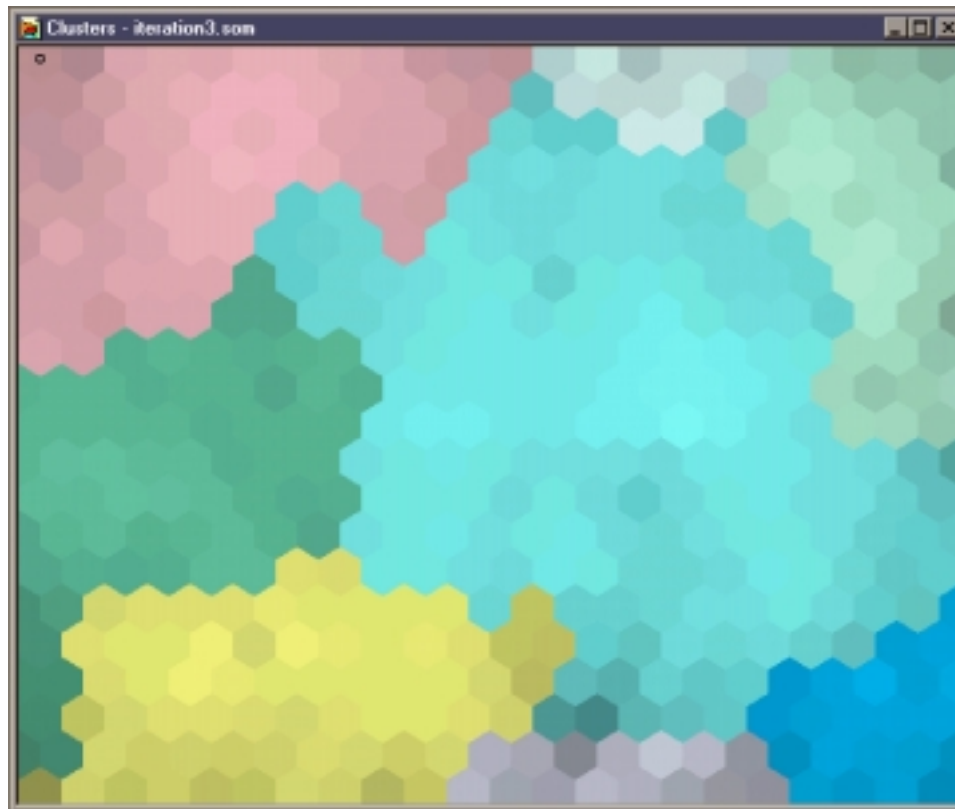# CREDIT RATING PREDICTION
USING SELF-ORGANIZING MAPS

Visually exploring and constructing a quantitative model

Roger P.G.H. Tan

CREDIT RATING PREDICTION
USING SELF-ORGANIZING MAPS

Visually exploring and constructing a quantitative model

Roger P.G.H. Tan

Studentnr. 140033

Erasmus University Rotterdam

Faculty of Economics

July 2000

contents

3 self-organizing maps 25

preface

This masters thesis forms the conclusion to my study of Econometrics, with specialization Business Oriented Computer Science, at the Erasmus University of Rotterdam. It was written during my internship at the Quantitative Research (QR) department of the Rotterdam based asset-manager Robeco Group. My time at the Robeco Group has been very enjoyable, and the combination of practical research and writing at the same time has proven to be a very relaxed and sure way of writing a thesis. I can recommend this to everyone in the final stage of his or her study.

This thesis is targeted at readers from two different scientific areas (computer science and financial econometrics), so some concepts are treated more extensively than first may seem necessary. Considerable time was also spent making this thesis into an attractive package, but at all times have I striven to keep looks and content in good balance.

Naturally I could not have written this without the comments and encouragement I received from many people, some of which I would like to especially mention: First and foremost I would like to thank dr.ir. Peter Ferket, my mentor at Robeco and head of QR, and dr.ir. Jan van den Berg and drs. Willem-Max van den Bergh, both associate professors at the faculty of Economics at the Erasmus University. They all provided invaluable comments on this thesis in its several stages of development. Furthermore my gratitude goes out to the members of the Credits research team, to my roommates and to the other colleagues at QR, for answering the many questions a Computer Science graduate inevitably has when acting like an econometrician. Finally I want to say thanks to dr. Guido Deboeck (Virtual Imagineer, U.S.A.) and dr. Gerhard Kranner (Eudaptics, Austria) for taking the time to answer my many emails, providing new insights and a better understanding of Self-Organizing Maps. Eudaptics also generously supplied me with the latest version of their Viscovery SOMine software, so that I could focus on the real research subject instead of having to devote time to programming.

As much as I have loved the past few years I spent partly studying, partly working and partly partying, I'm glad this stage of my life has come to a conclusion. I'm looking forward to put even more energy into my new job as I have put into this thesis.

Roger Tan, July 2000

1 introduction

In chapter 1 we introduce the main problem and the research topics for this thesis. Paragraph 1 gives a brief overview of the problem setting and paragraph 2 describes the domain of research. Paragraph 3 reviews the central question and several sub-questions to be answered in the remainder of this thesis.

1.1 Overview

When you lend a sum of money to someone you will most likely first estimate the probability of not being paid back. A correct assessment of this probability is based on the (observed) trustworthiness of the person in question and on your knowledge of her or his financial situation.

When investors lend money to companies or governments it is often in the form of a bond; a freely tradable loan issued by the borrowing company. The buyers of the bond have to make a similar assessment on creditworthiness of the issuing company, based on its financial statement (balance sheet and income account) and on expectations of future economic development.

Most buyers of bonds do not have the resources to perform this type of difficult and time-consuming research. Fortunately so-called rating agencies exist who specialize in assessing the creditworthiness of a company. The resulting credit or bond rating is a measure for the risk of the company not being able to pay an interest payment or redemption of its issued bond. Furthermore, the amount of interest paid on the bond is dependent on this expected chance of default[1]. A higher rating implies a less riskful environment to invest your money and less interest is received on the bond. A lower rating implies a more riskful company and the company has to pay more interest for the same amount of money it wants to borrow.

Unfortunately not all companies have been rated yet. Rating agencies are also often slow to adjust their ratings to new important information on a company or its environment. And sometimes different rating agencies assign different ratings to the same company. In all of these situations we would like to be able to make our own assessment of creditworthiness of the company, using the same criteria as the rating agencies. The resulting measure of creditworthiness should be comparable to the rating issued by the rating agencies.

This is more difficult than it first may seem, because the rating process is somewhat of a black box. Rating agencies closely guard their rating process; they merely state that financial and qualitative factors are taken into account when assigning ratings to companies. We would like to try to open this black box by describing the relationship between the financial statement of a company and its assigned credit rating. This might enable us to say how much of a company's rating is affected by the qualitative analysis performed by the rating agency. And the found knowledge could of course be used to classify companies that have not yet been rated or recently have substantially changed.

---

[1] A company 'defaults' when it has missed a redemption payment or an interest payment.

Several techniques have been developed for these kind of analyses. We will focus on a less common technique called Self-Organizing Maps, which is a combination of a projection and a clustering algorithm. Its main advantages are the insightful visualizations of large datasets and its flexibility.

1.2 Research domain

1.2.1 Bond ratings

Bond ratings are letter values on an ordinal scale, giving an opinion of creditworthiness of the issuer of a bond. The two most important rating agencies (issuers of ratings) are Standard & Poor's and Moody's. The ratings issued by these two agencies are comparable, but in this thesis we will focus on Standard & Poor's.

Examples of ratings are AA or B, the full rating scale is shown in table 1-1. A low rating (e.g. CC) corresponds to a high default risk, a high rating (e.g. AA) corresponds to a low default risk. A 'D' indicates an actual default on the bond. The scale is even more refined by appending '+' or '-' to the letter rating, indicating a slightly better or slightly worse rating.

Nowadays, more and more companies have been rated, but still most rated companies are based in the United States of America. Also more historical data is available for these companies. Therefore, our research will be conducted using only U.S. based companies.

1.2.2 Financial data and ratings

Rating agencies claim that the issued ratings are based on (1) a quantitative analysis of the financial statement of a company and (2) a qualitative analysis of the company and the environment of the company: What is the long term strategy, are there any impending threats on future profitability not expressible in the financial statement (like lawsuits), and what is the economic outlook for the sector as a whole? We will treat the credit rating process extensively in chapter 2, but suffice it to say that the contribution of qualitative factors to the rating is unclear. We can clarify the relationship between financial data and credit ratings using quantitative techniques like the Self-Organizing Map and indirectly give an assessment of the contribution of qualitative factors.

Financial statement data on most US companies is available in huge databases from datasources like Compustat and WorldScope. The information in these databases could help us gain a better understanding of the relationship between financial information and bond ratings. It might even provide us with a means to correctly

Table 1-1 Credit rating scale

Standard & Poor's
AAA
AA+
AA
AA-
A+
A
A-
BBB+
BBB
BBB-
BB+
BB
BB-
B+
B
B-
CCC+
CCC
CCC-
CC
C
D

predict bond ratings, based on the stored financial data alone. However, transforming the stored data into knowledge is no trivial task.

1.2.3 Self-Organizing Maps

A common problem is the complex nature of large amounts of data. Our universe contains a large number of companies, and for each company many financial characteristics are available. This hinders the inference of sensible relationships; to cope with the problem specific techniques have been developed[2]. In this thesis we will focus on the Self-Organizing Map technique.

Self-Organizing Maps (SOMs) use an advanced algorithm to form an as good as possible representation of the data. Clusters of similar companies are identified and displayed on a map, using colours to enhance the representation. The voluminous original dataset is compressed into a 2-dimensional, easily readable map. The contributions of individual characteristics are also part of the display, making it possible to visually infer relationships from the underlying data.

The Self-Organizing Map can be used as a visual exploration tool and as a classification model. Both functions will be illustrated using our bond rating problem.

---

[2] Fayyad, U.M., 1996, Chapter 1.

1.3 Research topics

The earlier sketched domain forms the background for the following central question in this thesis:

> In what way can we use Self-Organizing Maps to explore the relationship between financial statement data and credit ratings?

This question can be broken down into the following five sub-questions:

1. What are credit ratings and how is the credit rating process structured?

An analysis of the Standard & Poor credit rating process gives us a better understanding of the relation between credit ratings and financial statement data.

2. What are Self-Organizing Maps and how can they aid in exploring relationships in large data sets?

Before we can trust the results inferred from the SOM maps we first have to understand how the SOM gives a view on the underlying data. We provide an in-depth review of the algorithm itself and a guide on how to interpret the generated results.

3. Is it possible to find a logical clustering of companies, based on the financial statements of these companies?

First we would like to know if companies are discernible based on financial statement data alone.

4. If such a clustering is found, does this clustering coincide with levels of creditworthiness of the companies in a cluster?

We then compare the found clustering with the distribution of the ratings over the companies to determine to what extent they coincide.

5. Is it possible to classify companies in rating classes using only financial statement data?

Using previously found knowledge we set up a model specifically suited to the task of classifying new companies using financial statement data.

This thesis is divided into several chapters. Chapter 1 contains the introduction, a description of the research domain and this overview of the research topics. In chapter 2 we give a theoretical treatment of the credit rating process and in chapter 3 we provide an in-depth review of Self-Organizing Maps. Chapter 4 discusses the descriptive analysis after which chapter 5 focuses on the classification model. In chapter 6 we draw our conclusions and present some suggestions for further research.

2 credit ratings

This chapter provides a background on credits and credit ratings. Question 1 from the introduction is answered:

1. What are credit ratings and how is the credit rating process structured?

Paragraph 1 addresses the theoretical foundations of credits and credit ratings. Paragraph 2 reviews the rating process of Standard & Poor's, a well-known rating agency. Paragraph 3 evaluates the key financial ratios applicable to the economic sector under scrutiny in this thesis, Consumer Cyclicals.

2.1 Credits and credit ratings

2.1.1 Bonds

In its most simple form a bond is a loan from one entity to the other. The entity that receives the loan (this is often a government or a large company) is called the obligor or issuer, the loan itself is called a bond obligation or issue. The bond is freely tradable on the exchanges and split up into smaller parts, to make the bond more marketable.

Bonds belong to the group of fixed-income instruments, because they periodically pay a fixed amount (the coupon) to the buyer of the bond. Bonds differ from equity (or stockholders shares) in that buyers of bonds do not become owners of the company. When a company goes into bankruptcy, the owner of the bond is in a better position than the shareholder because first all the loans are redeemed, and from which is left (if any) the owners are repaid.

Characteristics

Each bond has certain characteristics, which fully describe the bond. The bond has to be redeemed on a fixed date, called the maturity date. Bonds with original maturities longer than a year are considered long-term, all bonds with maturities up to one year are considered short-term. Each period a certain interest percentage has to be paid in the form of the coupon. Often this percentage is fixed, but sometimes this percentage is dependent on the market interest rate (the coupon is floating). Other variations on the standard bond include sinking redemptions (periodically a part of the bond is redeemed), callable bonds (at certain dates the issuer has the right to prematurely redeem the bond), and of course special combinations leading to more exotic variants.

Value

The value of a bond depends largely on the coupon percentage and the current market interest rate. If the market interest rate rises, then the value of the bond lowers. The coupon percentage is fixed, and investors would rather buy a new bond with a coupon that is more in-line with the current market interest rate. If the market interest rate declines, then the value of the bond rises. Investors would rather buy our bond than new bonds with lower interest rates.

The value of the bond is determined in the market, by the forces of supply and demand. Using the market price the current yield of the bond can be calculated. This is the internal discount factor needed when discounting all future cash flows of the bond (coupon payments and redemption payment) to represent the current price. This

yield is often used when comparing bonds, as it is based on the current price of the bond, thus taking the coupon, the market interest rate and other factors into account. The difference between two yields is referred to as a spread. A whole range of appropriate government bonds (combining to a government curve[3]) is most often used as the benchmark, so the spread of the bond then means the difference between the yield of the bond and the yield of a comparable government bond on the government curve.

Default

Another factor influencing the value of the bond is the default risk associated with the bond. When an issuer is unable to meet one of the payments with respect to a specific bond, we say that the issuer has 'defaulted' on the bond. This does not necessarily mean that the issuer has gone bankrupt, a missed or delayed interest payment also counts as a default. If the issuer settles the payment a few days later the issuer has 'recovered'. How the spread is influenced by the default risk is explained in the next paragraph.

2.1.2 Credits

We use the term 'credits' for all bonds not issued by central governments in their own currency. All bonds issued by companies (also known as 'corporate bonds') are good examples of credits. Governments in emerging markets often issue their bonds denominated in U$, so these are credits too. All credits are inherently riskier than issues by stable governments in developed markets like the United States or the Netherlands: Because the company or the unstable government is prone to financial problems we cannot say with 100% certainty that all payments on the bond will be fulfilled.

Credit spread

When investors buy credits, they want something in return for the extra risk involved; this is known as the credit spread. Some extra yield is received to compensate for the default risk. Naturally, this credit spread is larger when the risk of default is larger. Conversely, the credit spread is smaller when the perceived risk is lower. Some credits are more eligible for repayment than others (they are 'senior' to other issues from the same issuer). Also sometimes issues are secured by e.g. a parent company. As this all reduces the risk involved with the credit, the accompanying credit spread also narrows.

The credit spread is only part of the difference between the yield of a bond and a comparable government bond. Other factors are the liquidity of the bond (large issues are more easily traded than smaller issues) and the inclusion of the bond in a bond index (bonds included in indices composed by e.g. J.P. Morgan are more in-demand by investors and thus more valuable).

---

[3] See Fabozzi, F.J., 1993, Chapter 13.

2.1.3 Credit ratings

According to Standard & Poor's (S&P), "the bond or credit rating is an opinion of the general creditworthiness of an obligor with respect to a particular debt security or other financial obligation, based on relevant risk factors."[4] All rating agencies seem to support this definition.

Rating agencies

A rating agency, of which S&P is one of the best known examples, assesses the relevant factors relating to the creditworthiness of the issuer. These include the quantitative factors like the profitability of the company and the amount of outstanding debt, but also the qualitative factors like skill of management and economic expectations for the company. The whole analysis is then condensed into a letter rating[5]. Standard & Poor's and Moody's both have been rating bonds for almost a century and are the leading rating agencies right now. Other reputable rating institutions are Fitch and Duff & Phelps.

Ratings interpretation

The types of assigned ratings are comparable for most agencies, and for S&P and Moody's there is a direct

Table 2-1 Credit ratings and interpretation

S&P	Moody's	Interpretation
AAA	Aaa	Highest quality
AA+	Aa1	High quality
AA	Aa2	
AA-	Aa3	
A+	A1	Strong payment capacity
A	A2	
A-	A3	
BBB+	Baa1	Adequate payment capacity
BBB	Baa2	
BBB-	Baa3	
BB+	Ba1	Likely to fulfil obligations; ongoing uncertainty
BB	Ba2	
BB-	Ba3	
B+	B1	High risk obligations
B	B2	
B-	B3	
CCC+	Caa	Current vulnerability to default, or in default (Moody's)
CCC		
CCC-		
CC	Ca	
C		Bankruptcy filed
D	D	Defaulted

---

[4] Standard & Poor's, 2000, page 7.

relation[6]. The ratings for S&P and for Moody's and the accompanying interpretation is shown in table 2-1.

The letter rating is sometimes augmented by a '+' or '-' (for S&P) or a '1', '2', or '3' (for Moody's). These indicate sub-levels of creditworthiness within a specific rating class. The difference between the sub-levels is called a 'notch', so an 'A+' and an 'A-' rating differ two notches. In practice, this is also used over the rating classes, so a 'B+' and an 'A' rating are also said to differ two notches.

Comparing ratings

The differences between regions, countries or even economic sectors can be so large that it is difficult to arrive at a certain rating when using the same criteria. To make comparisons possible, the rating agencies use different criteria and special risk characteristics for companies in different sectors and different countries or regions.

A good example is the qualitative assessment of an industrial company; business fundamentals then include technological change, labour unrest or regulatory actions. For a financial institution we would be looking at the reputation of the institution and the quality of the outstanding debt.

Issuer credit ratings

An issuer credit rating forms an opinion of the obligors overall capacity to meet its financial obligations, even when there is no public debt outstanding. This does not take into account the specific nature or provisions of any particular obligation. Issuer credit ratings are requested by companies to facilitate the negotiation of loans and long-term leases: A letter of credit is superfluous when a rating has been assigned to the company.

Companies issue several kinds of bonds. Some are more eligible for repayment (more senior) than others, leading to higher ratings for these specific bonds. There is a fixed relation between the ratings for different types of bonds issued by the same company: The rating for the senior unsecured debt is equal to the rating for the company as a whole, the issuer rating. Subordinated debts are rated one or two notches (subclasses) lower than the senior unsecured debt rating. In this thesis we will mainly focus on issuer ratings.

2.1.4 Ratings and default risk

The rating provides a relative rank ordering of creditworthiness. If we want this rank ordering to be of practical value, then it should also provide a guideline for the absolute risk involved; the chance of default. If the rating is a good assessment of the default risk, then the percentage of defaults should increase for the lower rating classes and decrease for the higher rating classes.

---

[5] Please refer to paragraph 2.2 for a comprehensive review of the credit rating process of Standard & Poor's.
[6] Cantor, R. and Packer, F., 1994, page 6.

Figure 2-1 shows the default rates corresponding to Moody's rating classes for 1999[7]. As is to be expected, the lower rating classes have corresponding higher default rates.

Default rates for 1999
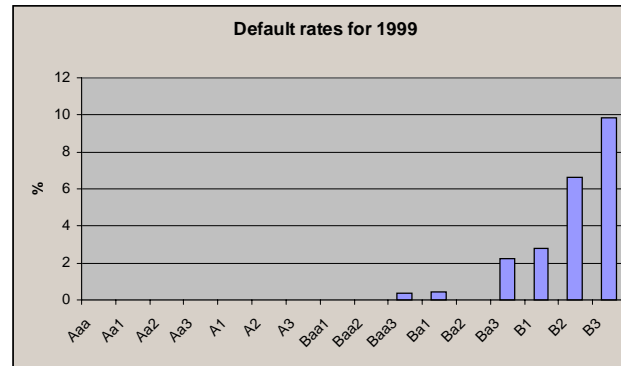
Figure 2-1 Default rates for 1999

Investment grade versus speculative grade

Credits with an assigned rating from AAA to BBB- are known as investment grade credits. Lower rated issues are known as speculative grade credits, high yield issues or junk bonds. The spreads on these high yield issues are relatively wide, thus providing an interesting investment opportunity. This is even more so after finding an average recovery rate of 42%[8] (for every U$ 100 worth of defaults on average U$ 42 recovers).

Sometimes fundmanagers are restricted to purchasing investment grade issues, to avoid speculative investments. However, the absolute default rates do not remain stable over the years. For example, restricting the fund managers to purchase at least BBB- grade issues does not guarantee lower than 1% default rates.

---

[7] Moody's, 2000, page 26.
[8] Moody's, 2000, page 17.

2.2 The S & P credit rating process

"The rating experience is as much an art as it is a science." – Solomon B. Samson, Chief Rating Officer at Standard & Poor's[9].

This paragraph describes the credit rating process of Standard & Poor's. Most information contained in this paragraph was taken from the "Corporate Ratings Criteria" document, on-line published at the S&P website. In this document, the distinction between the qualitative and the quantitative analysis is less clear. The qualitative analysis is most extensively treated and thus most emphasized. The descriptive analysis in chapter 4 will try to uncover whether this depiction reflects the actual rating practice of S&P.

2.2.1 Process steps
The Standard & Poor's credit rating process can be broken down into several steps. The process is summarized in figure 2-2.
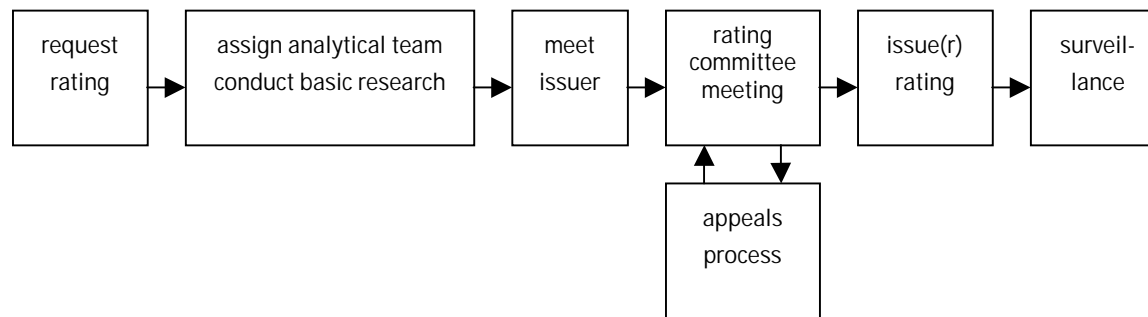
Figure 2-2 The Standard & Poor's credit rating process

Request rating
Companies themselves often approach Standard & Poor's to request a rating. In addition to this, it is S&P's policy to rate any public corporate debt issue larger than U$ 50 million, with or without request from the issuer.

Basic research
When the rating is requested a team of analysts is gathered. The analysts working at S&P each have their own sector specialty, covering all risk categories in the sector.

---

[9] Standard & Poor's, 1999.

The appropriate analysts are chosen and a lead analyst is assigned, who is responsible for the conduct of the rating process.

Some basic research is conducted, based on publicly available information and based on information received from the company prior to the meeting with the management[10]. The information requested prior to the meeting should contain:

- five years of audited annual financial statements (balance sheet and profits and losses account),

- the last several interim financial statements (this is mostly applicable to US companies, as they are required by law to provide quarterly financial statements),

- narrative descriptions of operations and products,

- relevant industry information.

As some of this may be sensitive information, S&P has a strict policy of confidentiality on all the information obtained in a non-public fashion. Any published rationale on the realization of the assigned rating only contains publicly available information.

Meeting the issuer

In the next step a part of the team meets with management of the company to review key factors that have an impact on the rating. This meeting covers the operating and financial plans of the company and the management policies but it is also a qualitative assessment of management itself. The meeting is scheduled well in advance so ample time for preparation is given.

The specific topics discussed at the meeting are:

- the industry environment and prospects,

- an overview of the major business segments, including operating statistics and comparisons with competitors and industry norms,

- management's financial policies and financial performance goals,

- distinctive accounting practices,

- management's projections, including income and cash flow statements and balance sheets, together with the underlying market and operating assumptions,

---

[10] So called 'public information ratings' are the exception to this rule; they are solely based on the annual publicly available financial statement.

- capital spending plans,

- financing alternatives and contingency plans.

Standard & Poor's does not base its rating on the issuers financial projections, but uses them to indicate how the management assesses potential problems and future economic developments.

Rating committee and appeals process

Shortly after the meeting with the management of the issuer the rating committee convenes. The rating committee consists of five to seven voting members, who will decide on the rating using information presented by the lead analyst. His presentation covers:

- an analysis of the nature of the company's business and its operating environment,

- an evaluation of the company's strategic and financial management,

- a financial analysis,

- and finally a rating recommendation.

After a discussion about the rating recommendation and the facts supporting it the committee votes on the recommendation. The issuer is notified of the rating and the major considerations supporting it. An appeal is possible (the issuer could possibly provide new information), but there is no guarantee that the committee will alter its decision.

Publishing the rating

For public issues the new rating is published using several media, e.g. the Internet site or the "CreditWeek" publication by Standard & Poor's. For ratings assigned on request by the issuer, the company itself may determine if they want the rating to be publicly available or not. This will often be the case, because rating requests are expensive and a public rating facilitates the negotiations for loans and leases.

Surveillance

The rated issues and issuers are being monitored on an ongoing basis. New financial or economic developments are reviewed and often a meeting with the management is scheduled annually. If these developments might lead to a rating change, this will be made known using the CreditWatch listings. A more thorough analysis is performed, after which the rating committee again convenes and decides on the rating change.

2.3 Financial statement analysis

2.3.1 Financial statement

The financial statement of a company comprises the balance sheet and the profits and losses account. There are strict accounting regulations the financial statement must adhere to, which vary for different countries. The financial statements for companies in different sectors also diverge: We would expect a factory to have a raw materials inventory on its balance sheet, but not a bank. The most important differences occur between Financial companies and Industrial companies, the next section describes the financial ratios that are most applicable to Industrial companies.

2.3.2 Financial ratios

The financial performance of a company can be analyzed by carefully examining the balance sheet and income statement for that company. To make these large quantities of data more comprehensible and to make comparisons between firms possible one often uses financial ratios.

There are several financial ratio classes:

- leverage ratios measure the debt level of a company,

- liquidity ratios measure the ease with which a company can acquire cash,

- profitability ratios measure the profits of a company in proportion to its assets.

In addition to these financial ratios a few other classes of variables can be observed to characterize a company:

- size variables measure the size of a company,

- stability variables measure the stability of the company over time in terms of size and income,

- market value ratios measure the value investors assign to a company.

Although financial ratios provide a means to quickly compare companies, some caution should be taken when using them. Companies often use different accounting standards, so two comparable companies can have very different values for certain ratios just because of different ways of valuing the items on the balance sheet. Furthermore, companies often want to present an as favourable as possible image, known as 'window dressing'. This also leads to ratios not fully representing the real financial state of the company.

2.3.3 Balance sheet and income statement

The financial ratios are calculated using elements from the balance sheet and from the income statement of a company. They are shown in table 2-2 and table 2-3.

