance

Variable	Coefficient	Association	p (two-tailed)
Industry: Manufacturing/Construction	η	0.101	0.039
Industries by the first two digits of CZ NACE	η	0.325	0.001

It is obvious that in both approaches the association with financial performance is statistically significant. For the less detailed classification of manufacturing and construction, the effect of the industry is not very large but still significant. We find a higher performance in construction (m = 0.468) than in manufacturing (m = 0.217). The effect is larger for the finer classification and it can be described verbally at least as medium. If we examined the differences between particular categories, we could find statistically significant differences only between certain categories. For example, in the field of "Production of electrical machines and equipment" or "Production of radio, television and communication equipment and devices", the average financial performance is relatively high, about m = 0.90. On the other hand, in the field of "Production of basic metals and fabricated metal products" and "Production of textiles and textile products" the average financial performance is relatively low, m = -0.386, or m = -0.267, respectively.

Interaction of company size and industry

The data in our sample clearly show that there is an association between control variables. In construction, the companies are considerably smaller (m = 173) than in manufacturing (m = 267), which is a statistically significant difference (p = 0.001). Naturally, the question arises whether the zero order relationships, i.e. the size affecting financial performance, and industry affects financial performance, are true. This can be verified by examining the conditional associations and correlations. Let us assume

that the industry affects the size but not vice versa. In this case, we must examine the authenticity of the association between the industry and financial performance by dividing the sample based on the categorized company size (if the variable describing industry was an interval, we could use partial correlations), and the authenticity of the association between the size and financial performance by dividing the sample based on the industry. The following table shows the former conditional associations.

Table 5-10: The association between the industry and the financial performance, controlling for the size of the company: statistics

Company size / industry => financial performance			Association	p
50–90	Nominal	η		n.s.
	Number of observations		120	
100–249	**Nominal**	***η***	**0.138**	**0.071**
	Number of observations		**171**	
250+	Nominal	η		n.s.
	Number of observations		120	

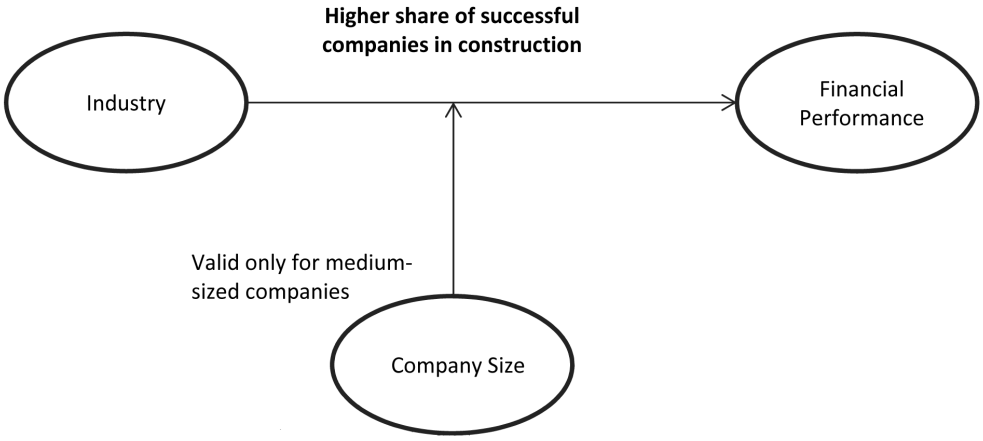
The zero order relationship examined is maintained only in the category of medium-sized businesses whereas in the remaining categories it is statistically insignificant. For medium-sized businesses it is slightly stronger than it seemed to be in the whole sample ($\eta = 0.138$ vs. $\eta = 0.101$). We can assess the substantive differences in the performance in various industries and size categories in the following table.

Table 5-11: The association between the industry and the financial performance, controlling for the size of the company: differences in CFP

Company size	Industry (Manufacturing /Construction)	N	Mean CFP	Std. Deviation	Std. Error Mean
50–99	Manufacturing	92	0.338	1.044	0.109
	Construction	28	0.552	0.751	0.142
100–249	Manufacturing	134	0.193	1.006	0.087
	Construction	37	0.531	0.991	0.163
>250	Manufacturing	105	0.144	0.944	0.092
	Construction	15	0.156	0.641	0.166

The biggest differences in performance based on industry were observed for medium-sized companies, while the difference for small companies is less than half compared to medium-sized enterprises, and there is almost none difference for large companies. Therefore, we cannot say that company size would explain the variability of financial performance better than the industry. More likely, the industry explains the financial performance only for medium-sized companies. We can illustrate this using a model in the chart below.

Figure 5-25: The effect of industry on financial performance controlled for the company size



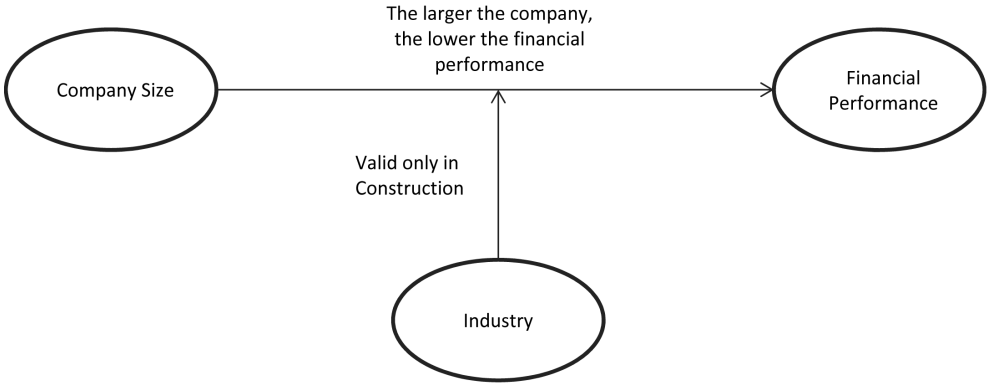
As mentioned above, it is also necessary to examine the validity of association between the company size and financial performance. In the entire sample, this associtation was rather weak. The following table shows the conditional association between company size and financial performance, if the sample is divided according to the industry.

Table 5-12: The influence of company size on financial performance, controlling for the industry: statistics

Industry / Company size => financial performance			Correlation	p
Manufacturing	Ordinal	τ_c	-0.031	0.386
	Number of observations		329	
Construction	**Ordinal**	**τ_c**	**-0.122**	**0.089**
	Number of observations		**79**	

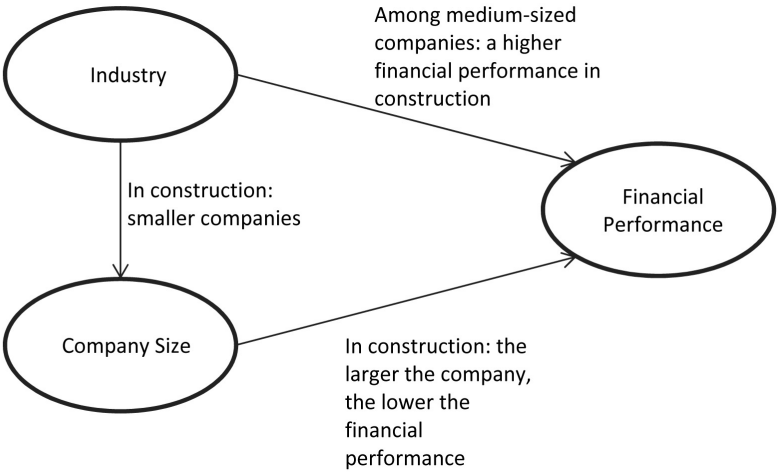
The company size, measured by the number of employees, shows a statistically significant relationship with financial performance only in the construction industry. Here, however, it is about twice as strong as it was originally throughout the sample. On the other hand, this relationship is statistically insignificant in the manufacturing industry. Again, we cannot say that the zero-order correlation is false. The third variable, the industry, only modifies the effect of the variable originally examined. The model in the chart below shows the relationships mentioned above.

Figure 5-26: The effect of company size on financial performance controlled for the industry



When we combine what was mentioned above, our findings then support the model of a direct effect of the company size and industry on financial performance, which is in both cases modified by a third variable. All the statistically significant associations found are shown by the model in the chart below.

Figure 5-27: The combined effect of company size and industry on financial performance



5.2 Internal competitiveness factors of a company

The first part of the questionnaire focused on how respondents subjectively perceive the potential internal factors of competitiveness. In nine closed questions, the respondents rated their company in comparison with its direct competitors, using a scale from 1 – a distinctively lower value – to 5 – a distinctively higher value. The following is a list of questions with their average values and correlation with the financial performance of the company:

Table 5-13: Variables describing the internal environment of a company: bivariate correlations with CFP

How do you rate, compared to your competitors, your...	Average value	Pearson's *r*	p (two-tailed)	N
Innovation Activity	**3.37**	**0.157**	**0.002**	**399**
Product Adjustment	**3.89**	**0.100**	**0.044**	**404**
Product Quality	**3.74**	**0.129**	**0.010**	**403**
Labour Costs	2.94	0.089	0.076	398
Other Costs	**2.95**	**-0.081**	**0.110**	**392**
Labour Qualification	**3.35**	**0.122**	**0.014**	**398**
Customer Care	3.59	0.048	0.343	400
Access to Funding	**3.33**	**0.150**	**0.000**	**379**
Company's Goodwill	**3.52**	**0.070**	**0.165**	**400**

Scale from 1 – a distinctively lower value – to 5 – a distinctively higher value.

The p (two-tailed) column in the table above shows p-values of two-tailed tests. However, the research in this part evaluates directional hypotheses that seek to demonstrate that the increasing value of the variable also increases the financial performance. Exceptions are Labour Costs and Other Costs, which were expected to demonstrate the negative direction of the correlation, i.e. that the financial performance increases with the decreasing values of these variables. Nevertheless, this trend was not confirmed for Labour Costs, so it is necessary to use the one-tailed p-value of $p = 0.962$ (1-p/2). On the contrary, for other variables, it is possible to divide the two-tailed p-value by two, which will show Other Costs and Company's Goodwill statistically significant at the level of $\alpha = 10\%$ also.

Regarding the size effect of relationships that can be generalized, it ranges from trivial/none (Other Costs $r = 0.081$) to low (Access to Funding $r = 0.180$). So, it is not possible to talk about any important factor of financial performance of a company at the level of simple bivariate correlation in this area of company characteristics.

Multidimensional model

To assess the degree of the joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model in this form: Financial performance = β_0 + β_1*Innovation Activity + β_2*Product Adjustment + β_3*Product Quality + β_4*Labour Costs + β_5*Other Costs + β_6*Labour Qualification + β_7*Customer Care + β_8*Access to Funding + β_9*Company's Goodwill + β_{10}*Industry + β_{11}*Company Size Medium-sized + β_{12}*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-14: Summary of the internal factors regression model

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.338	0.114	0.083	0.892	0.114	3.666	12	342	0.000

The value of multiple correlation could be considered moderate, as the model explains about 11% of the variability of financial performance. Still, the model is significant. Estimates of the parameters of this model are listed in the following table.

In this case there are fewer variables that statistically significantly affect the financial performance at the level of $\alpha = 10\%$. This is true for Innovation Activity, Other Costs, and Access to Funding; in addition, it includes all the control variables. The higher financial performance – if the values of the other variables remain unchanged – is related with by higher Innovation Activity, lower Other Costs, and better Access to Funding. Furthermore, the construction industry is more profitable than manufacturing (the reference value of zero is represented by the manufacturing industry) and the larger the company is, the lower its financial performance is

(measured by relative indicators of ROA and asset growth). Access to Funding and Company Size are the strongest predictors in this model, followed by Innovation Activity and Industry, being approximately at the same level, while the weakest predictor is the level of Other Costs.

Table 5-15: Coefficients of the internal factors regression model

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	-0.583	0.368		-1.587	0.113
Innovation Activity	**0.088**	**0.049**	**0.098**	**1.777**	**0.076**
Product Adjustment	0.053	0.061	0.049	0.873	0.383
Product Quality	0.082	0.074	0.068	1.111	0.267
Labour Costs	0.072	0.056	0.073	1.285	0.200
Other Costs	**-0.120**	**0.065**	**-0.102**	**-1.852**	**0.065**
Labour Qualification	0.062	0.072	0.051	0.867	0.386
Customer Care	-0.054	0.059	-0.053	-0.916	0.360
Access to Funding	**0.163**	**0.049**	**0.187**	**3.308**	**0.001**
Company's Goodwill	-0.061	0.059	-0.064	-1.037	0.300
Industry	**0.213**	**0.125**	**0.091**	**1.705**	**0.089**
Company size: Medium-sized vs. Small	**-0.275**	**0.116**	**-0.145**	**-2.377**	**0.018**
Company size: Large vs. Small	**-0.375**	**0.126**	**-0.184**	**-2.982**	**0.003**

5.3 External competitiveness factors of a company

The following set of eight variables focused on how respondents subjectively perceive the potential external competitiveness factors. The respondents used another eight closed questions to rate their company in comparison with its direct competitors, using a scale from 1 – a very low value – to 5 – a very high value. The following is a list of questions with their average rating and correlation with the financial performance of the company:

Table 5-16: Variables describing the external environment of a company: bivariate correlations with CFP

How do you rate…	Average value	Pearson's *r*	p (two-tailed)	N
Competitive Rivalry	**4.10**	**-0.133**	**0.008**	**404**
Power of Buyers	**3.87**	**-0.065**	**0.194**	**404**
Power of Suppliers	**3.27**	**-0.096**	**0.053**	**404**
Interest in Employment	**2.99**	**0.211**	**0.000**	**401**
Level of Corruption	2.28	-0.002	0.964	389
Support from the State	**1.67**	**0.133**	**0.008**	**403**
Support from Municipality/Local Administration	**2.01**	**0.067**	**0.181**	**400**
Market Progress	**3.60**	**0.151**	**0.003**	**386**

Scale from 1 – a very low value – to 5 – a very high value. For Market Progress: 1 – very narrowing to 5 – very widening.

The tested hypotheses were not directional for the variables Level of Corruption and Market Progress. However, for all the others the direction of the correlation found is in line as expected. For this reason, it is again possible to divide the p-values by two, and Power of Buyers and Support from Municipality/Local Administration will also become statistically significant correlations at the level of $\alpha = 10\%$. Only Level of Corruption shows no statistically significant correlation with financial performance.

We can also note that the factual relationship between both the variables with higher p-values and the financial performance is so weak that we cannot talk about influencing the financial performance by these variables. As for other variables, their effect can be interpreted as very low in the case of Power of Suppliers ($r = -0.096$), low for Competitive Rivalry ($r = -0.133$), Support from the State ($r = 0.133$) and Market Progress ($r = 0.151$), and low to medium for Interest in Employment ($r = 0.211$). A higher level of Competitive Rivalry or bargaining Power of Buyers and Suppliers hurts the company's financial performance, while higher Interest in Employment and Support from the State and from Municipality/Local Administration as well as the widening Market Progress help the financial performance. Again, we cannot talk about any important factor of financial performance at the level of a simple bivariate correlation in this area of company characteristics.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model in this form: Financial Performance = β_0 + β_1*Competitive Rivalry + β_2*Power of Buyers + β_3*Power of Suppliers + β_4*Interest in Employment + β_5*Level of Corruption + β_6*Support from the State + β_7* Support from Municipality/Local Administration + β_8*Market Progress + β_9*Industry + β_{10}*Company size Medium-sized + β_{11}*Company size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-17: Summary of the external factors regression model

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.343	0.118	0.090	0.849	0.118	4.213	11	347	0.000

As in the case of the internal factors, we could consider the value of multiple correlation of 0.343 moderate, as the model explains almost 12% of the variability of financial performance. Also, the model is significant. Estimates of the parameters of this model are listed in the following in the Table 5-18.

In this case, most of the variables statistically significantly affect financial performance. However, this is not true for Power of Buyers, Power of Suppliers, Level of Corruption, and Support from Municipality/Local Administration. As for the control variables, the impact of Industry is statistically insignificant.

The higher financial performance – if the values of the other variables remain unchanged – is related with lower Competitive Rivalry, higher Interest in Employment, higher Support from the State, and widening Market Progress. Increasing Company Size also means a lower financial performance (measured by relative indicators of ROA and asset growth). Relatively strong predictors are Interest in Employment, Competitive Rivalry, and a shift from a small business to a large enterprise. The lowest relative effect can be observed for the bargaining Power of Suppliers, followed by Support from the State. Market Progress and a shift from small

to a medium-sized business have a relatively moderate influence. As mentioned above, however, the explanatory power of the model as a whole is rather low or medium.

Table 5-18: Coefficients of the external factors regression model

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	0.152	0.421		0.361	0.719
Competitive Rivalry	**-0.162**	**0.060**	**-0.146**	**-2.714**	**0.007**
Power of Buyers	-0.003	0.060	-0.003	-0.052	0.959
Power of Suppliers	-0.090	0.059	-0.080	-1.533	0.126
Interest in Employment	**0.179**	**0.062**	**0.151**	**2.893**	**0.004**
Level of Corruption	0.012	0.042	0.016	0.290	0.772
Support from the State	**0.102**	**0.055**	**0.106**	**1.850**	**0.065**
Support from Municipality/Local Administration	-0.027	0.052	-0.030	-0.513	0.608
Market Progress	**0.153**	**0.058**	**0.137**	**2.646**	**0.009**
Industry	0.198	0.125	0.088	1.583	0.114
Company size: Medium-sized vs Small	**-0.242**	**0.110**	**-0.134**	**-2.198**	**0.029**
Company size: Large vs Small	**-0.282**	**0.121**	**-0.145**	**-2.342**	**0.020**

5.4 Stakeholder orientation and characteristics of an organizational structure

This part of the questionnaire examined a company as a whole; it did not go into the details of individual stakeholder groups, as was the case with the other sections. It rather focused on the structure or relations between stakeholder groups in terms of company management. If management were singled out as a separate stakeholder group, these questions would then belong to the section investigating just

management. The inclusion of this section was based on both the selected stakeholder approach to analysing an enterprise, and research into stakeholder management and company performance (Berman, 1999) or social and financial performance of a company (Preston, O'Bannon, 1997).

Table 5-19: Contents of the variables describing the stakeholder orientation and characteristics of an organizational structure

Question	Variable name	Values
How important is/are for your company...		
owners?	Importance of owners	1 – unimportant group/5 – very important group
employees?	Importance of employees	1 – unimportant group/5 – very important group
customers?	Importance of customers	1 – unimportant group/5 – very important group
suppliers?	Importance of suppliers	1 – unimportant group/5 – very important group
creditors?	Importance of creditors	1 – unimportant group/5 – very important group
the state?	Importance of the state	1 – unimportant group/5 – very important group
communities?	Importance of communities	1 – unimportant group/5 – very important group
What is the number of independent branches of your enterprise in the Czech Republic?	Number of Czech Branches	Interval
What is the number of independent branches of your enterprise abroad?	Number of foreign Branches	Interval
What is the number of relatively autonomous organizational units of your enterprise?	Number of autonomous units	Interval
What is the average autonomy of the organizational units within the defined categories?	Autonomy	Interval
What is the number of management levels in the main line of management?	Number of management levels	Interval
What is the span of control?	Span of control	Interval
What is the number of management levels per one employee?	Number of management levels per employee	Interval

The variables included here first assess the importance of the defined interest groups (stakeholders) for an enterprise, using a scale from 1 – an unimportant group – to 5 – a very important group. Furthermore, the enterprise is described with the number of independent branches in the Czech Republic, the number of independent branches abroad, the number of relatively autonomous organizational units and their average autonomy, the number of management levels, span of control, and the number of management levels per one employee. The autonomy of organizational units was rated on a scale from 1 – low autonomy to 5 – high autonomy in these areas: production program planning, dealings with customers, planning material inputs, dealings with suppliers, workforce development planning, and staff selection and recruitment. The following is a list of variables with their average rating and correlation with the financial performance of the enterprise:

Table 5-20: Variables describing stakeholder orientation of an enterprise and characteristics of an organizational structure: bivariate correlations with CFP

Variable	Average value	Kendall's τ_c	p (two-tailed)	N
Importance of Owners	4.48	0.023	0.510	408
Importance of Employees	4.05	0.043	0.271	410
Importance of Customers	4.68	0.013	0.706	410
Importance of Suppliers	3.79	-0.065	0.103	409
Importance of Creditors	2.73	-0.054	0.217	396
Importance of the State	**2.42**	**0.092**	**0.021**	**408**
Importance of Communities	2.58	-0.004	0.923	406
Number of Czech Branches[2]	1.23	0.027	0.390	400
Number of Foreign Branches[2]	0.25	-0.005	0.812	389
Number of Autonomous Units[2]	1.24	0.025	0.425	386
Autonomy[2]	1.20	0.036	0.252	403
Number of Management Levels[2]	3.18	-0.009	0.808	401
Span of Control[2]	13.22	0.017	0.612	396
Number of Management Levels per Employee[2]	0.02	0.051	0.116	401

[1] Scale from 1 – very low to 5 – very high.
[2] Variables with strongly abnormal two-dimensional distribution.