Table 2-2 Balance sheet

Assets	Liabilities
+ cash & equivalents	+ total short term debt
+ total net receivables	+ accounts payable
+ total inventory	+ other current liabilities
+ other current assets	+ income taxes payable
total current assets	total current liabilities
+ net property, plant & equipment	+ total long term debt
+ investment & advances	+ other non-current liabilities
+ intangibles	+ deferred income taxes & investment tax credit
+ other assets	+ minority interest
total assets	total liabilities
	+ preferred stock
	+ total common equity
	total liabilities & capital

Table 2-3 Income statement

Income statement
+ net sales
- cost of goods sold
- other expenses
earnings before interest, taxes, depreciation and amortization
- depreciation and amortization expense
earnings before interest and tax
- gross interest expense
+ special items (non-recurring)
pre-tax income
- total income taxes
- minority interest
net income
- preferred dividends
earnings applicable to common stock

2.3.4 Used ratios

Our preliminary selection yielded the following financial ratios.

Interest coverage ratios

These measure the extent to which interest or debt is covered by the earnings of a company.

EBIT interest coverage:

(earnings before interest and taxes) / (interest expenses)

EBITDA interest coverage

(earnings before interest, taxes, depreciation and amortization) / (interest expenses)

EBIT / total debt

(earnings before interest and taxes) / (total debt)

Leverage ratios

Financial leverage is created when firms borrow money. To measure this leverage, a number of ratios are available.

Debt ratio

(long term debt) / (long term debt + equity + minority interest)

Debt-equity ratio

This can be measured in several ways, two of which are:

(long term debt) / (equity)

and

(long term debt) / (total capital)

Net gearing

(total liabilities – cash) / (equity)

Profitability ratios

Profitability ratios measure the profits of a company in proportion to its assets.

Return on equity

This measures the income the firm was able to generate for its shareholders[11].

>   (net income) / (average equity)

Return on total assets

>   (earnings before interest and taxes) / (total assets)

Operating income / sales

>   (operating income before depreciation) / (sales)

Net profit margin

>   (net income) / (total sales)

Size variables

These measure the size of a company.

Total assets

>   The total assets of the company.

Market value

>   Price per share * number of shares outstanding.

Stability variables

Stability variables measure the stability of the company over time in terms of size and income.

Coefficient of variation of net income

>   (standard deviation of net income over 5 years) / (mean of net income over 5 years)

Coefficient of variation of total assets

>   (standard deviation of total assets over 5 years) / (mean of total assets over 5 years)

Market variables

Market variables are used to assess the value investors assign to a company.

---

[11] Note the use of the average of the equity (at the beginning and the end of the quarter). Averages are often used when comparing flow data (net income) with snapshot data.

Coefficient of variation of earnings forecasts (fiscal year 1)
This measures the risk encapsulated in the earnings forecasts (for fiscal year 1) of the several analysts. If the analysts do not agree with each other, that should be an indication for higher risk involved with this company.

(standard deviation of forecasts fiscal year 1 over analysts) / (mean of forecasts fiscal year 1 over analysts)

Market beta relative to NYSE
The beta is the sensitivity of the stock to market movements, in this case movements of the New York Stock Exchange[12]. A snapshot is taken on the last trading day of the quarter.

Earnings per share
This is calculated for the last month of the quarter.

(earnings applicable to common stock) / (total number of shares)

---

[12] Brealey, R.A. and Myers, S.C., 1991, chapter 7.

2.4 Summary

In this chapter we have reviewed some theoretical aspects of bonds and credits before exploring the ratings domain. The credits we are most interested in are bonds issued by companies (corporate bonds). We have seen the direct relation between creditworthiness, default probability and spread of a credit. If the perceived creditworthiness is better, then the assigned rating will be higher and the default probability will be lower. The difference in yield with a similar government bond (also known as the spread) will be subsequently lower.

The different process steps of the Standard & Poor's credit rating process emphasize the qualitative analysis performed by the agency. The quantitative analysis, based on financial statement data, is just a single step in the process. In the remainder of this thesis we will try to uncover whether actual rating practice reflects this depiction of matters, using the described financial ratios. These ratios form a means to summarize the balance sheet and income statement of a company and to compare the financial statements of different companies.

3 self-organizing maps

Chapter 3 reviews the Self-Organizing Map and its place in the knowledge discovery process. To provide a background for the SOM we will briefly discuss some related techniques before examining the Self-Organizing Map algorithm. Altogether this answers question 2 from the introduction:

2. What are Self-Organizing Maps and how can they aid in exploring relationships in large data sets?

Paragraph 1 describes the knowledge discovery process. Paragraph 2 describes some projection and clustering methods related to SOM. Paragraph 3 describes the classification techniques that we also use in the classification model of chapter 5. The remainder of the chapter is dedicated to an explanation of SOM and guidelines for the use of SOM.

3.1 Knowledge discovery

3.1.1 Introduction

These days it is quite common for corporations of all kinds and sizes to gather large amounts of data. This may vary from customer data (e.g. scanned purchase data for supermarkets) to data regarding some of the processes within a company (e.g. process states of a machine). On a meso-economic and macro-economic level a lot of data is available too, concerning the financial statements of individual companies or the financial statements of countries.

The volumes of these databases are often gigantic, making it impossible to retrieve sensible information just by looking at the raw data. To gain access to the knowledge contained in the stored data one has to rely on specific techniques, which extract information from the database in a systematic way. In the ICT sector these techniques are referred to as data-mining[13] techniques, and all the steps necessary to extract knowledge from databases is known as the knowledge discovery process.

3.1.2 Knowledge discovery process

The knowledge discovery process encompasses all the steps necessary to extract potentially useful information (knowledge) from the database[14].

The basic steps (displayed in figure 3-1) involve:

- Creating a target data set based on the available data, the knowledge of the underlying domain and the goals of the research.

- Pre-processing this data to account for extreme values and missing values.

- Applying any necessary transformations.

- 'Mining' the data so distinct patterns become available for interpretation and evaluation. In this thesis we will focus on visualization techniques, whereby specific patterns can be found in the resulting maps.

Figure 3-1 The knowledge discovery process
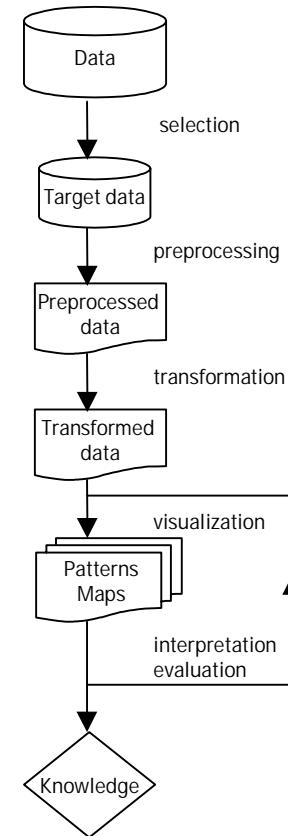
---

[13] Computer scientists use the term data-mining in a positive context (extracting previously unknown knowledge from large databases), econometricians use the term data-mining in a negative context (manipulating data and the used technique to support specific conclusions). This sometimes leads to confusion about the intended meaning.
[14] Fayyad, U.M., et al, 1996, chapter 2.

-     Interpreting and evaluating these maps, often repeating one or more steps of the process.

3.1.3 Description and prediction

The knowledge discovery process serves two main purposes: description and prediction. Descriptive knowledge discovery tries to correctly represent the data in a compact form. The new representation implicitly or explicitly shows relationships in the data. Not so obvious relationships emerge, thus attributing to a greater knowledge of the underlying domain. Obvious relationships are of course visible too, strengthening the image one has of the data based on preliminary research. Common used techniques are projection and clustering algorithms.

Predictive knowledge discovery is used to complement values for one or more characteristics (or variables) of observations in the data set. This is often in the form of a classification problem: A data set with known class memberships is used to build a model, and this model is used to predict the class membership for new observations. Common used techniques are linear regression based classifiers like ordered logit and artificial neural networks.

Of course this division is not strict. Some of the algorithms are combinations of techniques, and often the descriptive techniques are used as an intermediate step in large investigations. The output of the descriptive analysis then may serve as input for some of the prediction algorithms.

In the following sections we will highlight some of the available projection, clustering, and classification techniques. The Self-Organizing Map, treated extensively in the remainder of the chapter, is actually a neural network combining regression, projection and clustering!

3.2 Projection and clustering techniques

We use projection techniques to reduce the dimensionality of the data, making it easier to grasp the essence of the data. Projection techniques can be split into two groups, linear and non-linear projection methods. On the other hand, clustering techniques are designed to reduce the amount of data by grouping alike items together. The dimensionality of the data does not change. The several clustering methods can be split into two common types, hierarchical and non-hierarchical clustering.

3.2.1 Linear projection

Linear projection methods use a linear combination of the components of the original data to project the data onto a new co-ordinate system of lower dimensionality using a fixed set of scalar coefficients.

Principal component analysis (PCA) is a commonly used linear projection method. The PCA technique tries to capture the intrinsic dimensionality of the data by finding the directions in which the data displays the greatest variance. Often the data is stretched in one or more directions and has an intrinsic lower dimensionality than it first may seem (see figure 3-2). These directions in the data are called 'principal components'. The first principal component describes the direction of the largest variation in the data. The second principal component, orthogonal to the first, describes the direction of the second-largest variation in the data, etcetera. The variation in the data that has not been described by the first N principal components is called the residual variance.
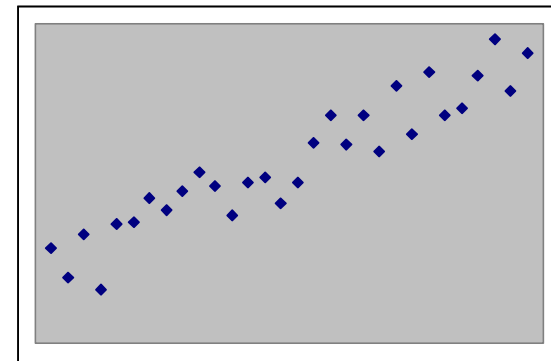
Figure 3-2 Two dimensional data stretched in one direction

The data is projected onto a new co-ordinate system spanned by the first two principal components, to give a more accurate view of the data. A drawback of linear projection methods is that they can not take non-linear or arbitrarily shaped structures in the data into account, possibly leading to incorrect projections.

In chapter 4, we compare the PCA technique with SOM. A full explanation of principal components can be found in Johnson and Wichern[15].

---

[15] Johnson, R.A., and Wichern, D.W., 1992, chapter 8.

3.2.2 Non-linear projection

Several techniques exist to project the non-linear structures in the data. They often focus on correctly displaying the differences between observations in the original data space.

Multi Dimensional Scaling (MDS)[16], developed by J.B. Kruskal during the sixties and seventies, actually denotes a whole range of techniques. It aims at placing the original, high dimensional data points on a lower dimensional display in such a way that the relative rank ordering of similarity between observations in the input space is preserved as much as possible. The new distance between the two least similar observations is largest, and vice versa the new distance between the two most similar observations is smallest.

The specification of the similarity measure defines the specific used version of MDS; metric MDS uses Euclidean distances[17] in the input space, non-metric MDS uses domain specific relative rank orderings.

One interesting application of non-metric MDS can be found in archaeology for the reconstruction of the geography of the Mycenaean kingdom of Pylos in Greece (circa 1200 BC)[18]. The found Palace archives (clay tablets) contain no direct geographical information, but relative distances between cities can be inferred from them. The MDS based map of the kingdom (figure 3-4) matches the map drawn by experts (figure 3-3) quite closely.
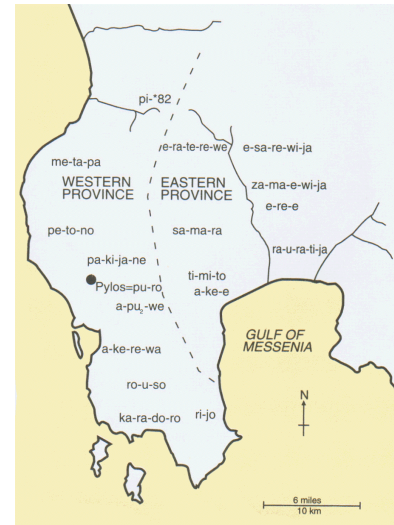

Figure 3-3 Expert map of Pylos kingdom

Figure 3-4 MDS map of Pylos kingdom

---

[16] Johnson, R.A. and Wichern, D.W., 1992, pages 602-608

[17] The Euclidean distance $d(x,y)$ between vectors x and y is defined as $d(x,y) = \sqrt{(x_1+y_1)^2+(x_2+y_2)^2+...+(x_n+y_n)^2}$

[18] See "http://www.archaeology.usyd.edu.au/~myers/multidim.htm" for more information.

3.2.3 Hierarchical clustering

Hierarchical clustering techniques group data items according to some measure of similarity in a hierarchical fashion. They can be divided into splitting and merging methods.

Splitting methods work top-down, starting with one big cluster. At each step the cluster is divided into two separate clusters thereby maximizing some inter-cluster distance measure *d*. The divisional process is stopped when *d* becomes too small. The found division of the data set is equivalent with a binary tree structure.

Figure 3-5 Clustering car brands using merging

Merging methods work bottom-up, starting with each case in a separate cluster. Clusters having the least inter-cluster distance d are merged, often the Euclidean distance is used for d. An example clustering of car brands is shown in figure 3-5.

3.2.4 Non-hierarchical clustering

Non-hierarchical or partitional clustering methods try to directly divide the data into a set of disjoint clusters. This is done in such a way that the intra-cluster distance is minimized and the inter-cluster distance is maximized.

K-means clustering is a non-hierarchical clustering method that is very much related to Self-Organizing Maps. A set of K reference vectors is chosen with the same dimensionality as the input data. Then for each reference vector a list is made of the observations lying most closely to the reference vectors. The reference vectors are then recomputed by taking the mean over the respective list. Each reference vector (also called 'centroid') thus represents the centre of the cluster. This is repeated until the reference vectors do not change much anymore.

3.3 Classification techniques

The techniques treated in this paragraph can all be used as classification methods. Linear regression and neural networks are more general methods that can also be used to solve other kinds of problems. The ordered logit model is specifically used for classification problems. All three techniques are used in chapter 5.

3.3.1 Linear regression

The multiple linear regression model is used to study the relationship between a dependent variable and several independent variables. The regression equation has the following form:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \, , \, i = 1,...n,$$

where y is the dependent or explained variable, $x_1,...,x_k$ are the independent or explanatory variables (also known as regressors), and i indexes the n sample observations. The disturbance ε is used to model external random influences that we can not capture with the model (e.g. errors of measurement). The coefficients of the independent variables ($\beta_1...\beta_k$) and the disturbance are most often estimated using the Ordinary Least Squares technique. Before we do this a number of assumptions have to be satisfied concerning amongst others the dependencies between variables and the distribution of the disturbances. A full overview of the multiple linear regression model is given in Greene[19].

3.3.2 Ordered logit

The ordered logit model is a so called ordered response model. It is an extension of the binary logit model, which is a regression-based technique: A latent variable is assumed to be the determining factor for class membership. This latent variable is linearly dependent on several regressors and a disturbance.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \, , \, i = 1,...n$$

We assume a logistic distribution for the disturbance ε, hence the name ordered logit. Although the classes have to be ordered they need not be of equal width. The classification is seen as a transformation of the latent variable and derived from y using

$$x_i \in c_1 \text{ if } y_i \leq \alpha_1$$

$$x_i \in c_j \text{ if } \alpha_{j-1} < y_i \leq \alpha_j \text{ for } j = 2,...,m\text{-}1$$

---

[19] Greene, W.H., 1997, Chapter 6.

$$x_i \in c_m \text{ if } \alpha_{m-1} < y_i$$

where c_j denotes class j, and α_j denotes the threshold for class j.

The class thresholds and the coefficients in the regression equation can be simultaneously estimated using Maximum Likelihood Estimators. More information on the ordered logit model can be found in Fok[20].

3.3.3 Artificial neural networks

An artificial neural networks is a conceptual network consisting of small simple processing elements called neurons. A short introduction to neural networks can be found in appendix I.

In many models the interconnected neurons are ordered in two or more layers. There should at least be an input and an output layer, any layers in between are called hidden layers. The extent to which each neuron reacts to its inputs (also called 'weight') is adjusted during a training phase. In this training phase observations of the required behaviour are sequentially presented to the network. At each observation the mathematical functions at the neurons are updated to better mimic the required behaviour. An example of a neural network is shown in Figure 3-6.
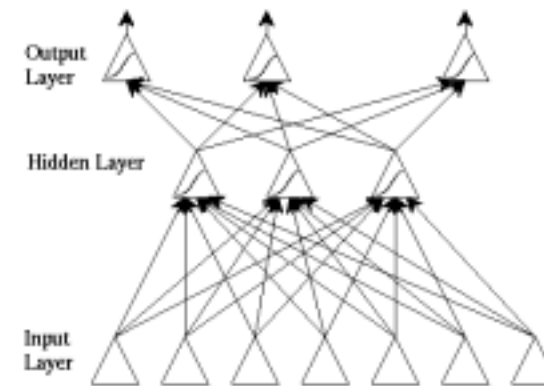
Figure 3-6 A neural network with one hidden layer

Feed forward and feed backward
In feed forward networks the input signal is propagated through any hidden layers to the output neurons. Feed backward (or backpropagation) networks also contain connections from neurons to preceding layers. Feed backward networks are more interesting for researchers (due to the more complex dynamics), but most real-world results have been achieved with feed forward networks.

Supervised and unsupervised
If a target value is taken into account when training the network we are using supervised learning. The network learns to correctly classify new observations. If the network is trained without taking a target value into account we are using unsupervised learning. The network is then trained to represent the distribution of the input in a compressed way. The Self-Organizing Map is a good example of an unsupervised learning neural network.

---

[20] Fok, D., 1999, Chapter 9.

3.4 Self Organizing Maps

3.4.1 Introduction

The self-organizing map (SOM) is a combination of a clustering and projection algorithm at the same time, driven by a neural network. The multi-dimensional input (e.g. companies with multiple financial ratios per company) is projected onto a 2-dimensional map, thereby preserving the local distances between the observations. The projected observations are subsequently merged into clusters, taking the placement on the map into account.

The model of the self organizing map was inspired by the human brain: The complex motoric and sensoric control of specific parts of the human body can be pinpointed to specific areas on a flat surface of the brain. More complex functions are appointed larger areas (or clusters) of brain tissue. The resulting man-like shape projected on the brain is known as the homunculus (figure 3-7).
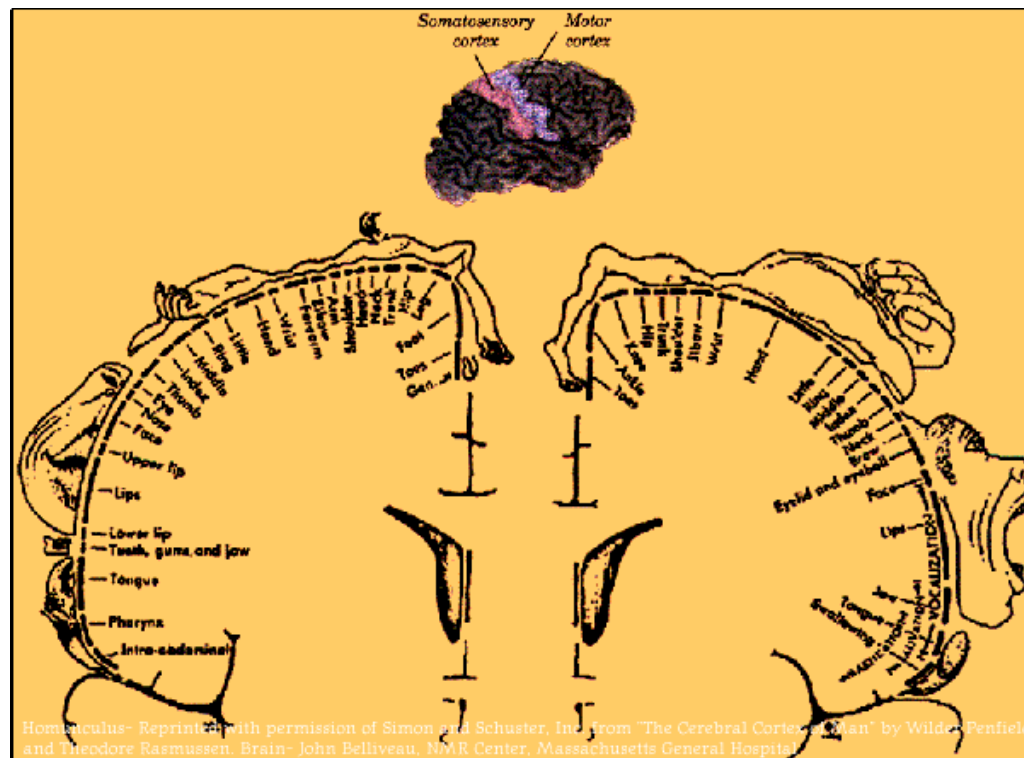
Figure 3-7 Picture of the homunculus in the brain, drawn by Wilder Penfield

3.4.2 Overview

The self-organizing map algorithm involves two steps. The first step projects the observations, the second step clusters the projected observations.

Projection

The first step of the algorithm involves projecting the observations onto a 2 dimensional, flexible grid composed of neurons or nodes. The grid is stretched and bended through the input space to form an as good as possible representation of the data. The projection on this grid is a generalization of simple projection (on the flat surface) and projection using Principal Component Analysis (PCA).

Simple projection simply projects the datapoints on the flat surface defined by the x and y axes. Projection using principal components is more advanced than simple projection (reflecting the intrinsic dimensionality of the data), but is still limited because the observations are projected on a flat plane. The flat plane is aligned according to the axes defined by the two directions inhibiting the largest variance of the data. The projection part of the SOM algorithm (also known as the self-organization process) can be thought of as a non-linear generalization of PCA[21]. The plane onto which the observations are projected can stretch and bend through the input space thus more thoroughly capturing the distribution of the observations in the input space.

The first two types of projections are often too restricted to fully capture the irregularities of the data. The three dimensional example in Figure 3-7shows this more clearly. The data is clustered in three distinct segments of the cube, simple projection projects the observations on the bottom of this cube (left picture). The flat plane
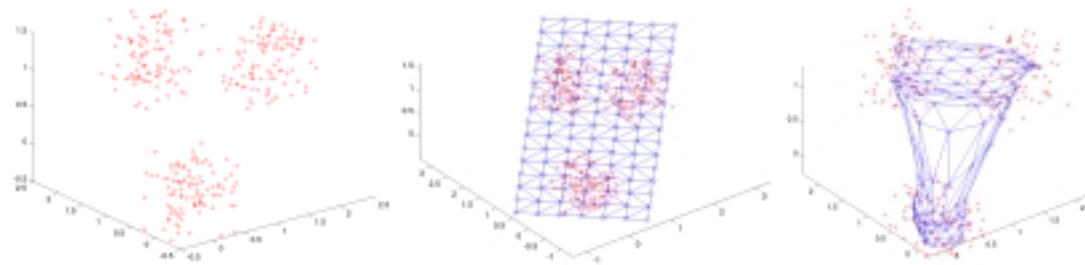
Figure 3-7 Plane of projection using the X-Y plane, using PCA and using SOM

shown in the middle picture is aligned along the first two principal components of the data. A projection on this surface gives a better representation of relative distances in the data set. The rightmost picture shows the flexible, bended and stretched grid used for SOM projection. By following the form of the data an even more

---

[21] Kaski, S., 1997.

accurate representation of relative distances in the data set is given. How the SOM achieves this projection is extensively treated in paragraph 3.5.

Clustering

The flexible grid, onto which the observations have been projected, is (for convenient output viewing) returned to a normal, unstretched flat plane and displayed as the map. The form of the grid in the input space remains fixed. The local ordering of the sample is preserved; neighbouring observations in the input space will be neighbouring observations on the map.

A bottom-up clustering method is used to cluster the projected observations: starting with each observation in a separate cluster, 2 clusters are merged if their relative distance (e.g. Euclidean distance) in the input space is smallest and if they are adjacent in the map. The number of shown clusters varies with the specific step of the algorithm we want to see. One step later in the algorithm means one less cluster shown (another cluster has merged), one step earlier means one more cluster shown.

Cluster are clear separations of the input space, so observations can only be member of one cluster (the clusters do not overlap). The clustering algorithm is discussed in paragraph 3.6.

3.5 SOM projection

3.5.1 The self-organization process

The observations are projected on the flexible grid using a special algorithm called the self-organization process. It involves changing the shape of the grid to conform to the shape of the data and projecting the observations on the shaped grid. The self-organization process accomplishes this concurrently as an iterative algorithm.

The flexible grid used as a basis for the SOM is usually considered to be a type of unsupervised feed forward neural network. The input consists of input vectors (observations), each consisting of often many variables. The input space is n-dimensional. Through this space a (usually) 2 dimensional grid is drawn, each grid point representing a neuron of the network. The grid, in an unstretched and flattened form, has an associated output representation as the map. The neurons are represented as grid points in the map.

The self-organization process assigns (projects), subsequently, each input vector to one of the neurons. The assignment of the input vector to the output neuron depends on the location of the input vector in the original input space: Input vectors situated nearby each other in the input space will be assigned to nearby (or even the same) neurons. They are thus placed nearby each other on the associated map grid points, this is known as 'topology preservation'. The neurons are not fixed in the input space; they move at each iteration of the algorithm, thereby stretching and bending the grid, to better accommodate the distribution of the input vectors.

Alternative interpretation

We can also interpret the self-organization process as a form of regression. Normal or parametrized regression tries to fit a line or curve through the observations using a presupposed form of the underlying function. Only the compression (coefficient) and the height (constant) of the curve are adjusted.

Instead of being tied to a fixed functional form the neural network of the SOM can vary over a whole class of functions to make an as good as possible approximate match of the data. However, the freedom of the network to find a functional form is restricted by the interconnections between the neurons, and the final form of the grid is fixed. Therefore it is called semi-parametrized regression.

When a priori a specific form of the function is not taken into account at all, we would refer to it as non-parametrized regression. Some kind of functional form can be found using representative reference vectors (e.g. averages over sets of observations), but these vectors do not influence each other, so any form is achievable.

3.5.2 A two dimensional example

For simplification, we first consider a 2-dimensional input space and a 1-dimensional output map. The neurons associated with the grid points in the output map can be seen as 'model vectors' for the data in the input space. They form a best representation for the data in that specific local part of the input space. Each neuron has associated values for each of the dimensions in the input space, and can be visualized in the input space. Of course a 2-dimensional input space is much easier to show than a 20-dimensional input space!

The input space with the observations is visualized in figure 3-8 , together with a line drawn through (or fitted to) these observations. This is the linear regression line, the observations are implicitly projected on this line.