Due to the distribution of some variables that strongly fail to meet the two-dimensional normality condition, Kendall's coefficient τ_c was used to assess the bivariate correlations. This coefficient is nonparametric, immune to tied ranks (deVaus, 2002), and according to Field (2002) it is a more appropriate estimate of the correlation in population than Spearman's Rho.

Out of all the fourteen variables tested, only Importance of the State has a statistically significant effect on the financial performance. However, its effect on financial performance is on the borderline between trivial to no effect and a very low effect. Again, we cannot talk about any important factor of financial performance at the level of a simple bivariate correlation in this area of enterprise characteristics.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model[3]:
Financial performance = β_0 + β_1*Importance of Owners + β_2*Importance of Employees + β_3*Importance of Customers + β_4*Importance of Suppliers + β_5*Importance of Creditors + β_6*Importance of the State + β_7*Importance of Communities + β_8*Number of Management Levels + β_9*Span of Control + β_{10}* Number of Management Levels per Employee + β_{11}*Industry

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors. Predictors Size Medium-sized and Size Large were removed because of collinearity with Number of Management Levels, Span of Control, and Number of Management Levels per Employee.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-21: Summary of the regression model describing stakeholder orientation of an enterprise and characteristics of its organizational structure

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.261	0.068	0.039	0.887	0.068	2.349	11	353	0.008

[3] It includes Number of Independent Branches in the Czech Republic, Number of Independent Branches Abroad, Number of Relatively Autonomous Organizational Units, Average Autonomy of the Organizational Units.

The multiple correlation value of 0.261 is below the medium effect size, as the model explains about 7% of the financial performance variability. Still, the model is significant. Estimates of the parameters of this model are listed in the following table.

Table 5-22: Coefficients of the regression model describing stakeholder orientation of an enterprise and characteristics of its organizational structure

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	-0.695	0.494		-1.406	0.161
Importance of Owners	0.059	0.054	0.057	1.092	0.276
Importance of Employees	0.022	0.062	0.021	0.346	0.729
Importance of Customers	**0.171**	**0.087**	**0.114**	**1.962**	**0.051**
Importance of Suppliers	-0.111	0.060	-0.113	-1.842	0.066
Importance of Creditors	-0.053	0.040	-0.072	-1.319	0.188
Importance of the State	**0.127**	**0.045**	**0.171**	**2.849**	**0.005**
Importance of Communities	-0.057	0.047	-0.070	-1.208	0.228
Number of Management Levels	-0.005	0.054	-0.005	-0.093	0.926
Span of Control	0.003	0.005	0.035	0.675	0.500
Number of Management Levels per Employee	**5.567**	**3.203**	**0.092**	**1.738**	**0.083**
Industry	**0.223**	**0.124**	**0.099**	**1.801**	**0.073**

Similarly to the zero order correlations, only a few variables statistically significantly affect financial performance. These include Importance of Customers, Importance of the State, Number of Management Levels per Employee, and Industry, with only Importance of the State at the level of statistical significance $\alpha = 0.05$. Both higher Importance of Customers and higher Importance of the State imply higher financial performance of a company in case the values of the other variables remain unchanged. Even the variable Number of Management Levels per Employee has a positive impact on financial performance. Furthermore, the construction industry is more profitable than manufacturing (the reference value of zero is represented by the manufacturing industry). The relatively strongest predictor is Importance of the State,

followed by Importance of Customers, while the impact of Industry and Number of Management Levels per Employee is similarly strong.

5.5 Owners

This part of the questionnaire dealt with the ownership and property structure of the company. As Šedová (2007) stated, "a company's ownership structure significantly affects the creation of a certain model of ownership. The individual ownership models, applied in countries with market economies, allow different mechanisms of promoting the interests of owners." The relationships and interests of the owners towards the company were examined at two levels, as proposed by Kučera (2005): the relationship between the owners and other stakeholders, and the relationship of owners with one another. The definition of the different types of ownership took into account the specifics of exerting ownership in the Czech Republic. The following is a list of variables with an explanation of their content:

Table 5-23: Contents of variables describing the ownership and property structure

Question	Variable	Values
Is your enterprise part of a concern?	Concern	Yes/No
What is the legal form of the owner?	Owner's Legal Form	Natural person/Legal person/Both
What is the ownership concentration?	Ownership Concentration	Sole owner/Majority owner/More big owners
Are there any foreign direct investments in the enterprise?	FDI	Yes/No
Are owners whose stake in the company exceeds 5% present in the board of directors?	Owners in Board of Directors	Yes/No
Are owners whose stake in the company exceeds 5% present in the supervisory board?	Owners in Supervisory Board	Yes/No
Are owners whose stake in the company exceeds 5% present in the top management?	Owners in Top Management	Yes/No
What is the level of the principal components of tangible assets?	Assets Level	1 – obsolete/5 – top-level

The characteristics of the variables are significantly heterogeneous here. Part of the variables are dichotomous, part are nominal variables of three values, and one of them is ordinal, which we will consider quasi-interval for the purpose of linear regression. Bivariate associations between the variables of this section and the financial performance of a company follow here, and tests of their significance using ANOVA and t-test:

Table 5-24: Variables describing the ownership and property structure: bivariate relationships with CFP

Variable	Coefficient	Association	p (two-tailed)
Concern	η	-	n.s.
Owner's Legal Form: Natural Person vs. Legal Person	$\boldsymbol{\eta}$	**0.138**	**0.028**
Owner's Legal form: Natural Person & Legal Person vs. Natural Person	η	-	n.s.
Owner's Legal Form: Natural Person & Legal Person vs. Legal Person	η	-	n.s.
Ownership Concentration	η	-	n.s.
FDI	η	-	n.s.
Owners in Board of Directors	$\boldsymbol{\eta}$	**0.104**	**0.035**
Owners in Supervisory Board	$\boldsymbol{\eta}$	**0.169**	**0.001**
Owners in Top Management	$\boldsymbol{\eta}$	**0.116**	**0.019**
Assets Level	$\boldsymbol{\tau_c}$	**0.190**	**0.001**

The table shows that there is a statistically significant difference between companies whose owner is a natural person (m = 0.386) and companies whose owner is a legal person (m = 0.128). Higher performance was found in enterprises owned by a natural person. The effect size is, however, low (η = 0.138). What also plays a role is how the owners are involved in the management of the companies. If the owners are present in the board of directors or the supervisory board, the performance of such enterprises is statistically significantly lower (m = 0.365 vs. m = 0.125 for the board of directors, and m = 0.376 vs. m = -0.005 for the supervisory board). Conversely, if the owners are present in top management, the financial performance of such enterprises is higher (m = 0.129 vs. m = 0.355). However, once again the effect size is rather low. The positive effect of Assets Level on financial performance is low to moderate. Again, it is not possible to talk about any important factor of financial performance in the area of enterprise characteristics at the level of a simple bivariate association.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model in this form: Financial performance = β_0 + β_1*Concern + β_2*Owner's Legal Form Legal Person + β_3* Owner's Legal Form Legal Person & Natural Person + β_4*Majority Owners + β_5*Big Owners + β_6*FDI + β_7*Board of Directors + β_8*Supervisory Board + β_9*Top Management + β_{10}*Assets Level + β_{11}*Industry + β_{12}*Company Size Medium-sized + β_{13}*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors. Predictors Size Medium-sized and Size Large were removed because of collinearity with Number of Management Levels, Span of Control, and Number of Management Levels per Employee.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-25: Summary of the regression model describing the ownership and property structure

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.431	0.186	0.153	0.851	0.186	5.644	13	321	0.000

In the case of this model, the multiple correlation value of 0.431 ranges from medium to a substantial effect. The model explains almost 19% of the variability of financial performance. Also, the model is significant. Estimates of the parameters of this model are listed in the Table 5-26.

Ownership concentration and the presence of the FDI are statistically insignificant in this model. Furthermore, Industry and the shift from natural persons as owners to the mix of both natural and legal persons are also statistically insignificant. In the case of the latter, it is probably caused by a small number of companies with this mix among limited liability companies and joint stock companies.

Even this model shows that the larger the company is, the worse its financial performance becomes (measured by the relative indicators of ROA and asset growth). Notions from bivariate associations that the presence of the owners in the board of directors hurts financial performance while the presence of the owners in top management helps it are confirmed here (in both cases the reference value of zero is

represented by the absence of the owners). In addition, financial performance is enhanced by higher Assets Level, inclusion in a Concern Structure (the reference value of zero means that the company is not part of a concern), and the ownership by natural persons only (the reference value of zero is represented by businesses owned by natural persons only). The relatively strongest predictor is Assets Level, followed at a considerable distance by the inclusion in the Concern structure, Company Size, Owners in Top Management, Owners in Board of Directors, and Natural Persons as Owners.

Table 5-26: Coefficients of the regression model describing the ownership and property structure

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	**-0.519**	**0.246**		**-2.111**	**0.036**
Concern	**0.372**	**0.146**	**0.181**	**2.547**	**0.011**
Owner's Legal Form: Natural Person vs Legal Person	**-0.255**	**0.148**	**-0.135**	**-1.722**	**0.086**
Owner's Legal form: Natural Person & Legal Person vs Natural Person	-0.161	0.189	-0.047	-0.851	0.395
Ownership Concentration: Majority Owner vs. Sole Owner	0.014	0.136	0.006	0.101	0.919
Ownership Concentration: Multiple Big Owners vs. Sole Owner	-0.099	0.140	-0.053	-0.704	0.482
FDI	-0.041	0.144	-0.020	-0.283	0.777
Owners in Board of Directors	**-0.208**	**0.114**	**-0.112**	**-1.817**	**0.070**
Owners in Supervisory Board	-0.169	0.123	-0.084	-1.373	0.171
Owners in Top Management	**0.300**	**0.135**	**0.158**	**2.215**	**0.027**
Assets Level	**0.277**	**0.054**	**0.268**	**5.162**	**0.000**
Industry	0.105	0.122	0.046	0.861	0.390
Company Size: Medium-Sized vs. Small	**-0.238**	**0.114**	**-0.126**	**-2.085**	**0.038**
Company Size: Large vs. Small	**-0.304**	**0.134**	**-0.150**	**-2.261**	**0.024**

5.6 Employees

This section of the questionnaire focused on several framing questions: what is the structure of employees, what is their turnover and what are the reasons for the

Table 5-27: Contents of variables describing the stakeholder group of employees

Question	Variable	Types and values
What is the number of employees?	Number of Employees	Interval
What is the ratio of women?	Ratio of Women	Interval
What is the ratio of university graduates?	Ratio of Graduates	Interval
What is the ratio of technical and administrative staff?	Ratio of Technical and Administrative Staff	Interval
What is the ratio of workers?	Ratio of Workers	Interval
What is the rate of staff turnover?	Staff Turnover Rate: Low vs. High	Below 2 %/2 – 10 %/above 10 %
What is the average frequency of these reasons for staff turnover: reorganization, low wage or salary, uninteresting work, lack of career development opportunities, poor workplace relationships, or personal reasons?	Reasons for Staff Turnover	Interval
Do you monitor systematically the reasons for employees' leaving?	Staff Turnover Monitoring	Yes/No
What is the proportion of the variable component of salary for top management?	Motivation of Top Management	Interval
What is the proportion of the variable component of wage for workers?	Motivation of Workers	Interval
What is the impact of the variable component of wage or salary on staff motivation?	Effect of Motivation Component of Wage	1 – insignificant/5 – very high
What is the average intensity of using motivation devices, such as stock options, transport allowances, etc.?	Intensity of Using Motivation Devices	Interval
What resources does the company spend on employee benefits (as a percentage of labour costs)?	Expenses on Employee Benefits	Interval
What is the impact of these employee benefits on staff motivation?	Effect of Employee Benefits	1 – insignificant/5 – very high
What resources does the company spend on employee training (as a percentage of labour costs)?	Expenses on Employee Training	Interval

turnover, and how are they motivated? The previous Table 5-27 presents a list of variables with an explanation of their contents.

Most of the variables are interval, three of them are ordinal, and one is dichotomous. Interval variables paired with financial performance for the most part fail to meet the condition of two-dimensional normality of distribution; therefore we use Spearman's rs coefficient. Only the effect of the Staff Turnover Rate on financial performance (this variable takes three values) will be analysed using ANOVA, and the effect of Staff Turnover Monitoring on financial performance (this variable takes two values) will be analysed with the t-test. The following are bivariate associations between variables of this area and financial performance of a company:

Table 5-28: Variables describing the stakeholder group of employees: bivariate relationships with CFP

Variable	Coefficient	Association	p (two-tailed)	N
Number of Employees	r_s	**-0.101**	**0.041**	**405**
Ratio of Women	r_s	**-0.229**	**0.001**	**390**
Ratio of Graduates	r_s	**0.129**	**0.010**	**393**
Ratio of Technical and Administrative Staff	r_s	**0.114**	**0.023**	**394**
Ratio of Workers	r_s	**-0.114**	**0.023**	**394**
Staff Turnover Rate: Low vs. High	η	**0.125**	**0.039**	**152**
Reasons for Staff Turnover	r_s	**-0.138**	**0.007**	**375**
Staff Turnover Monitoring	η	-	n.s.	394
Motivation of Top Management	r_s	-0.023	0.676	345
Motivation of Workers	r_s	0.065	0.231	345
Effect of Motivation Component of Wage	r_s	**0.094**	**0.064**	**391**
Intensity of Using Motivation Devices	r_s	**0.146**	**0.003**	**408**
Expenses on Employee Benefits	r_s	0.010	0.859	334
Effect of Employee Benefits	r_s	0.018	0.715	393
Expenses on Employee Training	r_s	0.065	0.235	336

Most of the variables show a statistically significant relationship with the financial performance of a company. This is true for Number of Employees, Ratio of Graduates, Ratio of Technical and Administrative Staff, Ratio of Workers, Staff Turnover Rate, Reasons for Staff Turnover, Effect of Motivation Component of Wage, and Intensity of Using Motivation Devices. Although these variables have a statistically significant effect on financial performance, these effects are factually weak (r_s = 0.094 to 0.146). Only in the case of Ratio of Women we can talk about an almost medium effect on financial performance. This effect is negative (r_s = -0.229). As far as Staff Turnover Rate is concerned, a lower Staff Turnover Rate implies a higher average financial performance than a higher Staff Turnover Rate (m = 0.358 vs. m = 0.081). Again, we cannot talk about any important factor of financial performance at the level of a simple bivariate association in this area of enterprise characteristics.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model. The variable Staff Turnover Rate, which is ordinal, will be treated as nominal since it takes only three values. The model will be as follows:

Financial Performance = β_0 + β_1*Ratio of Women + β_2*Ratio of Technical and Administrative Staff + β_3*Low Staff Turnover Rate + β_4*High Staff Turnover Rate + β_5*Reasons for Staff Turnover + β_6*Staff Turnover Monitoring + β_7*Motivation of Top Management + β_8*Effect of Motivation Component of Wage + β_9*Intensity of Using Motivation Devices + β_{10}*Effect of Employee Benefits + β_{11}*Industry + β_{12}*Company Size Medium-sized + β_{13}*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors. The predictor Number of Employees was removed because of heterogeneity of variance, and it was replaced by dummy variables Company Size Medium-sized and Company Size Large. Predictors Ratio of Graduates, Ratio of Technical and Administrative Staff, Ratio of Workers, Motivation of Workers, Expenses on Employee Benefits, and Expenses on Employee Training were also removed because of variance heterogeneity.

The summary of the model obtained from the forced entry linear regression is shown in the Table 5-29.

In the case of this model, the multiple correlation value of 0.335 ranges from medium to substantial. The model explains about 11% of the variability of financial performance and is significant. Estimates of the parameters of this model are listed in the Table 5-30.

Table 5-29: Summary of the regression model describing the stakeholder group of employees

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.335	0.112	0.077	0.901	0.112	3.177	13	326	0.000

Table 5-30: Coefficients of the regression model describing the stakeholder group of employees

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	-0.437	0.518		-0.844	0.399
Ratio of Women	**-0.007**	**0.003**	**-0.160**	**-2.827**	**0.005**
Ratio of Technical and Administrative Staff	**0.864**	**0.321**	**0.145**	**2.687**	**0.008**
Staff Turnover Rate: Low	0.032	0.123	0.015	0.261	0.794
Staff Turnover Rate: High	-0.017	0.158	-0.006	-0.106	0.916
Reasons for Staff Turnover	-0.198	0.163	-0.070	-1.220	0.223
Staff Turnover Monitoring	-0.090	0.105	-0.047	-0.854	0.394
Motivation of Top Management	0.000	0.003	-0.007	-0.130	0.896
Effect of Motivation Component of Wage	0.082	0.055	0.081	1.493	0.136
Intensity of Using Motivation Devices	**0.531**	**0.194**	**0.157**	**2.741**	**0.006**
Effect of Employee Benefits	-0.043	0.055	-0.045	-0.789	0.431
Industry	0.036	0.135	0.015	0.268	0.789
Company Size: Medium-sized vs. Small	-0.115	0.125	-0.060	-0.926	0.355
Company Size: Large vs. Small	**-0.226**	**0.136**	**-0.110**	**-1.653**	**0.099**

Only four predictors are statistically significant in this model, and one of them – the shift in Company Size to a large enterprise – on the borderline of an acceptable p-value ($p = 0.099$). The other three statistically significant predictors are Ratio of Women, Ratio of Technical and Administrative Staff, and Intensity of Using Motivation Devices. If the values of other variables remain constant, a higher Ratio of Women decreases financial performance of a company, whereas a higher Ratio of Technical and Administrative Staff and higher average Intensity of Using Motivation Devices increases it. Company size has, as usual, a negative impact on financial performance. As far as the relative effect on financial performance is concerned, Ratio of Women, Intensity of Using Motivation Devices, and Ratio of Technical and Administrative Staff have a very similar effect. On the other hand, Company Size has a lower relative effect.

5.7 Customers

Customers as a stakeholder group are mentioned in all the literature of stakeholder theory. In the vast majority of empirical studies they are referred to as the most important stakeholder group (e.g. Šimberová, 2008), which is also one of the findings of this research. As Klapalová (2008) noted, many authors have worked on customer orientation as a factor of long-term business success since 1950s, including Peter Drucker among others. As for more recent authors focusing on market-orientation, i.e. including customer-orientation, we can name e.g. Kohli et al. (1993) or Narver and Slater (1990).

In this section, the questionnaire provided us with 70 variables focused on the following areas: business strategy, long-term relationships with customers and their stability, their territorial structure, and specificity of company products. It can also be repeated from the other sections of the questionnaire that the researchers were determining the importance of customers for a company, level of customer care, their bargaining power, degree of flexibility in adjusting products to customer demands, level of innovation activity, quality assessment, and evaluating goodwill or brand of a company.