Using SOM a neural network is drawn through the observations in the input space. At first the network is randomly initialized, as shown in figure 3-9. The self-organization process projects the observations on the neurons and shapes the network to conform to the observations by means of the algorithm in the following section.
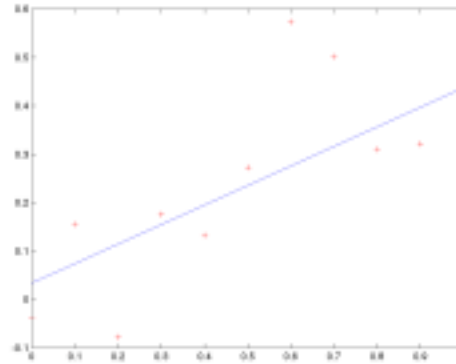
Figure 3-8 Linear regression line through the data points (red plusses)
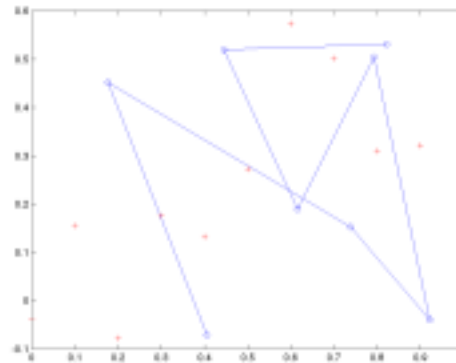
Figure 3-9 Random initialization of a SOM neural network in the same input space

Stepping through the algorithm

For every observation, the algorithm performs the following steps:

1. The winning neuron, most closely resembling the current observation, is identified. The most common used measure of likeliness is the Euclidean distance.

2. We now say that the current observation is projected on the winning neuron and this winning neuron is adjusted to more closely resemble the current observation. The neurons are interconnected, so some of the neighbours (in the grid) of the winning neuron will also be adjusted. The farther away from the winning neuron in the output grid, the less adjustment is made to the neuron.

To stabilize the algorithm a function is included that reduces the adjustments to the neurons over the performed iterations. Appendix II contains a complete example showing the adjustments in each iteration .

The final fit, after several iterations of the algorithm, is shown in figure 3-10. The observations have been projected on the neurons and the neurons have been adjusted to form best representations of these observations. In less populated areas the distances between the neurons is larger than in more populated areas.

The output map is identical to the shown grid; the observations are projected on the same neurons (gridpoints). The interconnections in the input space also hold true in the map: Neighbouring units in the input space are neighbouring units in the map. The whole available range of visualizations will be treated in paragraph 3.6.
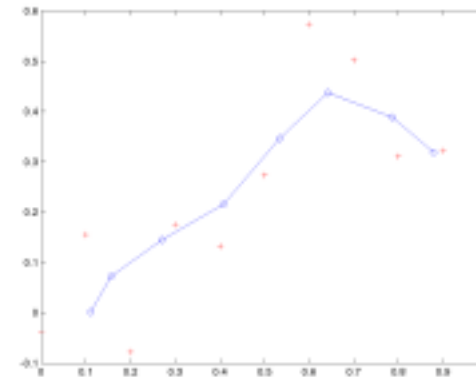
Figure 3-10 Fit of the SOM neural network after self-organization

3.5.3 Mathematical description

The self-organization process can be described in mathematical form. The input consists of a sample of n-dimensional observations

$$x(t) = [x_1(t), x_2(t), ..., x_n(t)] ,$$

where t is regarded as the index of the observations in the sample ($t = 1, 2, ..., T$).

The goal of the algorithm is to determine the values for a set of n-dimensional neurons,

$$m_i(T) = [m_{i1}(T), m_{i2}(T), ..., m_{in}(T)] ,$$

where the *i* denotes the index of the current neuron in the output map ($i = 1, 2, ..., I$). The neurons are first initialized to arbitrary values. The placement of the neurons in the output map is fixed, so the index *i* does not change.

For every *t*, the algorithm performs the following steps:

1. The winning neuron $m_c(t)$ most closely resembling the current observation *x(t)* is selected (c denotes the winning and i denotes the current neuron):

$$\|x(t) - m_c(t)\| = \min_i \{\|x(t) - m_i(t)\|\} .$$

2. The m_i are updated:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t) [x(t) - m_i(t)] .$$

The adjustment is monotonically decreasing as the number of iterations increases. This is controlled by the learning rate factor $\alpha(t)$ ($0 < \alpha(t) < 1$), which is usually defined as a linearly decreasing function over the iterations. The neighbours of the winning neuron are also adjusted, but the adjustment is decreasing as the distance from the winning neuron in the output grid increases. This adjustment is determined by the neighbourhood function $h_{ci}(t)$. Different specific forms of the neighbourhood function can be found in paragraph 3.6.5.

Multiple stages

Instead of performing the self-organization process only once, the map is trained in multiple stages (also called epochs) in which the algorithm reiterates over all the observations in the train set. This way the map converges to a more stable situation while improving statistical accuracy. Any differences in initialization or ordering of the observations are also cancelled out.

The Viscovery implementation (see paragraph 3.6) uses a specific method called 'batch training' to accelerate the train process while keeping the same results. For more information on the batch train process please refer to Deboeck[22].

3.5.4 A three dimensional example

An example using a three dimensional input space is more representative of a real world application of the SOM: A high-dimensional input space mapped to a two dimensional output grid. In Figure 3-11 the neurons are placed in a three dimensional input space with three groups of data. Please note that the network is not random but linearly initialized according to the first two principal components of the dataset.
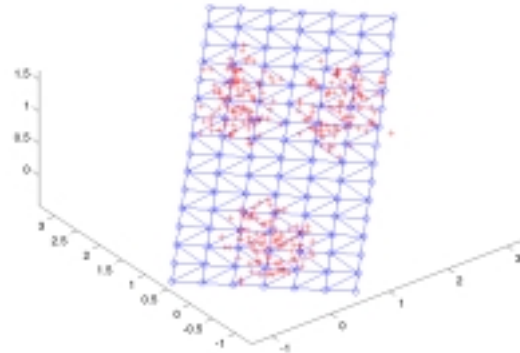
Figure 3-11 Linearly initialized network in a 3D input space
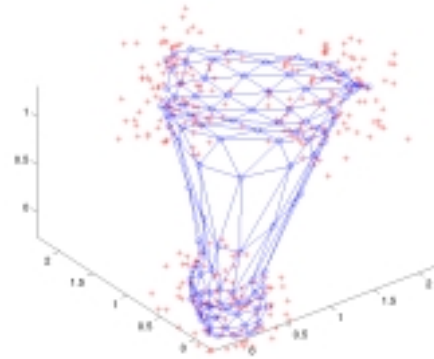
Figure 3-12 Distribution of the neurons after self-organization

The distribution of the neurons after the self-organization process is shown in Figure 3-12. The network, still a 2 dimensional lattice, has curved and stretched to form an as good as possible fit to the original data. The neurons are concentrated in those areas of the input space containing the most observations. The largest separation occurs between the cluster of observations in the bottom half of the cube and the two clusters of observations in the upper half of the cube.

---

[22] Deboeck, G., 1998, page 167.

3.6 SOM visualization and clustering

The previous treatment of the inner workings of SOM are generic for most implementations, but the available visualizations of the final map vary for each software package. We have made use of the Viscovery SOMine 3.0 Enterprise edition program, generously supplied to us by Eudaptics in Austria[23]. Some of the shown visualization and cluster capabilities can not be found in other programs[24].

3.6.1 Maps

The visible output of the algorithm consists of the map, which is an unstretched, flattened representation of the grid in the input space. Observations mapped to a specific neuron in the input space appear on the same specific neuron (grid point) in the map. Neighbouring observations in the input space are neighbouring observations on the map.

The map has several manifestations:

- Clusters: to view the clustering of neurons[25].

- U-matrix: to view relative distances between neurons (in the input space).

- Component planes: to view distributions of separate variables over the map.

It is important to remember that for each map manifestation the distribution of observations over the map does not change. We are looking at the same map, but each time different information is shown.

Unified distance matrix

The Unified distance matrix (U-matrix) can be used to assess relative distances between neurons in the input space. When translating the grid in the input space to the output map, distance information is lost (the grid is returned to an unstretched, flattened state). This information is re-introduced by colour coding the map. Greater
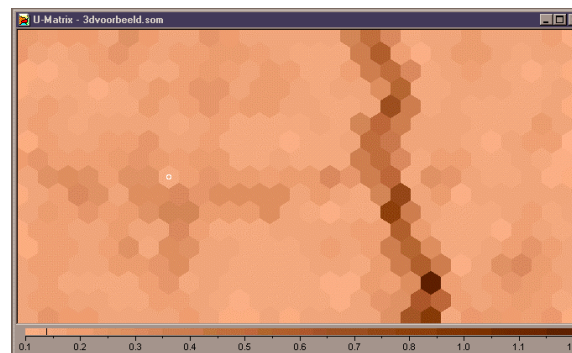
Figure 3-13 U-matrix

---

[23] Eudaptics, 1999.
[24] In addition to this, the intuitive interface and the ability to work with Excel files make it an attractive package.
[25] The clusters and specific clustering algorithms will be treated in paragraph 3.6.3.

differences between the neurons in the input space translate to darker colours in the map.

The U-matrix for the earlier used three dimensional example is shown in figure 3-13. The implicit clustering is visible as groups of neurons having almost equal colour separated by nodes with distinctly different colours. In this U-matrix one very clear cluster at the right of the map can be found. The two clusters at the left half, separated in the middle, are less clear. This agrees with the placement of the three clusters of observations, as can be checked in figure 3-12.

Component planes

A component plane is a manifestation of the map whereby the values for only one of the variables (a component) are shown. In this way the distribution of this separate variable over the map can easily be inspected. When comparing two different component planes of the same map highly correlated variables would stand out because of the likeliness of their component planes. Components not contributing much to the distribution of the observations show a more random pattern in their component planes, they are only contributing noise to the clustering.

Often a display of the U-matrix surrounded by the component planes of all the variables is created. Figure 3-14 shows such a display for our three-dimensional example. The three component planes represent the X, Y and Z variables.
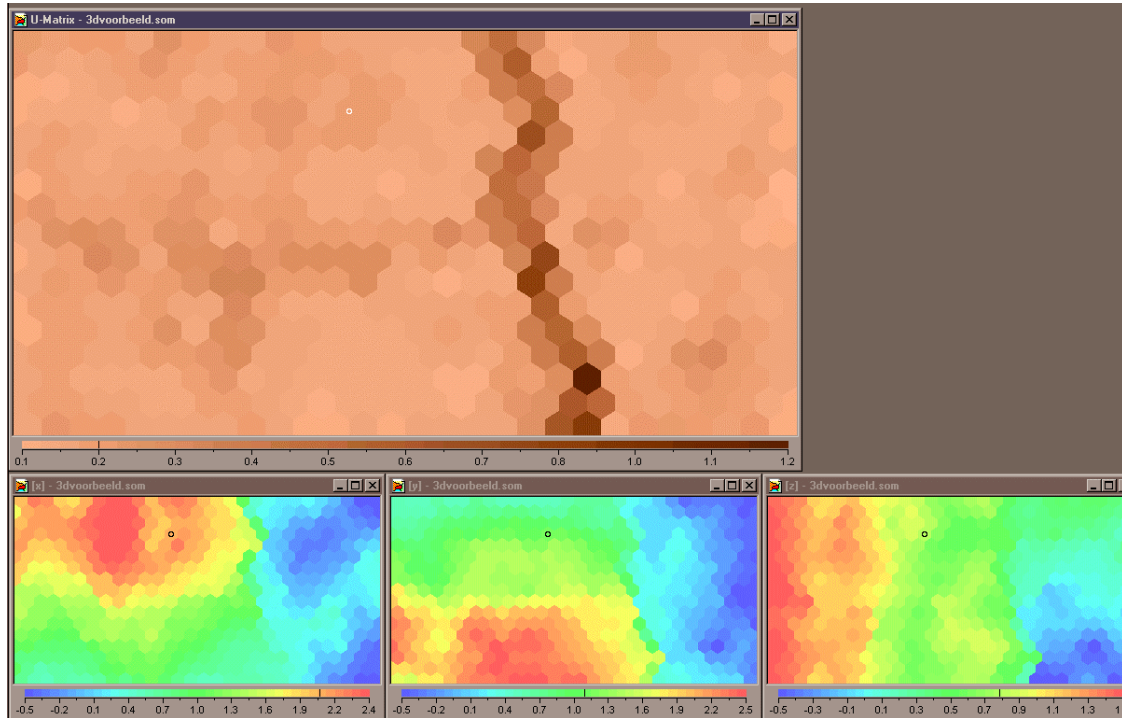
Figure 3-14 U-matrix and component planes for all three variables

The display shows that no two variables are highly correlated. The right cluster is characterized by small values for all variables. The top-left cluster is characterized by high values for X and Z, the bottom-left cluster displays high values for Y and Z. This also agrees with the placement of the clusters of observations in Figure X.

3.6.2 Map quality

We can discern two types of map quality:

- The data representation accuracy.

- The data set topology representation accuracy.

Both make use of the 'Best Matching Unit' concept.
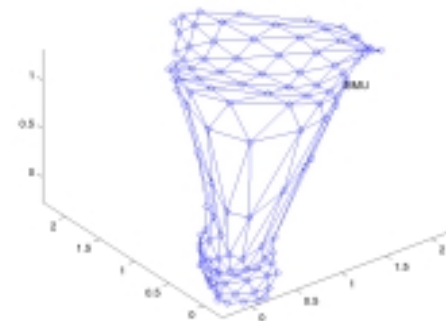
Figure 3-15 Best matching unit for vector [2, 0, 1]

Best Matching Unit

In the input space for each input vector a so-called Best Matching Unit can be found. This is the neuron that most closely resembles the input vector in the input space, the input vector is 'matched' to this neuron. In Figure 3-15 the BMU for vector [2, 0, 1] is shown.

Data representation

The data representation accuracy is most often measured using the average quantization error

$$\frac{\sum_x d(x, m_c)}{T} \text{ where } d(x, m_c) = \min_i \{d(x, m_i)\},$$

Figure 3-16 Quantization error map

c denotes the index of the best matching unit, x is the current input vector (observation) and m_i is the current neuron. The distance measure used is once again the Euclidean distance. The average quantization error decreases as the number of neurons increases; every sample is more likely to be projected on a separate neuron. The quantization error increases as the width of the neighbourhood function increases because then more neighbouring neurons are likely to be adjusted.

A visual representation of the quantization error of the map is provided as the quantization error map. This is a manifestation of the map displaying the quantization error per neuron. Darker neurons indicate larger quantization errors. A good map shows low and equally distributed quantization errors (Figure 3-16).

Data set topology representation

The data set topology representation accuracy can be measured in several ways. One error function often used is the topographic error measure: The percentage of first and second best matching units of a sample vector that are not adjacent to each other. This also measures the smoothness of the mapping.

A more visual tool for evaluating the data set topology representation accuracy is the frequency map. This manifestation of the map displays the number of matched observations per neuron (a darker colour means more matched neurons). A good map should show equally distributed frequencies on the frequency map (Figure 3-17).
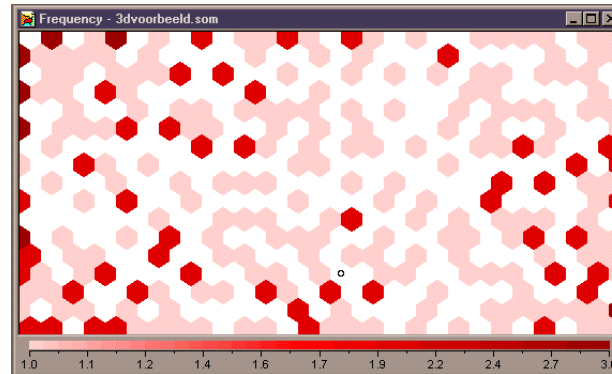
Figure 3-17 Frequency map

3.6.3 Clusters

It is left to the user to find any clustering of observations based on the U-matrix and the component planes. This so-called implicit clustering can be complemented with other clustering techniques to find an explicit clustering. Most software implementations of the Self-Organizing Map do not incorporate any explicit clustering algorithms. The Viscovery SOMine package includes up to three different clustering methods.

The clustering algorithm frees the user from the difficult task of identifying clusters in the U-matrix. However, by altering parameters of the clustering algorithm the number of shown clusters may vary. The user still has to select the most adequate clustering based on all available information.

The three clustering methods implemented in Viscovery SOMine are Ward's clustering, SOM single linkage and a combination of these two, called SOM-Ward. Instead of directly clustering the original observations these algorithms perform a clustering on the neurons (grid points) in the map, on which the observations are projected. As these neurons form 'best representations' for the observations in the input space there is no qualitative difference. The clustering of the observations can be found by retrieving the projected observations for each neuron in each cluster.

Distance measure

Two of the implemented clustering algorithms make use of a specific distance measure, called the Ward distance. It is defined as:

$$\text{Ward distance } d_{x,y} = \frac{n_x \cdot n_y}{n_x + n_y} \cdot \left\| mean_x - mean_y \right\|^2$$

Table 3-1 Ward distances for different cluster sizes

where x and y are clusters, n_x is the number of neurons in cluster x and $mean_x$ is the vector with averages over all components of the neurons in cluster x, also known as the cluster centroid. Distances between clusters with an evenly distributed number of neurons are enlarged in comparison with distances between clusters with an uneven distribution of the numbers of neurons (see table 3-1). This accelerates the merging of stray small clusters.

n_x	n_y	$\dfrac{n_x \cdot n_y}{n_x + n_y}$
1	10	0.91
2	9	1.64
3	8	2.18
4	7	2.55
5	6	2.73

Ward's clustering

This is one of the classic bottom-up methods. It starts with all the neurons in a separate cluster, in each step merging the clusters having the least Ward distance. This distance is calculated without taking the ordering of the map into account, only distances between neurons in the input space are used. When the found clustering is shown on the map, the clusters may appear disconnected: In the input space the neurons are close-by warranting the inclusion in one cluster, but the grid may be bended through the input space in such a way that the neurons are far apart on the map.

SOM single linkage

This clustering method concentrates on the ordering of the neurons on the map. For each neuron the distance with it's neighbour is calculated, when this distance exceeds a certain threshold a separator is set between the neurons in the grid. If the separators form a closed loop the neurons within the loop are marked as a cluster. Because the forming of the clusters only depends on the smallest possible distances between clusters this clustering method is known as a single linkage method.

SOM-Ward

This clustering method is essentially the same as Ward's clustering, but this time the ordering of the neurons on the map is taken into account. Only clusters that are direct neighbours in the map can be merged together to form a larger cluster. The SOM-Ward clustering technique is primarily used in our research. An example of SOM-Ward clustering (using the same 3 dimensional data set) is shown in figure 3-18.

Shown clusters

The number of shown clusters may vary according to user specified settings. For the Ward and the SOM-Ward clustering this implies fixing the formed clusters at a specific step of the clustering algorithm. For the SOM single linkage clustering this means setting the threshold, generating more or less separators and clusters.
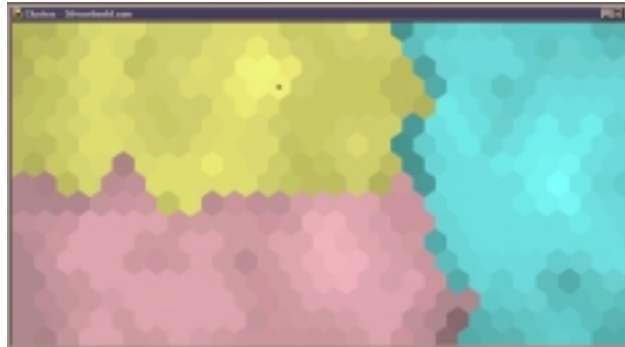
Figure 3-18 Clusters according to SOM-Ward clustering

3.6.4 Cluster quality

For each specific clustering algorithm a quantitative quality measure for the current clustering is calculated. For SOM-Ward clusters this cluster indicator subtracts the observed distance levels at all steps in the cluster algorithm from an exponential increasing 'standard' distance level. If this deviation at the next clustering step (from c to c-1 clusters) is more positive than the deviation at the current clustering step (from c+1 to c clusters) then the current cluster configuration is better.

$$\text{Cluster Indicator } I(c) = \left(\frac{\mu(c)}{\mu(c+1)} - 1 \right) \cdot 100 \text{ if } I(c) > 0 \text{, else } I(c) = 0$$

where $\mu(c) = d(c) \cdot c^{-\beta}$, $d(c)$ is the SOM-Ward distance for the step in the algorithm from c to c-1 clusters, $3 \leq c \leq$ number of neurons.

The β is the coefficient found by linear regression through the [(ln(c), ln(d(c)))] data points, where $2 \leq c \leq$ number of neurons. This exponential curve conforms to the observed 'standard' exponential increase of distance levels as the number of clusters decreases.

I(c) = 0 if c = 1, c = 2 or if d(c+1) > d(c); a smaller than normal distance level at the next clustering step (from c to c-1 clusters) than the distance level at the current clustering step (from c+1 to c clusters) means a worse clustering with c clusters.

3.6.5 Map settings

The results of the map can be influenced by a number of settings, adjusted before the training process is started. The more important ones will be summarized here.

Number of neurons

One of the main settings to choose when training a map is the number of output neurons.

A small number of neurons (smaller than the total number of observations in the train set) means a more general fit is made. The map is better at generalizing and is less sensitive to noise in the data. Figure 3-19 shows the underlying function (y = sin(x)), the train data with some uniform distributed random noise added, and a 1-dimensional 5 neuron grid.

A large number of neurons (larger than the total number of observations in the train set) means a more precise fit is made, but the map is more sensitive to noise in the data. The neurons do not precisely match the original observations, but almost all observations are mapped to separate neurons. Figure 3-20 shows the same data, now with a 20-neuron grid.

Clearly, the fit of the network to the original data is better in this second case, but the error in respect to the underlying function is also greater. Notice that the network is not completely 'attracted' to outliers, due to the learning rate factor and the neighbourhood function. Although the network has more neurons it still is a fairly good generalizer for the underlying function. Compare this to polynomial fitting; higher order polynomials often lead to large errors!

Figure 3-19 Fitting a 5 neuron network to datapoints with underlying function y = sin(x)
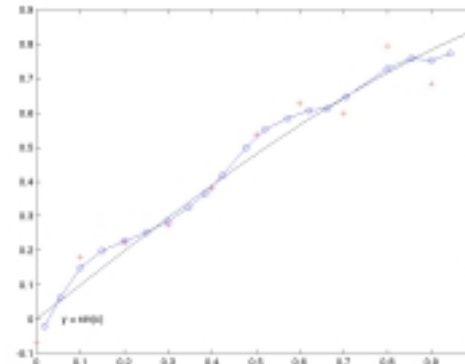
Figure 3-20 Fitting a 20 neuron network to datapoints with underlying function y = sin(x)

The number of neurons should be chosen in proportion to the trust one places in his or her data: If a lot of noise is to be expected, then a relatively small number of neurons should be chosen. If the distribution of the sample data very closely resembles the underlying distribution of the population, then a relatively large number of neurons can be initialized. The extra neurons then warrant a more refined representation of the data by the network.

Initialization

Instead of random initialization one often uses linear initialization. Both can be used, but linear initialization provides a better starting point for the organization of the map. The map is often linear initialized along the axes provided by the first two principal components of the data set.

Choice of learning rate factor and neighbourhood function

The learning rate factor $\alpha(t)$ is normally a linearly decreasing function over the iterations, but can also be specified as an inverse-time function:

$$\alpha(t) = \frac{A}{(B+t)} \, ,$$

where A and B are constants. Earlier and later samples will now be taken into account with approximately similar average weights[26].

The neighbourhood function often has the Gaussian form

$$h_{ij}(t) = \exp\left[-\frac{\|r_i - r_j\|^2}{2\sigma(t)^2} \right] \, ,$$

where r_i denotes the place of this neuron in the map and $\sigma(t)$ is some monotonically decreasing function over the iterations. Sometimes a simpler form of the neighbourhood function is used, e.g. the bubble function which just denotes a fixed set of neurons around the winning neuron (in the map). The Gaussian form ensures a global best ordering of the map (the quantization error arrives at a global minimum instead of a local minimum)[27].

---

[26] Kohonen, T., 1997, page 117.
[27] Kohonen, T., 1997, page 118.

3.7 SOM interpretation and evaluation

In the knowledge discovery process the SOM maps are mainly used for two reasons: describing the data set and predicting values for certain aspects of the data. Each of these applications demands a specific way of evaluating and interpreting the map.

3.7.1 Description

When a map has been created the user has to evaluate the map, determine a good clustering and possibly improve on the clustering so that a clear understanding of the underlying data set emerges.

Determining a good clustering is a non-trivial task. Of course the variables used for map creation have to be suitable for the research setting. Then each specific setting for the used clustering algorithm renders a different number of clusters visible. The map quality measures and the quantitative cluster quality measure form a starting point for determining a good clustering. It is up to the expert user to choose a clustering suitable for the task at hand, specifically by taking any domain knowledge into account.

Improving the clustering

Often one tries to improve on the results (clustering or readability of the display) by reducing the number of variables used in the creation of the map. Removing a variable is warranted only under certain conditions, if these conditions hold then the variable does not contribute much to the generated map and can safely be removed:

- With or without the variable the distribution of the companies over the map remains equal.

- With or without the variable the clustering remains the same (same size and same characteristics in terms of individual variables).

Two strong visual clues lead us to these kinds of variables:

- The component plane of the variable shows a random distribution (Figure 3-21). The component only adds noise to the formation of the map, it does not contribute to the distribution of companies over the map. For instance, this could happen when the variance of the normalized variable is significantly lower than the variance of the other normalized variables.

- The component plane of the variable bears a close resemblance with the component plane of another variable (Figure 3-21). The variables are then highly correlated (not necessarily in a linear fashion). The

dependent variable does not contribute to the distribution of companies over the map, because the same information is already contained in the other variable.
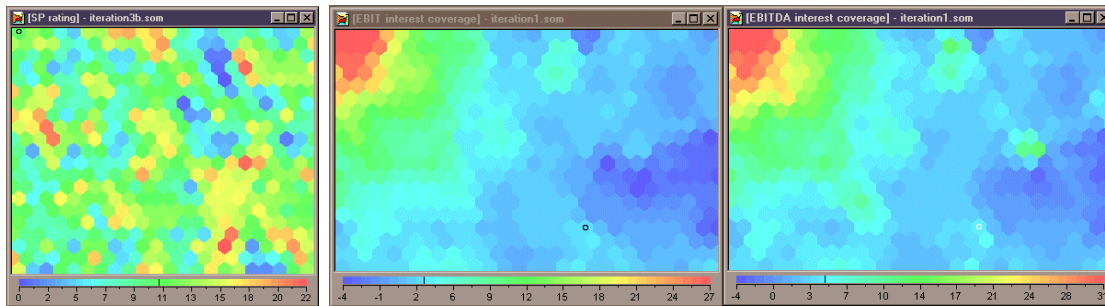
Figure 3-21 Random and highly correlated component planes

A less strong visual clue also leads us to spurious variables:

- The distribution of the high and low values of the component plane does not coincide with one or more specific clusters (Figure 3-22). A strong characterization of the clusters (regarding this variable) can not be given. It is most likely that the variable does not contribute to the clustering, so we choose to remove the variable.
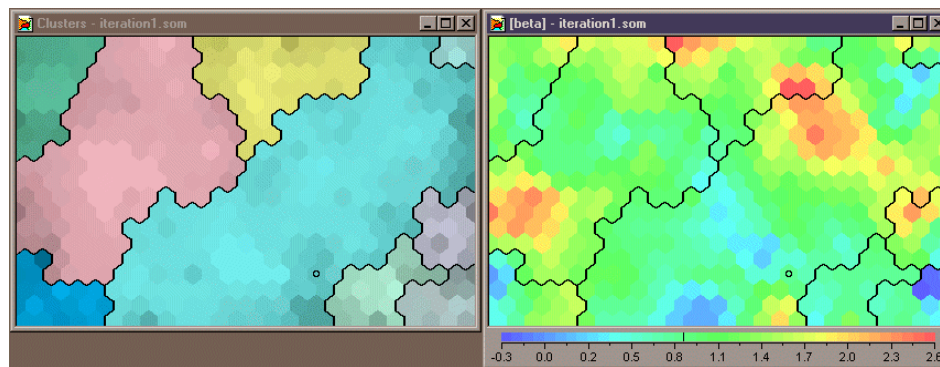
Figure 3-22 Distribution of variable does not coincide with clustering

Examples

In appendix III and IV two examples of descriptive SOM use can be found, one on a medical domain and the second on a data based marketing domain. Chapter 4 also uses SOM in a descriptive way to evaluate the link between credit ratings and financial ratios.

3.7.2 Prediction

The SOM can be used to predict values for any of the variables of new observations. We are then not so much interested in the found clustering as we are in the form of the neural network in the input space. The final form of the neural network is found using the self-organization process (this is a form of semi-parametric regression), and remains fixed.

The network can now be used to predict values of one or more variables for previously unknown observations, just as we would use the regression line to predict values for new observations in a standard linear regression model. It is also possible to do this for observations used to create the map, but this would of course lead to an artificially good prediction.

The values for specific variables are predicted as follows:

1. A neighbourhood of K neurons of the current new observation is determined. The user can set K.

2. A weighted average of the variable is taken over this neighbourhood, where close-by neurons are weighted more strongly.

When K is set to 1 the prediction is based on only one neuron, also called the 'best matching unit'.

Model set-up

To correctly assess the prediction power of the model we have to make a distinction between an in-sample train set and an out-of-sample test set. The test set is reserved for a final assessment of prediction capabilities after the model has been constructed. When creating and iteratively improving the model we want to test the prediction power of the created map without using the test set. We also do not want to use the same observations as used for training the map, so we have to make an extra distinction in the in-sample data set: a train set to train the map, and a validation set to tune its parameters.

Improving the prediction

Improvements of the prediction can be found in the removal or addition of variables and in the size of the neighbourhood used for prediction. We can safely delete variables if the prediction results do not worsen after removing the variable from the model.

Often the variables resulting from a descriptive analysis are a good starting point for the prediction, but other combinations should be tried too. When adding variables and each time re-evaluating the validation test results there is a possibility that some non-linear relationships with other, not yet added variables would be overlooked. On the other hand, because the starting point is a set of variables that has already proven to contain most information (in the descriptive analysis) we can expect to arrive at a reasonably good prediction in

a relatively short time span. When using all the variables for map creation, and then subsequently removing variables not contributing much to the prediction power, we can be certain that all contributing combinations are found. Unfortunately this strategy is more time consuming.

Using target variable as a train variable

For classification purposes most often multi-layered backpropagation networks are used. For these networks it is possible to train the network based on the train variables and the target variable. For each observation the state of the train variables is shown to the network. The network gives a prediction for class membership, and this prediction is compared with the real class membership (the target variable), leading to adjustments in the network to account for any deviations (the backpropagation step). This is also known as supervised training, the network adapts to better distinguish the differences between the classes the observations can belong to.

For the SOM as a feed forward network, it is not possible to directly match the real value of the target variable with the predicted value of the target value. But we can simulate it by using the target variable as a train variable during map creation, this is known as semi-supervised training[28]. How this can be beneficial to a distinction between observations in different clusters is illustrated in the following figures. Without using the target variable as a train variable, the map in figure 3-23 (consisting of just two neurons) is created using only 1 variable or 1 dimension. A distinction between the observations is difficult to make, it is hard to see to which
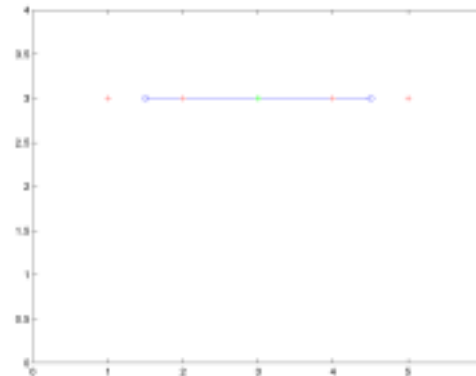
Figure 3-23 SOM network when only 1-dimensional (x-axis) information about the datapoints (red plusses) is available. The best matching neuron for the new observation (green star) is difficult to measure.

neuron the new observation (green) is matched in the one-dimensional final map (the distance to either neuron is equal).

When using the target variable as a train variable, the map is created using two dimensions (figure 3-24). The placement of the neurons shifts, it is much clearer that the new observation matches the rightmost neuron. Remember that we do not have the value of the target variable for the new observation, so we can still only use the x-dimension to determine the best matching unit for this new observation.
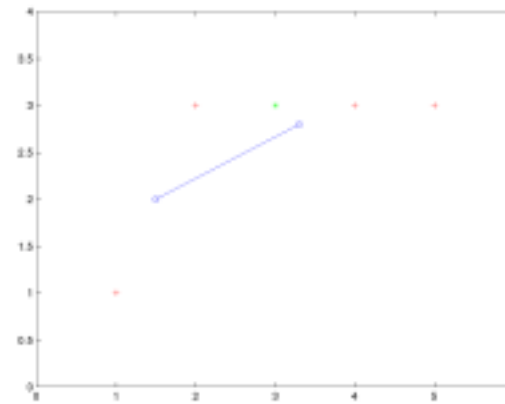
Figure 3-24 SOM network when target variable (y-axis) is also used when training the network. The best matching neuron for the new observation (green star) is easy to measure.

Of course this particular example only illustrates one possible outcome of using the target variable as a train variable. A deeper investigation into the effects of this technique lies outside the scope of this thesis.

Examples

An example of the use of SOM as a prediction model can be found in chapter 5: Financial ratios are used to classify companies according to creditworthiness.

---

[28] Kohonen, T., 1997.

3.8 SOM questions and answers

Q: Is it a neural network?

A: Yes, but a very special one; a feed forward neural network with no hidden layers. The inner workings of the SOM are relatively simple (see paragraph 3.5) and therefore much clearer than for networks using multiple layers and backpropagation.

Q: Is it a blackbox?

A: No, the SOM is nothing more than the projection on a non-linear plane drawn through the observations. The form of the plane is set using a very strict and clear algorithm, and the form of the plane is fixed after the algorithm has completed. The component planes give us insight into the contribution of individual variables to the clustering. Other neural nets use multiple layers and backpropagation, making the inner workings of the network more difficult to comprehend.

Q: How can the neural network be flattened and unstretched for output viewing (the map) but still keep the fixed form in the input space (fixed after completing the algorithm)?

A: It is not really the grid in the input space that is flattened and unstretched, rather a direct representation of this grid in 2 dimensions. Each neuron in the input space directly corresponds with a grid point in the 2 dimensional map.

Q: Is there a chance of overfitting the neural network when using a large number of neurons (larger than the number of observations)?

A: This depends on your definition of overfitting. The SOM algorithm includes automatic 'dampening' functions in the form of the learning rate factor and the neighbourhood function. When using a large number of neurons the network more precisely represents the underlying dataset, some would consider this overfitting. However, thanks to the dampening functions the neurons are not completely attracted by the specific observations.

Q: Does the order in which the observations are being processed by the self-organization process make any difference for the final results?

A: No, because instead of processing the observations just once, often multiple iterations are used. Together with the used dampening functions the map converges to a stable form.

Q: What is the statistical significance of results found with SOM?

A: The SOM can be used in two ways, (1) to give an accurate description of the data set, and (2) to predict values for one or more variables. For descriptive use several SOM and cluster quality measures exist (see paragraph 3.6), but (like other visualization techniques) no general statistical 'goodness' indicator exists.

For predictive use we should see the SOM as a form of non-linear regression, without a presupposed form of the fitted function. Because of the non-linearity of the model the direct contributions of the individual variables are difficult to assess. The total performance of the model can be measured and validated using common statistical techniques.

3.9 Summary

Chapter 3 covered the theoretical foundations of SOM. We viewed the place of Self-Organizing Maps in the knowledge discovery process, and we described some projection, clustering and classification techniques related to SOM. The SOM is a combination of non-linear projection and hierarchical clustering, driven by a simple feed forward neural network. The observations are projected on a flexible grid of neurons that stretches and bends to accommodate to the distribution of the data in the input space. After the network has found its final form, it is displayed in a flattened state as a map. The observations projected on this map are then clustered, according to similarity of the used variables.

A Self-Organizing Map can be used in two ways: As a descriptive analysis tool, and as a prediction model. For use in a descriptive setting the map display and the clustering is most important. Visually comparing the clusters and other parts of the SOM display provides a good and insightful overview of the underlying data set. When deploying the SOM as a prediction model, we are more interested in the distribution of the companies over the map (or equivalently, the form of the map) than the clustering. The SOM then functions as a semi-parametric (possibly non-linear) regression model.

4 descriptive analysis

The paragraphs in chapter 4 form an account of our descriptive analysis, using the SOM as a visual exploration tool. We answer question 3 and 4 from the introduction:

3. Is it possible to find a logical clustering of the companies, based on the financial statements of these companies?

4. If such a clustering is found, does this clustering coincide with levels of creditworthiness of the companies in a cluster?

Paragraph 1 covers the basic data analysis. Paragraph 2 explores the possibility of clustering companies based on financial data. In paragraph 3 we then compare the found clustering with the credit ratings of the clustered companies. Paragraph 4 reviews the performed sensitivity analysis and in paragraph 5 we benchmark the SOM results to a principal components analysis.

4.1 Basic data analysis

Our basic data analysis comprises the first three steps of the knowledge discovery process, namely data selection, data pre-processing and data transformation.

4.1.1 Data selection

The data selection step involves the selection of interesting financial ratios, the selection of evaluation period and history length, the selection of the company universe and the selection of the data provider.

Financial ratios

Our preliminary selection in chapter 2 yielded several types of financial ratios for industrial companies. We can distinguish the following kinds of ratios or variables:

- Interest coverage ratios: these measure the extent to which the earnings of a company cover debt or interest.

- Leverage ratios: these measure the financial leverage created when firms borrow money.

- Profitability ratios: profitability ratios measure the profits of a company in proportion to its assets.

- Size variables: these measure the size of a company.

- Stability variables: stability variables measure the stability of the company over time in terms of size and income.

- Market variables: market variables are used to assess the value investors assign to a company.

An overview of the initial selection is displayed in Table4-1.

Table 4-1 Overview of selected financial ratios

Type	Name	Description
Interest coverage	EBIT interest coverage	(earnings before interest and taxes) / (interest expenses)
	EBITDA interest coverage	(earnings before interest, taxes, depreciation and amortization) / (interest expenses)
	EBIT / total debt	(earnings before interest and taxes) / (total debt)
Leverage	Debt ratio	(long term debt) / (long term debt + equity + minority interest)
	Debt-equity 1	(long term debt) / (equity)
	Debt-equity 2	(long term debt) / (total capital)
	Net gearing	(total liabilities – cash) / (equity)
Profitability	Return on equity	(net income) / (average equity)
	Return on total assets	(earnings before interest and taxes) / (total assets)
	Operating income / sales	(operating income before depreciation) / (sales)
	Net profit margin	(net income) / (total sales)
Size	Total assets	The total assets of the company
	Market value	Price per share * number of shares outstanding
Stability	Coefficient of variation of net income	(standard deviation of net income over 5 years) / (mean of net income over 5 years)
	Coefficient of variation of total assets	(standard deviation of total assets over 5 years) / (mean of total assets over 5 years)
Market	Coefficient of variation of earnings forecasts FY1	(standard deviation of forecasts fiscal year 1 over analysts) / (mean of forecasts fiscal year 1 over analysts)
	Market beta relative to NYSE	snapshot taken on the last trading day of the quarter
	Earnings per share	(earnings applicable to common stock) / (total number of shares)

Rating classification

The ratings are classified according to the S&P rating classification scale. These letter ratings have been transformed to numerical rating codes; the transformation is shown in table 4-2. The rating code represents an ordering between the different rating classes. A higher rating corresponds to a lower default risk, a lower rating corresponds to a higher default risk. Note that this numerical scale also seems to imply rating classes of equal width, which not necessarily has to be true. At this point we choose to use this particular transformation because we do not know the exact width of the rating classes. Furthermore, because the SOM can also model non-linear relationships this is less of a disadvantage than it first may seem.

Table 4-2 S&P Rating classification scale

S & P Rating	Rating Code
AAA	22
AA+	21
AA	20
AA-	19
A+	18
A	17
A-	16
BBB+	15
BBB	14
BBB-	13
BB+	12
BB	11
BB-	10
B+	9
B	8
B-	7
CCC+	6
CCC	5
CCC-	4
CC	3
C	2
D	1

Publication lag

Rating agencies try to react as soon as possible on any news that may affect the rating of a company. A significant part of this information (the financial statement of a company) is only available after a certain time lag. This publication lag is often 3 to 6 months. Thus the rating for the fourth quarter of 1998 is based on financial figures for the second or third quarter of 1998. Taking this into account we downloaded the ratings with a 2-quarter offset; the figures of the fourth quarter of 1998 were matched with the ratings of the second quarter of 1999.

Except for defaults, the ratings do not change twice within two quarters. Actually, they do not change much at all (generally less than once per year). This is consistent with the general policy of rating agencies to keep the ratings as stable as possible.

Evaluation period and history length

We opt to use quarterly data, as this provides a detailed and up-to-date view on the company's current financial position. A 5-year history (20 quarters) is downloaded for each variable, but the initial analysis will be based on the figures of the fourth of quarter 1998. Later on we will determine if and how much more history needs to be taken into account.

Company universe

The constructed universe consists of US companies with at least one Standard & Poor's credit rating in the evaluated time period. This time period starts at the first quarter of 1994 and ends at the fourth quarter of 1998. The total number of companies in our universe is 1677.

Sector classification

Each company is classified according to the S&P sector classification[29]. The distribution of companies over S&P sectors is shown in the following figure:
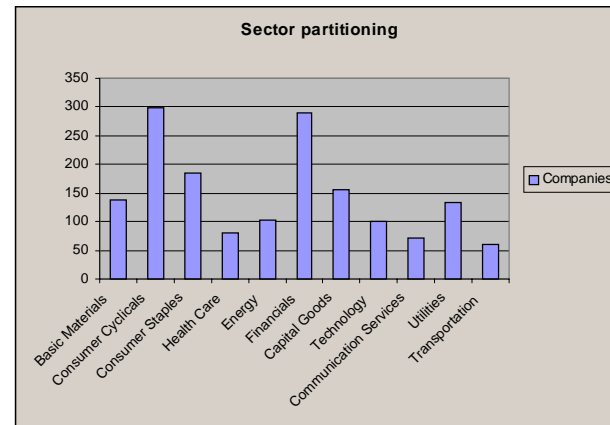
Figure 4-1 S&P sector classification

We have chosen to perform our analysis on a per sector basis, because of the following reasons:

- Companies in different sectors can display very different values for the same variable (the long-term debt of a bank will on average be much larger than the long term debt of a steel factory).

- Often a variable is not applicable to companies in one sector but very applicable to companies in another (a bank does not have a raw materials inventory like a steel factory has).

In this thesis we will analyze sector Consumer Cyclicals (294 companies).

Data provider

Most of the data was downloaded from the Compustat Quarterly historical database, available through the Factset datavendor. The Compustat database is actually owned by Standard & Poor's, this leads us to believe

---

[29] We do not use the new MSCI sector classification because that classification scheme does not yet cover the whole company universe.

that we at least partially use the same data for our model as S&P uses for their rating decisions. Forecasts are available from the IBES database, a well-known data provider for earnings forecasts.

For each variable the underlying components were downloaded (instead of downloading just the ratio when available). This guarantees a ratio calculated according to our specifications and it facilitates checking individual values.

4.1.2 Pre-processing & transformation

Summary statistics

The first pre-processing step involves calculating some summary statistics and viewing scatter plots per variable. The scatter plots, not displayed here, show that most variables contain one or more extreme values. The summary statistics for the fourth quarter of 1998 can be found in table A-1 in appendix VI. The summary statistics include the mean, the standard deviation, the median and the median standard[30] deviation. The median and the median standard deviation are included because of the observed extreme values. Extreme values greatly affect the mean and the standard deviation to the point that no conclusions can be drawn from these figures. The median and the median standard deviation do not have these drawbacks, they are called *robust* estimators.

Other calculated statistics include the minimum, the maximum, the number of not-availables, the percentage of values greater (smaller) than the mean plus (minus) 3 times the standard deviation and the percentage of values greater (smaller) than the median plus (minus) 3 times the median standard deviation. Finally, to conduct some tests on the normality of the variables, we calculate the skewness, the kurtosis and the Jarque-Bera statistic[31]. These tests show that a few extreme values are greatly offsetting the distribution per variable. The mean and the median differ a lot, as do the standard deviation and the median standard deviation.

Stability of variables over time

To test the stability of variables over time we have evaluated the summary statistics per variable for all downloaded quarters[32]. We can only justify the merging of cross-sections (to enlarge the data set leading to higher statistical significance) when the characteristics of the data do not significantly change over time. This holds true for the median and the median standard deviation; differences between periods are small. When evaluating the mean and the standard deviation, great differences between periods can be detected for almost all variables. This once again confirms our observation of extreme values off-setting the distributions.

---

[30] A complete description of the median standard deviation can be found in the appendix V.
[31] A complete description of these tests can be found in appendix V.
[32] These figures are not displayed here but are available upon request.

Missing Values

The data contains a substantial amount of missing values per variable, as is often the case with financial data. For the same company we often see a sequence of missing values. This is due to the fact that some variables consist of one or more of the same components. If one of these components is missing, the derived variables can not be calculated.

Extreme values

The scatter plots show that almost all variables contain one or more extreme values. We checked these extreme values by decomposing the variable and checking the values for the underlying components. This tells us that most extreme values represent accurate values. For example, in the "Total Assets" variable some companies are several times as big as most other companies. Often the same companies arise in different time periods as extreme values, this is also an indication that the values are structurally higher or lower (thus correct) instead of mere data errors. Therefore we should explicitly not call them outliers.

Coping with extreme values

The found extreme values present us with some problems. We do not want to lose this information, but we also want to capture all of the information contained in the 'normal' range of the variable. When not removing the extreme values it might result in a loss of resolution in this range.

To cope with the extreme values we proceed in the following way: For every variable we calculate a cut-off, so that approximately 2,5 percent of the observations is situated above the median plus this cut-off or below the median minus this cut-off. Then the observations having values larger (smaller) than the median plus (minus) the cut-off are replaced by this upper (lower) value. The histograms in figure A-13 in appendix VI clearly exhibit much more evenly distributed variables after the cut-off. For comparison the found model can at a later time be tested using non-edited data.

Normality of the data

The calculated values for the tests on normality show that (except for the beta variable) the data is not normal distributed, before and after the cut-off. Therefore any model we consider should not assume a normal distribution of the variables. The significant improvement of the normality measures after the cut-off clearly reflects the large influence extreme values have on the distribution. Note that the distribution of the variables after the cut-off would never fully return to normality because of the nature of the cut-off. The cut-off values are concentrated at the outer edges of the histogram, as is visible in figure A-13 in appendix VI.

Correlation

To find possible collinearities between variables it is insightful to calculate a correlation matrix; the matrix is displayed in table A-3 in appendix VI. The correlations higher than a best-practice value of 0.70[33] are highlighted in the matrix. For linear models these high correlations can lead to problems involved with multi-collinearity. To circumvent these problems we will test our models with different sets of variables, every time selecting other combinations of non multi-collinear variables.

Transformation

The histograms and the summary statistics show rather skewed size variables. Therefore these are transformed using a logarithmic transformation with a base 10 logarithm.

Other transformations are unnecessary. The Viscovery program automatically re-scales the values per variable so they are comparable when creating the map. This way no specific preference (that could influence the forming of the map) is shown for any of the variables. For convenient output viewing these values are scaled back to their original values.

---

[33] Geluk, I. and van der Hart, J., 1996, page 52.

4.2 Clustering companies

In our case the visualization and evaluation steps of the knowledge discovery process boil down to creating and evaluating maps. This is often an iterative process. After creating a map, the new insights provided by interpreting the map and evaluating the results may force us to create another map using different settings or variables.

4.2.1 Creating suitable maps

After the training process (using the variables resulting from the previous steps in the knowledge discovery process) a map is created and displayed. However, we still have to determine the optimal clustering for our specific problem domain.

Determining a good clustering is a non-trivial component of the visualization step. When a map is created the companies used to create the map are implicitly clustered: the companies are distributed over the map, whereby similar companies can be found near each other on the map.

To make this implicit clustering explicit, Viscovery applies an advanced algorithm to determine the cluster membership for all the companies. It calculates the clustering using a combination of a bottom-up method (Ward) and the traditional SOM single linkage clustering. All possible clusterings are formed and for each possibility a cluster indicator is calculated[34]. The cluster indicator points to possibly good cluster configurations; the differences between companies in separate clusters are larger when the cluster indicator is higher.

The user is presented with a choice between several possible clusterings, for each possibility displaying the number of clusters and the value of the cluster indicator. The user has to combine this quantitative measure with available information like summary statistics per cluster, the distribution of individual variables over the overall clustering (using component planes) and specific domain knowledge. In the end the found clustering should accurately reflect key (and not necessarily linear) relationships between clusters and individual variables.

---

[34] The used clustering algorithms and the cluster indicator are explained in chapter 3

4.2.2 Intermediate results

First iteration

In the first iteration we use all variables to create a 500 neuron self-organizing map, but we explicitly do not use the rating code when creating the map. This first step partly functions as a pre-processing step; by using all variables we can infer inter-variable relations visually. The map is displayed in figure A-14 in appendix VI. We can see that some variables are highly correlated[35], because their component planes (displaying the distribution of the variable over the map) are very much alike. Also some rather odd relationships are visible, as well as some variables that do not seem to have any relationship with the found clustering in the main map.

Inferred relations

The following variables are highly correlated and thus candidate for removal:

- EBITDA interest coverage with EBIT interest coverage and with EBIT / total debt.

- Debt-equity ratio 2 with debt ratio.

- Net gearing with debt-equity ratio 1.

- Log total assets with log market value.

The net gearing variable shows a strange correlation with debt-equity ratio 2. For extreme values they are negatively correlated, while we would expect a positive correlation. The raw data tells us that this only happens when a company has negative equity[36]. Due to the definition of net gearing this variable increases when equity decreases, until equity turns negative. The value for the variable then turns negative which shows up on the map as a relatively low value.

This also holds true for the debt-equity ratio 1. As both variables try to convey the same message as the debt equity ratio 2 and in view of the lesser contribution to the overall clustering we see no need to keep them.

Adjustments

We removed the following variables: EBIT interest coverage, EBIT / total debt, Debt ratio, Debt-equity ratio 1, Net gearing and Log market value.

---

[35] The correlation matrix can function as an extra confirmation, but due to the necessarily linear nature of the correlation coefficient it might not capture all dependencies between the variables.
[36] Negative equity occurs when a company has to compensate for so many losses that the sum of its reserves and stockholders equity turns negative.

Second iteration

In the second iteration a new map was drawn, after performing the adjustments of the first iteration. The map can be found in appendix VI as figure A-15.

Inferred relations

At this point the map is formed using four profitability ratios, whereas we are using just one interest coverage ratio, one leverage ratio, one size variable, two stability variables and two market risk variables. This way we are giving extra weight to the profitability measure of a company. For this analysis we would like to have a representative overview of the financial statement of each company, an extra weight for profitability is undesirable. Therefore we remove two profitability ratios: return on equity and the net profit margin. The return on equity variable approximates the return on assets variable but contributes less to the overall clustering. Likewise, the net profit margin (net income / total sales) mimics the "operating income / sales" variable, but contributes less to the overall clustering.

On this map the EPS variable shows a correlation with the return on assets variable, and is thus candidate for removal. The beta variable does not contribute much to the clustering at all, and therefore we will remove it.

Adjustments

We removed the following variables: Return on equity, Net profit margin, EPS and Beta.

Third iteration

In the third iteration the final map is created based on the following variables:

- EBITDA interest coverage

- Debt-equity ratio 2

- Return on assets

- Operating income / sales

- Log total assets

- Coefficient of variation of net income

- Coefficient of variation of total assets

- Coefficient of variation of forecasts

As in the previous iterations we do not use the S & P senior unsecured debt rating when creating the map. The map is displayed in the appendix VI as figure A-17.

4.2.3 Results

The final clustering is shown in figure 4-2. The inter-cluster distance indicator is highest for the shown eight clusters, so on a quantitative basis this clustering is a good starting point. When evaluating the component planes we notice for each variable a concentration of extreme high values in one or two clusters and a concentration of extreme low values in one or two clusters. An even distribution of the high and low values over the map would mean that this specific variable is only adding noise to the clustering. As this is clearly not the case, we are confident that the found clustering is adequate for this data set.

Figure 4-2 Found clusters

The summary statistics[37] per cluster re-enforce this image. Per cluster the following statistics are calculated:

- Number of matching companies (number and in terms of percentage).

- For each variable the mean, minimum, maximum and standard deviation

The companies are evenly distributed over the clusters, and the statistics for the variables per cluster differ enough to make meaningful characterizations of the clusters.

---

[37] These can be found in table A-4 in appendix VI

When visually inspecting the map and the distribution of individual variables over the map the following characterization of the clusters can be made (in order of descending creditworthiness):

C 2 - Healthy companies with high interest coverage, low leverage, high profitability, very stable companies and low perceived market risk. Remarkable: these are not always the biggest companies.

C 4 - Large stable companies with a high profit margin. Remarkable: not so high interest coverage.

C1, C3 and C8 - Average companies with no real outstanding features.

C 5 - Small companies with low interest coverage and high leverage. Remarkable: a stable coefficient of variation of net income, these companies do not grow much.

C 6 - Underperformers: very low interest coverage, very low or even negative profitability, negative earnings forecasts.

C 7 - Unstable companies: very unstable and a very high perceived market risk.

4.3 Comparing S&P ratings

4.3.1 Associating ratings

By associating the Standard & Poor's ratings with the map we can view the distribution of the ratings over the companies in the map. Figure 4-3 shows a concentration of high ratings in the upper left corner, gradually fading to low ratings in the lower right corner. Two distinct spots of extreme low ratings appear on the map, but nowhere near the high ratings.

From this we infer that the SOM algorithm has found a relation between financial ratios and creditworthiness, *without using the credit ratings as input*. Furthermore, when superimposing the clusters on the ratings component (in figure 4-4) the general relationship between specific financial profiles and creditworthiness is confirmed. Healthy and stable companies receive high ratings, whereas underperformers and unstable companies receive low ratings.

Diversity within clusters

The observed rating diversity within each cluster prohibits us from assigning specific credit rating levels (e.g. AA-) to specific clusters. The summary statistics (table A-4 in appendix VI) show a standard deviation between 2 and 3 for the S&P rating per cluster.