The above-mentioned variables from this section of the questionnaire were transformed into six key variables, whose meaning is explained in the Table 5-31.

Three of the variables are nominal: Strategy, Strategy Focus and Exports. The other variables are interval: Stability Buyers, Share of Exports and Product Specificity. Because the distribution of their values is very different from normal (bimodal for Product Specificity, multimodal for Stability of Customers, and very positively skewed for Exports), we measure their relationship with financial performance using the τ_c coefficient. Since a large number of companies do not export, we also test the effect on financial performance based on the division of businesses into two groups: exporting and non-exporting. In this case and in the case of the variable Strategy we will use the t-test and ANOVA. The following are bivariate associations between the variables from this field and financial performance of a company:

Table 5-31: Contents of variables describing customers of a company

Question	Variable	Types and values
What business strategy does your company follow?	Strategy	Differentiation/Differentiation focus/Cost leader/Cost leader focus
Strategy Focus Yes/No	Strategy Focus	Yes/No
What is the stability of your customers?	Stability of Customers	Interval
What is the share of your foreign customers?	Share of Exports	Interval
Does your company have any foreign customers?	Exports	Yes/No
What is the ratio of specific products to all your products?	Product Specificity	Interval

Table 5-32: Variables describing customers: bivariate relationships with CFP

Variable	Coefficient	Association	p (two-tailed)	N
Strategy: Cost Leadership vs. Differentiation Focus	η	**0.157**	**0.057**	**182**
Strategy: Differentiation vs. Differentiation Focus	η	**0.157**	**0.085**	**204**
Strategy: Cost Focus vs. Differentiation Focus			n.s.	
Strategy: Focus Yes/No	η	**0.138**	**0.010**	**351**
Strategy: Cost/Differentiation			n.s.	
Stability of Customers	τ_c	**0.136**	**0.001**	**393**
Share of Exports	τ_c	-0.056	0.124	376
Exports: Yes/No	η	**0.128**	**0.013**	**376**
Product Specificity	τ_c	**0.119**	**0.001**	**387**

Besides the original form of the variable measuring the Share of Exports on all products, all the other variables show a statistically significant association with financial performance of the company. It can be considered even in the case of Share of Exports if it is measured only Yes/No (the number of completely non-exporting

companies is 89 out of 376, for which this information is available). The average performance of non-exporting companies (m = 0.454) is significantly higher than the average performance of exporting companies (m = 0.170). As far as Strategies are concerned, the highest performance can be observed in companies that pursue the strategy of Differentiation Focus (m = 0.464), followed by companies pursuing the strategy of Cost Focus (m = 0.261), then companies with the Differentiation Strategy (m = 0.144), and the lowest performance in our sample was found in companies pursuing the strategy of Cost Leadership (m = 0.103). There are statistically significant differences between strategies Cost Leadership and Differentiation Focus, and between strategies Differentiation and Differentiation Focus. No statistically significant difference was found between the two strategies oriented on costs and the two strategies oriented on differentiation; however, there is a statistically significant difference between businesses pursuing attention-focused strategies Focus, which have a higher average performance (m = 0.384) than the businesses pursuing strategies without this focused attention, whose average performance is lower (m = 0.126). Other variables that affect financial performance positively include Stability of Customers and Product Specificity. Nevertheless, it should be noted that although these associations are statistically significant, their factual significance, effect size, is very low. Again, we cannot talk about any important factor of financial performance at the level of simple bivariate associations in this area of enterprise characteristics.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model. The variable Exports will be included in its dichotomous form. The variable Strategy will be included in the form of Strategy Focus Yes/No, since although its impact on financial performance is lower, it is statistically significant at the more broadly accepted statistical significance level ($\alpha = 10\%$). Hence, the model will be as follows:
Financial Performance = $\beta_0 + \beta_1$*Strategy Focus + β_2*Stability of Customers + β_3*Exports Yes/No+ β_4*Product Specificity + β_5*Industry + β_6*Company Size Medium-sized + β_7*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-33: Summary of the regression model describing the stakeholder group of customers

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.311	0.097	0.076	0.903	0.097	04.641	7	303	0.000

The multiple correlation value of 0.311 is medium in the case of this model. The model explains less than 10% of the financial performance variability. Still, the model is significant. Estimates of the parameters of this model are listed in the following table.

Table 5-34: Coefficients of the regression model describing the stakeholder group of customers

Model	Unstandardized Coefficients		Standardized-Coefficients	t	p (two-tailed)
	B	Std. Error	Beta		
(Constant)	-0.327	0.299		-1.094	0.275
Strategy Focus Yes/No	-0.137	0.108	-0.073	-1.271	0.205
Stability of Customers	**0.201**	**0.064**	**0.180**	**3.146**	**0.002**
Exports Yes/No	**-0.278**	**0.155**	**-0.125**	**-1.788**	**0.075**
Product Specificity	**0.003**	**0.001**	**0.118**	**2.091**	**0.037**
Industry	0.182	0.161	0.077	1.128	0.260
Company Size: Medium-sized vs. Small	-0.162	0.123	-0.086	-1.312	0.190
Company Size: Large vs. Small	-0.205	0.140	-0.097	-1.469	0.143

Statistically significant predictors in this model include Stability of Customers, Exports and Product Specificity. Both Stability of Customers and Product Specificity have a positive impact on the financial performance of a company if the values of the other variables remain unchanged. Conversely, if a company exports, its financial performance is hurt (the reference value of zero is represented by non-exporting companies). The relative power of the individual predictors is highest for Stability of Customers and lowest for Exports and Product Specificity.

5.8 Suppliers

Supplier relationships management is very important in strategies aimed at costs (Gaddle, Hakansson, 2002) as well as strategies aimed at quality. There are studies examining the relationship of supplier stability and corporate competitiveness (O'Toole, Donaldson, 2000), supplier specialization and competitiveness (Hope, Spencer, 2001), or the method of selecting suppliers and company competitiveness (Kannan, Tan, 2004). Therefore, the variables verified here examine – as an analogy to the group of customers – the long-term nature of the relationship, its stability, territorial structure of suppliers, and specificity of supplies. Furthermore, criteria for selecting new suppliers were also examined. This section of the questionnaire provided us with the total of 97 variables that were transformed into 11 variables listed below, whose meaning is explained in the following table.

Table 5-35: Contents of variables describing suppliers of a company

Question	Variable	Types and values
What is the stability of your suppliers?	Stability of suppliers	Interval
What is the percentage of supplies from abroad?	Foreign Supplies	Interval
What is the proportion of specific supplies to total supplies?	Specificity of Supplies	Interval
How important is price as a criterion for selecting suppliers?	Supplier Selection Criterion: Price	1 – totally unimportant/5 – very important
How important are the payment terms as a criterion for selecting suppliers?	Supplier Selection Criterion: Payment Terms	1 – totally unimportant/5 – very important
How important are other terms of delivery as a criterion for selecting suppliers?	Supplier Selection Criterion: Other Terms of Delivery	1 – totally unimportant/5 – very important
How important is product quality as a criterion for selecting suppliers?	Supplier Selection Criterion: Product Quality	1 – totally unimportant/5 – very important
How important is a quality certificate as a criterion for selecting suppliers?	Supplier Selection Criterion: Quality Certificate	1 – totally unimportant/5 – very important
How important is a supplier's history as a criterion for selecting suppliers?	Supplier Selection Criterion: Supplier's History	1 – totally unimportant/5 – very important
How important are references as a criterion for selecting suppliers?	Supplier Selection Criterion: Reference	1 – totally unimportant/5 – very important
How important is the compatibility of CSR activities as a criterion for selecting suppliers?	Supplier Selection Criterion: CSR Compatibility	1 – totally unimportant/5 – very important

The first three variables are interval while the others are ordinal. Because the distribution of values of the interval variables is very different from normal, similarly to the variables describing the stakeholder group of customers (bimodal for Specificity of supplies, multimodal for Stability of Suppliers, and very positively skewed towards Foreign Supplies), we will use the τ_c coefficient to measure their correlation with financial performance. Since many businesses have no supplies from abroad, we will also test the relationship of financial performance and businesses classified into two groups: with foreign supplies and without them. In this case, the t-test will be used. The following are bivariate associations between the variables of this area and financial performance of a company:

Table 5-36: Variables describing suppliers of a company: bivariate relationships with CFP

Variable	Coefficient	Association	p (two-tailed)	N
Stability of Suppliers	τ_c	**0.064**	**0.058**	**402**
Foreign Supplies	τ_c	-0.022	0.517	390
Foreign Supplies: Yes/No	η		n.s.	
Specificity of Supplies	τ_c	0.013	0.707	390
Supplier Selection Criterion: Price	r	**-0.110**	**0.027**	**403**
Supplier Selection Criterion: Payment terms	r	**-0.162**	**0.001**	**401**
Supplier **Selection Criterion**: Other Terms of Delivery	r	-0.037	0.455	401
Supplier Selection Criterion: Product Quality	r	**0.083**	**0.098**	**403**
Supplier **Selection Criterion**: Quality Certificate	r	-0.010	0.846	401
Supplier **Selection Criterion**: Supplier's History	r	0.049	0.331	401
Supplier Selection Criterion: Reference	r	**0.108**	**0.031**	**401**
Supplier Selection Criterion: CSR compatibility	r	**0.147**	**0.003**	**398**

Financial performance statistically significantly correlates with most of the Supplier Selection Criteria, i.e. Price, Payment Terms, Reference, and CSR Compatibility. A worse but still acceptable p-value is found in Product Quality as a Supplier Selection Criterion, and in Stability of Suppliers. Product Quality is the least significant even in the factual aspect (r = 0.083), which also applies to Stability of

Suppliers (τ_c = 0.064). Other associations can be described as at least factually weak. Unlike in the stakeholder group of customers, if we recode the interval variables of Stability of Suppliers, Foreign Supplies and Specificity of Supplies into Yes/No or High/Low, we do not reach statistically significant associations for these recoded variables describing the stakeholder group of suppliers. Again, we cannot talk about any important factor of financial performance at the level of a simple bivariate association in this area of enterprise characteristics.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model. The variable Foreign Supplies will be included in a dichotomous form, as the original variable does not show a linear correlation to financial performance. For the same reason, the variable Specificity of Supplies will be included in the form of Low/High[4]. Hence, the model will be as follows:

Financial performance = β_0 + β_1*Stability of Suppliers + β_2*Foreign Supplies + β_3*Specificity of Supplies + β_4*Supplier Selection Criterion: Price + β_5*Supplier Selection Criterion: Payment Terms + β_6*Supplier Selection Criterion: Other Terms of Delivery + β_7*Supplier Selection Criterion: Product Quality + β_8*Supplier Selection Criterion: Quality Certificate + β_9*Supplier Selection Criterion: Supplier's History + β_{10}*Supplier Selection Criterion: Reference + β_{11}*Supplier Selection Criterion: CSR Compatibility + β_{12}*Industry + β_{13}*Company Size Medium-sized + β_{14}*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-37: Summary of the regression model describing the stakeholder group of suppliers

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.387	0.149	0.093	0.968	0.149	2.636	14	210	0.001

[4] Low specificity = 0 – 10 % of specific supplies, High specificity = 90 – 100 % of specific supplies.

The multiple correlation value of 0.387 ranges from medium to substantial in the case of this model, which explains little bit less than 15% of financial performance variability. Also, the model is significant. Estimates of the parameters of this model are listed in the following table.

Table 5-38: Coefficients of the regression model describing the stakeholder group of suppliers

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	-1.054	0.787		-1.339	0.182
Stability of Suppliers	0.155	0.094	0.112	1.648	0.101
Foreign Supplies: Yes/No	-0.155	0.174	-0.067	-0.891	0.374
Specificity of Supplies	-0.044	0.139	-0.021	-0.315	0.753
Supplier Selection Criterion: Price	**-0.182**	**0.101**	**-0.133**	**-1.805**	**0.073**
Supplier Selection Criterion: Payment Terms	-0.078	0.081	-0.072	-0.955	0.341
Supplier selection criterion: Other Terms of Delivery	**-0.143**	**0.084**	**-0.126**	**-1.714**	**0.088**
Supplier Selection Criterion: Product Quality	**0.384**	**0.143**	**0.190**	**2.684**	**0.008**
Supplier Selection Criterion: Quality Certificate	-0.020	0.066	-0.021	-0.310	0.757
Supplier Selection Criterion: Supplier's History	-0.015	0.080	-0.013	-0.182	0.856
Supplier Selection Criterion: Reference	**0.139**	**0.078**	**0.125**	**1.792**	**0.075**
Supplier Selection Criterion: CSR Compatibility	**0.150**	**0.057**	**0.181**	**2.646**	**0.009**
Industry	0.246	0.205	0.090	1.203	0.230
Company size: Medium-sized vs. Small	-0.247	0.157	-0.120	-1.576	0.117
Company size: Large vs. Small	**-0.317**	**0.177**	**-0.139**	**-1.792**	**0.075**

Similarly to the bivariate associations, the effect of these supplier selection criteria is statistically significant: Price, Product Quality, Reference, and CSR Compatibility. The effect of Payment Terms as a selection criterion is no longer statistically significant; instead, the effect of Other Terms of Delivery is statistically significant. These Other Terms of Delivery and Price have a negative impact on financial performance, in contrast to e.g. Product Quality, which is relatively the strongest predictor of financial performance. Stability of Suppliers, which was statistically significant (even though at a lower level of statistical significance, $p = 0.064$), is now narrowly unacceptable even at the level of $\alpha = 10\%$. Out of the general characteristics of enterprises, financial performance is influenced in this model also by a shift of the company size from medium-sized to large, while even here the direction is negative. In terms of the relative power of predictors, the strongest predictor is Quality as a supplier selection criterion, followed closely by CSR compatibility as a supplier selection criterion, and the remaining statistically significant predictors follow after approximately the same interval.

5.9 Corporate social responsibility

The last section of the questionnaire examined what forms of social responsibility enterprises use to become engaged in what areas, and also what codes (e.g. ethical) they have. This was actually a way of expressing corporate social performance, which has been demonstrated many times as a factor of competitiveness (e.g. a meta-analysis by Allouche, Laroche, 2005).

The variables from this section were transformed into two, i.e. how many forms of CSR activities a company is engaged in, and whether a company has at least one code. No company indicated that it was not engaged in any form of CSR activities; the maximum was six different CSR activities. Whether a company has at least one code is a dichotomous variable; therefore, a t-test will be used.

Table 5-39: Variables describing CSR activities of a company: bivariate relationships with CFP

Variable	*Coefficient*	*Association*	*p (two-tailed)*	*N*
CSR	**r**	**0.153**	**0.002**	**392**
Ethical code	η		n.s.	394

Only the number of CSR activities is statistically significantly related to financial performance. The effect of the number of CSR activities on financial performance can be described as weak, so even in this area there is no important factor of corporate competitiveness.

Multidimensional model

To assess the degree of joint effect of the aforementioned independent variables on financial performance, we can try to formulate a linear regression model. The model will be as follows:
Financial Performance = $\beta_0 + \beta_1$*CSR + β_2*Ethical Code + β_3*Industry + β_4*Company Size Medium-sized + β_5*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors.

The summary of the model obtained from the forced entry linear regression is shown in the following table.

Table 5-40: Summary of the regression model describing corporate social performance

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.235	0.055	0.043	0.928	0.055	4.388	5	375	0.001

The multiple correlation value of 0.235 ranges from low to medium in the case of this model, with the model explaining less than 6% of financial performance variability. Still, the model is significant. Estimates of the parameters of this model are listed in the following table.

Table 5-41: Coefficients of the regression model describing corporate social performance

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	-0.009	0.150		-0.060	0.952
CSR	**0.158**	**0.046**	**0.182**	**3.431**	**0.001**
Ethical Code	-0.029	0.100	-0.015	-0.287	0.774
Industry	0.178	0.120	0.075	1.478	0.140
Company Size: Medium-sized vs. Small	**-0.194**	**0.114**	**-0.101**	**-1.694**	**0.091**
Company Size: Large vs. Small	**-0.375**	**0.128**	**-0.178**	**-2.922**	**0.004**

Similarly to the bivariate associations, only the effect of the number of CSR activities is statistically significant. If the values of the other variables remain unchanged, a higher number of CSR activities has a positive effect on the financial performance of a company. The effect of Company Size is also statistically significant, since it is still true that the larger a company is, the lower its financial performance becomes. The number of CSR activities is a relatively slightly stronger predictor than a shift in Company Size to large and considerably stronger than a shift in Company Size to medium-sized.

5.10 Comprehensive model

In the preceding section we examined the sub-models of financial performance that always focused on one area of a business. Now we will try to construct a model of the predictors that were statistically significant in the individual sub-models. The number of such predictors exceeds 33, which is a relatively high number for a regression model using the available number of measurements. Therefore, we will treat missing values excluding missing values pairwise and not list wise as we have thus far. We will also rely on assertions by Green (1991) concerning the minimum accepted sample size. Green postulates that for an overall fit of a regression model, the minimum sample size is 50 + 8*p, where p is the number of predictors. In our case this would mean the need for 314 observations. Another rule set by Green is that if we want to test individual predictors, which is our case, there is a need for 104 + p observations. This would make the need for 137 observations in our case. A more sophisticated rule is set by Miles and Shevlin (2001). This takes into account the expected effect of the model. A higher expected effect results in a need of a lower number of observations. We expect a rather large effect, which would require approximately 100 observations, according to Miles and Shevlin. However, the actually obtained effect is rather medium-sized, resulting in a need for approximately 250 observations. All of these rules are met, as there are between 351 and 409 observations in this regression model.

The discussed model will therefore be tested as follows:

Financial performance = β_0 + β_1*Innovation Activity + β_2*Other Costs + β_3*Access to Funding + β_4*Competitive Rivalry + β_5*Interest in Employment + β_6*Support from the State + β_7*Market Progress + β_8*Importance of Customers + β_9*Importance of the State + β_{10}*Concern + β_{11}*Owner's Legal Form: Legal Person + β_{12}*Owner's Legal Form: Natural Person & Legal Person + β_{13}*Owners in Board of Directors+ β_{14}*Owners in Top Management + β_{15}*Assets Level + β_{16}*Ratio of Women + β_{17}*Ratio of Technical and Administrative Staff + β_{18}*Intensity of Using Motivation Devices + β_{19}*Strategy Focus + β_{20}*Stability of Customers + β_{21}*Exports Yes/No + β_{22}*Product Specificity + β_{23}*Supplier Selection Criterion: Price + β_{24}*Supplier Selection Criterion: Other Terms of Delivery + β_{25}*Supplier Selection Criterion: Product Quality + β_{26}*Supplier Selection Criterion: Reference + β_{27}*Supplier Selection Criterion: CSR

Compatibility + β_{28}*CSR + β_{29}*Industry + β_{30}*Company Size Medium-sized + β_{31}*Company Size Large

The following assumptions were assessed: appropriate variable type, non-zero variance of predictors, no strong multicollinearity, homoscedasticity, independent errors, normally distributed errors, independent values of outcome, and linear relationship between the outcome and predictors. The predictor Number of Management Levels per Employee was removed because of collinearity with Company Size. Two observations were removed because of an undue large influence on the regression model. Their standardized residual was more than 3.5.

The summary of the model obtained by the forced entry linear regression is in the next table.

Table 5-42: Summary of the comprehensive regression model

				Change statistics				
R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	p
0.551	0.304	0.235	0.832	0.304	4.399	31	312	0.000

The multiple correlation value of 0.551 is substantial in the case of this model. The model explains over 30% of financial performance variability. Also, the model is significant. When compared with the sub-models, it is obvious that this comprehensive model can explain a much larger proportion of the variability of the response variable. Estimates of the parameters of this model are listed in the Table 5-43.

Approximately half of the predictors tested have statistically significant effects on financial performance of a company. In all of them, the directions of action are consistent with those found in the sub-models. Hence, if the values of the other variables remain unchanged, greater financial performance is caused by lower Competitive Rivalry, widening Market Progress, greater Importance of the State, Owner's Legal Form as a Natural Person, the absence of Owners in Board of Directors, higher Assets Level, lower Ratio of Women, higher Ratio of Technical and Administrative Staff, higher Intensity of Using Motivation Devices, emphasis on Strategy Focus (either Cost Leadership Focus or Differentiation Focus), higher Stability of Customers as long as the company does not export its products, if it does not consider Other Terms of Delivery important as a supplier selection criterion, and if it is rather smaller than greater.

The relatively strongest predictor is clearly Assets Level. The other predictors, whose relative strength is approximately equal, follow after a considerable interval.

Table 5-43: Coefficients of the comprehensive regression model

Model	Unstandardized Coefficients		Standardized Coefficients	t	p
	B	Std. Error	Beta		
(Constant)	**-1.845**	**0.743**		**-2.485**	**0.013**
Innovation Activity	-0.014	0.049	-0.016	-0.288	0.773
Other Costs	-0.098	0.060	-0.080	-1.623	0.106
Access to Funding	0.048	0.047	0.055	1.025	0.306
Competitive rivalry	**-0.130**	**0.061**	**-0.112**	**-2.131**	**0.034**
Interest in Employment	0.071	0.064	0.057	1.100	0.272
Support from the State	0.061	0.054	0.058	1.137	0.256
Market progress	**0.113**	**0.058**	**0.098**	**1.945**	**0.053**
Importance of Customers	0.023	0.079	0.015	0.290	0.772
Importance of the state	**0.069**	**0.041**	**0.088**	**1.675**	**0.095**
Concern	0.156	0.130	0.074	1.199	0.232
Owner's legal form: Natural Person vs. Legal Person	**-0.221**	**0.131**	**-0.115**	**-1.681**	**0.094**
Owner's Legal Form: Natural Person & Legal Person vs. Natural Person	-0.004	0.195	-0.001	-0.020	0.984
Owners in Board of Directors	**-0.183**	**0.100**	**-0.096**	**-1.838**	**0.067**
Owners in Top Management	0.187	0.121	0.097	1.548	0.123
Assets Level	**0.189**	**0.062**	**0.175**	**3.065**	**0.002**
Ratio of Women	**-0.004**	**0.003**	**-0.097**	**-1.748**	**0.081**
Ratio of Technical and Administrative Staff	**0.685**	**0.307**	**0.114**	**2.231**	**0.026**
Intensity of Using Motivation Devices	**0.330**	**0.185**	**0.096**	**1.787**	**0.075**
Strategy FOCUS	**-0.169**	**0.098**	**-0.089**	**-1.714**	**0.088**
Stability of Customers	**0.116**	**0.059**	**0.106**	**1.979**	**0.049**
Exports Yes/No	**-0.268**	**0.138**	**-0.120**	**-1.936**	**0.054**
Product Specificity	0.001	0.001	0.029	0.551	0.582
Supplier Selection Criterion: Price	0.030	0.071	0.022	0.415	0.679
Supplier Selection Criterion: Other Terms of Delivery	**-0.113**	**0.056**	**-0.107**	**-2.013**	**0.045**
Supplier Selection Criterion: Product Quality	0.078	0.094	0.043	0.831	0.406
Supplier Selection Criterion: Reference	0.040	0.054	0.039	0.741	0.460
Supplier Selection Criterion: CSR Compatibility	0.040	0.044	0.050	0.910	0.363

6 Interpretation of the results achieved

This chapter focuses on the characteristics of the sample and its individual subgroups depending on the selected variables. There are 10 variables that were selected on the basis of the application of the DAF 74 method as factors that have the greatest impact on the financial performance of companies. These variables were supplemented by an additional 6 variables with regard to the interpretation needs. The order of these sixteen variables was adjusted with respect to the logic of interpretation. They are listed in the Table 6-44

As mentioned in Chapter 2, the sample of companies analysed was evaluated using the summation of two indicators, i.e. ROA and Assets Growth. Out of the 432 companies for which data were obtained from the questionnaire survey, 408 could be evaluated, while the remaining 24 had insufficient data. **We consider this set of 408 companies a sample** for the needs of this chapter. These businesses were ranked using the sum value of the indicators, ranging from companies with the highest value, i.e. businesses with the highest financial performance, to companies with the lowest financial performance. The sample was then divided into quartiles comprising 102 companies each. In the following text, these quartiles are indicated as a group of companies A, B, C and D. Companies with the highest financial performance are in group A, while companies with the lowest financial performance are included in group D.

Table 6-44: Overview of variables

Variable	Note	Variable type
P Industry	1= manufacturing, 2 = construction industry	nominal
P Legal form	1 = Ltd., 2 = joint-stock company	nominal
P Company size	1 = 50-99, 2 = 100-249, 3 = 250 or more employees	ordinal
P Concern	1 = no, 2 = yes	dichotomous
F4 FDI	1= without foreign capital, 2 = with foreign capital	nominal
P Ownership concentration	1=the only owner, 2=majority owner, 3=several big owners	nominal
F1 Strategy	1 = cost leadership, 2 = differentiation, 3 = cost focus, 4 = differentiation focus	nominal
F5 Share of exports	the share of exports on sales (%)	interval
F6 Span of control	an average number of employees per manager (%)	interval
F2 Motivation of workers	the ratio of the average upper limit of the motivation component of workers' wage to their base wage (%)	interval
F3 Motivation of top management	the ratio of the average upper limit of the motivation component of top managers' salary to their base salary (%)	interval
F8 Ratio of technical and administrative staff	the ratio of technical and administrative staff to the total number (%)	interval
F9 Ratio of workers	the ratio of workers to the total number of staff (%)	interval
F7 Ratio of women	the ratio of women to the total number of staff (%)	interval
F10 Reasons for staff urnover rate	1= less than 2%, 2 = 2-10%, 3 = more than 10%	ordinal
P Assets level	1= low (obsolete machinery), 2, 3, 4, 5 = high (top-of-the-range machinery)	ordinal

6.1 Overall characteristics of the sample

In this section, the characteristics of the sample as a whole is described partially, i.e. according to each of the sixteen variables mentioned above; moreover, groups A, B, C and D are described within the sample.

Industry

The data were obtained from 408 companies, i.e. 100% of the sample.

Companies from the manufacturing industry represent 80.4% of the sample, while companies from the construction industry account for 19.6% of the sample.

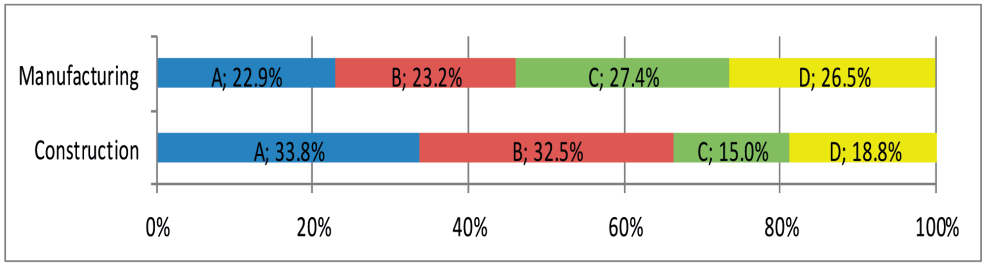
As shown in Graph 6-3, companies from the construction industry show a significantly higher financial performance than companies from the manufacturing industry.

Out of the total number of manufacturing companies, 22.9% of them belong to group A, 23.2% to group B, 27.4% to group C and 26.5% to group D.

Out of the total number of construction companies, 33.8% of them belong to group A, 32.5% to group B, 15.0% to group C and 18.8% to group D.

In other words: the percentage of manufacturing companies that are included in the two top quartiles, i.e. in groups A and B, is only about 46% of their total number. On the other hand, the percentage of construction companies included in groups A and B exceeds 66%.

Graph 6-3: Relative frequency of quartiles in manufacturing and construction



Legal form

The data were obtained for 408 companies, i.e. 100% of the sample.

The sample includes 55.4% of private limited companies and 44.6% of joint-stock companies.

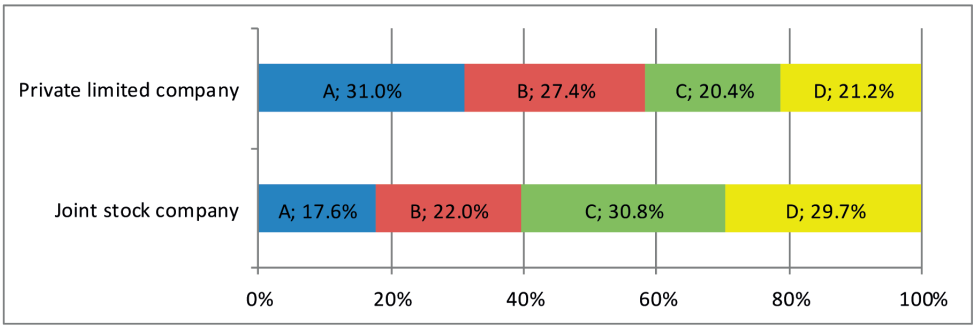
As shown in Graph 6-4, private limited companies show a significantly higher financial performance than joint-stock companies.

Out of the total number of private limited companies whose financial performance was analysed, 31.0% of them belong to group A, 27.4% to group B, 20.4% to group C and 21.2% to group D.

Out of the total number of joint-stock companies, 17.6% of them belong to group A, 22.0% to group B, 30.8% to group C and 29.7% to group D.

In other words: the percentage of private limited companies that are included in the two top quartiles, i.e. in groups A and B, exceeds 58% of their total number. On the other hand, the percentage of joint-stock companies included in groups A and B accounts for less than 40%.

Graph 6-4: Relative frequencies of quartiles in private limited companies and in joint stock companies



Company size

The data were obtained from 408 companies, i.e. 100% of the sample.

The sample includes 29.2% of small companies with 50 to 99 employees, 41.7% of medium-sized companies with 100 to 249 employees, and 29.2% of large companies with at least 250 employees.

As shown in Graph 6-5, smaller companies show a higher financial performance than larger ones.

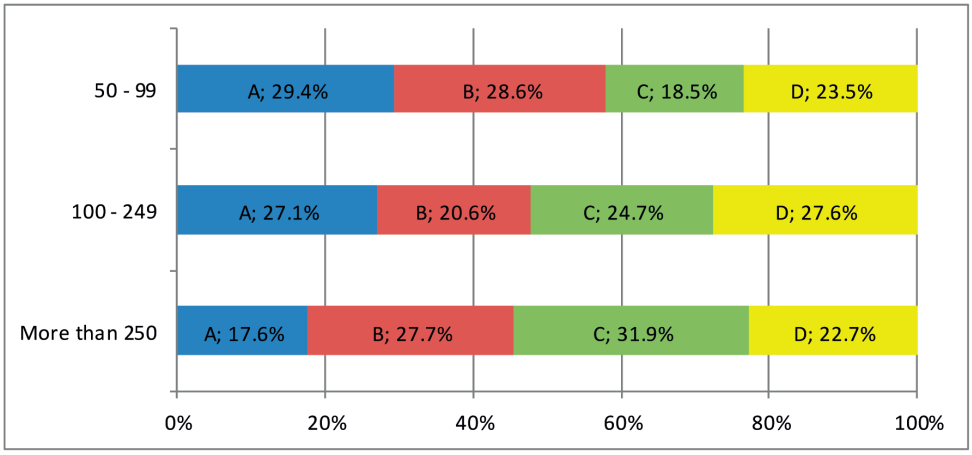
Out of the total number of companies with 50 to 99 employees, 29.4% of them belong to group A, 28.6% to group B, 18.5% to group C and 23.5% to group D.

Out of the total number of companies with 100 to 249 employees, 27.1% of them belong to group A, 20.6% to group B, 24.7% to group C and 27.6% to group D.

Out of the total number of companies with at least 250 employees, 17.6% of them belong to group A, 27.7% to group B, 31.9% to group C and 22.7% to group D.

The percentage of small businesses, which are included in the two top quartiles, i.e. in groups A and B, accounts for 58% of the total number, while the percentage of medium-sized companies is less than 48%, and the percentage of large companies is only about 45%.

Graph 6-5: Relative frequencies of quartiles according to company size



Concern

The data were obtained from 401 companies, i.e. 98.3% of the sample.

The percentage of companies from the sample that belong to a concern is less than a third – namely 28.7% – while companies that are not part of a concern represent 71.3% of the sample.

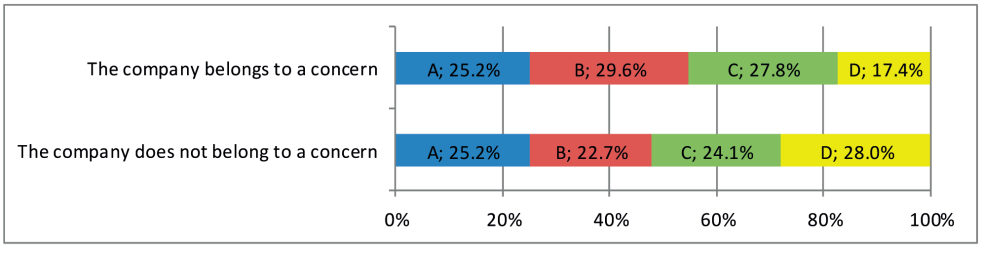
As shown in Graph 6-6, companies that belong to a concern show a slightly higher financial performance.

Out of the total number of companies that are part of a concern, 25.2% of them belong to group A, 29.6% to group B, 27.8% to group C and 17.4% to group D.

Out of the total number of companies that are not part of a concern, 25.2% of them belong to group A, 22.7% to group B, 24.1% to group C and 28.0% to group D.

The percentage of businesses belonging to a concern, which are included in the two top quartiles, i.e. in groups A and B, accounts for almost 55% of the total number; on the other hand, the percentage of companies from groups A and B that are not part of a concern is less than 48%.

Graph 6-6: Relative frequencies of quartiles according to concern membership



FDI

The data were obtained from 347 companies, i.e. 85.0% of the sample.

72.2% of companies from the sample are owned locally while in the remaining 27.4% of companies foreign capital is involved in their ownership; in most cases, these situations include big majority or even 100% ownership.

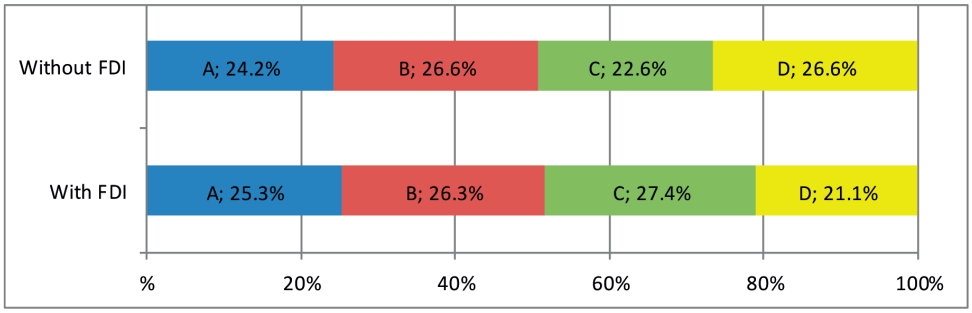
As shown in Graph 6-7, there is almost no difference in financial performance between companies that are owned locally and companies with foreign owners.

Out of the total number of companies that are owned only locally, 24.2% of them belong to group A, 26.6% to group B, 22.6% to group C and 26.6% to group D.

Out of the total number of companies with foreign owners, 25.3% of them belong to group A, 26.3% to group B, 27.4% to group C and 21.1% to group D.

The percentage of locally owned businesses, which are included in the two top quartiles, i.e. in groups A and B, accounts for approximately 51% of the total number, while the percentage of companies from groups A and B with foreign owners is about 52%.

Graph 6-7: Relative frequencies of quartiles in companies with and without FDI



Ownership concentration

The data were obtained from 395 companies, i.e. 96.8% of the sample.

The sample includes 38.5% of companies with a single owner, 21.8% of companies with a majority owner and 39.7% of companies owned by several big owners.

As shown in Graph 6-8, companies with several big owners show the highest financial performance while companies with a majority owner show the lowest financial performance.

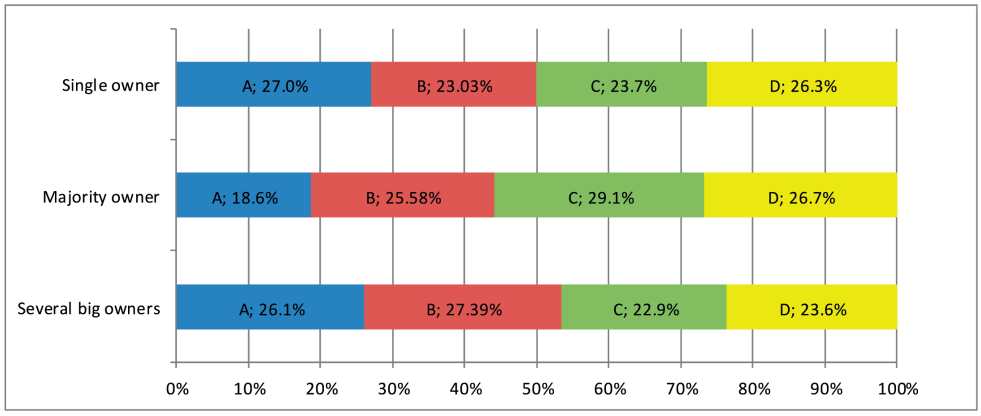
Out of the total number of companies with a single owner, 27.0% of them belong to group A, 23.0% to group B, 23.7% to group C and 26.3% to group D.

Out of the total number of companies with a majority owner, only 18.6% of them belong to group A, 25.6% to group B, 29.1% to group C and 26.7% to group D.

Out of the total number of companies owned by several big owners, 26.1% of them belong to group A, 27.4% to group B, 22.9% to group C and 23.6% to group D.

The percentage of companies with a single owner that are included in the two top quartiles, i.e. in groups A and B, accounts for 50% of their total number; the percentage of companies with a majority owner included in groups A and B accounts for only about 44%. On the other hand, the percentage of companies owned by several big owners included in groups A and B exceeds 53%.

Graph 6-8: Relative frequencies of quartiles according to the ownership concentration



Strategy

The data were obtained from 351 companies, i.e. 86.0% of the sample.

The sample includes 23.1% of companies using the cost-leadership strategy, 29.3% of companies using the differentiation strategy, 18.8% of companies using the cost-focus strategy, and 28.8% of companies using the differentiation-focus strategy.

As shown in Graph 6-9, an association between an applied strategy and financial performance is very weak. Yet, there is some effect of strategy: companies using the differentiation strategy show lower financial performance. As for the companies using one of the three remaining strategies, their financial performance is higher and they do not differ very much.

Out of the total number of companies using the cost-leadership strategy, 18.5% of them belong to group A, 33.3% to group B, 17.3% to group C and 30.9% to group D.