If we assume the model of the data the SOM creates to be correct and if we assume the S&P ratings to be correct then this means that the values for financial ratios of companies having a rating within a range of approximately 4 to 6 notches do not substantially differ (the exact range varies per cluster).
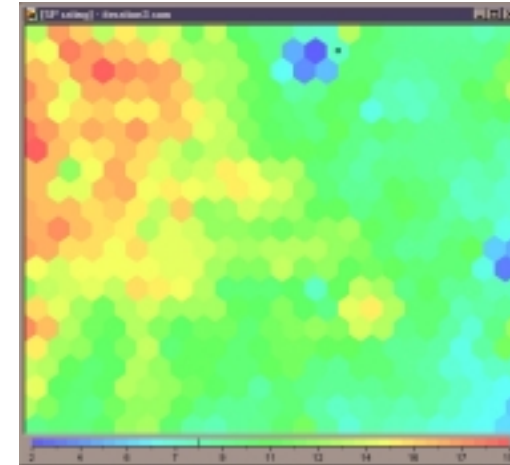
Figure 4-3 Distribution of S&P ratings over the map
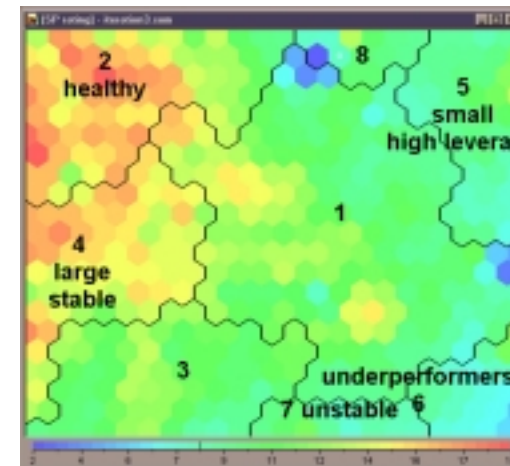
Figure 4-4 Clusters superimposed on ratings distribution

Starting from the same assumptions we can most likely attribute the observed rating diversity within each cluster to external, non-financial factors. According to its financial statement the company belongs in a certain cluster, but an unknown external factor contributed to a higher or lower rating.

4.3.2 Measuring the goodness of fit

Different scenarios

We would like to have a more quantitative measure for the goodness of fit of the ratings mapping. In other words, we are trying to measure the similarity between the found clusters and the ratings in the clusters. To achieve this we first define the scenarios from 'poor fit' to 'perfect fit'.

- A poor fit is a fit where the ratings are randomized over the map. When randomizing the ratings we take the original form of the distribution of the ratings into account. Ratings occurring less frequently on the original map still do not appear very often. We have created this situation by shuffling the existing ratings over the companies. It is portrayed in figure 4-5 a.

- For this specific research a perfect fit would be when the model solely based on financial ratios perfectly describes the dataset. The found clustering makes a perfect distinction between companies with different levels of creditworthiness, and companies with exactly the same level of creditworthiness are perfectly alike. Each cluster only contains companies with an equal rating. This purely hypothetical situation is shown in figure 4-5 c.

- The observed fit (when forcing the map to display 22 clusters) is shown in figure 4-5 b.
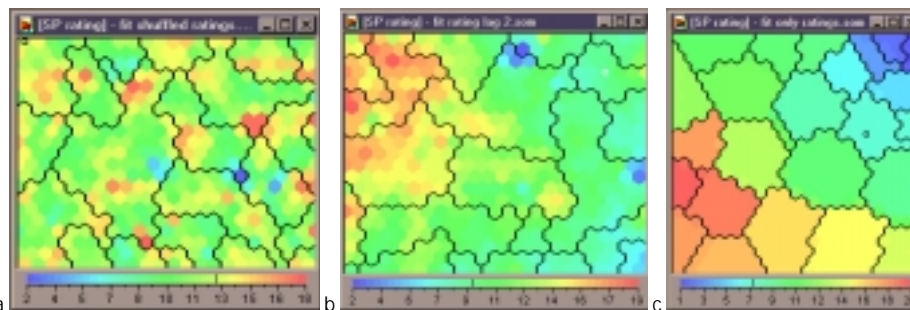
Figure 4-5 Rating mapping from poor to perfect

Visually the three different scenarios are very distinguishable, this can be mathematically verified using the cluster coefficient of determination.

Cluster coefficient of determination

In the context of a standard linear regression model, the well-known coefficient of determination R^2 measures the proportion of the total variance in y that is accounted for by variance in the used variables[38]. In our case we define a special variant called the cluster coefficient of determination $R^2_{cluster}$. This measures the proportion of the total variance in the ratings that is accounted for by variance over the clusters. For a standard multi-linear regression model, the following equation holds:

$$SST = SSR + SSE$$

where SST is the total variance in the ratings, SSR is the variance in the regressors and SSE is the variance in the residual errors. Equivalently we define

$$SST = SSC + SSE$$

where SST is the total variance in the ratings, SSC is the variance of the ratings over the clusters and SSE is the residual variance of the ratings in the clusters. The variance over the clusters is difficult to measure, we can however easily measure the residual variance in the clusters. The cluster coefficient of determination is then mathematically defined as

$$R^2_{cluster} = \frac{SSC}{SST} = 1 - \frac{SSE}{SST}$$

The total variance in the ratings is simply the variance of the original ratings distribution (shown in figure 4-6).

The residual variance in the clusters can be estimated as the rating minus the observed cluster average of the rating, squared (because otherwise these variations sum to zero) for each company.

$$SSE_{clusters} = \frac{\sum_{i=1}^{N}\left(r_i - \bar{r}_{cluster(i)}\right)^2}{N-1}$$

'
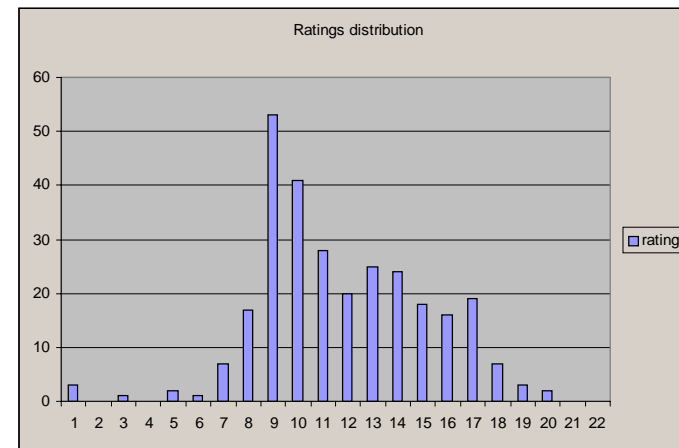
Figure 4-6 Distribution of the ratings in sector Consumer Cyclicals
variance = 11.25, standard deviation = 3.35

---

[38] Greene, W.H. 1997, pages 250-253

where r_i is the rating of company i and $\bar{r}_{cluster(i)}$ is the average rating of the cluster company i belongs to. We are trying to estimate the variance based on a sample of N companies (instead of the whole population), therefore we divide by N - 1.

$R^2_{cluster}$ is a measure for the fit of the ratings mapping to the current clustering, when keeping the number of clusters constant. A small $R^2_{cluster}$ would indicate that the ratings mapping is poor (a high residual variance of the ratings within each cluster), a high $R^2_{cluster}$ would indicate that the ratings mapping is good (a small residual variance of the ratings within each cluster). The number of clusters must be suitably chosen, otherwise an artificially high $R^2_{cluster}$ can easily be obtained by using a lot of clusters.

4.3.3 Results

Fit of observed mapping

We have found the current clustering using the self-organizing map and a fixed set of financial ratios, without using the S&P ratings. So when we assume the following:

- the used financial ratios are a representative financial characterization of the companies in this sector,

- the SOM creates a good model of the underlying data,

- the data does not contain any major errors,

then the $R^2_{cluster}$ represents the variance of the ratings that can be explained by a model based solely on financial ratios. The residual variance can then most likely be attributed to the qualitative factors rating agencies take into account when assigning a rating to a company. Table 3 shows the found values of $R^2_{cluster}$ for a poor fit, an observed fit, a good fit and a perfect fit.

The perfect situation can only occur when ratings are solely influenced by financial ratios. We know that rating agencies also take qualitative factors into account when determining a rating. As of yet we do not know the precise contribution of the qualitative factors, but we will try to simulate them using deviations from the real ratings based on a standard normal distribution (mean 0 and standard deviation 2 notches, or one rating class). This is the good fit also displayed in table 4-3.

The $R^2_{cluster}$ for the observed mapping indicates that approximately 60% of the rating of a company can be explained by its financial statement. Without forgetting the above mentioned assumptions we could attribute the

Table 4-3 Goodness of fit of ratings mapping

Mapping	Poor	Observed	Good	Perfect
$R^2_{cluster}$	0.07	0.61	0.95	1

other 40% to (amongst others) the qualitative analysis performed by the rating agency. The difference between the $R^2_{cluster}$ for the observed mapping and for our 'good' scenario also indicates that the influence of the qualitative analysis reaches further than a simple one or two notches adjustment of the assigned rating.

Fit of different publication lag lengths

Earlier on we argued to use a 2 quarter publication lag for the assigned ratings. Using the $R^2_{cluster}$ we can see the effects of this decision. We calculate the fit for the 2-quarter lag, a 1-quarter lag and a 0-quarter lag. The results are shown in table 4-4.

The fit marginally improves when using a smaller lag period. As a lag period of 2 quarters is better justifiable we will continue to use a 2 quarters lag period.

Table 4-4 Comparison between different rating lags

Mapping	2-quarter lag	1-quarter lag	0-quarter lag	Perfect
$R^2_{cluster}$	0.61	0.64	0.67	1

Drawbacks

Some of the drawbacks of the $R^2_{cluster}$ figure are:

- We have to assume that the data is free of errors and representative for the underlying domain.

- We have to assume that the clustering found using SOM is representative of the underlying dataset.

- The $R^2_{cluster}$ can only be compared for maps with approximately equal clusters. More clusters reduce the variance of the ratings within clusters and thus improve the $R^2_{cluster}$. This is equivalent to using more variables in a standard regression model: The $R^2_{cluster}$ can only improve, more of the variance will be explained by the extra variables.

Alternative use

This last drawback can actually be used to indicate a suitable number of clusters for the current map and the associated ratings. When the $R^2_{cluster}$ does not improve after selecting an extra cluster to view, then the last cluster division did not contribute to a better division of the ratings. Before and after the cluster division the $R^2_{cluster}$ is the same, so the ratings in that specific cluster are even distributed and the first situation is (locally) optimal.

This is not unlike the way we use the eigenvalues in PCA to determine the intrinsic dimensionality of the data. As less as possible intrinsic dimensions are selected that reasonably capture the variance in the dataset. Where

PCA compares the variance of the principal components with the variance of the dataset, our method compares the variance of the clusters with the variance of the variable we want to see explained (in this case the rating).

This differs in that with PCA the dataset is used to calculate the principal components, while we do not use the ratings to determine a suitable clustering. We should thus not try to find a number of clusters that fully explains the variance in the ratings, as it may well be that the variance can not be totally explained. We can however use the rate of improvement of the $R^2_{cluster}$ on a sensible range of number of clusters to determine how good the mapping really is.

If the $R^2_{cluster}$ is plotted against the number of clusters, this shows up as flat spots in the graph. In figure 4-7 we have plotted the $R^2_{cluster}$ for 1 to 24 clusters, this is about the range that is of interest to us (the maximum number of real rating classes is 22). We clearly see a plateau starting at 6-8 clusters and one at about 14 clusters. Using 6 to 8 clusters already explains 50% of the variance in the ratings, and this figure increases to 60% at 14 clusters.
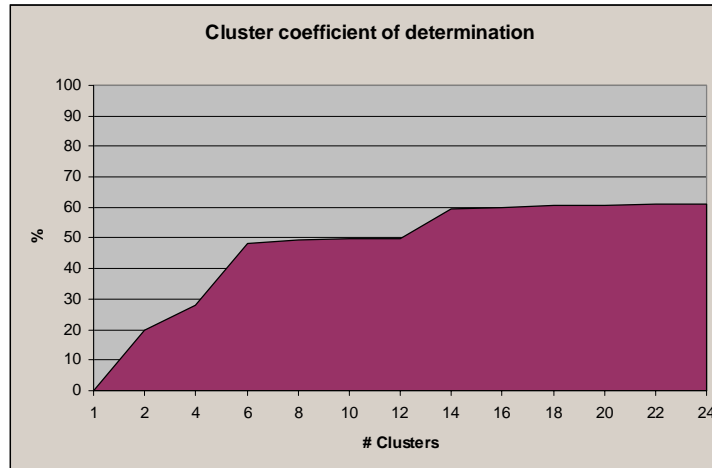
Figure 4-7 Cluster coefficient of determination for 1 to 24 clusters

Our initial choice of 8 clusters seems to be reasonably good; most of the explainable variance in the ratings is captured, no direct improvement can be found when using one extra cluster and it is still possible to easily infer relationships from the maps.

When using 14 clusters almost all possibly explainable variance of the clusters has been captured. This directly corresponds with the distribution of the ratings over the companies. Most companies are concentrated in but 14 of the 22 rating classes.

4.4 Sensitivity analysis

To justify the found results we will perform some sensitivity analysis. Much of our sensitivity analysis involves proving that two maps are equal. When determining the likeliness of two different maps constructed using the SOM algorithm, two criteria are evaluated:

- The results of the qualitative analysis performed on the first map should match results of the qualitative analysis performed on the second map.

- The distribution of the companies over the map should *locally* stay the same.

When the inferred relationships from the two maps are alike then the maps are qualitatively equal. To show that the local distribution of companies over the map does not change we make use of cluster coincidence plots.

4.4.1 Cluster coincidence plots

The distribution of companies over clusters in the first map is plotted against the distribution of companies over clusters in the second map. The area of the bubble shows the number of the companies in the cluster of the second map that coincide with a single cluster of the first map. The cluster coincidence chart of two equal maps should show a few relatively big bubbles as opposed to a lot of small bubbles. The cluster coincidence plot for iteration 3 versus iteration 2 is shown in figure 4-8. Some of the more common situations are presented in the following paragraphs.
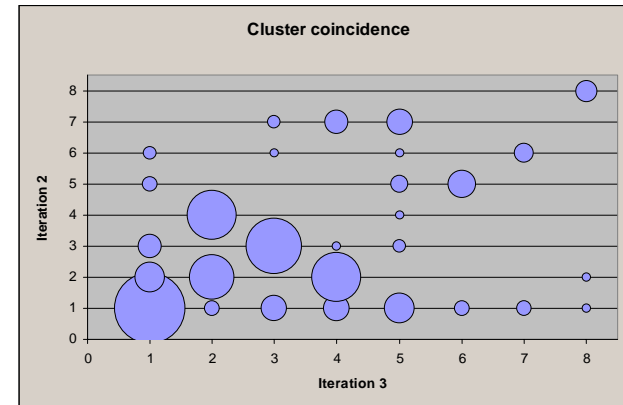
Figure 4-8 Cluster coincidence of iteration 3 vs iteration 2, bigger bubbles means better coincidence

One large and a few small bubbles in a column

In the first column of the plot we see that most companies belonging to cluster 1 in iteration 3 belong to cluster 1 in iteration 2. Some of the companies of cluster 1 in iteration 3 are distributed over other clusters in iteration 2, but these other clusters are all adjacent to cluster 1 in the previous map. A possible explanation would be that these companies all lie on the borders between these clusters and that they tend to 'jump' between clusters. As

long as the main bulk of the cluster is concentrated in one cluster in the previous iteration there is nothing to worry about.

A different cluster number

Sometimes the cluster numbering does not match. The companies of cluster 4 in iteration 3 can mostly be found in cluster 2 in iteration 2. Once again, as long as the main bulk of the cluster is concentrated in one cluster in the previous iteration then the local distribution of the companies over the map does not change.

Separate clusters in the previous iteration

When the companies in one cluster in iteration 3 are divided over 2 clusters in the previous iteration (as is the case for cluster 2) then one should check if the two clusters are adjacent in the previous map. This is often true for relatively similar clusters (unstable and underperformers are relatively similar, unstable and healthy are not). The difference between the two clusters is apparently not that big; companies in separate clusters in the previous iteration belong to the same cluster in the current iteration.

4.4.2 Results

Two kinds of sensitivity analysis can be distinguished. The tests on sensitivity of the algorithm aim to show that the qualitative results stay the same, regardless of the chosen settings for map creation. The tests on data sensitivity try to show that the results remain equal, regardless of any of the specific choices we made in each step of the knowledge discovery process.

Ordering of samples

Using a different ordering of the companies generates exactly the same map. We tried using a random and an inverted order of the companies, both show that the algorithm is stable.

Number of neurons

Maps built using 100, 250 and 1000 neurons are displayed in figures A-19 to A-23 in appendix VI. For each map the results of the qualitative analysis remain unchanged, so we are inclined to say that the maps strongly match each other. The cluster coincidence plots, displayed below the maps, re-enforce this image. As they show a few relatively big bubbles we are confident that the local distribution of companies stays the same.

Eliminating variables

During the course of the analysis we gradually reduce the number of variables from 18 to 8. Although the maps are not exactly the same, we did not lose much important information when discarding the spurious variables. The results of the qualitative analysis stay the same, some relationships are even clearer.

Figures A-14 and A-15 in appendix VI show that cluster coincidence for iteration 1 vs. iteration 2 and for iteration 2 vs. iteration 3 is high. Companies that appear in one cluster in a previous iteration appear in a single or just a few clusters in the current iteration.

Using non-edited data

To see the effect of the used cut-off we create a map using non-edited data, shown in appendix VI in figure A-25. When using non-edited data to create the map the qualitative analysis is much harder to perform. Most relations still hold, but they are often unclear. The extreme values per variable reduce the contained information, leading to an inconsequential variable (a uniform coloured component plane with a small 'hot spot'). Cluster coincidence (shown in figure A-26) is high; this shows that the distribution of the companies over the maps approximately stays the same. So although the interpretation of the clustering is more difficult, the actual projection and clustering does not change much!

Merging four quarters of 1998

The previous analysis was performed for the fourth quarter of 1998. We do not expect the companies to change much over the course of a year, so when merging the four cross-sections of 1998 (first, second, third and fourth quarter) we expect to find the same clustering. The resulting map is shown in figure A-27 in appendix VI. Using companies from all four quarters of 1998 (1098 data points) generates a map with the same global ordering of the clusters as the previous map. Although the map is rotated 90 degrees compared to the original map, the same relations can be inferred.

Because the merged map contains approximately 4 times as much companies as our final map, we can not compute the cluster coincidence between these maps. But we can compute the cluster coincidence between our final map and the placement on the merged map of companies from one specific quarter, this is shown in figure A-28 in appendix VI. Cluster coincidence is high when comparing only the fourth quarter companies of the merged map with the final map (found in iteration 3). Cluster coincidence however slightly deteriorates when comparing companies from the older cross-sections (quarters 3, 2 and 1) with the final map.

Differences between 1998, 1997, 1996, 1995 and 1994

Comparing maps for 1998, 1997, 1996, 1995 and 1994 (figures A-27 to A-32 in appendix VI) provides an insight into any fundamental differences between the characteristics of the clusters over the years. The self-organizing maps of the different years show a varying image, but most relationships hold over the years.

- In all years the companies with high interest coverage, low leverage, high profitability and high stability represent healthy companies.

- The characteristics of unstable companies and underperformers are also preserved.

- The relative placement of the clusters on the map does not change, indicating the same global ordering of the companies in the input space. Healthy and large, stable companies are situated near each other, as are unstable companies and underperformers.

The characteristics for two specific clusters seem to have significantly changed from 1996 to 1997:

- The large and stable companies show a medium return on assets in the years 1994, 1995 and 1996. In 1997 and 1998 these companies show a high return on assets.

- The small companies show a high return on assets in the years 1994 through 1998, and a low return on assets in 1997.

Table 4-5 shows that the return on assets of large companies significantly improved when comparing 1996 and 1997, and this remained high for 1998. It also shows that the return on assets of small companies significantly worsened in 1997.

Standard & Poor's did not noticeably shift their ratings from 1996 to 1997, so either the rating agency changed their criteria (regarding return on assets) for obtaining a specific rating or they never put much emphasis on return on assets at all.

Table 4-5 Return on assets over the years

Cluster		1998	1997	1996	1995	1994
large, stable	mean return on assets	0.033	0.036	0.021	0.024	0.022
	mean SP rating	14.957	15.341	15.914	16.292	15.169
small	mean return on assets	0.031	0.01	0.028	0.024	0.034
	mean SP rating	8.857	8.15	8.68	9	10.66

4.5 Benchmark

4.5.1 Principal Component Analysis

To provide a means for comparison we will rework part of our analysis using the principal components technique, previously discussed in chapter 3. The PCA technique tries to capture the intrinsic dimensionality of the data by finding the directions in which the data displays the greatest variance. The data can then be projected on the plane spanned by the first two of these directions.

Data

We once again use the data set containing company values for the fourth quarter of 1998, and all the financial ratios gathered at the start of our original analysis (eighteen in total). The used software package is XLStat, a Microsoft Excel add-in containing numerous statistical analysis tools.

Missing values

For a complete principal components analysis no values may be missing from the data. All records containing at least one missing value are deleted. Missing values are quite common for financial statement data; for our analysis this means deleting 158 records out of 287! Normally we would try to find a work-around for the problem (e.g. insert averages for the missing values), but at this stage we accept the lesser significance of the results. The Self-Organizing Map technique does not have this drawback, the SOM algorithm uses as much of the available data as possible to create the map.

Correlations matrix

First step in the analysis is the creation of a correlations matrix, displayed in table A-5 in appendix VI. Based on this matrix the application finds the uncorrelated principal components and corresponding eigenvalues (table A-6), which are equal to the variances of the principal components. The total population variance due to each principal component can be calculated using these eigenvalues (this is shown in table A-7 in appendix VI).

4.5.2 Results

The first eight principal components cover 86% of the variance in the data. Furthermore, the correlations between original variables and the principal components lose their significance after the first eight principal components (the correlations do not exceed 60% anymore). We therefore conclude that the linear relations in the data set can adequately be described using just the first eight principal components. The dimensionality of the data set has been reduced from 18 original variables to 8 principal components.

The principal components are shown in table 4-6. The characterization of each principal component is based on the variables having the highest correlations per component.

Table 4-6 Principal components

Principal component	Covered variance / cumulative	Characterization	Variables	Correlation with principal component
1	0.28 / 0.28	Interest coverage, earnings	EBIT interest coverage	0.92
			EBITDA interest coverage	0.90
			EBIT / total debt	0.90
			return on assets	0.84
2	0.13 / 0.41	Leverage	net gearing	0.83
			debt-equity ratio 1	0.80
3	0.12 / 0.53	Leverage	debt-equity ratio 2	0.72
			debt ratio	0.66
4	0.08 / 0.61	Profitability	net profit margin	0.63
5	0.08 / 0.69	Size	log total assets	0.69
			log market value	0.63
6	0.07 / 0.76	Market	beta	0.81
7	0.05 / 0.81	Stability	c.o.v. net income	0.66
8	0.05 / 0.86	Perceived risk	c.o.v. forecasts	0.76

The PCA has grouped the original financial ratios according to the broad classification described in the beginning of this chapter. Leverage has been divided over two principal components (2 and 3), but the division is the same as the one found in the self-organizing map analysis. Debt-equity ratio 1 and net gearing are highly correlated, and so are debt-equity ratio 2 and debt ratio.

Another similarity with SOM is the reduction of the number of variables from 18 to 8. The exact set of variables found using PCA is slightly different.

- PCA selected the market variable (beta) where SOM selected another stability variable (c.o.v. total assets).

- PCA selected two instead of one leverage ratios.

- PCA selected one profitability ratio instead of two.

And where SOM of course selected only original variables, most of the variables found using PCA are combinations of other variables. We could simplify this by removing all but one of the highly correlated variables within a principal component.

Visualization

The projection of the data on the plane spanned by the first two principal components is shown in figure 4-9. For high dimensional data like our data set this projection unfortunately has little added value. We can not infer any relations with respect to higher dimensions of the data from this picture.

Clustering

As PCA does not include a clustering algorithm, we are unable to cluster the observations. PCA could be augmented by one of the standard clustering techniques, but it would still be difficult to directly (visually) infer relationships between the clustering and one or more variables, as is possible with SOM.
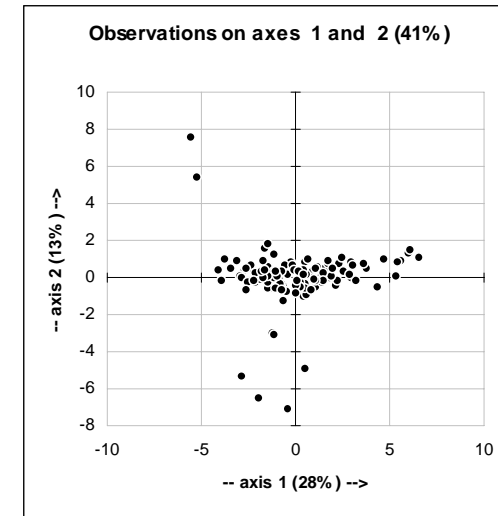
Figure 4-9 Projection on first principal plane

4.5.3 Comparison with SOM

We have summarized the differences and similarities between SOM and PCA in table 4-7. PCA and SOM can both be used to reduce the dimensionality of large data sets, the found compressed data sets are very much alike. Using SOM is especially advantageous when one suspects non-linear relationships in the data, as PCA can not handle these. The added visualization and clustering techniques of the used SOM implementation provide an insight in the data that PCA lacks.

Table 4-7 Comparison of SOM versus PCA

	SOM	PCA
Software package	Viscovery SOMine 3	XLStat (MS Excel add-in)
User friendliness	high	high
Missing values allowed	yes	no
Dimensionality reduction	from 18 to 8	from 18 to 8
Spread of variables over variable classes	broad	broad
Found relationships	linear and non-linear	linear
Projection of observations	On flexible plane through all dimensions	On flat plane spanned by first two principal components
Added value of projection	high	low
Clustering of observations	yes	no

4.6 Summary

In this chapter we have used the knowledge discovery process and specifically the Self-Organizing Map technique to perform a descriptive analysis of the credit rating domain. We started with a basic data analysis, to get a general 'feel' of the data and already making some important decisions. A single sector (Consumer Cyclicals) from the available universe of US companies was selected, and for each company we computed the financial ratios already mentioned in chapter 2. Next to these figures we also downloaded the Standard & Poor's credit ratings. The size variables were log transformed and we used a cut-off for all variables, to take care of extreme values.

We then proceeded to create a SOM clustering of the observations, using only financial statement data to train the map. A clustering was found, whereby the clusters can be characterized by the average values for the financial ratios of the companies in the cluster. Furthermore, when comparing the distribution of the S&P ratings over the companies in the clusters with the characterizations of the clusters there appears to be a positive correlation: Companies in the 'Healthy' cluster received high ratings, whereas companies in the 'Underperformers' cluster received low ratings.

We have made this visual coincidence somewhat more quantifiable using the cluster coefficient of determination. Our descriptive model based on financial statement data alone explains about 60% of the variance in the ratings. If we presume the SOM model to be accurate and if the data does not contain any major errors, then we could possibly attribute the other 40% to the qualitative analysis performed by S&P.

Tests on sensitivity of the algorithm (specific settings of the SOM during training) and sensitivity of the data (different cross-sections) show that the model and the found results are stable. An analysis performed using Principal Components Analysis gives similar results, but without the benefit of the insightful visualizations and clusterings specific to SOM.

5 classification model

Chapter 5 describes our efforts to build a classification model based on financial statement data. Question 5 from the introduction will be answered:

5.   Is it possible to classify companies in rating classes using only financial statement data?

Paragraph 1 describes the general model set-up. Then the construction of our SOM model is extensively reviewed in paragraph 2. The model is validated in paragraph 3, and in the next paragraph we compare the SOM model with our two benchmark models, linear regression and ordered logit. The final out-of-sample test is conducted for all three models in paragraph 5.

5.1 Model set-up

5.1.1 Training and prediction

The models are all set-up according to the following template:

1. The sample of companies in the Consumer Cyclicals sector is randomly divided in a train and validation set (in-sample) and a test set (out-of-sample).