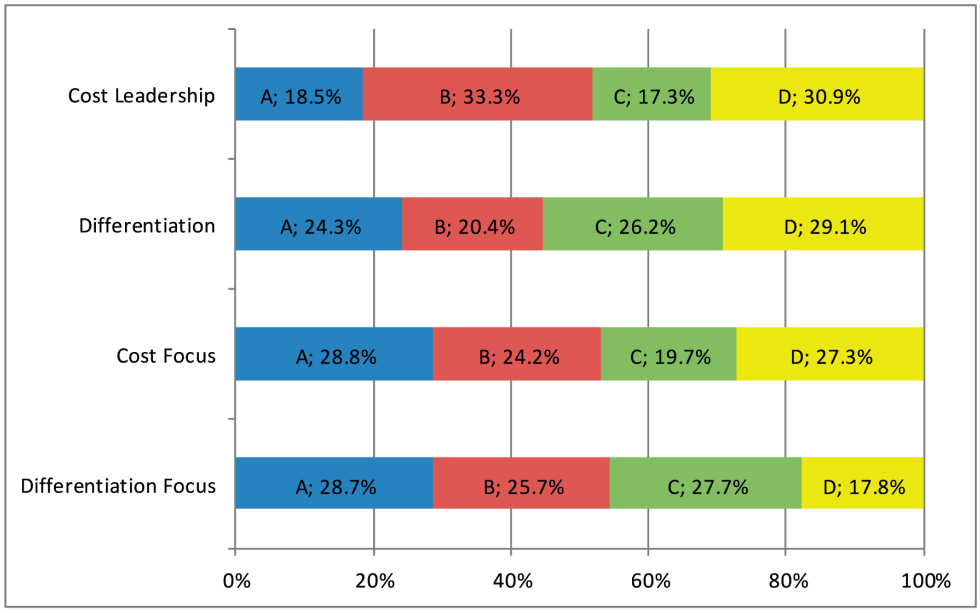
Out of the total number of companies using the differentiation strategy, 24.3% of them belong to group A, 20.4% to group B, 26.2% to group C and 29.1% to group D.

Out of the total number of companies using the cost-focus strategy, 28.8% of them belong to group A, 24.2% to group B, 19.7% to group C and 27.3% to group D.

Out of the total number of companies using the differentiation-focus strategy, 28.7% of them belong to group A, 25.7% to group B, 27.7% to group C and 17.8% to group D.

The percentage of companies using the cost-leadership strategy that are included in the two top quartiles, i.e. in groups A and B, accounts for almost 52% of their total

Graph 6-9: Relative frequencies of quartiles according to generic strategy



number. The percentage of companies using the differentiation strategy included in groups A and B accounts for less than 45%. The percentage of companies using the cost-focus strategy included in groups A and B accounts for 53%. Finally, the percentage of companies using the differentiation-focus strategy included in groups A and B accounts for approximately 54%.

Share of exports

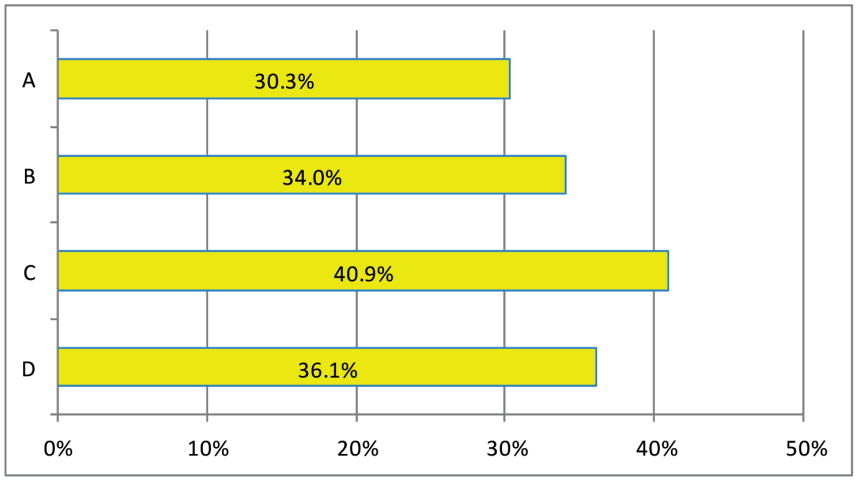
The data were obtained from 376 companies, i.e. 92.2% of the sample.

Companies with predominant exports represent a minority in the sample. The percentage of companies whose share of exports on sales is more than fifty per cent accounts for 34.3% in the sample. There are only 4% of companies that export all their production, while the percentage of companies that do not export at all is 23.6%.

It should be pointed out that there are striking differences between the manufacturing industry and construction: in the manufacturing industry, the percentage of companies with more than a 50% export share on sales is 43.1% of their total number, there are no such companies in the construction industry. The largest share of exports on sales is 30% here and only three companies show this figure.

It appears that the export orientation of companies is generally in opposition to their financial performance. As shown in Graph 6-10, group A includes only 30.3% of companies with more than a 50% share of exports on sales, group B includes 34.0% export-oriented companies, in group C their representation is highest – 40.9%, and group D includes 36.1% of such companies.

Graph 6-10: Share of companies with predominant exports in the quartiles



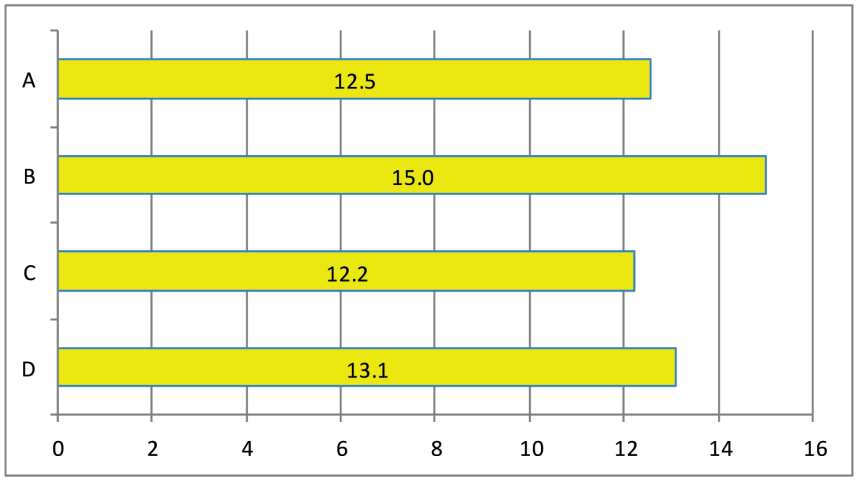
Span of control

The data were obtained from 393 companies, i.e. 96.3% of the sample.

The average number of employees who report to one manager in a company is 13.1 for all companies from the sample.

The average number of employees who report to one manager in a company is 12.5 for all companies from group A, 15.0 for companies from group B, 12.2 for companies from group C and 13.1 for companies from group D.

Graph 6-11: The average span of control in the quartiles



We can conclude that there is no zero-order relationship between the span of control and financial performance.

Motivation of workers

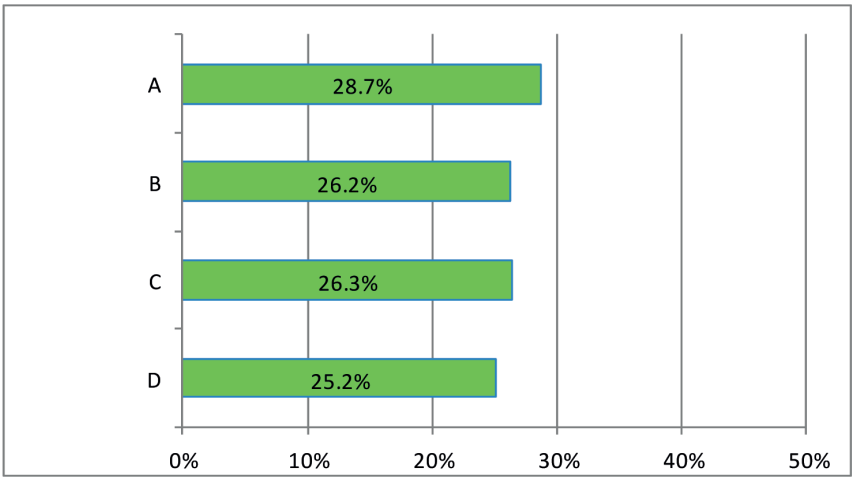
The data were obtained from 345 companies, i.e. 84.6% of the sample.

The average upper limit of the motivation components of workers' wages is 26.7% of the basic wage for all companies from the sample.

The results shown in Graph 6-12 imply a certain, though very weak correlation: companies with a higher motivation component of workers' wages tend to be the companies with higher financial performance.

The average upper limit of the motivation components of workers' wages is 28.7% of the basic wage in companies from group A, 26.2% in companies from group B, 26.3% in companies from group C and 25.2% in companies from group D.

Graph 6-12: The average upper limit of the motivation components of workers` wages in the quartiles



Motivation of top management

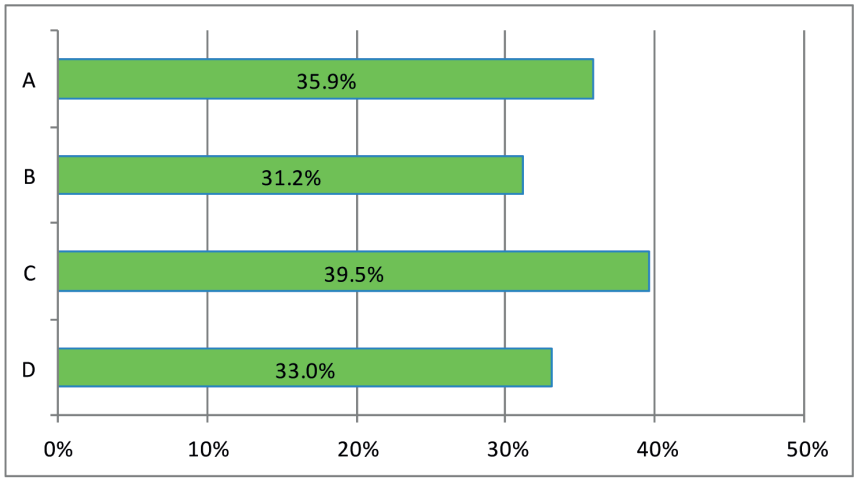
The data were obtained from 345 companies, i.e. 84.6% of the sample.

The average upper limit of the motivation components of top-managers' salaries is 34.4% of the basic salary for all companies from the sample.

The average upper limit of the motivation components of top-managers' salaries is 35.9% of the basic salary in companies from group A, 31.2% in companies from group B, 39.5% in companies from group C and 33.0% in companies from group D.

We can conclude that there is no zero-order relationship between the motivation component of the salaries and financial performance.

Graph 6-13: The average upper limit of the motivation components of top-managers' salaries in the quartiles



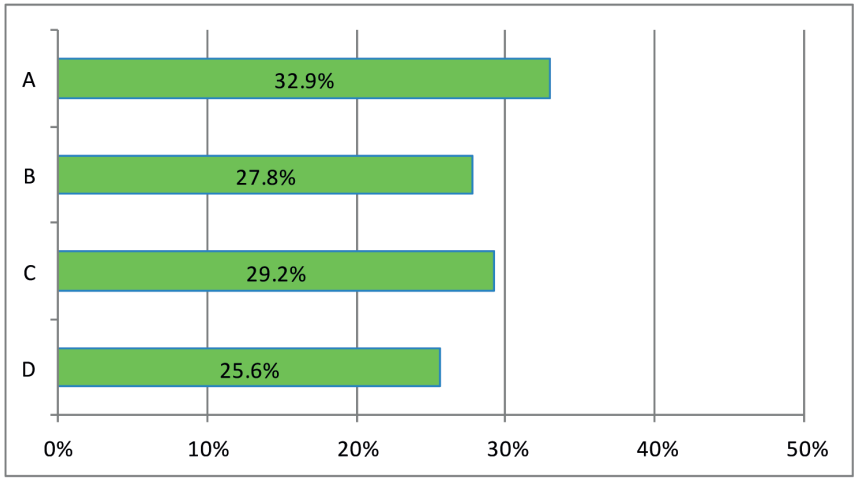
Ratio of technical and administrative staff

The data were obtained from 394 companies, i.e. 96.6% of the sample.

The average percentage of technical and administrative staff from the total number of employees is 28.7% for all companies from the sample.

Using the perspective of a particular industry, the average percentage of technical and administrative staff is lower (28.0%) in manufacturing companies than in construction ones (31.6%).

Graph 6-14: The average percentage of technical and administrative staff in the quartiles



The results shown in Graph 6-14 imply a certain, though very weak correlation: companies with a higher ratio of technical and administrative staff tend to be the companies with higher financial performance.

The average percentage of technical and administrative staff is 32.9% in companies from group A, 27.8% in companies from group B, 29.2% in companies from group C and 25.6% in companies from group D.

Ratio of workers

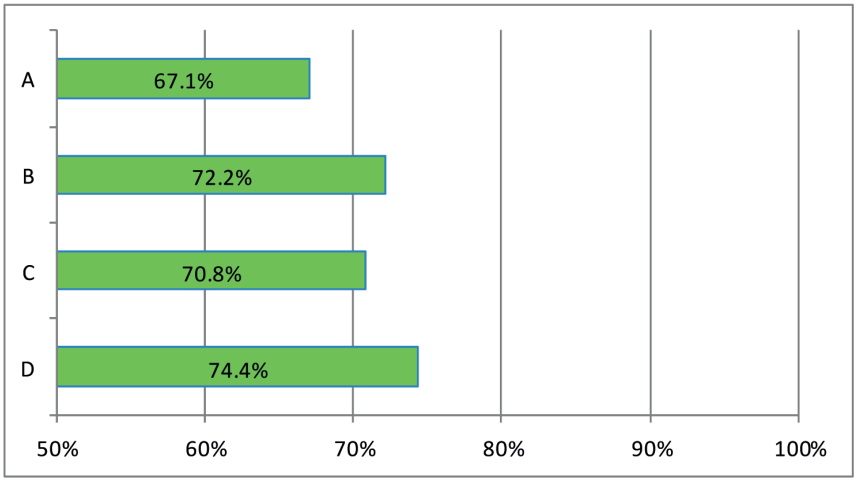
The data were obtained from 394 companies, i.e. 96.6% of the sample.

The average percentage of workers from the total number of employees is 71.3% for all companies from the sample.

The results shown in Graph 6-15 imply a certain, though very weak correlation: companies with a lower ratio of workers tend to be the companies with higher financial performance.

The average percentage of workers is 67.1% in companies from group A, 72.2% in companies from group B, 70.8% in companies from group C and 74.4% in companies from group D.

Graph 6-15: The average percentage of workers in the quartiles



Ratio of women

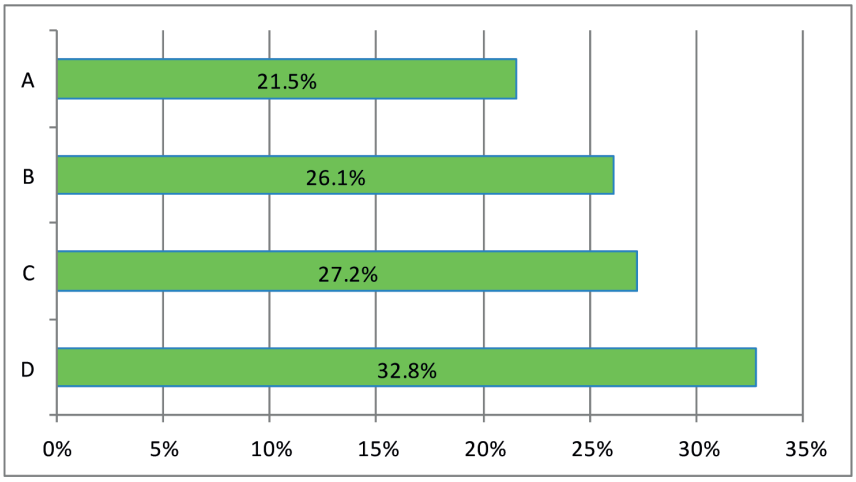
The data were obtained from 390 companies, i.e. 95.6% of the sample.

The average percentage of women from the total number of employees is 26.9% for all companies from the sample.

The results showed in Graph 6-16 suggest that companies with a lower ratio of women tend to be the companies with higher financial performance.

The average percentage of women is 21.5% in companies from group A, 26.1% in companies from group B, 27.2% in companies from group C and 32.8% in companies from group D.

Graph 6-16: The average percentage of women in the quartiles

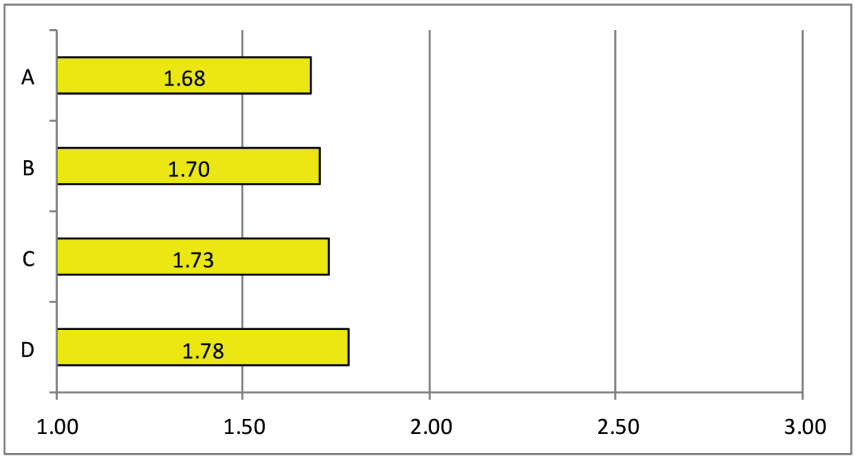


Reasons for the staff turnover rate

The data were obtained from 375 companies, i.e. 91.9% of the sample.

The average frequency of reasons for the staff turnover rate (from the employee`s initiative) is 1.73 points for all companies from the sample.

Graph 6-17: The average frequency of reasons for the staff turnover rate (from the employee`s initiative) in the quartiles



The results showed in Graph 6-17 suggest that there is a lower frequency of reasons for the staff turnover rate (from the employee`s initiative) in companies with higher financial performance.

The average frequency of reasons for the staff turnover rate (from the employee`s initiative) is 1.68 points in companies from group A, 1.70 points in companies from group B, 1.73 points in companies from group C and 1.78 points in companies from group D.

Assets level

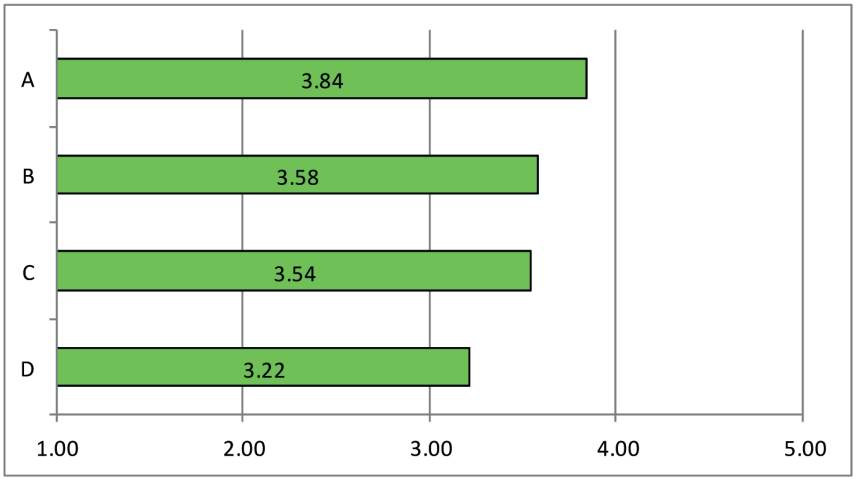
The data were obtained from 400 companies, i.e. 98.0% of the sample.

The average tangible assets level is 3.54 points for all companies from the sample.

The results in Graph 6-18 suggest that companies with a higher tangible assets level show higher financial performance.

The average tangible assets level is 3.84 points in companies from group A, 3.58 points in companies from group B, 3.54 points in companies from group C and 3.22 points in companies from group D.

Graph 6-18: The average tangible assets level in the quartiles



6.2 Identification of typical combinations of factor values leading to certain types of financial performance

As we mentioned in the previous chapter, the entire sample of companies analysed was divided according to their financial performance into four quartiles containing 102 companies each, labelled as either group A, B, C or D. In the subsequent considerations, we assume that the financial performance typical for each group can be achieved in several ways, i.e. combining different values of those variables that crucially affect the financial performance of companies. In order to identify these typical combinations, businesses in each of the groups A, B, C and D were subjected to a cluster analysis, performed using sixteen of the above-mentioned variables.

While respecting the reasonable range of the text, we focus only on one of the groups when presenting the interpretation of the obtained results. With regard to the principle of "best practices", we chose group A. We want to use it to demonstrate what different combinations of variable values (characteristics) lead to high financial performance of a company.

First, in relation to the previous subchapter, we characterize group A as a whole, then the individual subgroups A 1 to A 5, which result from the application of the cluster analysis.

Group A as a whole

Group A shows a relatively high proportion of companies from the construction industry. Out of their total number in the sample, nearly 34% of them belong here, while the proportion of companies from the manufacturing industry is only less than 23% of their total number.

Limited liability companies also have a relatively higher representation in this group. Out of their total number, 31% of them are represented in this group, while the proportion of joint-stock companies is less than 18% of their total number in group A.

Companies with 50–99 employees have the relatively highest representation in group A: out of their total number more than 29% of businesses belong here. In the case of companies with 100–249 employees, about 27% from their total number belong here. As for companies with at least 250 employees, less than 18% belong to group A.

The representation of companies in group A does not differ depending on whether a company is or is not part of a concern. In both cases the percentage is the same and amounts to slightly more than 25%.

The representation of companies depending on whether they are owned locally or by a foreign owner is very similar. Out of the total number of businesses that are owned only locally, approximately 24% belong to group A, while a little more than 25% of companies that are partly or fully owned by a foreign owner belong here.

However, after assessing the ownership structure more comprehensively, the relationships sought turn out to be clearer. In group A, two types have the largest relative proportion, namely:

a) companies that have a single owner and are part of a concern (75% of the total number of companies that are part of a concern),
b) companies that have several large owners and are not part of a concern (almost 56% of the total number of enterprises that are not part of a concern).

These two types of ownership structure are present in nearly 60% of the companies included in group A.

In group A there are relatively more frequently companies exercising the cost focus and differentiation focus strategies. In the case of cost focus, this group makes up nearly 29% of the total number of companies from the sample that exercise this strategy; in the case of differentiation focus, the figure is also almost 29%. On the contrary, companies using the cost leadership and differentiation strategies achieve lower financial performance on average. As for cost leadership, group A contains less than 19% of the total number of businesses applying this strategy, while in the case of differentiation, it is about 24%.

The dominant strategies of cost focus and differentiation focus are closely related with a strong concentration on specific products. Of the total number of companies in group A, 71% has a prevailing focus on specific products.

Companies in group A are characterized by low export orientation, which is somewhat surprising. As for group A, the representation of companies in which the share of exports in sales is predominant, i.e. higher than 50%, is the lowest compared with the other three groups (B, C and D). Only 30% of the companies from group A have a predominant share of exports in sales, i.e. they supply more products on the foreign market than the domestic one. This situation could be explained by the fact that in the period analysed there was a long-term appreciation of the domestic currency against the euro and the dollar. A certain influence can also be attributed to the relatively high representation of companies from the construction industry, whose exports are limited for technical reasons, typical for this sector. The construction companies from the sample do not contain a single business whose share of exports in sales would exceed the 50% threshold.

The value of span of control is 12.5 for businesses from group A. It is almost equal to the value valid for group C and it is lower than in the other two groups.

Systems of motivation in companies from group A provide a relatively large space for the motivational wage component both for workers – nearly 29% in average – and top management – almost 36% in average. In other words, according to the wage regulations, the upper limit of workers' wages is on average about 1.29 times the base salary; in the case of top management, it is 1.36 times the base salary.

Businesses represented in group A have – in comparison with those in groups B, C and D – a relatively higher proportion of technical and administrative staff of the total number of employees (almost 33%). It can be assumed that this situation is

related to the specificity of products, the predominant focus on differentiation strategies, and consequently to the higher proportion of intellectual work (research, development, marketing, and management). The second option may be relatively high labour productivity in manufacturing, caused by the fact that blue-collar occupations are largely equipped with technology.

Companies from group A show a relatively low proportion of women in the total workforce. The fact that the figure is on average less than 22% can be explained by a relatively high proportion of construction companies in group A, where the employment of women is low. Similar reasons can be seen also in the structure of the categories in the manufacturing industry, where companies operating in the fields traditionally dominated by women, such as textile manufacturing, food production, etc., are included only to a limited extent.

Compared with the other groups (B, C and D), companies in group A show the relatively lowest staff turnover initiated by employees. It amounts to 1.68 points on a three-point scale (1 – lowest to 3 – highest). This is undoubtedly related to the excellent economic results of companies in group A and the consequent pay conditions.

The level of the significant components of tangible assets amounts to the value of 3.84 points on a five-point scale (1 – lowest to 5 – highest) in companies from group A, and it is the highest compared with the other three groups. This suggests an influence of this factor on the financial performance of the companies analysed.

Now, let us focus our attention in more detail on the particular groups of A 1 to A 5.

Group A 1

It includes 15 companies.

They all belong to the manufacturing industry.

This group includes 60% of private limited companies and 40% of joint-stock companies.

All size categories are represented: small companies (50-99 employees), accounting for approximately 47%, medium-sized companies (100-249 employees), accounting for 20%, and large companies (at least 250 employees) with a share of about 33%.

Most companies (about 73%) are not part of a concern.

About half of the companies in this group are owned by local owners, while the other half is partly or fully owned by foreign owners.

As for the business strategies, the vast majority of these businesses implement strategies aimed at low cost (cost leadership and cost focus). These companies account for about 73% in group A 1.

Only a small proportion of companies from group A 1 (about 15%) is export-oriented, i.e. their share of exports in sales is higher than 50%. The average share of exports in sales in a company amounts to approximately 33%.

The span of control varies significantly from about four to almost twenty-two in the individual companies.

In companies from group A 1, the average motivation component of workers' wages is almost 20%, which is less than the average for the whole group A (about 29%). Similar findings hold for the average motivation component of top management salaries. The share of this component is about 32% in companies from group A 1, while the average for the whole group A is nearly 36%.

The ratio of technical and administrative staff to the total number of employees differs substantially for each company. On average it reaches nearly 37% and it is higher than the average of group A (about 33%) as well as the entire sample (less than 29%).

Considerably volatile values are also observed for the indicator ratio of workers to the total number of employees. The nature of the calculation shows that it is complementary to the previous indicator: on average it amounts to about 63% and it is lower than the average of group A (approximately 67%) as well as the entire sample (about 71%).

The ratio of women to the total number of employees is very high in the individual companies from group A 1. It ranges from 33% to 80%, and the average reaches to almost 55% of the total number of company employees. It is significantly higher than the average for the whole group A (about 22%) as well as the entire sample (27%). Industry specifics seem to have the main effect: the highest ratio of women (75 to 80%) occurs here in four companies, operating in the field of baking, toy manufacturing, textile production, and book publishing.

Staff turnover initiated by employees, measured on a scoring scale of 1 – lowest to 3 – highest, reaches the average value of 1.73 points. It is equal to the average value of the entire sample and it is higher than the average value for the whole group A (1.68).

The level of the significant components of tangible assets, measured on a scoring scale of 1 – lowest to 5 – highest, reaches the average value of 3.67 points. It is lower than the average value in group A (3.84), but higher than the value of the entire sample (3.54).

Group A 2

It includes 28 companies.

They all belong to the manufacturing industry.

In terms of the legal form of business, group A 2 contains significantly more private limited companies. There are 20 of them; thus, their proportion is more than 71%. The rest are joint-stock companies.

Size categories are represented equally: small companies (50-99 employees) with a share of about 32%, medium-sized companies (100-249 employees), accounting for approximately 36%, and large companies (at least 250 employees) with a share of about 32%.

The share of companies that are not part of a concern is higher, but not significantly. There are 16 companies (more than 57%) compared with 12 companies (less than 43%) that are part of a concern.

Local owners own approximately two-thirds of companies from this group, and one-third is fully or partly owned by foreign owners. Foreign owners own almost all the businesses that are part of a concern.

Businesses in group A 2 predominantly apply strategies aimed at the specific nature of their products with the option to demand higher prices. Application of strategies aimed at low cost is rare. Specifically: of the twenty-six companies (two did not provide information), 16 (i.e. approximately 62.5%) pursue the strategy of differentiation focus, and 3 companies (approximately 11.5%) apply the differentiation strategy. In contrast, 6 companies apply the cost focus strategy (about 23%) and only one company (less than 4%) applies the cost leadership strategy.

Group A 2 is typical for its exceptionally high export orientation. The average share of exports in sales is nearly 85% for businesses in this group. All 23 companies that answered this question report an at least 50% share of exports in sales. Of these companies, 12 report that their export share is 90% or more; 4 of these businesses export their entire production.

Similarly to the previous group, the span of control is very volatile in these companies, ranging from about three to almost twenty-seven.

In companies from group A 2, the average motivation component of workers' wages is almost 24%. This is less than the average for group A as a whole (approximately 29%) and also less than the average of the sample (nearly 27%). Similar findings also hold for the average motivation component of top management salaries. The share of this component is about 29% in companies from group A 2, while the average for the whole group A is almost 36% and the average for the entire sample is about 34%.

Similarly to the previous group, even here the ratio of technical and administrative staff and the ratio of workers to the total number of employees in each company differ substantially. On average, the ratio of technical and administrative staff is higher than 31%. This value is lower than the average of group A (about 33%), but higher than the average of the entire sample (less than 29%). The nature of the calculation shows that the ratio of workers is complementary to the previous indicator and it amounts to about 69% for companies in group A 2. It is higher than the average of group A (approximately 67%), but lower than the average of the entire sample (about 71%). A more detailed analysis shows that companies that are part of a concern have a smaller share of technical and administrative staff than companies that are not part of a concern. This is undoubtedly related to the centralization within the concern when a number of administrative, technical, or other similar activities that enterprises not included in a concern must perform themselves, are performed by the headquarters or a specialized organization for the concern units.

The average ratio of women to the total number of employees is about 25% in the companies from group A 2. This value is significantly lower than the average value for the companies in group A 1 (nearly 55%). However, it is higher than the average for the whole group A (about 20%), but slightly lower than the average of the entire

sample (27%). The highest ratio of women (80%) was found in a company engaged in the production of textiles.

Staff turnover initiated by employees, measured on a scoring scale of 1 – lowest to 3 – highest, reaches an average of 1.7 points. It is higher than the average for the whole group A (1.68), but lower than the average value of the entire sample (1.73).

The level of the significant components of tangible assets, measured on a scoring scale of 1 – lowest to 5 – highest, reaches an average value of 4.11 points. It is the highest value in group A as well as in the remaining three groups B, C and D. For the sake of comparison, the average value of the companies in group A is 3.84 points, while the average value of the entire sample is 3.54 points. Upon closer analysis we find that the average value of the level of tangible assets in companies that are part of a concern is 4.16 points. It can be assumed that this is due to inflows of investment from the concerns' headquarters.

Group A 3

It includes only three companies.

They all belong to the manufacturing industry and related categories: two belong to category 28 – Manufacture of metal constructions and fabricated metal products, and one to category 29 – Manufacture and reconstruction of machinery and equipment.

In terms of their legal forms of business, they are private limited companies.

One of the companies belongs to the size group of 50 to 99 employees and two companies belong to the size group of 100-249 employees.

The smaller company is not part of a concern and it is owned by a local owner, while both larger businesses are part of a concern and they have foreign owners.

The companies mainly apply strategies focused on the specific nature of their products, with the possibility of achieving higher prices (differentiation and differentiation focus strategies).

All the businesses have a strong export orientation. The share of exports in sales amounts to 70% for one of the companies and 95% for the two remaining companies.

The span of control is clearly higher (16.8) for the company that is not part of a concern than for the companies that are part of a concern (7.48 and 9.92).

The ratio of technical and administrative staff amounts to 16%, 23% and 24% in these companies. The values are significantly lower than the average of group A (about 33%), and the average of the entire sample (less than 29%).

The ratio of women to the total number of employees is low, which is undoubtedly caused by the industry focus (manufacture of fabricated metal products). In one of the companies it is only 5%, while in the remaining two companies it is about 13% and 15%.

Staff turnover initiated by employees, measured on a scoring scale of 1 – lowest to 3 – highest, which amounts to 1.4, 1.2 and 2.2 points in the individual companies, can be considered average, on the whole.

The level of the significant components of tangible assets, measured on a scoring scale of 1 – lowest to 5 – highest, reaches an average value of 3 points. It is lower than both the average value in group A (3.84) and the average value of the entire sample (3.54).

Group A 4

It includes 43 companies.

While in the previous three groups, there were no businesses from the construction industry, there are many of them in this group. The number of construction companies is 23, which represents 27.7% of their total number in the sample (83) of. There are 20 companies from the manufacturing industry, but this represents only 5.7% of the total number of these companies in the sample (349).

In terms of the legal form of business, this group includes 27 private limited companies and 16 joint-stock companies.

These businesses are mainly small and medium-sized companies. The group includes 15 companies from the category of 50 to 99 employees (about 35%), 22 companies from the category of 100 to 249 employees (about 51%), and only 6 companies with at least 250 employees (14%).

The share of companies that are part of a concern is not big: there are 10 companies of this type from the total number of 43 companies included in this group, i.e. about 23%.

Only a few enterprises (6) are owned by foreign owners. These are the businesses that are part of a concern.

As in the case of companies from groups A 2 and A 3, the strategies of companies in this group focus on the specific nature of their products with the possibility to demand higher prices rather than on achieving low cost. However, compared with the previous two groups, this orientation is slightly less significant in group A 4. 14 companies (i.e. about 35.9%) of 39 (4 did not respond) pursue the differentiation strategy while 10 companies (about 25.6%) follow the differentiation focus strategy. In contrast, the strategy of cost leadership is followed by 8 companies (approximately 20.5%) and 7 companies (about 17.9%) implement the strategy of cost focus. If we compare this strategic orientation in manufacturing and construction companies, we find that this orientation is more significant in the construction industry. While approximately 61.5% of the total number of companies in group A 4 pursue the strategies of differentiation and differentiation focus, the percentage of manufacturing companies applying these strategies is less than 59%; on the other hand, the percentage of construction companies amounts to nearly 64%. This fact can be attributed to industry specificities.

Export orientation of companies in group A 4 is very low; more than half of the businesses do not export at all. Of the 40 companies that answered the respective question, there are only 9 (i.e. 22.5%) whose share of exports in sales is 10% or more. The maximum export share is 40% and it is reported by one company. “Non-export” orientation is significant especially for construction companies, which is typical for

the construction industry. Zero export is reported by 16 of the 23 construction companies. In contrast, there are only two construction companies whose share of exports in sales amounts to 10% or more.

The span of control is very volatile in the individual businesses, even more than in groups A 1 and A 2. It can be said that it has virtually no explanatory value as a characteristics of the companies from this group.

The average motivation component of workers' wages is lower than 25% in the companies of group A 4. This is less than the average for the whole group A (approximately 29%) and also less than the average of the sample (nearly 27%). Similar figures are found even for the average motivation component of top management salaries. In companies from group A 4, the share of this component is less than 33%, while the average for the whole group A is almost 36% and for the entire sample it is about 34%.

Similarly to the previous groups, even here the ratio of technical and administrative staff and the ratio of workers to the total number of employees differ substantially in the individual companies; on average, they reach more than 36%. This value is higher than the average for group A (about 33%) and significantly higher than the average for the entire sample (less than 29%). Eight companies from the total number of 42 in group A reported that their ratio of technical and administrative staff is 50% and higher; three of these businesses are manufacturing companies and five are construction companies. In addition to their production activities, these companies are to a large extent engaged in engineering and design activities.

Businesses in group A 4 have on average a low ratio of women to the total workforce; the value is about 13%. This value is lower than the average for the whole group A (about 22%), and significantly lower than the average for the entire sample (27%). It is interesting that there is no big difference between the manufacturing and construction companies. The average ratio of women is about 14% in the manufacturing companies and about 12% in the construction companies. A company engaged in the production of computers and medical devices reported the highest ratio of women, reaching the value of 43%.

Staff turnover initiated by employees, measured on a scoring scale of 1 – lowest to 3 – highest, reaches an average value of 1.66 points. Is approximately the same as the average value for the whole group A (1.68), but lower than the average value for the entire sample (1.73).

The level of the significant components of tangible assets, measured on a scoring scale of 1 – lowest to 5 – highest, reaches an average value of 3.81 points. This value is similar to the average for the whole group A (3.84). However, it is significantly higher than the average value for the entire sample (3.54).

Group A 5

It includes 13 companies.

Similarly to the previous group of A 4, this group includes both manufacturing companies (9 businesses) and construction companies (4). It represents about 2.6%

of the total number of manufacturing companies in the sample and approximately 4.8% of the total number of construction companies.

With the exception of two businesses that are joint-stock companies, they are private limited companies (11 businesses).

They are mostly medium-sized companies, i.e. they belong to the category of companies with 100 to 249 employees; their number in this group is 9 (about 70%). The number of small businesses, i.e. the category of companies with 50 to 99 employees, is 3 (about 23%). There is only one large company (about 7%), whose number of employees is at least 250.

Only one of the companies in this group is part of a concern. Most businesses are owned by local owners.

The differentiation strategy is the most commonly applied business strategy; of the ten companies that responded to this question, six pursue it. As for the other strategies, two businesses apply the cost focus strategy and two businesses use the cost leadership strategy.

Export orientation of the companies is low. Of the ten companies that responded to this question, 4 companies do not export at all, 3 of which belong to the construction industry. For the remaining six companies, the share of exports in sales is up to 50%.

The number of workers who report to one superior in the businesses of group A 5 is considerably volatile, like in the previous groups. It ranges from about five to about thirty-six. Nevertheless, the explanatory power of this indicator is problematic.

The average motivation component of workers' wages is extremely high in the businesses of group A 5 as it amounts to more than 55%. This is substantially more than the average for the whole group A (approximately 29%) as well as the average for the sample (nearly 27%). Similarly, the average motivation component of top management salaries is also extremely high. In companies from group A 5, the amount of this component is more than 68%, while the average for the whole group A is about 36% and for the entire sample it is about 34%.

The ratio of technical and administrative staff to the total workforce is not considerably volatile, as it was in the previous group; it amounts to 25% on average. This value is significantly lower than the average for group A (about 33%), and lower than the average for the entire sample (less than 29%).

Similarly to the businesses in group A 4, the average ratio of women to the total number of company employees is also low in the companies from group A 5; the value is less than 13%. This value is lower than the average for the whole group A (about 22%), and significantly lower than the average for the entire sample (27%). The ratio of women in construction companies is lower than it is in manufacturing companies as it ranges from 5 to 10%.

Staff turnover initiated by employees, measured on a scoring scale of 1 – lowest to 3 – highest, reaches an average value of 1.68 points. It is identical with the average for the whole group A (1.68), but lower than the average value for the entire sample (1.73).

The level of the significant components of tangible assets, measured on a scoring scale of 1 – lowest to 5 – highest, reaches an average of 3.7 points. This value is lower than the average for the whole group A (3.84); however, it is higher than the average value for the entire sample (3.54).

6.3 Summary

Based on the conducted analysis and its subsequent interpretation, we will now focus on those variables that best describe the specifics of companies included in the above groups. They are listed in table 6-45 and table 6-46.

In the following explanation to table 6-45 and table 6-46, we deal with only four of the five groups listed above, namely groups A 1, A 2, A 4 and A 5. Group A 3, containing only three businesses, will not be dealt with due to its insignificant size. It may be said, though, that this group is very close to group A 2 in terms of the variables that we are focusing on here.

It is typical for **group A 1** that it includes only businesses from the manufacturing industry. However, there is considerable branch heterogeneity within this industry.

Compared with the other three groups, it has the highest share of businesses owned by foreign owners (46% of companies).

Unlike in the other three groups, businesses in this group, very significantly prefer those strategies that are aimed at low cost (72% of companies).

These are mainly domestic-market suppliers. The average share of exports in company sales amounts to 33%.