2. The map is trained using the train set, then the ratings are predicted for the validation set. This in-sample training and validating is repeated (using different settings and variables) until we are satisfied with the found model. The test set is reserved for the final out-of-sample test.

3. The predicted ratings are compared with the real ratings and several measures of likeliness are computed. The relative prediction error (or classification performance) of the model can thus be ascertained.

In the following paragraphs we will more thoroughly review each of the model steps.

5.1.2 Data

Equivalent to our descriptive analysis in chapter 4 the used sample consists of companies from the Consumer Cyclicals sector. We use exactly the same data, so we do not need to repeat our basic data analysis. This time we merge the 8 cross-sections of years 1997 and 1998 (4 quarters in each year) to gain a larger sample, which should improve the statistical accuracy of the model. This is also consistent with results found in our descriptive analysis.

We randomly divide the sample into a train, a validation and a test set. The train set covers approximately half of all the companies, whereas the validation and test sets each cover a quarter of all the companies. The train and validation set together form our in-sample dataset, while the test set is reserved for our out-of-sample test. The train set is used to train the map. We then use the validation set to predict the ratings for the companies in this set and compare them with the real ratings. Iteratively different settings and sets of variables are tested, each time re-training the map on the train set and predicting ratings for the validation set. The test set is reserved for the final out-of sample test, and is not used until we are completely satisfied with the found model.

When dividing the sets we make sure that multiple instances of the same company (from multiple cross-sections) remain in the same set. We are thus assured that the map does not base the prediction for a company solely on a previous or later instance of the same company. We also make sure that all classes are as good as possible represented in all three sets. This is not always possible because of the very few companies in some

classes. Please refer to paragraph 5.1.4 for more information on the ratings distribution and the corresponding implications.

5.1.3 The prediction process

The self-organizing map is trained using the train set. In the input space, the flexible grid with neurons takes on the form of the data (the train set)[39]. After training, the form of the grid remains fixed and can be used to predict values for new observations. This is very similar to linear regression, where first the function parameters are estimated and then the functional form found is used to predict values for new samples.

To predict the rating for a new company we look up the neuron in the grid nearest to this company (according to the Euclidean distance measure). This neuron is also called the Best Matching Unit. The rating associated with this neuron is then assigned to the new company as its predicted rating. Please note that although the ratings themselves are strict integers, the predicted ratings are not. When two or more companies are assigned to the same neuron (during the training process), and the ratings of the companies differ, then the predicted value is a non-integer number[40].

Alternative interpretation

We can provide an alternative interpretation for the prediction process: The neurons form best representations for small groups of similar companies from the train set. The extent to which the companies are regarded similar is only dependent on the used financial ratios. All the neurons together (the grid) form an as good as possible representation of the whole train set. When we want to evaluate a new company (e.g. a company from the validation set), we first match it with its most similar neuron. We are in a way looking for the companies in the train set that are most similar to the new company. Then the averaged rating for these companies is assigned to the new company, presupposing that companies in the same sector in a similar financial situation are granted the same rating by Standard & Poor's. The neurons, based on all the companies in the train set, function as proxies for companies in specific financial situations in this particular sector. Their associated ratings convey the common credit outlook S&P employs for these kinds of companies.

5.1.4 Ratings distribution

We use the same rating classification scale as in our descriptive analysis, which is shown in table 5-1. The direct translation of the letter ratings to an equi-distant numerical scale presupposes that rating classes are of equal width. The risk gain when going from an AA to an AA- company is as large as the risk gain when going from a

---

[39] Please refer to chapter 2 for an extensive treatment of the SOM algorithm and the prediction process.
[40] As the rating of each neuron is updated when training the map this also leads to deviations from integer values.

BBB- to a BB+ company. This might not be true. However, the non-linear relationships in the SOM model compensate for these restrictions, making it possible to represent unequal class widths.

By carefully examining the ratings distribution of all the companies, and per train, validation and test set, we can get a clearer picture of what results to expect. The rating distributions are shown in figures 5-1.

The histogram of the overall ratings distribution shows that certain rating classes are under-represented: Only a few defaults occur (0.36 percent or 7 companies), and no C, AA+ or AAA companies are selected in our universe. The contribution of the CC to B-, AA- and AA rating classes is very low, so effectively only the classes from B through A+ (or 8 through 18) are correctly represented. The average rating is BB+ or 12.

The same image holds true for the train, validation and test set alone. Additionally, the validation set exhibits an under-representation of the BB+ class and an over-representation of the B+ class. The test set shows an under-representation of the B and A- classes.

The lack of extreme high rated companies can be explained from the choice of sector: The Consumer Cyclicals sector consists of companies in a volatile market with lots of risks and high demands on companies. As the sets were randomly chosen the aberrations in the distributions of the validation and the test set can be attributed to pure coincidence. A relatively high percentage of a certain class in one set is off-set by a relatively low percentage of the same class in the other sets.

Table 5-1 S & P Rating classification scale

S & P Rating	Rating Code
AAA	22
AA+	21
AA	20
AA-	19
A+	18
A	17
A-	16
BBB+	15
BBB	14
BBB-	13
BB+	12
BB	11
BB-	10
B+	9
B	8
B-	7
CCC+	6
CCC	5
CCC-	4
CC	3
C	2
D	1

Implications

The lack of sufficient examples of extreme low ratings or extreme high ratings makes it improbable for any model to correctly predict these classes. And even if the ratings are predicted correctly we can not verify the classification results for these classes.

Using ratings as a train variable

In our descriptive analysis we explicitly did not use the S&P rating when training the map. We wanted first to find a suitable clustering of companies based on financial ratios and then match this clustering with the creditworthiness of companies as expressed in the ratings.

Now we have a more direct goal: we want to predict the S&P rating as good as possible based on all available information. The descriptive analysis supported the observation that the ratings contain extra information next to an assessment of credit risk purely based on financial ratios. This information could be beneficial to our model, so we should somehow represent it in our model. We achieve this by using the rating as a train variable during map training. The companies are then clustered on financial ratios and qualitative information. The rationale behind this kind of semi-supervised training is covered in paragraph 3.7.2.
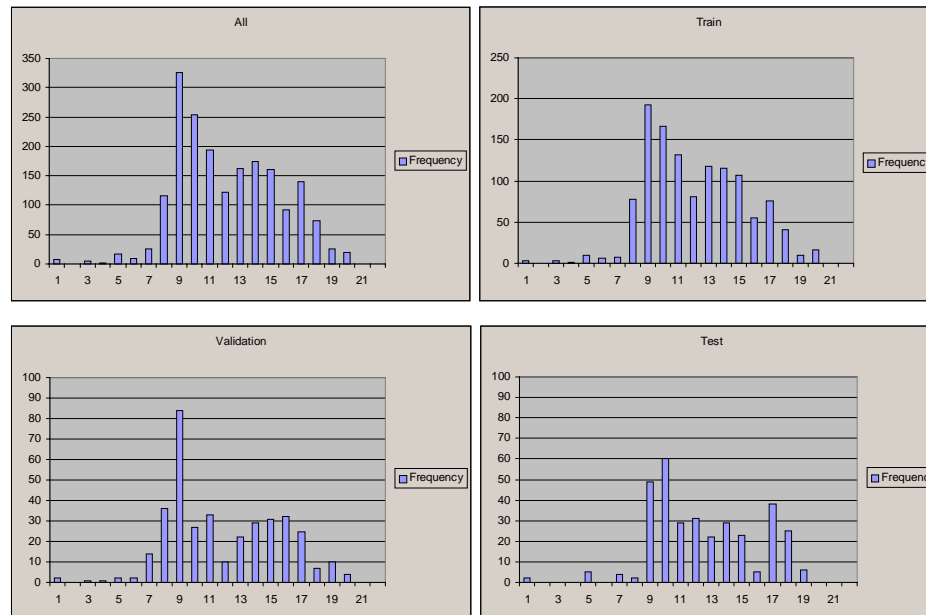
Figure 5-1 Ratings distributions for all companies, train set, validation set and test set.

5.1.5 Measuring performance

The performance of a model can be measured using several criteria. They all have in common that real ratings are compared with predicted ratings, leading to some kind of classification error. Using these classification errors we can compare models to find the best performing model.

We are comparing predicted ratings mainly based on financial ratios with real ratings based on financial ratios and a qualitative analysis. The qualitative analysis adds considerable 'noise' to the rating, so our model can never be perfect.

Success ratio

An obvious criterion for the classification error of a model is the success ratio: The percentage of the validation set for which the map predicts ratings within a specified maximum number of notches deviation from the real rating. The SOM algorithm necessarily predicts non-integer ratings so we have to convert the predicted ratings to an integer valued scale. Rounding the numbers to the closest integer number most easily does this. We can now compute success ratios for 0 notches deviation, 1 notch deviation, 2 notches deviation, and so on. Please note that the deviation is not restricted to one side of the real rating and that the success ratios are reported cumulatively. A 78% success ratio for 2 notches deviation would thus mean that predictions for 78% of the companies are at most 2 notches in error (e.g. from BBB to BB+ or to A-).

Plotting a histogram of the success ratios gives a clear indication of how widespread the errors are. Figure 5-2 shows the plot for our initial model.
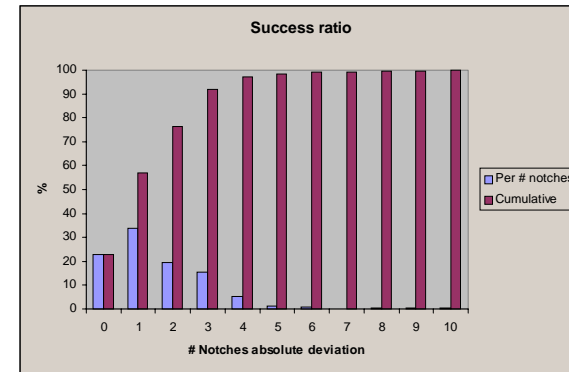
Figure 5-2 Example of success ratios histogram

Mean Absolute Deviation

A related useful error measure is the Mean Absolute Deviation, mathematically defined as

$$MAD = \frac{\sum_{n=1}^{N} \left| R_n - \tilde{R}_n \right|}{N} ,$$

where R_n is the real rating for company n, $\tilde{R}_n$ is the predicted rating for company n and N is the total number of companies in the sample. This shows how much the predicted ratings deviate from the real ratings, without stressing extreme deviations (as opposed to a measure like the standard deviation).

R^2

The coefficient of determination or R^2 is an often used measure for the performance of a linear regression model. It shows the variance in the predictions of the model that can be explained by the variance in the variables. A perfectly classifying model is characterized by an R^2 of 1. The non-linearity of the SOM model prohibits us from directly calculating this R^2, but we can calculate a simulated R^2 by assuming a linear model to have generated the found results.

We hereto first create a scatterplot of the real versus predicted ratings. Ideally the points should lie on the diagonal; all predicted ratings are the same as the real ratings and a higher real rating coincides with a higher predicted rating. The scatterplot for our initial model is shown in figure 5-3.

It is now very easy to execute a linear regression through these points and calculate the accompanying R^2 measure. Although this R^2 is somewhat artificial and cannot be compared to the R^2 in a direct linear regression classification model, we can use it to compare

Figure 5-3 Example of ratings scatterplot

between models, linear or non-linear. The found constant and coefficient of the regression convey less meaning, we explain this in the next subchapter.

Statistical validation

The R^2 describes the fit of the predicted ratings to the real ratings. This is used as a measure for the performance of the model. In normal linear regression models the coefficient of a variable represents the contribution of this variable to the model result. The validity of the model is verified by statistically testing that this contribution was not a chance occurrence.

When we assume (amongst others) the residual errors of the regression to be normal distributed around zero, then the ratio of the coefficient and its standard regression error follow a students t distribution[41]. We can test whether the coefficient is unequal to zero and the contribution of this variable is statistically significant. If we use a zero coefficient as our null hypothesis, we can reject this hypothesis when the observed t-value exceeds a certain threshold. For an often used significance of 5% this threshold is 1.96.

For non-linear models it is not possible to directly relate the individual components of the model to the prediction. We can compute the t-value of the coefficient in our assumed linear model between predictions and real ratings, but the resulting high significance is in our case rather trivial. It is easy to see that a strong relationship between predicted and real ratings exists, and the large number of observations only serves to strengthen this relationship. The contribution of individual variables remains clouded.

[41] Greene, W.H., 1997, pages 264 and 265.

To verify the validity of the model as a whole we can compare its performances with a number of naïve and random models. A good model will perform better than these models, regardless of chosen settings.

5 classification model

94

5.2 Model construction

Our search for a good SOM model for predicting S&P ratings starts with an initial model. Starting from this initial model we first try to reduce the number of used variables, leading to a less complicated model without sacrificing significant classification performance. Then using this smaller set we test the initial model assumptions, possibly leading to better classifications. We also explore some interesting paths like emphasizing the less-frequently occurring extreme ratings. Finally we will compare the best model with a number of suitable random models, a constant prediction model and our benchmark models; linear regression and ordered logit.

Every time when we evaluate a model we not only look at absolute scores, but also at the practical usability of the used variables and settings. We do not want a model that is perfectly tuned (overfitted) to the current validation set, but we want a robust model that is easy to grasp and use.

5.2.1 Initial model

Our initial model is partly based on the results found in the first part of our research. To be more specific:

- We use a two-year or eight-quarter historical period, as this proved to be the longest period for which no important changes occurred.

- A two quarter publication lag is used (the time between the realization and the publishing of the financial figures).

- The number of used neurons (1000) is approximately equal to the number of used samples in the train set (1200 companies), this proved to be adequate for describing the data set in our previous research.

- The original 18 financial ratios and the S&P ratings are used to train the model.

Model performance

The performance of our initial model is shown in figure 5-4 and in table 5-2.
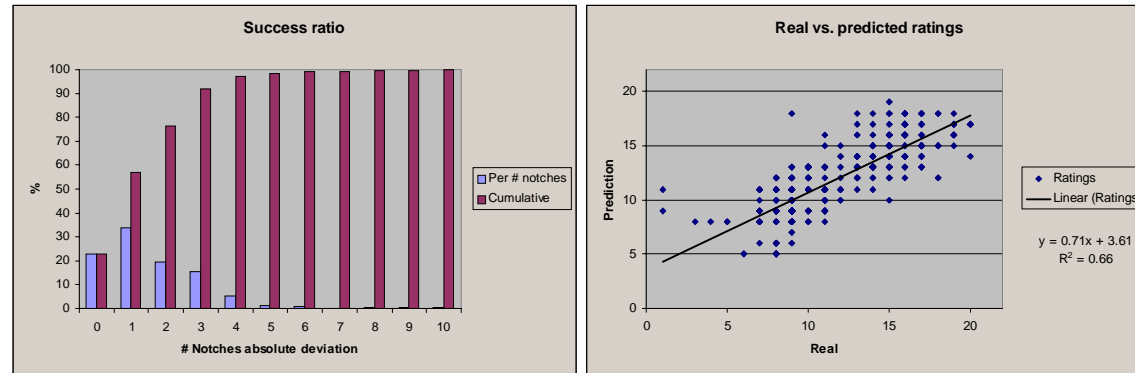
Figure 5-4 Success ratios and ratings plot for our initial model

The success ratios show that 23 percent of the companies in the validation set are perfectly classified. Approximately 76 percent of the companies have been classified with an error of at most 2 notches.

Table 5-2 Model performance for initial model

Model	MAD	Success ratio			R^2
		0	1	2	
Initial	1.59	0.23	0.57	0.76	0.66

5.2.2 Variable reduction

The initial number of variables (18) is extensive and contains redundant information, as we inferred from our descriptive analysis. A model with fewer variables presents a less chaotic display making it easier to understand the model and the relationships between the variables and the predicted rating. Such a model is also easier to use for future classifications, as less additional data has to be downloaded and transformed or checked.

As presented in chapter 2, the variables have been grouped in six major classes:

- Interest coverage ratios: these measure the extent to which the earnings of a company cover debt or interest.

- Leverage ratios: these measure the financial leverage created when firms borrow money.

- Profitability ratios: profitability ratios measure the profits of a company in proportion to its assets.

- Size variables: these measure the size of a company.

- Stability variables: stability variables measure the stability of the company over time in terms of size and income.

- Market variables: market variables are used to assess the value investors assign to a company.

We aim to select the most promising variables in each class. We therefore try all major combinations in each class, while keeping all other variables and settings equal. The variables performing best (classification performance of the model) and conveying the most information (component plane of the variable) are selected. We do not try all possible combinations of variables, as this would mean trying 2^{18} - 1 or about 262143 combinations.

The selected variables in each class are shown in table 5-3 . In table A-8 in appendix VII a full overview of the model scores per variable combination can be found.

For some classes another variable or combination of variables actually performs better. In these cases the selected variable is chosen because of a more clear definition of the variable or because of a definition more suitable for practical use. The model using the chosen variable is always the top or second-best performer in its class and the differences between these two are always very small.

Table 5-3 Selected variables per variable class

Variable / financial ratio class	Selected variable
Interest coverage	EBITDA interest coverage
Leverage	Debt ratio
Profitability	Return on equity
Size	Log total assets
Stability	Coefficient of variation of total assets
Market	Coefficient of variation of forecasts year 1

After selecting the variables per class we want to examine the effects of removing a variable class. If removing the variable class does not lead to a significant drop in classification results then it apparently does not contribute to the model and we see no need to include it in our model.

The prediction results for the found model and the results when subsequently removing each of the variable classes are shown in Table x. From the effect on classification performance when removing a variable we have deduced the relative importance of the variable, these are shown in the last column of Table 2. Two things are most noticeable:

1. The size variable contributes considerably to an accurate account of the creditworthiness of companies. This is visually evident from the component planes, as the ratings and size component planes are most alike (shown in figure 5-5). When we leave out the size variable the prediction results significantly drop.

2. The market variable does not contribute much to the form of the map, as is shown in figure 5-5. Leaving out this variable improves the prediction results for all possible performance measures, what leads us to believe that the market variable only adds noise to the prediction.

Table 5-4 Model performances when removing variable classes and relative importance of each variable class

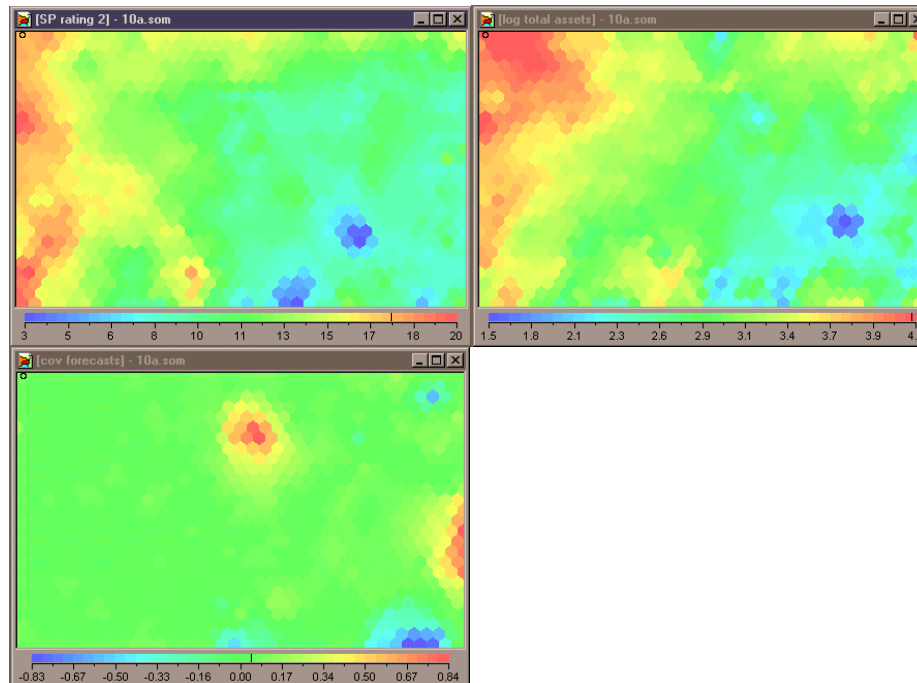| Model (used classes)	MAD	Success ratio			R^2	Relative Importance
		0	1	2		
All	1.74	0.24	0.56	0.73	0.59	n.a.
without interest coverage	1.90	0.21	0.53	0.71	0.54	++
without leverage	1.75	0.22	0.51	0.74	0.62	+
without profitability	1.73	0.22	0.51	0.74	0.60	+
without size	2.13	0.19	0.49	0.67	0.41	++++
without stability	1.68	0.23	0.56	0.74	0.64	+
without market	1.66	0.26	0.59	0.74	0.61	-

Figure 5-5 Component planes for S&P Rating, Log total assets and Coefficient of variation of Forecasts

Relationship with previously found financial ratios and clusters

The financial ratios found in this chapter do at first sight not exactly agree with the variables used in our descriptive analysis. The differences are shown in table 5-5.

Table 5-5 Comparison between selected variables in descriptive and classification analysis

Variable class	Selected for description	Selected for classification
Interest coverage	EBITDA interest coverage	EBITDA interest coverage
Leverage	Debt-equity ratio 2	Debt ratio
Profitability	Return on assets Operating income / sales	Return on equity
Size	Log total assets	Log total assets
Stability	Coefficient of variation of total assets Coefficient of variation of net income	Coefficient of variation of total assets
Market	Coefficient of variation of forecasts	-

In both analyses a broad selection is made, to represent all financial aspects of a company. Within a variable class the choices for specific financial ratios may slightly differ, and the market variable class is not represented anymore in our prediction analysis. One possible explanation for these differences is the different variable selection procedure: In our descriptive analysis at we have selected the ratios based solely on their contribution to a good clustering, without taking the ratings into account. Now we are directly relating the financial ratios to S&P ratings containing quantitative and qualitative information, leading to different choices.

5.2.3 Sensitivity analysis

Now that we have found an adequate smaller subset of variables we reconsider some of the earlier specific choices for the model parameters and test what impact changing these settings has on the model results. The model parameters for which we try different settings are:

- History length: The length of history or the number of quarterly cross-sections to use for training of the model.

- Number of neurons: The number of neurons used in the map.

- Prediction neighbourhood K: The size of the neighbourhood to take into account when predicting values from the map.

- Using ratings as a train variable: The contribution of the extra qualitative information in the ratings.

The classification performances for all evaluated model parameter settings are displayed in table 5-6. The ultimately selected settings are displayed in red.

Table 5-6 Classification performances for different parameter settings

Model	MAD	Success ratio			R^2
		0	1	2	
Used historical period					
1998 Q4	1.49	0.35	0.56	0.75	0.69
1998	1.63	0.26	0.56	0.74	0.61
1997 & 1998 (default)	1.66	0.26	0.59	0.74	0.61
1996 & 1997 & 1998	1.79	0.23	0.57	0.79	0.56
Used number of neurons					
250 neurons	1.50	0.26	0.65	0.78	0.67
500 neurons	1.60	0.25	0.58	0.76	0.64
1000 neurons (default)	1.66	0.26	0.59	0.74	0.61
2000 neurons	1.68	0.26	0.55	0.76	0.62
Size of prediction neighb. K					
K = 1 (default)	1.66	0.26	0.59	0.74	0.61
K = 10	1.57	0.26	0.62	0.74	0.65
K = 25	1.48	0.28	0.63	0.77	0.67
K = 50	1.43	0.28	0.63	0.77	0.68
K = 75	1.42	0.31	0.64	0.78	0.69
K = 100	1.44	0.29	0.64	0.77	0.69
Taking ratings into account					
No ratings	1.50	0.27	0.64	0.78	0.65
1 x ratings (default)	1.66	0.26	0.59	0.74	0.61
2 x ratings	1.88	0.20	0.51	0.72	0.55

History length

A longer historical period means a larger sample and a statistically more sound model. But a too long historical period could also obscure some relationships in the data because of changed environments for companies and changed measures for extending ratings to companies.

Initially a two year history was used (the eight quarterly cross-sections of 1997 and 1998 merged). Using only one quarter of data generates a better classifying model than using two full years (or eight quarters) of data. Because of the higher statistical significance (due to the larger sample) we do not change our initial two-year historical period.

Number of neurons

The number of neurons should be tuned to the way we want to use the SOM. A large number of neurons (>= number of samples) gives a more accurate description of the data. A smaller number of neurons (<< number of samples) produces a more general map, which predicts better in a multitude of cases.

As the number of observations in our sample equals 1200, we have used about 1000 neurons to train the map in our initial model. When we try maps using 250, 500 and 2000 neurons, we find that less detail (250 - 500 neurons) builds a better generalizing map.

Prediction neighbourhood K

In the previous models we always used the single best matching neuron for the current company to extract a prediction from the map. A common variant of the prediction algorithm is to take a weighted average of the rating over the K neighbouring neurons (in the input space). The less nearby neurons in the neighbourhood contribute less to the prediction, in a linear fashion. There is a correspondence between choosing a larger K for prediction and a smaller number of neurons when training the map. Both have the effect of generalizing the predicted values for the ratings, so for clarity in our final model we should choose one of the two methods to enhance the generalizing capabilities of the map, and not use both.

The results in Table X show that the generalizing effect of using larger K's gives better classifications. We opt to only use the K neighbourhood as a generalizing instrument, as this is more flexible than using less neurons. The map still accurately represents the underlying dataset (no information is lost), while we can vary the generality of the predictions.

Using ratings as a train variable

We have used the ratings to train the map on the premises that the ratings contain valuable information that is not contained in the financial ratios. In SOM literature this is known as semi-supervised training. The target variable (in our case the rating) is trained along with the normal variables to create a better distinction between clusters for this variable. Please note that this does not necessarily improve the classification performance of the model: if companies in similar financial situations have very different economic expectations then the assigned ratings will be very diverse. This qualitative information would (in terms of our SOM model) add noise to the map, not leading to improvements in classification results.

This is exactly what happens in our case. The classification performances improve when we do not use the ratings as a train variable, and worsen when we attribute more influence to the non-financial component in the ratings (double the weight). There seems to be additional information contained in the ratings, otherwise the results would stay the same, regardless of the weight used during training. However, this information does not

contribute to a better clustering of the companies resulting in better classifications. The additional information in the ratings is contradicting the information contained in the financial ratios.

If we do not use the ratings when clustering the companies we can be sure that only financial information is taken into account when classifying a company. The assigned rating is an average rating for companies in similar financial situations. If we do use the rating as a train variable, then the clustering is based on financial information and the qualitative information as expressed in the ratings. Some companies, financially speaking belonging to a cluster of e.g. AA rated companies, are clustered with BBB rated companies because of a rating downgrade based on qualitative factors. The average values for the financial ratios for these BBB rated companies are off-set by these AA companies 'in disguise', leading to worsened predictions for true BBB companies. If the clustering is more dependent on the ratings (larger weight for the ratings variable during training), then this effect is more noticeable in worsened classification performances.

The qualitative up- and downgrades of companies are not systematic for companies in certain financial situations, they are more random-like. This is understandable, if these qualitative rating changes were systematic then they would be expressed in a higher or lower rating average for companies in these financial situations. Using the ratings as a train variable is only profitable when we are certain that the additional information in the rating does not contradict the information contained in the other variables. More generally speaking we should restrict the use of the target variable as a train variable to these models where the target variable is not contradicting the other variables of the model.

5.2.4 Results

Based on our analyses we can summarize our model so far:

- The historical period used is 2 years of length (1997 and 1998).

- The variables used to train the map are:

    - EBITDA interest coverage,

    - Debt ratio,

    - Return on equity,

    - Log total assets,

    - Coefficient of variation of total assets.