The level of the significant components of tangible assets is slightly below average; compared with companies in the other three groups it is the lowest.

It is also typical for **group A 2** that it includes only businesses from the manufacturing industry. Branch heterogeneity is slightly lower than in the case of group A 1.

Compared with the other three groups, the share of businesses that are part of a concern owned by foreign investors is the highest here.

Companies in this group typically pursue the differentiation-type strategies, namely the differentiation focus (62.5%) and differentiation (11.5%) strategies. Companies whose strategies are aimed at low cost represent a minority.

Businesses in group A 2 have an extremely high export orientation. The average share of exports in company sales in this group amounts to as much as 85%.

A high level of the significant components of tangible assets relates with the implementation of the differentiation-type strategies and the ability to succeed in foreign markets; this level is significantly higher than in the other three groups. It can be assumed that in many cases these are companies included in concerns that were equipped with high-quality machinery and equipment by their headquarters.

Unlike groups A 1 and A 2, **group A 4** includes construction companies as well as manufacturing companies.

Table 6-45: Summary of frequencies of categorical variables in the sample, group A and its subgroups

		A 1		A 2		A 3		A 4		A 5		A		Whole Sample	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
INDUSTRY	Manufacturing	15	100	28	100	3	100	20	47	9	69	75	74	328	80
	Construction	0	0	0	0	0	0	23	53	4	31	27	26	80	20
FDI	Without FDI	7	54	16	64	1	33	30	83	6	75	60	71	252	73
	With FDI	6	46	9	36	2	67	6	17	2	25	25	29	95	27
CONCERN MEMBERSHIP	Yes	4	27	12	43	2	67	10	23	1	8	29	29	115	29
	No	11	73	16	57	1	33	33	77	11	92	72	71	286	71
STRATEGY	Low Cost (CL+CF)	8	73	7	27	0	0	15	38	4	40	34	39	147	42
	Differentiation (DIF+DF)	3	27	19	73	2	100	24	62	6	60	54	61	204	58

Table 6-46: Summary of average values of chosen variables in the sample, group A and its subgroups

		A 1		A 2		A 3		A 4		A 5		A		Whole Sample	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
EXPORT	Share on Sales (%)	33		85		87		5		20		34		37	
ASSETS LEVEL 1 - lowest, 5 - highest		3.67		4.11		3.00		3.81		3.70		3.83		3.54	

A large majority of the companies (83%) are owned by a local owner, which is the biggest share of all the groups analysed.

The share of companies in the group that are part of a concern (23%) can be considered slightly below average.

Companies tend to prefer the differentiation-type strategies. The differentiation focus strategy is applied by 35.9% of the companies, while the differentiation strategy is used by 25.6% of companies.

In most cases the companies are domestic-market suppliers. The average share of exports in company sales is only 5% in this group. None of the exporting companies shows a higher share of exports in sales than 40%.

The level of significant components of tangible assets is close to the average for the whole group A.

Group A 5 also includes both manufacturing and construction companies.

Most companies (75%) are owned by a local owner.

With the exception of one company, the businesses in this group are not part of a concern.

Slightly more than half of the companies implement the differentiation-type strategies while the remaining businesses focus on low-cost strategies.

The companies tend to be domestic-market suppliers. The average share of exports in company sales in this group is 20%.

The level of the significant components of tangible assets is lower than the average for the whole group A.

The conducted analyses suggest that the characteristics of groups A 1 and A 2 are distinct and mutually very different. The business philosophy of most companies in group A 1 could be **"Keeping costs low and using the know-how of foreign owners".** In contrast, the business philosophy for most businesses in group A 2 might be **"Meeting the specific requirements of foreign customers. Focusing on quality, using high-quality machinery and equipment".**

The characteristics of groups A 4 and A 5 are neither as distinct nor different. These groups are quite similar, and to some extent they can be located between groups A 1 and A 2. Their common business philosophy can be formulated as follows: **"Using namely good knowledge and contacts in the domestic markets in business activities".**

7 Conclusion

The experience stemming from the research conducted confirmed the justified assumption that the task has been nontrivial and the methodology examined for its solution rather ambitious. We focus on the following areas in the final recapitulation, including problems associated with their solution.

a) Input data

Input data were obtained from two sources: financial statements published in the Albertina database, and questionnaires that the participating companies completed during the empirical investigation.

Data from the financial statements can be considered relatively accurate. However, the term "relatively" refers to the fact that in some cases all the data in the respective time series were not available; moreover, even various inaccuracies and distortions which the accounting practice can never completely exclude could occur here. Principally, the character of the information was objective.

Nevertheless, the situation was different for the data obtained from the questionnaires. Although the respondents were asked for objective information, such as the number of employees in each category, the ratio of each type of customers on sales, etc., it was clear that the data will be predominantly estimates, given the difficulty of obtaining or calculating such indicators. However, respondents were also asked in ways that anticipated response subjectivity, as it resulted directly from the questions asked. For instance, one of these questions was "What is the level of tangible assets of your business ?" and the answer was supposed to be recorded on a scale of 1 to 5, where "1" meant obsolete equipment and "5" state-of-the-art equipment. It was left up to the respondents how they would assess the equipment on the scale. Another important question concerned, for example, the business strategy applied by a company. Four strategies were characterized very briefly (Cost Leadership, Differentiation, Cost Focus and Differentiation Focus). It was left up to the respondents under which of these generic strategies they will subsume the strategy of their company.

It is obvious that in many cases the researchers had to rely on experience, qualifications and the respondents' responsibility. However, it should also be noted that the answers to those questions that were of a subjective nature eventually produced nominal and ordinal variables; these variables proved to be more useful for further processing with statistical methods and for the final interpretation than the numeric figures in the form of interval variables.

b) Preparation of variables for entering the process of identifying financial performance factors based on learning approaches

Using a questionnaire, the researchers obtained data which were expected to describe the business in a complex way if possible. The questionnaire was prepared in previous research. On the one hand, it was a compromise between the complexity and detail of the data, and on the other hand, between acceptability of the questionnaire length in terms of its completion by respondents and the subsequent analysis of the data obtained.

After determining what specific requirements the methods of feature selection from the field of statistical pattern recognition impose on the input variables, a substantial modification of the set of these variables started. The needed size of the so-called "training set" is a problem that occurs when using learning approaches, with respect to the dimensionality of the space of features, i.e. their number. In our particular case, the ratio is the number of observations that are available to the number of tested (or explanatory in the case of a regression model) variables. A frequently used rule of thumb says that this ratio should be at least 10:1. There are more sophisticated rules formulated for specific methods and for different uses of these methods; but even if we accept this general rule, it is true that if we wanted to use all of the almost 700 variables obtained from the questionnaires, it would be necessary to have questionnaires from at least 7,000 companies. In a situation where the population included less than half this number of companies, it was, of course, an impossible requirement. Therefore, within a series of experiments that were conducted during the research, repeated selection and in many cases also re-codification of the original variables was gradually used. This measure undoubtedly had a very positive impact on solving the given task.

c) Identification of financial performance factors through statistical methods

The original idea was to look for factors of corporate performance without an a priori model. In literature we commonly read about bivariate tests of correlations between potential performance factors and performance. In this case, however, the correlations found tend to be rather weak, which reflects the reality where these factors do not act separately. Thus, effects that can occur only with the involvement of the influence of other factors, which may even have no direct influence on performance, are not taken into account. Although it is possible to use these bivariate tests for the search of conditional or partial correlations, it is feasible only for a small number of

input variables and/or when testing a priori notions about these relationships. Another option is to use multidimensional models, with the most common ones being multiple regression models (general linear models), path analysis or structural equation models. They already take into account the interaction of multiple variables and are able to test complex causal models. Nevertheless, they still require a precise a priori idea of the tested correlations. Therefore, most studies of corporate performance focus on a specific area of a business, such as HRM, IT, or ownership structure etc. Although this covers the complicated links within the explanatory variables and between the explanatory variables and the dependent variables, it is always only for this one area and not across the areas. However, even variables from an "adjacent" area may play a role in explaining the effects of another variable, which also occurred in the results of our research. Thus, we pursued an exploratory type of research investigating the characteristics of businesses from various areas without the pre-formulated way of their anticipated effect. This, however, led to a vast increase in the number of variables required to sufficiently describe each company, i.e. in the dimensionality of the problem solved. Without an a priori model, which would at least partially reduce the number of possible interactions between the input variables, it was impossible to solve such a task in its completeness using the above-mentioned methods. This was the cause of the need to use different methodology. Methods of feature selection from the field of statistical pattern recognition proved suitable.

Feature selection methods typically consist of an optimization procedure and criterion of quality of the examined subset of features. The optimization procedure searches the space of the subsets gradually with respect to the criterion values evaluated on the previously investigated configurations. Criterial functions may be based on parametric models of data; however, they can also be defined without the need to include model assumptions. In our context, we primarily examined non-parametric modelling; in this case, the criterial function worked directly with the data in the data file instead of selecting a model explicitly. This was achieved by introducing a criterial function evaluating only the statistics of metric distances of samples above the currently evaluated subspace. The defined solution was on the borderline between the "nearest neighbour" methods and methods of kernel estimates, and it made it possible to elegantly solve the problem of occasionally missing values of some features.

By applying the feature selection methods, we were able to reduce the input variables to those that could either assign a company into the correct performance group in the best way, or predict this performance best. The disadvantage of the process bringing the highest accuracy and stability of this prediction (combination of the DAF method and a non-linear regression model) was the absence of guidelines to interpret the variables selected in this way. While the correlation coefficient shows the direction of the effect, or a linear regression model offers an equation explaining the dependent variable, a non-linear regression model, which we used as a measure of prediction accuracy, provided no such global guideline. The nonlinear model offered higher accuracy thanks to the ability to capture context-dependent

relationships between the features, i.e. different in various parts of the space. Observation of local dependencies between the features did not allow easy generalization leading to global conclusions about the features; it turned out that the same features could often manifest themselves as significant, insignificant, or even counterproductive, always depending on other currently selected features.

d) Economic interpretation

Interpretation of variables selected as "most informative" or best predicting corporate performance became a separate subsequent step. It is clear from the nature of the task solved that the purpose of interpretation is not to quantify the individual locally-dependent characteristics of the variables, but a meaningful generalization of this information towards an easily interpretable result. It involves a description of the space of variables, which has the potential to further clarify the features or feature groups defined by variables of companies, and in particular their mutual relations.

For this purpose, the companies were grouped into smaller groups within the existing company groups formed according to their financial performance; the grouping was based on the similarity of values of the identified factors of corporate performance. Consequently, these groups, which were created using the quantitative approach, had to be interpreted using a qualitative approach. It was necessary to synthesize the values of the individual factors and other explanatory variables, taking into account the variability of these values, so that a representative of each of these groups of companies could be characterized in the terminology of business studies. Such a process is obviously challenging in terms of the substantive knowledge of the issues, inventiveness and imagination. Nevertheless, it was this step that finished the implementation of the entire task, and it was possible to economically assess the correctness and relevance of the results achieved.

In conclusion, we can say that even with the great complexity of the task whose solution was very laborious, the results obtained can be regarded as beneficial. The objective laid down in the project can be considered satisfied, as a series of experiments verified the effective applicability of feature selection methods from the field of statistical pattern recognition in identifying the factors of financial performance of companies.

8 Bibliography

[1] ABOR, J., BIEPKE, N. (2007). Corporate Governance, Ownersip Structure and Performance of SMEs in Ghana: Implications for Financing Opportunities. Corporate Governance, Vol. 3, Issue 7, pp. 288–300. DOI: 10.1108/14720700710756562

[2] ALEXE, G., ALEXE, S., HAMMER, P. L., VIZVARI, B. (2006). Pattern-based feature selection in genomics and proteomics. Annals of Operations Research, Vol. 148, Issue 1, pp. 189–201. DOI: 10.1007/s10479-006-0084-x

[3] ALLOUCHE, J., LAROCHE, P. (2005). A Meta-Analytical Investigation of the Relationship Between Corporate Social And Financial Performance. Revue de Gestion des Ressources Humaines, Vol. 2005, Issue 57, pp. 18–41.

[4] AMBASTHA, A., MOMAYA, K. (2004). Competitiveness of firms: Review of theory, frameworks, and models. Singapore Management Review, Vol. 26, Issue 1, pp. 45–61.

[5] ANDREWS, R., BOYNE, G. A. (2010). Capacity, leadership, and Organizational Performance: Testing the Black Box Model of Public Management. Public Administration Review, Vol. 2010, Issue May/June, pp. 443–454.

[6] ARTIACH, T., LEE, D., NELSON, D., WALKER, J. (2010). The determinants of corporate sustainability performance. Accounting and Finance, Vol. 50, Issue 1, pp. 31–51. DOI: 10.1111/j.1467-629X.2009.00315.x

[7] BERANOVÁ, M. (2008). Modelling of Knowledge as an Instrument to Improve Retail Business Competitiveness. Trendy ekonomiky a managementu, Vol. 2, Issue 2, pp. 13–19.

[8] BERMAN, S., WICKS, A., KOTHA, S., JONES, T. (1999). Does stakeholder orientation matter: The relationship between stakeholder management models and firm financial performance. Academy of Management Journal, Vol. 42, Issue 5, pp. 488–506. DOI: 10.2307/256972

[9] BLAŽEK, L. et al. (2007). Konkurenční schopnost podniků (Primární analýza výsledků empirického šetření). Brno: Masarykova univerzita, 302 p. ISBN 978-80-210-4456-2.

[10] BLAŽEK, L. et al.. (2008). Konkurenční schopnost podniků: Analýza faktorů hospodářské úspěšnosti. (in Czech), Brno: Masarykova univerzita, 211 p. ISBN 978-80-210-4734-1.

[11] BLAŽEK, L. et al.. (2009). Konkurenční schopnost podniků. Analýza faktorů hospodářské úspěšnosti. Druhá etapa. Brno: Masarykova univerzita, 349 p. ISBN 978-80-210-5058-7.

[12] BLAŽEK, L. et al. (2011). Nadnárodní společnosti v České republice II. Brno: Masarykova univerzita, 345 p. ISBN 978-80-210-5677-0.
[13] BLUM, A., LANGLEY, P. (1997). Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence, Vol. 97, Issue 1–2, pp. 245–271. DOI: 10.1016/S0004-3702(97)00063-5
[14] BOTTAZZI, G., SECCHI, A., TAMAGNI, F. (2008). Productivity, profitability and financial performance. Industrial and Corporate Change, Vol. 17, Issue 4, pp. 711–751. DOI: 10.1093/icc/dtn027
[15] CAGWIN, D., BARKER, K. J. (2006). Activity-based costing, total quality management and business process reengineering: their separate and concurrent association with improvement in financial performance. Academy of Accounting and Financial Studies Journal, Vol. 10, Issue. 1, pp. 49–77.
[16] CARTER, J. R. (1977). In search of synergy – a structure-performance test. The Review of Economics and Statistics, Vol. 59, Issue 3, pp. 279–279.
[17] COLES J., LEMMON, M., MESCHKE, F. J. Structural models and endogeneity in corporate finance: The link between managerial ownership and corporate performance. Journal Of Financial Economics, Vol. 103, Issue, pp. 149–168. DOI: 10.2307/1925046
[18] COVER, T. M., HART, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, Vol. 13, Issue 1, pp. 21–27. DOI: 10.1109/TIT.1967.1053964
[19] DAS, S. (2001). Filters wrappers and a boosting-based hybrid for feature selection. In: ICML '01: Proc. 18th Int. Conf on Machine Learning, pp. 74–81.
[20] DASH, M., CHOI, K., SCHEUERMANN, P., LIU, H. (2002). Feature selection for clustering - a Filter solution. In: ICDM '02: Proc. 2002 IEEE Int. Conf on Data Mining, IEEE Computer Society, pp. 15–22.
[21] DE VAUS, D. (2002). Analyzing Social Science Data. 50 Key Problems in Data Analysis. 1st edition. Sage: London, 401 p. ISBN 0-7619-5937-8.
[22] DEVIJVER, P. A., KITTLER, J. (1982). Pattern Recognition: A Statistical Approach, London: Prentice-Hall. ISBN 0-136542360.
[23] DUDA, R. O., HART, P. E., STORK, D. G. (2000). Pattern Classification, 2nd edition, Wiley-Interscience, 41p. ISBN 0-471-05669.
[24] DUNNE, K., CUNNINGHAM, P., AZUAJE, F. (2002). Solutions to Instability Problems with equential Wrapperbased Approaches to Feature Selection. Technical Report TCD-CS-2002-28, Dublin: Department of Computer Science, Trinity College.
[25] EVROPSKÁ KOMISE (2011). European Competitiveness Report 2011. 289 p., [cit. 12. 7. 2014], available on WWW <http://ec.europa.eu/enterprise/policies/industrial-competitiveness/industrial-policy/files/ecr_2011_en.pdf>.
[26] FERRI, F. J., PUDIL, P., HATEF, M., KITTLER, J. (1994). Comparative Study of Techniques for Large-Scale Feature Selection. Machine Intelligence and Pattern Recognition, Vol. 1994, Issue 16, pp. 403–413.
[27] FIELD, A. (2009). Discovering statistics using SPSS. London: Sage Publications Ltd. 850 p. ISBN 978-1-84787-906-6.
[28] FORMAN, G. (2003). An extensive empirical study of feature selection metrics for text classification. The Journal of machine learning research, Vol. 3, Issue 1, pp. 1289–1305.
[29] FRAJ-ANDRÉS, E., MARTINEZ-SALINAS, E., MATUTE-VALLEJO, J. (2009). A multidimensional approach to the influence of environmental marketing and orientation

on the firm's organizational performance. Journal of Business Ethics, Vol. 88, Issue 2, pp. 263–286. DOI: 10.1007/s10551-008-9962-2

[30] FUKUNAGA, K. (1990). Introduction to Statistical Pattern Recognition. San Diego: Academic Press. 598 p. ISBN 0-12-269851-7.

[31] GADDE, L. E., HAKANSSON, H. (2001). Supply network strategies. Chichester: John Willey & Sons, Ltd.

[32] GREEN, S. B. (1991). How many subjects does it take to do a regression analysis?. Multivariate Behavioral Research, Vol. 26, Issue 3, pp. 499–510. DOI: 10.1207/s15327906mbr2603_7

[33] GUYON I., ELISSEEFF, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, Vol. 2003, Issue 3, pp. 1157–1182.

[34] HANSEN, G. S., WERNERFELT, B. (1989). Determinants of firm performance: The relative importance of economic and organizational factors. Strategic Management Journal. Vol. 10, Issue 5, pp. 399–411. DOI: 10.1002/smj.4250100502

[35] HOMBURG, C., KROHMER, H., WORKMAN, J.P. (1999). Strategic consensus and performance: The role of strategy type and market-related dynamism. Strategic Management Journal, Vol. 20, Issue 4, pp. 339–357. DOI: 10.1002/(SICI)1097-0266(199904)20:4<339::AID-SMJ29>3.0.CO;2-T

[36] HOPE, R. D., SPENCER, C. (2001). SRM is not yet a suite spot. Gartner Group. Vol. 2001, Issue 10. 5 s. [cit. 13. 7. 2012], available on WWW: <http://www.gartner.com/id=342114>.

[37] HULT, G. T. M. ET AL. (2008). An assessment of the measurement of performance in international business research. Journal of International Business Studies, Vol. 39, Issue 6, pp. 1064–1080. DOI: 10.1057/palgrave.jibs.8400398

[38] HUSSEIN, F., WARD, R., KHARMA, N. (2001). Genetic algorithms for feature selection and weighting, a review and study. In: Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on. IEEE. pp. 1240–1244.

[39] JAIN, A. K., DUIN, R. P. W., MAO, J. (2000). Statistical Pattern Recognition: A Review. Pattern Analysis and Machine Inteligence. IEEE Transactions on, Vol. 22, Issue 1, pp. 4–37.