- The size of the map is 1000 neurons, or approximately as much as the number of observations.

- We predict ratings based on a neighbourhood of 50 neurons.

The model results are shown in figure 5-6 and table 5-7.

The performance figures show that most of the classification performance of the model can be captured using just a subset of five of the original eighteen variables. The performance loss is relatively small. After adjusting some of the parameters the performance of the model clearly improves. For most parameters the initially chosen settings were adequate, the greatest performance increase is found in the larger prediction neighbourhood size and in not using the ratings as a train variable.
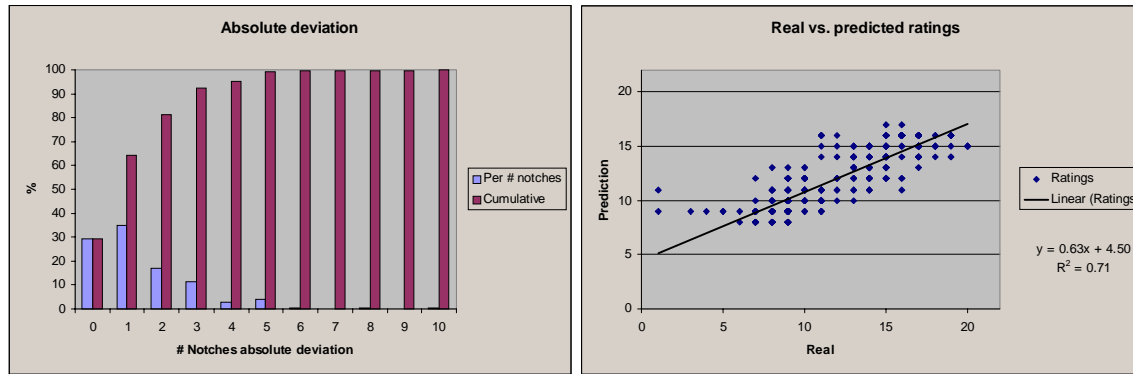
Figure 5-6 Success ratios and ratings plot for SOM model

Table 5-7 Model performances

Model	MAD	Success ratio			R^2
		0	1	2	
Initial	1.59	0.23	0.57	0.76	0.66
Reducing variables	1.66	0.26	0.59	0.74	0.61
Adjusting parameters	1.40	0.29	0.64	0.81	0.71

5.3 Model validation

To establish the validity of the found model we will compare it with a number of naïve and random models. We then scrutinize the classifications per rating class to obtain a more complete image of the classification power of the model.

5.3.1 Comparison with constant prediction

Our first comparison involves a constant prediction. As the average rating equals 12, we evaluate the success ratio when constantly predicting a rating of 12. The results are shown in figure 5-7 and in table 5-8.

As we expected, the mean absolute deviation from this constant prediction is approximately equal to the standard deviation of the original ratings distribution (3.19 versus 3.35).
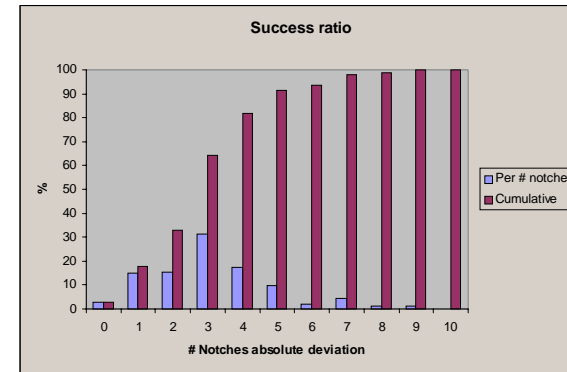
Figure 5-7 Success ratios for constant prediction

5.3.2 Comparison with random prediction

We can simulate a random prediction by randomly assigning ratings to the companies in the validation set. As each rating then has an equal chance of occurring, it would not be a fair to compare these predictions with the normal model. We therefore restrict the predictions to random ratings from the original ratings distribution of the validation set[42]. The success ratios histogram and ratings plot for one of the random models with a Mean Absolute Deviation of 4.19 is shown in figure 5-8. The model results can also be found in table 5-8.

---

[42] This distribution is shown in figure 5-1.
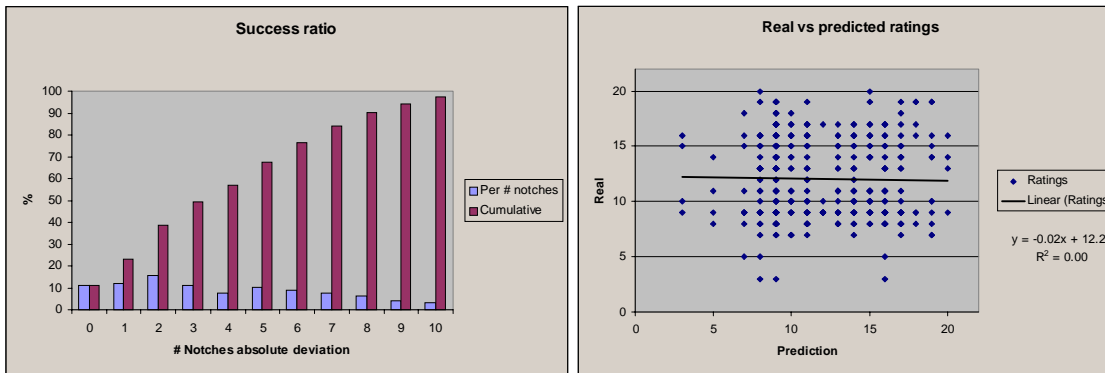
Figure 5-8 Performance figures for random model

The figures show that the random model performs poorly. To verify that we did not just happened to stumble upon a relatively bad predicting random model we have simulated 100 random models (this is also known as bootstrapping). The histogram of the Mean Absolute Deviation for these models is shown in figure 5-9. The MAD of the models seems to be normal distributed around a mean of 4.11.
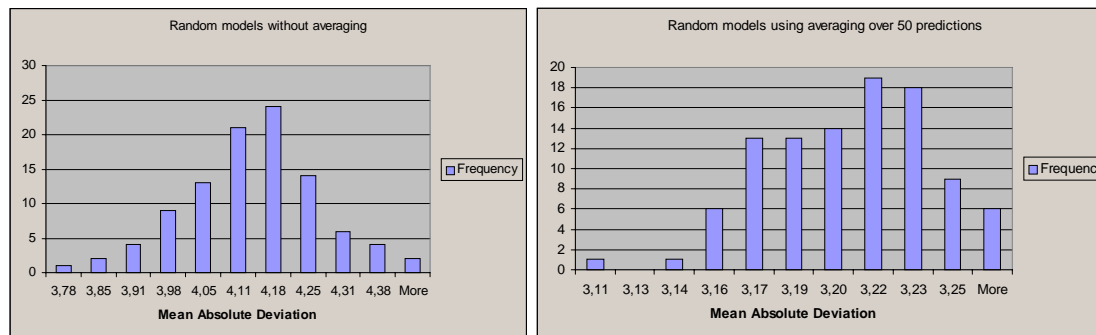
Figure 5-9 Distribution of MAD for 100 random models with and without averaging

We have not yet accounted for the averaging the SOM uses when predicting ratings, so we repeat the simulations using an average over 50 predictions as the predicted rating. The results are also shown in figure 5-9. The MAD of the models converges on the same MAD as the constant prediction (3.20). Furthermore, the spread of the MAD is much smaller.

Comparing the MAD of our SOM model (1.40) with the distribution of the random models shows that it is highly unlikely that we have struck upon a good model by chance.

5.3.3 Classifications per rating class

Scrutinizing the average positive and negative deviation per class and the maximum and minimum deviation per class gives a more detailed image of the classifications. The middle classes are relatively well predicted with an occasional peak. The outer edges of the ratings distribution are not so well predicted. This is most likely due to the relatively small number of observations in these parts of the sample.

This classification bias is visible in figure 5-10. Low ratings (1 through 7) are classified too high, high ratings (17 through 20) are classified too low. Two possible explanations for this behaviour are:

1. The main bulk of the sample has an average rating, so the model will be best fitted for these kind of companies. Classifications of extreme rated companies will always be biased towards the average, which is higher for low ratings and lower for high ratings.

2. Lower ratings are difficult to classify too low, as there are hardly any lower classes. Vice versa the same holds for high ratings.

Table 5-8 Performance comparison

Model	MAD	Success ratio			R^2
		0	1	2	
SOM	1.40	0.29	0.64	0.81	0.71
Constant	3.19	0.03	0.18	0.33	-
Random	4.11	0.10	0.27	0.41	0.00
Equalized SOM	1.46	0.22	0.60	0.76	0.71

5.3.4 Equalized ratings distribution

We try to eliminate the classification bias by equalizing the ratings distribution of the sample. We therefore remove observations exceeding a sample of 75 observations per class, starting with the oldest observations. The observations in the sparse classes are replicated until these classes are also filled with 75 observations. The map will now allocate more neurons to these sparse classes, or equivalently, give less weight to the middle classes. The distribution between the outer and the middle classes is more even, but this procedure does not add new information to the map. We expect predictions in the outer rating classes to still contain large errors.

A comparison of the deviation per class before and after equalization is given in figure 5-10. The average prediction errors are more constant, especially for the positive deviations. The classification bias seems to have been removed for the middle classes, at the expense of larger positive and negative peaks. The overall classification performance slightly deteriorates, as is displayed in table 5-8.
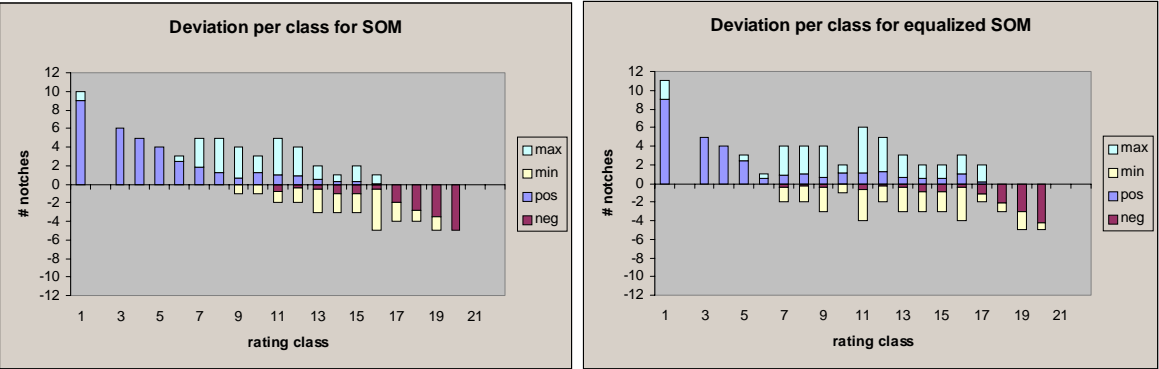

Figure 5-10 Classification bias before and after equalizing

5.4 Benchmark

We use two other models as a benchmark for the classification results of our SOM model. The first is a standard linear regression model, the second is a more advanced technique called ordered logit.

5.4.1 Linear regression

The linear regression model uses the Ordinary Least Squares estimation method to fit the ordinal ratings to the 18 variables used in our SOM analysis. We standardize each variable by subtracting the average and dividing the standard deviation of the variable from each instance, this is also known as the z-score. We substitute 0 (zero) for the not-availables as the linear regression model can not handle these. This is equal to replacing the not-availables with the averages of the variable. We also remove the highly correlated variables, to make the regression more stable. The significance of the found coefficients for the variables can directly be tested using t-values. Non-contributing variables ($-1.96 < t < 1.96$) are removed from the model. Please note that (similar to the SOM model) we are presuming even rating class widths.

5.4.2 Ordered logit

The ordered logit model is a so called ordered response model. It is an extension of the binary logit model and has the same foundation: A latent variable is assumed to be the determining factor for class membership. The value of this latent variable is determined by the used variables, for which the coefficients are estimated. This value and the class boundaries on the linear scale determine the class membership. The class boundaries are also estimated so the classes need not be of equal width[43], only the ordering of the classes is prescribed. This could more accurately reflect the structure of the credit rating classes and the differences between rating classes.

The ordered logit analysis again starts with the same 18 variables as for our SOM analysis. As with the linear regression model we standardize the variables, substitute 0 (zero) for the not-availables and remove the highly correlated variables. The boundaries and coefficients are estimated, and t-values are calculated to determine the statistical significance of the coefficients of the variables. Non-contributing variables ($-1.96 < t < 1.96$) are removed and the analysis is repeated until all coefficients are significant.

---

[43] More information on the ordered logit model can be found In Chapter 2 and in Fok, D., 1999. We would like to thank Dennis for the use of his ordered logit application and for his help with the interpretation of the results.

5.4.3 Results & comparison with SOM

The selected variables for linear regression and ordered logit are shown in table 5-9, along with the coefficients and the t-values. For comparison we have also displayed the corresponding SOM variables and their relative importance.

Table 5-9 Selected variables for SOM and ordered logit

Variable class	SOM variable	Imp	Linear regression variable	Coef	T	Ordered Logit variable	Coef	T
Interest coverage	EBITDA interest coverage	++	EBITDA interest coverage	0.68	9.94	EBITDA interest coverage	0.63	7.62
Leverage	Debt ratio	+	Debt ratio	-0.45	-6.81	Debt ratio	-0.65	-7.94
Profitability	Return on equity	+	Return on total assets	0.24	3.44	Return on total assets	0.31	4.28
			Operating income / sales	0.24	3.64	Operating income / sales	0.33	4.51
			Net profit margin	-0.23	-3.71	Net profit margin	-0.13	-2.07
Size	Log total assets	+++ +	Log total assets	2.06	34.51	Log total assets	2.19	26.48
Stability	CoV total assets	+	CoV total assets	-0.45	-8.24	CoV total assets	-0.45	-8.43
Market			CoV forecasts	0.13	2.21			
			Beta	-0.17	-3.09	Beta	-0.18	-3.16

Selected variables

The variables selected in the linear regression or ordered logit analysis do not substantially differ from the variables selected in the SOM analysis. Furthermore, the relative importance, coefficients and signs are similar in all three models.

The linear regression / ordered logit variable combination for the Profitability class has also been investigated in our SOM analysis, but did not lead to better performances. Likewise the Market class was dropped from the SOM model, the small coefficient in our linear regression and ordered logit analysis confirms this.

The net profit margin has a negative sign in the linear regression and ordered logit model, while we would expect a positive sign. This is probably due to the somewhat flawed definition of the variable as was observed in our SOM descriptive analysis. This is one of the reasons why we opted not to use this variable in the SOM model. The relative small coefficient shows that the variable does not contribute much to the linear regression and ordered logit model, either.

Rating scale

The linear regression model presumes an ordered rating scale with equal class widths. The ordered logit model only presumes an ordered rating scale, the boundaries and thus the class widths are estimated. The estimated boundaries for the ordered logit model are displayed in table A-9 in appendix VII.

The estimated scale shows that all classes are of approximately equal width, they do not differ much from the class widths in the linear regression model.

Performance

The performance of SOM, linear regression and ordered logit are compared in table 5-10, deviations per class are shown in figure 5-11. The performances for all three models are similar, especially in the middle classes. The classification bias is present in all three models.
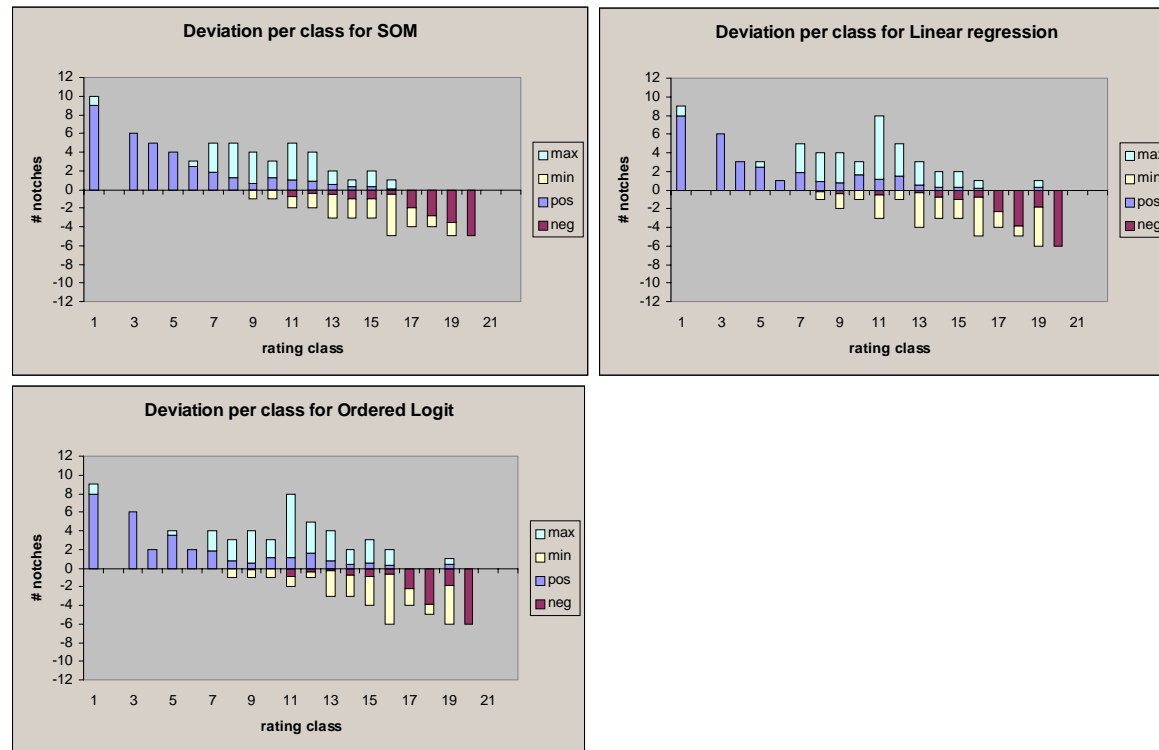
Figure 5-11 Deviation per class for SOM, linear regression and ordered logit in-sample

The good scores for the linear regression model indicate that the possible non-linearities in the data might not be that influential at all. Also the conversion of the letter rating scale into an (equally spaced) numerical scale does not seem to cause much problems. This may be due to the large number of classes (22) giving a close approximation to a continuous scale. The similarities in deviations for linear regression and ordered logit strengthen the image that the ordered logit model approximates a pure linear model.

Table 5-10 Performance comparison

Model	MAD	Success ratio			R^2
		0	1	2	
SOM	1.40	0.29	0.64	0.81	0.71
Linear regression	1.52	0.27	0.59	0.82	0.65
Ordered logit	1.44	0.28	0.64	0.83	0.67

5.5 Out-of-sample test

The out-of-sample test consists of classifying the companies in the test set, for SOM, for linear regression and for ordered logit. To gain the best possible results we use all available in-sample data (train and validation set) to construct the models. Unfortunately some classes can not be tested as the test set does not contain observations in these classes (3, 4, 6 and 20).

The test set is a subset of the sample of companies from 1997 and 1998. To test the stability of the found results we will also perform an out-of-sample test on older data (1996 and 1995). Finally the predicted ratings are linked to spreads to evaluate any possible matches of our ratings with the market point of view.

5.5.1 Results for test set

The results are displayed in table 5-11 and figure 5-12. The classification results for SOM are out-of-sample slightly worse than for the in-sample data set, while the other models perform slightly better out-of-sample. The relatively constant performances for the in- and out-of-sample datasets stems from the uniformity of the train, validation and test set: The test set is a subset of the sample of companies from which the train and validation set were also derived, so we would not expect large qualitative differences between these sets.

During the re-estimation of the linear regression model the 'Coefficient of variation of forecasts' variable was

Table 5-11 Out-of-sample performances

Model	MAD	Success ratio			R^2
		0	1	2	
SOM out-of-sample	1.48	0.25	0.60	0.82	0.64
SOM in-sample	1.40	0.29	0.64	0.81	0.71
Linear regression out-of-sample	1.48	0.21	0.59	0.84	0.65
Linear regression in-sample	1.52	0.27	0.59	0.82	0.65
Ordered logit out-of-sample	1.38	0.28	0.60	0.85	0.66
Ordered logit in-sample	1.44	0.28	0.64	0.83	0.67

removed from the model because of a too small t-value. Likewise the ordered-logit algorithm removed the 'Net profit margin' variable. In view of our earlier comments on this variable with respect to the in-sample models this comes as no surprise.

Classifications per rating class

The deviations per class are shown in figure 5-12. The linear regression and ordered logit model seem to classify slightly better than the SOM model. The middle classes (9 through 17) are classified fairly well in all

three models, larger errors seem to be somewhat magnified by the SOM. We again notice striking similarities between linear regression and ordered logit.
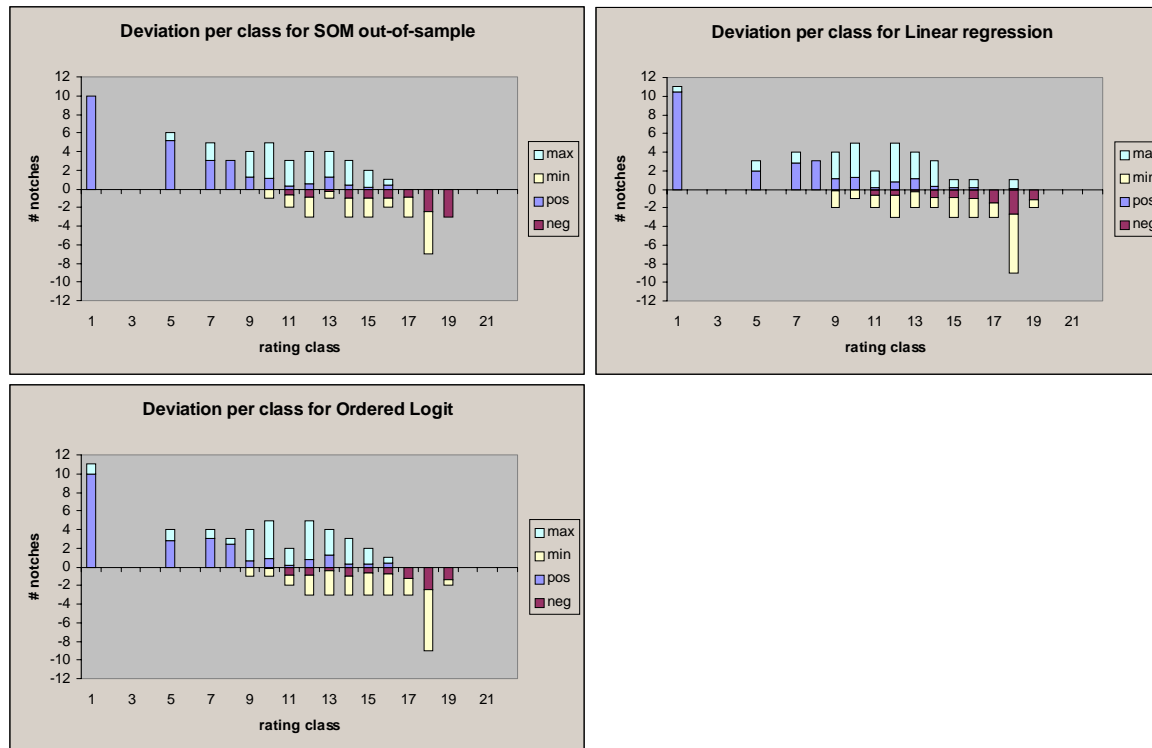
Figure 5-12 Deviations per class for SOM, linear regression and ordered logit out-of-sample

5.5.2 Results for older historical periods

To test the stability over time of the found results we use the SOM model to classify companies in older historical periods. The model is not retrained, as we want to observe the applicability of the current model to older data. If the contribution of the variables remains relatively stable over the years then the performances should not change much. The results are shown in table 5-12.

The performances for 1995 and 1996 are comparable to the current performances (1997 and 1998). The classification performances deteriorate for 1994, the

Table 5-12 Performance comparison

Model	MAD	Success ratio			R²
		0	1	2	
1996	1.46	0.25	0.65	0.86	0.66
1995	1.56	0.23	0.59	0.82	0.67
1994	1.78	0.24	0.54	0.76	0.53

environment has apparently too much changed for the model to still be accurate. This indicates that we can not keep the model constant. We should periodically retrain the map to be sure that the SOM incorporates the latest insights and valuations of the companies in specific financial situations.

5.5.3 Linking spreads

Up until now we have mainly looked at the rating of a company as a whole. In real-life situations we always want to know the risk involved when buying a bond from the company. Next to the rating this risk is expressed in the spread of the credit, which is (roughly said) the difference between its yield and the yield of a comparable government bond. A government bond has the least chance of default so it is regarded as the benchmark for other bonds. Lower rated bonds are more riskful to invest in, so the yield on the bond should be higher than higher rated bonds that are less riskful to invest in. When 'the market' regards a specific bond as more riskful than before, the spread on the bond widens. If the market regards a specific bond as less riskful than before, the spread on the bond narrows[44].

When new information regarding a company becomes available, rating agencies are sometimes slow to update the rating while the market has already processed this new information. If our model does not experience as much lag as the rating agency, then we should see a higher than average spread when our model rates the bond lower than S&P. Vice versa, if our model assigns a higher rating to a bond then the spread should be smaller than average.

The following analysis is a first attempt to model this relationship. No definite conclusions should be drawn from these results. We should also keep in mind that the market is not always correct. It is possible that rating agencies uncover previously unknown information during their qualitative analysis and that the spread (the market) reacts upon the resulting rating change.

Data

We use bond data from Lehman Brothers, a well known broker and dataprovider, selected from their universe of bond indices. The senior unsecured bonds are all chosen from the Consumer Cyclicals sector, and from all possible rating classes. The bonds are linked to our data using the CUSIP code, a code containing a general part identifying companies and a specific part identifying individual bonds.

Two problems immediately arise regarding the linking of individual bonds to companies:

---

[44] More information on the valuation of bonds can be found in chapter 2.

1. Lehman Brothers uses another definition for the Consumer Cyclicals sector. Some of the companies belonging to Consumer Cyclicals according to our definition are not included in the LB universe, and some of the companies in the LB Consumer Cyclicals sector belong to other sectors in our universe.

2. For a lot of bonds the company has merged with other companies or has otherwise disappeared, while the bond still exists for the old company. The CUSIP codes for the bond and the company do not match, even when the same company is involved.

This results in 50% of the bonds not being matched to earlier downloaded company data. It would probably be possible to reduce this figure, but at a relatively large time-expense.

Table 5-13 Mapping of LB bonds over sectors

	Original sector
33.5%	Consumer Cyclicals
9.3%	Consumer Staples
5.1%	Financials
2.3%	Capital Goods
0.4%	Technology
49.4%	Unmatched

The bonds are distributed over the original sectors as shown in table 5-13. As most bonds reside in the correct or an almost equivalent sector we are confident that the results are still representative.

We downloaded the following characteristics for each bond.

- Time to maturity: The remaining time until the bond expires

- Option adjusted spread: The spread of the bond, adjusted for specific bond types like callables. The spreads were downloaded for the last day of the fourth quarter in 1998.

- S&P rating

Pre-processing

We have grouped the bonds according to maturity into buckets of 1 to 5 years and 5 to 10 years. The outer rating classes (1 to 7 and 18 to 22) have been removed, because the predictions for these classes are severely biased. Within the buckets we have standardized[45] the spreads per rating class, thus making a comparison over the classes possible. These standardized spreads are compared with the deviation of the predicted from the real rating.