[40] JAIN, A. K., ZONGKER, D. (1997). Feature Selection: Evaluation, Application and Small Sample Performance. Pattern Analysis and Machine Intelligence, IEEE Transactions on, Vol. 19, Issue 2, pp. 153–158.

[41] JIRÁSEK, J. A. (2000). Konkurenčnost: Vítězství a porážky na kolbišti trhu. Praha: Professional Publishing, ISBN 80-86419-11-8.

[42] KALOUSIS, A., PRADOS J., HILARIO, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. In Knowledge and information systems, Issue 12, No. 1, 2007, pp. 95–116.

[43] KANNAN, V. R., TAN, K. CH. (2004). Supplier alliances: differences in attitudes to supplier and quality management of adopters and non-adopters. Supply Chain Management, Vol. 9, Issue 4. DOI: 10.1108/13598540410550028

[44] KAPLAN, R.S., NORTON, D.P. (2004). Strategy maps: converting intangible assets into tangible outcomes. Boston: Harvard Business School. ISBN 1-59139-134-2.

[45] KESSLER, A. (2007). Success factors for new businesses in Austria and the Czech Republic. Entrepreneurship and regional development, Vol. 19, Issue 5, pp. 381–403. DOI: 10.1080/08985620701439959

[46] KIRBY, J. (2005). Toward a Theory of High Performance. Harvard Business Review, Vol. 83, Issue 7/8, pp. 30–39.
[47] KLAPALOVÁ, A. (2008). Význam zákazníka pro konkurenceschopnost podniku. Vývojové tendence podniků IV (Svazek I). Brno: Masarykova univerzita, p. 111–160, ISBN 978-80-210-4723-5.
[48] KOHAVI, R., JOHN, G. H. (1997). Wrappers for Feature Subset Selection. Artificial Intelligence, Vol. 97, Issue 1–2, pp. 273–324.
[49] KOHLI, A. K. ET AL. (1993). Markor: a Measure of Market Orientation. Journal of Marketing Research. Vol. 30, Issue 4, pp. 467–477. DOI: 10.2307/3172691
[50] KONONENKO, I. (1994). Estimating attributes: Analysis and extensions of Relief. In: Machine Learning: ECML-94. Springer Berlin Heidelberg, pp. 171–182.
[51] KRIŠTOF, M. (2006). Měření ekonomické efektivnosti podniku. Vývojové tendence podniků (Svazek I). Brno: Masarykova univerzita, pp. 263–280, ISBN 80-210-4133-1.
[52] KRIVOGORSKY, V., GRUDNITSKI, G. (2010). Country-specific institutional effects on ownership: concentration and performance of continental European firms. Journal of Management & Governance, Vol. 14, Issue. 2, pp. 167–193.
[53] KŘÍŽEK, P., KITTLER, J., HLAVÁČ, V. (2007). Improving stability of feature selection methods. In: Proc. 12th Int. Conf on Computer Analysis of Images and Patterns, Springer-Verlag, Vol. 4673, pp. 929–936.
[54] KUČERA, R. (2005). Vztahy k vlastníkům (akcionářům). Vývojové tendence podniků (Svazek I). Brno: Masarykova univerzita, pp. 255–270. ISBN 80-210-3847-0.
[55] KUDO, M., SKLANSKY J. (2000). Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recognition, Vol. 33, Issue 1, pp. 25–41. DOI: 10.1016/S0031-3203(99)00041-2
[56] KUNCHEVA, L. I. (2007). A stability index for feature selection. In Artificial intelligence and applications, 2007, pp. 421-427.
[57] LIU, P.-L., CHEN, W.-CH., TSAI, CH.-H. (2004). An empirical study on the correlation between knowledge management capability and competitiveness in Taiwan's industries. Technovation, Vol. 24, Issue 12, pp. 971–977. DOI: 10.1016/S0166-4972(03)00061-0
[58] LIU, H., YU, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. Knowledge and Data Engineering, IEEE Transactions on, Vol. 17, Issue 4, pp. 491–502.
[59] MARILL, T., GREEN, D. (1963). On the effectiveness of receptors in recognition systems. Information Theory, IEEE Transactions on, Vol. 9, Issue 1, pp. 11–17. DOI: 10.1109/TIT.1963.1057810
[60] MAYER, H.A., SOMOL, P., HUBER, R., PUDIL, P. (2000). Improving Statistical Measures of Feature Subsets by Conventional and Evolutionary Approaches. In: Advances in Pattern Recognition. Springer Berlin Heidelberg. ISBN 978-3-540-44522-7.
[61] MCLACHLAN, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. 2nd edition. New York: John Wiley & Sons, ISBN 0-471-69115-1.
[62] MILES, J. N. V., SHEVLIN, M. (2001). Applying regression and correlation: a guide for students and researchers. London: Sage Publications Ltd. 272 p. ISBN 978-0761962304.
[63] MOLINA, M. A., DEL PINO, I. B., RODRIGUEZ, A. C. (2004). Industry, Management Capabilities and Firms Competitiveness: An Empirical Contribution. Managerial and decision economics, Vol. 25, Issue 5, pp. 265–281.

[64] NADARAYA, E. A. (1964). On Estimating Regression. Theory of Probability and its Applications, Vol. 9, Issue 1, pp. 141–142. DOI: 10.1137/1109020

[65] NARVER, J. C., SLATER, S. F. (1990). The effect of a market orientation on a business profitability. Journal of Marketing, Vol. 1990, Issue 10, pp. 20–35. DOI: 10.2307/1251757

[66] NOVOVIČOVÁ, J., PUDIL, P. (1997). Feature selection and classification by modified model with latent structure. In: Dealing With Complexity: Neural Network Approac. Springer Verlag, Vol. 1997, pp. 126–140.

[67] NOVOVIČOVÁ, J., PUDIL, P., KITTLER, J. (1996). Divergence Based Feature Selection for Multimodal Class Densities. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, Issue 2, pp. 218–223. DOI: 10.1109/34.481557

[68] O'TOOLE, T., DONALDSON, B. (2000). Managing buyer-supplier relationship archetypes. Irish Marketing Review, Vol. 13, Issue 1, pp. 12–20.

[69] POKORNÁ, J., ČÁSTEK, O. (2013). How to measure organizational performance in search for factors of competitiveness. Acta universitatis agriculturae et silviculturae Mendelianae Brunensis, Vol. LXI, Issue 2, pp. 451–461. DOI: 10.11118/actaun201361020451

[70] POWELL, T, DENT-MICALLEF, A. (1997). Information Technology as Competitive Advantage: The Role of Human, Business, and Technology Resources. Strategic Management Journal. Vol. 18, Issue 5, pp. 375-405. DOI: 10.1002/(SICI)1097-0266(199705)18:5<375::AID-SMJ876>3.0.CO;2-7

[71] PRESTON, L. E., O'BANNON, D. P. (1997). The Corporate Social-Financial Performance Relationship. Business and Society. Vol. 36, Issue 4, pp. 419–420. DOI: 10.1177/000765039703600406

[72] PUDIL, P., BLAŽEK, L., SOMOL, P., ČÁSTEK, O., GRIM, J. (2013). Identification of Corporate Competitiveness Factors – Comparing Different Approaches. In: Proceedings of the International Conference on Management, Leadership and Governance. Bangkok, 07.02.2013 – 08.02.2013. Reading: Academic Conferences and Publishing International Limited, pp. 259–267. ISBN 978-1-909507-01-2.

[73] PUDIL, P., BLAŽEK, L., SOMOL, P., POKORNÁ, J., PIROŽEK, P. (2012). Searching Factors of Corporate Competitiveness Using Statistical Pattern Recognition Techniques, In: Proceedings of the 8th European Conference on Management Leadership and Governance. Pafos, 08.11.2012 – 09.11.2012. Reading: Academic Conferences and Publishing International Limited, pp. 556–559. ISBN 978-1-908272-76-8.

[74] PUDIL, P., NOVOVIČOVÁ, J., CHOAKJARERNWANIT, N., KITTLER, J. (1995). Feature selection based on approximation of class densities by finite mixtures of special type. Pattern Recognition, Vol. 28, Issue 9, pp. 1389–1398. DOI: 10.1016/0031-3203(94)00009-B

[75] 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 Application, IASTED, ACTA Press, pp. 224–229.

[76] PUDIL, P., NOVOVIČOVÁ, J. (1998). Novel Methods for Subset Selection with Respect to Problem Knowledge. IEEE Transactions on Intelligent Systems, Special Issue on Feature Transformation and Subset Selection, Vol. 453, pp. 66–74.

[77] PUDIL, P., NOVOVIČOVÁ, J., KITTLER, J. (1994). Floating Search Methods in Feature Selection. Pattern Recognition Letters, Vol. 15, Issue 11, pp. 1119–1125. DOI: 10.1016/0167-8655(94)90127-9

[78] PUDIL, P., SOMOL, P. (2008). Identifying the most Informative Variables for Decision-Making Problems – a Survey of Recent Approaches and Accompanying Problems. Acta Oeconomica Pragensia, Vol. 2008, Issue 4, pp. 37–55.
[79] RAUDYS, Š. (2006). Feature Over-Selection. Lecture Notes in Computer Science LNCS Springer, Vol. 4109, pp. 622–631. DOI: 10.1007/11815921_68
[80] REUNANEN J. (2003). Overfitting in making comparisons between variable selection methods, Journal of Machine Learning Research. Vol. 3, 2003, pp. 1371–1382.
[81] RICHARD, P.J. ET AL. (2009). Measuring organizational performance: towards methodological best practice. Journal of Management, Vol. 35, Issue 3, pp. 718–804. DOI: 10.1177/0149206308330560
[82] RIPLEY, B. (1996). Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press, Massachusetts, ISBN 0-521-46086-7.
[83] RIPLEY, B. (2005). Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press, Massachusetts, ISBN 978-052171770.
[84] SAEYS Y., INAKI I. I., LARRANGA P. L. (2007). A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), 2007, pp. 2507–2517. DOI: 10.1093/bioinformatics/btm344
[85] SALAPPA, A., DOUMPOS, M., ZOPOUNIDIS, C. (2007). Feature selection algorithms in classification problems: an experimental evaluation. Optimization Methods and Software, Vol. 22, Issue 1, pp. 199–214. DOI: 10.1080/10556780600881910
[86] SÁNCHEZ-BALLESTA, J. P., GARCÍA-MECA, E. (2007). A Meta-Analytic Vision of The Effect of Ownership Structure on Firm Performance. Corporate Governance, Vol. 15, Issue 5, pp. 879-892.
[87] SCHWAB, K. (2012). The Global Competitiveness Report 2012 – 2013. Ženeva: World Economic Forum. 545 s. [cit 10. 7. 2014], available on WWW : <http://www3.weforum.org/docs/WEF_GlobalCompetitivenessReport_2012-13.pdf>.
[88] SEBASTIANI, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, Vol. 34, Issue 1, pp. 1–47. DOI: 10.1145/505282.505283
[89] SEBBAN, M., NOCK, R. (2002). A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recognition, Vol. 35, Issue 4, pp. 835–846. DOI: 10.1016/S0031-3203(01)00084-X
[90] ŠEDOVÁ, J. (2007). Vlastníci a jejich vztah k podniku. Vývojové tendence podniků III (Svazek I). Brno: Masarykova univerzita, pp. 137–180, ISBN 978-80-210-4466-1.
[91] SIEDLECKI, W., SKLANSKY, J. (1993). On automatic feature selection. World Scientific Publishing Co., Inc.: River Edge, NJ, USA. pp. 63–87.
[92] ŠIMBEROVÁ, I. (2008). Řízení vztahů se stakeholdry na průmyslových trzích v kontextu současných marketingových koncepcí. Vědecké spisy vysokého učení technického v Brně. Brno: VUTIUM. Sv. 251. p. 38. ISSN 1213-418X.
[93] SIMONOFF, J. S. (1996). Smoothing Methods in Statistics. Springer. ISBN 0-387-94716-7.
[94] ŠIŠKA, J. (2012). Analýza výsledků dotazníkového šetření v nadnárodních společnostech: Bachelor diploma thesis. Brno: Masarykova univerzita, Head of the thesis: KRÁLOVÁ, M.
[95] ŠIŠKA, L. (2008). Skupiny podniků vytvořené dle ukazatelů hospodářské úspěšnosti. Konkurenční schopnost podniků (Analýza faktorů hospodářské úspěšnosti). Brno: Masarykova univerzita, pp. 51 – 66, ISBN 978-80-210-4734-1.

[96] SKURICHINA, M. (2001). Stabilizing Weak Classifiers. PhD thesis, Pattern Recognition Group, Delft University of Technology, Netherlands.
[97] SOMOL, P., BAESENS, B., PUDIL, P., VANTHIENEN, J. (2005). Filter- versus wrapper-based feature selection for credit scoring. International Journal of Intelligent Systems, Vol. 20, Issue 10, pp. 985–999. DOI: 10.1002/int.20103
[98] SOMOL, P., GRIM, J., PUDIL, P. (2011). Fast Dependency-Aware Feature Selection in Very-High-Dimensional Pattern Recognition. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Anchorage, 09. 10. 2011 – 12. 10. 2011. San Diego: IEEE Computer Society, ISBN 978-1-4577-0652-3.
[99] SOMOL, P., NOVOVIČOVÁ, J. (2008a). Evaluating the stability of feature selectors that optimize feature subset cardinality. In Structural, Syntactic, and Statistical Pattern Recognition, Vol. 5342, pp. 956–966.
[100] SOMOL, P., NOVOVIČOVÁ, J., GRIM, J., PUDIL, P. (2008b). Dynamic oscillating search algorithms for feature selection. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, pp. 1–4.
[101] SOMOL, P., NOVOVIČOVÁ, J., PUDIL, P. (2006). Flexible-hybrid sequential floating search in statistical feature selection. In: Structural Syntactic and Statistical Pattern Recognition, Vol. 4109, pp. 632-639.
[102] SOMOL, P., PUDIL, P. (2000). Oscillating search algorithms for feature selection, In: Proceedings of the 15th IAPR Int. Conference on Pattern Recognition, Conference B: Pattern Recognition and Neural Networks, pp. 406–409.
[103] SOMOL, P., PUDIL, P., ČÁSTEK, O., POKORNÁ, J. (2014). Improved Model for Attribute Selection on High-Dimensional Economic Data. In: Proceedings of the 2nd International Conference on Management, Leadership and Governance [CD]. Boston, 20.03.2014 – 21.03.2014. Reading: Academic Conferences and Publishing International Limited, pp. 276–285. ISBN 978-1-909507-99-9.
[104] SOMOL, P., NOVOVIČOVÁ, J. (2010). Evaluating Stability and Comparing Output of Feature Selectors that Optimize Feature Subset Cardinality. IEEE Transactions on PAMI Vol. 32, Issue 11, pp. 1921–1939. DOI: 10.1109/TPAMI.2010.34
[105] SOMOL, P., PUDIL, P. (2000). Oscillating Search Algorithms For Feature Selection, Proc. 15th IAPR International Conference on Pattern Recognition, Barcelona, Spain, pp. 406–409.
[106] 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, Issue 7, pp. 900–912. DOI: 10.1109/TPAMI.2004.28
[107] SOMOL, P., PUDIL, P., NOVOVIČOVÁ, J., PACLÍK, P. (1999). Adaptive Floating Search Methods in Feature Selection. Pattern Recognition Letters. Vol. 20, Issue 11, 12, 13, pp. 1157–1163. DOI: 10.1016/S0167-8655(99)00083-5
[108] ŠPALEK, J., ČÁSTEK, O. (2010). Přínos učících se metod statistického rozpoznávaní obrazu při hledáni konkurenceschopnosti českých podniku (in Czech). Journal of Economics, Vol. 58, Issue 9, pp. 922–937.
[109] STEARNS, S. D. (1976). On selecting features for pattern classifiers. In: Proceedings of the 3rd International Joint Conference on Pattern Recognition. pp. 71–75.
[110] SUCHÁNEK, P. (2005). Hodnocení konkurenceschopnosti podniku. Vyvojové tendence podniků (Svazek I.). Brno: Masarykova univerzita, pp. 271–278, ISBN 80-210-3847-0.
[111] THEODORIDIS, S., KOUTROUMBAS, K. (2006). Pattern Recognition. USA: Academic Press, 3rd edition.

[112] TSAMARDINOS, I., ALIFERIS, C. (2003). Towards Principled Feature Selection: Relevancy, Filters, and Wrappers. In: Proceedings of the ninth international workshop on Artificial Intelligence and Statistics. Morgan Kaufmann Publishers: Key West, FL, USA.

[113] VAFAIE, H., IMAM, I. (1994). Feature Selection Methods: Genetic Algorithms vs. Greedy-like Search, In: Proceedings of the International Conference on Fuzzy and Intelligent Control Systems.

[114] WEBB, A. (2002). Statistical Pattern Recognition, 2nd ed., Chichester: John Wiley &Sons, 2002. ISBN 0-470-84514-7.

[115] WHITE, R. E. (1986). Generic Business Strategies, Organizational Context and Performance: An Empirical Investigation. Strategic Management Journal, Vol. 7, Issue 3, pp. 217–231. DOI: 10.1002/smj.4250070304

[116] WHITNEY, A. W. (1971). A direct method of nonparametric measurement selection. IEEE Trans. Comput. Vol. 20, Issue 9, pp.1100–1103. DOI: 10.1109/T-C.1971.223410

[117] WILLIAMSON, P. J., VERDIN, P.J. (1992). Age, experience and corporate synergy: When are they sources of business unit advantage?. British Journal of Management, Vol. 3, Issue 4, pp. 221–235. DOI: 10.1111/j.1467-8551.1992.tb00047.x

[118] XING, E. P. (2003). Feature Selection in Microarray Analysis. Springer, 2003, pp. 110-129.

[119] YANG Y., PEDERSEN J. O. (1997). A comparative study on feature selection in text categorization. In: ICML '97: Proc. 14th Int. Conf on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco 1997, pp. 412–420.

[120] YANG, J., HONAVAR, V. (1998). Feature Subset Selection Using a Genetic Algorithm. IEEE Intelligent Systems. Vol. 453, Issue 13, pp. 44–49. DOI: 10.1109/5254.671091

[121] YILMAZ, C., ALPKAN, L., ERGUN, E. (2005). Cultural determinants of customer - and learning-oriented value systems and their joint effects on firm performance. Journal of Business Research. Vol. 58, Issue 10, pp. 1340–1352. DOI: 10.1016/j.jbusres.2004.06.002

[122] YOUNDT, M. A., SNELL, S. A., DEAN, J. W., LEPAK, D. P. (1996). Human resource management, manufacturing strategy and firm performance. Academy of Management Journal, Vol. 39, Issue 4, pp. 836–866.

[123] YU, L., LIU, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, pp. 56–63.

[124] YUSTA, S. C. (2009). Different metaheuristic strategies to solve the feature selection problem. Pattern Recogn. Lett., Vol. 30, Issue 5, pp. 525–534. DOI: 10.1016/j.patrec.2008.11.012

List of Figures

List of Tables

List of Graphs

Pavel Pudil, Ladislav Blažek, Ondřej Částek, Petr Somol, Jana Pokorná, Maria Králová

Identifying Corporate Performance Factors Based
on Feature Selection in Statistical Pattern Recognition.
METHODS, APPLICATION, INTERPRETATION

Cover design: RNDr. Petr Somol, Ph.D.
Layout and typeset: Lea Novotná
Published by Masaryk University in cooperation with University of Economics, Prague
Brno 2014
First Edition
Printed by Tribun EU s.r.o., Cejl 892/32, 602 00 Brno

ISBN 978-80-210-7557-3
ISBN 978-80-210-7672-3 (online : pdf)
DOI: 10.5817/CZ.MUNI.M210-7557-2014

Faculty of Management
University of Economics, Prague
&
Faculty of Economics and
Administration, Masaryk University

muni PRESS 2014

DOI: 10.5817/CZ.MUNI.M210-7557-2014

ISBN 978-80-210-7557-3