If the predicted rating is a better measure for the risk perceived by the market, then the spread should be proportional to the deviation of the predicted rating from the real rating. A lower than average spread, or negative standardized spread, should be accompanied by a higher predicted rating (a positive rating deviation). A higher than average spread, or positive standardized spread, should be accompanied by a lower predicted

---

[45] Standardization means subtracting the average of the spreads in the rating class from the current spread and dividing by the standard deviation of the spreads in the rating class.

rating (a negative rating deviation). An average spread, or a zero standardized spread, should be accompanied by an equal predicted rating (no rating deviation).

In a scatterplot the data would be distributed as a line from the upper left quadrant to the lower right quadrant, with a possible concentration of datapoints around zero.

Results

The scatter plots for the two maturity buckets are shown in figure 5-13. The sought-after relationship does not show in the figures, the points seem to be randomly scattered in the plot. A negative or positive standardized spread is as much correctly predicted as it is not.
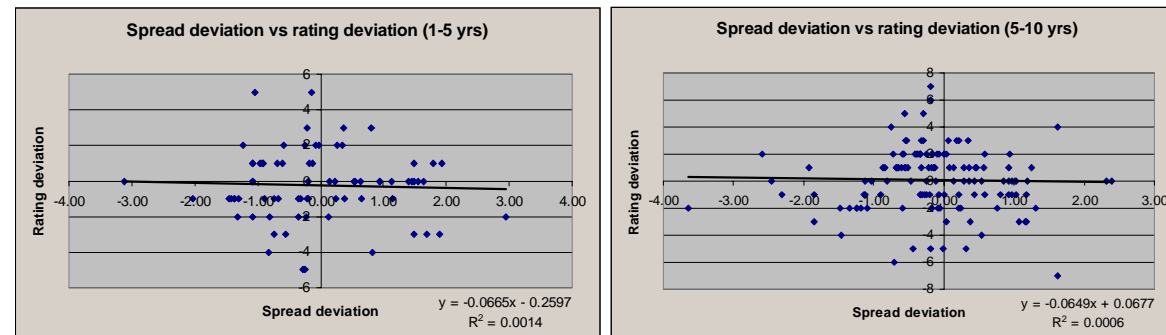
Figure 5-13 Standardized spreads vs rating deviations

5.6 Summary

In this chapter we have constructed a classification model, using the same data as in our descriptive analysis of chapter 4. Before arriving at the final model, several variations have been tested. Each variation is evaluated using three different performance measures; the success ratio, the mean absolute deviation and the R^2.

First we created our initial model. We improved upon this model by removing variables and adjusting parameters, after each change testing the effect on model performance. In the final model the original 18 variables have been reduced to 5 variables without sacrificing too much classification performance.

To validate the model we have compared it with a constant predicting model and with 100 random predicting models. The comparison with our benchmark models, linear regression and ordered logit, provide another way to test the validity of the model. The in-sample and out-of-sample tests show comparable results for all the models. The selected variables are similar, and so are the classification results.

6 conclusions

In this chapter we draw our conclusions. The central question from chapter 1 is revisited, and the answers to the sub-questions (given in the previous chapters) are summarized. Finally some directions for further research are given.

6.1 Conclusions

In this thesis we tried to answer the following central question:

> In what way can we use Self-Organizing Maps to explore the relationship between financial statement data and credit ratings?

We have broken down this question into five sub-questions, each answered in separate chapters of this thesis.

1. What are credit ratings and how is the credit rating process structured?

Chapter 2 provided a theoretical background on bonds, credits and credit ratings. We have seen that the credit rating is basically an opinion on creditworthiness of the credit issuer, often a company or government. The number of defaults in each rating class shows that the credit rating is a good measure of creditworthiness.

The credit rating is determined by two main factors: A quantitative analysis of the balance sheet and income account, and a qualitative analysis concerning the management of the company, the economic expectations of the sector and other non-quantitative elements that could affect creditworthiness. A review of the Standard & Poor's credit rating process shows that rating agencies put much more weight on the qualitative analysis than on the quantitative analysis.

2. What are Self-Organizing Maps and how can they aid in exploring relationships in large data sets?

In chapter 3 we showed that Self-Organizing Maps are an innovative way to visualize the used data set. The observations in the input space are projected on a surface (a neural network) that can stretch and bend to better accommodate the distribution of the data in the input space. This projection is then visualized as a two-dimensional map, surrounded by components representing the used variables. Additionally, the observations are clustered according to likeliness of the underlying variables.

The full SOM display contributes to a better understanding of the underlying domain. Relationships between variables, linear and non-linear, are clearly visible. The clustering of the mapped observations provides an insight into the likeliness of the observations. The SOM thus functions as a descriptive tool. The SOM can also

be used as a prediction model. The found clustering is then of less importance; we use the stretched and bended map surface as a form of non-linear regression.

3. Is it possible to find a logical clustering of companies, based on the financial statements of these companies?

4. If such a clustering is found, does this clustering coincide with levels of creditworthiness of the companies in a cluster?

These two questions were answered in the descriptive analysis of chapter 4. Using a selection of financial ratios we created a Self-Organizing Map display of the US companies in sector Consumer Cyclicals. Qualitative information was not taken into account when creating this SOM. The resulting display showed that a clustering of companies based on financial ratios is very well possible. The found segmentation grouped companies in an intuitively logical manner. Furthermore, when we compared the clustering with actual credit rating levels we found a strong relation. Approximately 60 % of the variance in the ratings was matched by the found clustering. If we assume the used variables and the model to be correct, then we might attribute the 40% residual variance to be due to the qualitative analysis performed by S&P.

5. Is it possible to classify companies in rating classes using only financial statement data?

In chapter 5 we used the results of our descriptive analysis to construct a classification model. The final results show that it is possible to predict credit ratings, but only to a certain extent. About 80% of the companies in the sample are classified with an error of at most two notches. Once again it is most likely that this is due to the qualitative analysis, which can not be duplicated by a quantitative model.

Even if the model is not suitable for classifying companies exactly correct, the model still gives an improved insight into the credit rating process. The qualitative analysis is less important than S&P might lead us to believe and the most important variables determining creditworthiness are size and interest coverage. Furthermore, the classification of the model can be used as a first approximation when the real credit rating is currently not available. The stability of the performances when comparing different techniques show that we have found a stable model for this sector, in which the selected variables are most important.

6.2 Further research

Of course much research still remains to be done on the domain of credit ratings. We highlight some of the more obvious directions:

The research in this thesis was performed on a single sector of US companies. It is interesting to see if a classification model using only financial ratios performs better in other sectors. A better model performance could indicate that qualitative factors are less determining for creditworthiness. The chosen financial ratios are likely to vary for different sectors. A comparison between the selected ratios would be insightful.

Another interesting venue of research is the change in credit ratings. While the model has some flaws when predicting current ratings, it might be a better predictor for rating changes. If the predicted rating actually precedes a change in rating then the model is more practical usable.

Related to this is the comparison of the rating with the spread. We have briefly touched the subject in chapter 5, but it would be beneficial to devote more time to studying the relationship between spreads and predicted ratings.

On a more computer science related note, it would be interesting to further explore and explain the use of semi-supervised learning (using the ratings as a train variable) with the Self-Organizing Map. A comparison with normal supervised learning would give an insight into the special characteristics of semi-supervised learning and effects on model performance.

7 bibliography

Bishop, C.M., 1995. "Neural networks for Pattern Recognition", New York, United States of America, Oxford University Press Inc. .

Brealey, R.A. and Myers, S.C., 1991. "Principles of corporate finance", Fourth edition, United States of America, McGraw-Hill Inc. .

Cantor, R. and Packer, F. "The credit rating industry", FRBNY Quarterly Review / Summer-Fall 1994

Deboeck, G., 1998. "Visual explorations in finance with Self-Organizing Maps", London, Great Britain, Springer-Verlag.

Eudaptics, 1999. "Viscovery SOMine 3.0 User's Manual", www.eudaptics.com.

Fabozzi, F.J., 1993. "Fixed Income Mathematics", Revised edition, Chicago, Illinois, United States of America, Probus Publishing Company.

Fayyad, U.M., Piatetsky-Shapiro, G. , Smyth, P. and Uthurusamy, R., 1996. "Advances in knowledge discovery and data mining", Menlo Park, California, United States of America, American Association for Artificial Intelligence Press.

Fok, D., 1999. "Risk profile analysis of Rabobank investors", masters thesis, Erasmus University of Rotterdam.

Geluk, I. en Van der Hart, J., 1996. "Cursus econometrie in de praktijk", Rotterdam, Robeco Group.

Greene, W.H., 1997. "Econometric Analysis", Third edition, Upper Saddle River, New Jersey, United States of America, Prentice-Hall.

Johnson, R.A. and Wichern, D.W., 1992. "Applied Multivariate Statistical Analysis", Third edition, Englewood Cliffs, New Jersey, United States of America, Prentice-Hall.

Kaski, S., 1997. "Data exploration using self-organizing maps", www.cis.hut.fi/~sami/thesis/thesis_tohtml.html.

Kohonen, T., 1997. "Self-Organizing Maps", Second Edition, Heidelberg, Germany, Springer-Verlag.

Moody's, 2000. "Historical default rates of corporate bond issuers, 1920-1999", www.moodys.com.

Standard & Poor's, 2000. "Corporate ratings criteria", www.standardandpoors.com.

appendix

I   Artificial neural networks

Overview

Artificial neural networks are conceptual models of networks consisting of small simple processing elements called neurons. These neurons are inspired by real neurons (figure A-1) and the way they function in the human brain: The neurons are interconnected (hence the term network), the output of a neuron serves as one of the inputs for another neuron. The output of each neuron is based on its inputs using some kind of mathematical function, all neurons operate in parallel.

Neural networks originated in artificial intelligence research, which in the sixties and seventies produced 'expert systems' that somehow failed to capture certain key elements of human intelligence. These expert systems were based on a model of the high-level reasoning processes. Thus some of the research focused on mimicking the lower level structure in the brain, hoping that this would yield better results.
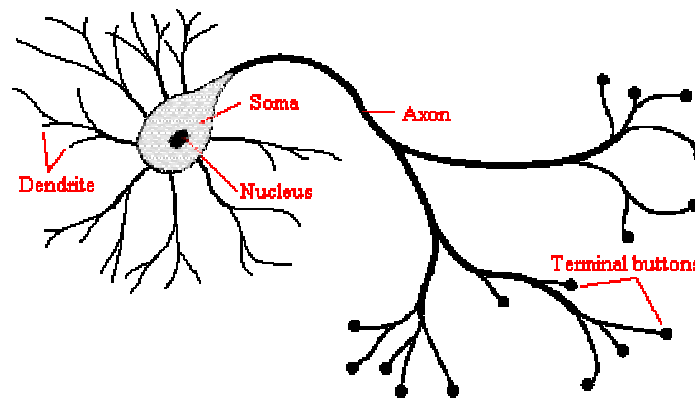
Figure A-1 Schematic of biological neuron

Of course the human brain is much more complex than the relatively simple mathematical model we use. One obvious distinction is the size; our neural networks (the word 'artificial' is often omitted) use hundreds or maybe thousands of neurons, whereas the number of neurons the brain contains lies in the order of 10^{11}. Furthermore there are a number of other processes that greatly affect the way electrochemical pulses propagate through the neurons in the brain. Therefore neural networks should not be regarded as accurate representations of (parts of) the brain but more as a general model of the behavioural processes in the brain.

Use of neural networks

Neural networks are most often used in applications for which an algorithmic solution can not easily be formulated, but a large number of observations of the required behaviour is available. Furthermore neural networks are often used when one wants to find a beforehand unknown, non-linear structure in the data.

- Statisticians use neural networks as non-linear regression and classification models.

- Engineers use neural networks for signal processing and automatic control.

- Cognitive scientists view neural networks as a way to model thinking and consciousness (higher brain functions).

- Neuro-physiologists use neural networks to model sensory systems, memory and motorics (medium level brain functions).

- Biologists use neural networks for the interpretation of nucleotide sequences.

II   Iterations of the SOM algorithm

This simple example shows the adjustments made to the neural network in each iteration of the self-organization process. Our model consists of a three-neuron (a, b and c) network that will be fitted to five observations in the two-dimensional input space.

The network is random initialized as is shown in figure A-2. We subsequently iterate over all the observations to adjust the network to better fit these observations.
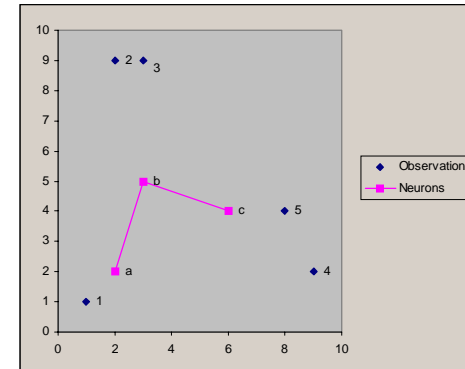
Figure A-2 A three-neuron network randomly initialized in the two-dimensional input space

Iteration 1

We first identify the winning neuron that most closely resembles the first observation. This is of course neuron a. The placement of this neuron in the input space is adjusted to match observation 1 (figure A-3). We now say that observation 1 is mapped to neuron a.
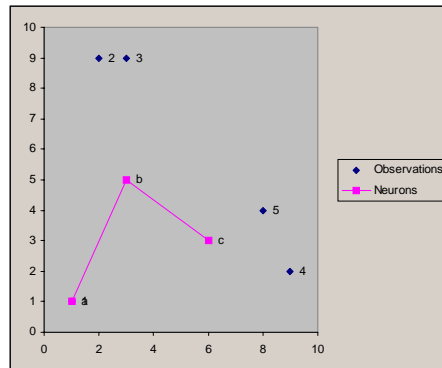
Figure A-3 Adjusting the winning neuron in iteration 1
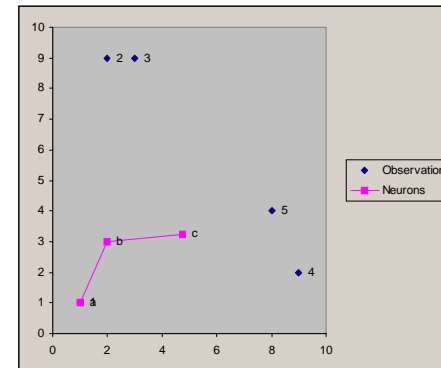
Figure A-4 Adjusting the neighbourhood in iteration 1

Besides the winning neuron the neighbours of the winning neuron are also adjusted, but to a lesser degree depending on the neighbourhood function. This is shown in figure A-4.

Iteration 2 through 5

Figures A-5 through A-8 show how the neurons are updated to match each of the observations . The learning rate factor reduces the adjustments for the later iterations.
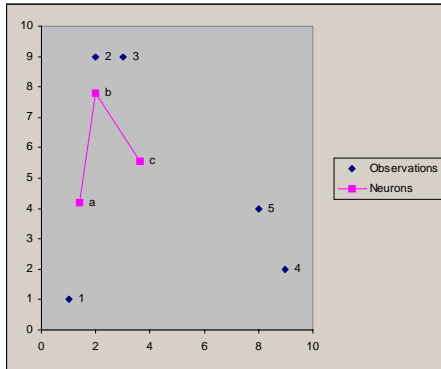
Figure A-5 Iteration 2, winning neuron = b
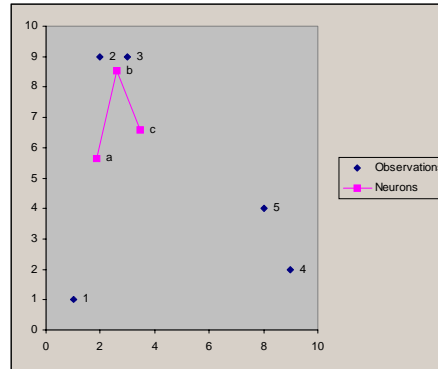
Figure A-6 Iteration 3, winning neuron = b
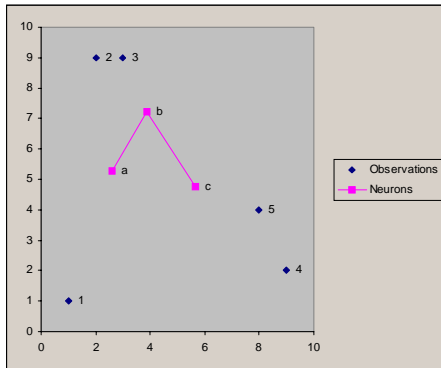
Figure A-7 Iteration 4, winning neuron = c
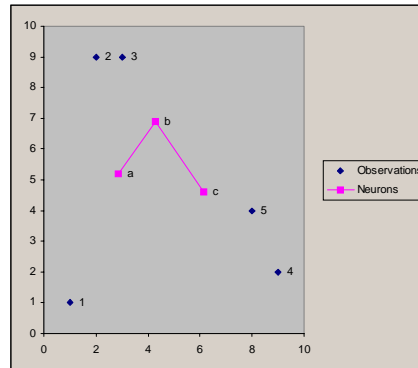
Figure A-8 Iteration 5, winning neuron = c

Output map

The final map after the adjustments in iteration 5 is shown in Figure A-8. The associated output map is shown in figure A-9. This output map is a representation of the neural network in the input space. The neurons and their associated observations are displayed, but the absolute distance information is lost. This is later reintroduced by colour coding the map.
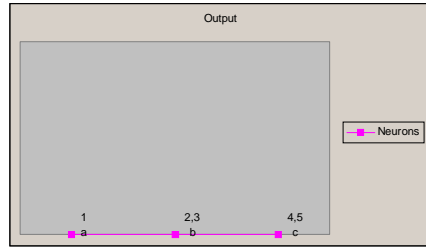
Figure A-9 Output map after the self-organization process

III   SOM example: Rectal muscle sizes

In this example the SOM is used as a descriptive tool in the medical domain. In medical research the sample size is often necessarily small, so statistical inference is more difficult. We show how the SOM can visually aid in providing a good understanding of the data at hand.

Data

A Self-Organizing Map is used to display the characteristics of persons for whom rectal muscle sizes were measured using ultrasound images. The scans were performed at the Academic Hospital of Maastricht by Dr. Regina Beets-Tan for her PhD research.

The sample consists of a group of 60 test subjects, 46 females and 14 males. The age varies from 19 to 72, and some of the women have given birth while others have not. The test subjects are chosen in such a way that no bias towards age or number of births is present in the sample.

For each test subject the following five muscles in the rectal area were measured:

-   Internal sphincter muscle

-   Longitudinal muscle

-   External sphincter muscle

-   Total sphincter thickness

-   Perineal body

For each test subject the following additional information was recorded:

-   Sex

-   Age

-   Number of births (Partus x or Px)

-   Weight

-   Length

Training

The SOM is created using the five muscle sizes as train variables. The additional information is not used to train the map. We can now answer the following two questions:

1. Can clusters of persons be distinguished by their rectal muscle sizes?

2. Do these clusters coincide with groupings based on sex, age, number of births, weight or length?

The Self-Organizing Map is displayed in Figure A-10. The five train variables are displayed at the bottom (Internal sphincter, Longitudinal muscle, External sphincter, Total sphincter thickness, Perineal body). The clusters and independent variables are displayed above (in the order Sex, Age, Partus x, Weight and Length) .
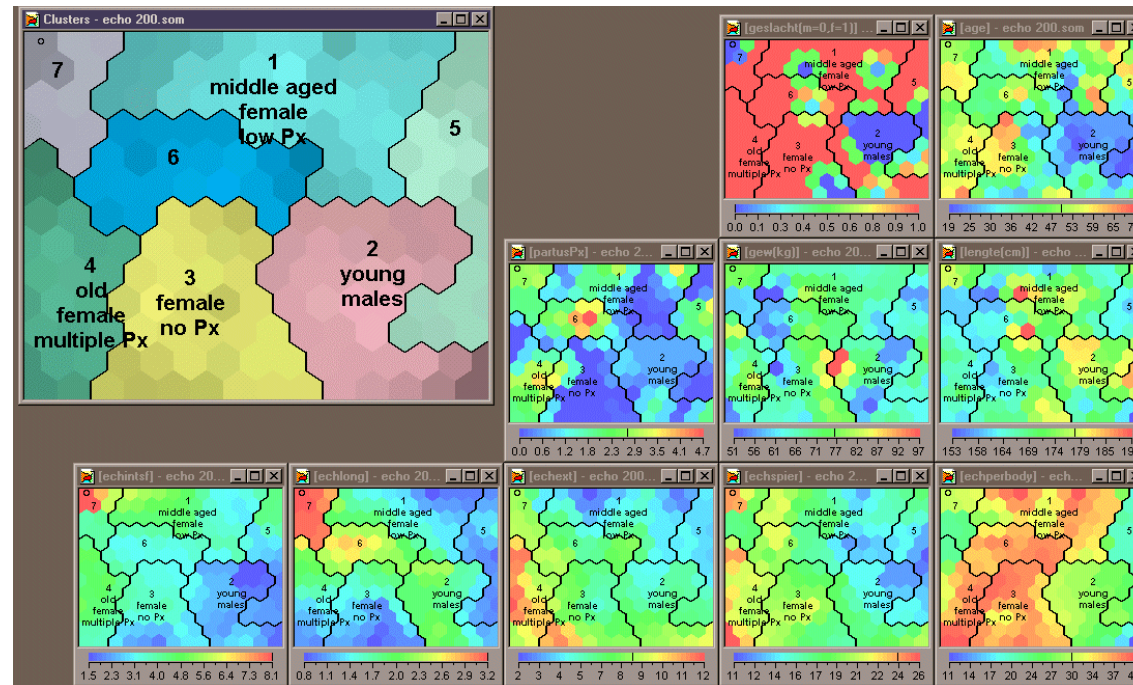
Figure A-10 SOM of rectal muscle sizes

The SOM has formed 7 clusters, purely based on rectal muscle sizes. More important than the exact boundaries of the clusters are the overall relationships we can infer from this display.

Inferred relationships

Looking at the Sex component we clearly see most of the males grouped together. This means that based on measurements of the rectal muscles we should be able to distinguish a man from a woman. Furthermore, when comparing the Age component with the Sex component it becomes clear that mostly younger males participated. A third fact about the male test subjects is captured in the Perineal body component, the values are low for all clusters containing males, and high for almost all females. The Perineal body can not be measured for males, hence the difference. In a more extensive analysis we would have to correct for this difference.

The Partus x component reveals a clustering of women not having given any births at all. These women are characterized by a small internal sphincter and a small longitudinal muscle. The relative random pattern for age, weight and size in this cluster confirms that this relationship holds for all women in the sample, not just the young or small ones. As the weight and size components show a definite resemblance, we know that they are correlated. This is what we would expect, taller people are heavier than short people.

Conclusions

For medical appliances the Self-Organizing Map can serve as a valuable tool to enhance the understanding of the underlying domain. The dataset is accurately represented, the relatively small sample size does not form an insurmountable problem. However, the inferred relationships have to be used with care.

IV   SOM example: Customer segmentation

The Self-Organizing Map is especially suitable for use in the marketing domain. Large data sets containing (often non-linear) customer data are more and more common for corporations of all sizes. Finding relationships in these databases and using them to optimize the relationship with the customers is known as Customer Relationship Management.

In this example we will show how customers of the Rabobank can be grouped according to their investment preferences, as expressed in a short survey. We then compare these expressed preferences with some real characteristics of their investment portfolios over the previous year.

Data

The sample consists of 1000 investing customers of the Rabobank. Each customer has been asked to fill in a survey, consisting of 24 questions. For each question five answers are possible, ranging from "Fully disagree", "Disagree", "Disagree nor agree", "Agree", to "Fully agree". The full questionnaire can be found at the end of this chapter.

Next to these questions we also recorded the number of transactions, the use of the Internet or the "Rabo Orderlijn" (direct telephone contact with a Rabobank broker), the age, the total size of investments, and the length of the relationship between the customer and the Rabobank.

The use of the Internet or the Rabo Orderlijn is used by the Rabobank marketing department as a dependence variable: Clients are considered independent if they have used the Internet or the Rabo Orderlijn at least once to make a transaction. The 'size of investments' and 'number of transactions' variables have been log transformed, to equalize the sometimes large differences in total invested assets or number of transactions.

Training

The SOM is created using the answers to the 24 questions as train variables. The additional information was not used to train the map. We can now answer the following questions:

1.   Is it possible to make clusters of customers (a customer segmentation) with different investment profiles based on answers from the survey?

2.   Does the observed behaviour of customers (number of transactions and independence) coincide with these investment profiles?

The SOM is displayed in figure A-11. None of the train variables are displayed, the components that are displayed are the Log size of investments, Log number of transactions, Relationship length, Age and Independence.
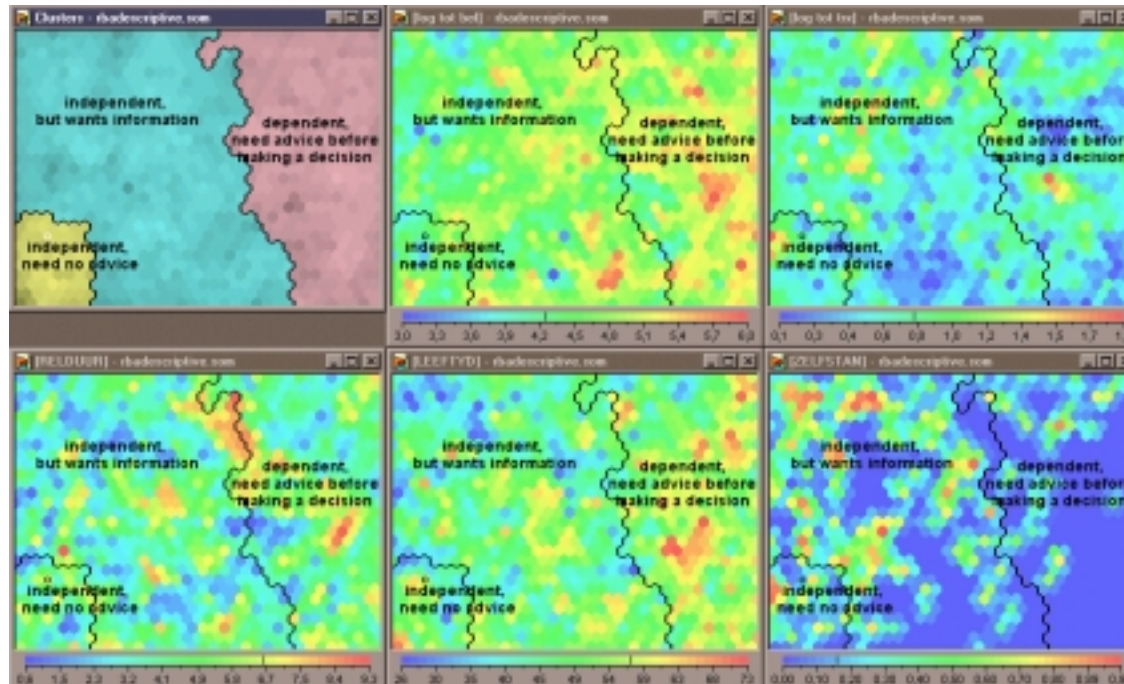
Figure A-11 SOM of Rabobank customers survey

Inferred relationships

If we only take the answers given at the survey into account the customers can be segmented into three distinct groups. The most independent customers are located in the bottom left corner of the map. They say to have an active investment style, and they do not need advice from the bank before making a decision. The somewhat less independent customers are situated in the upper left and the middle of the map. Although they clearly want to make their own decisions, they still seem to benefit from consultation with the bank. The most dependent group can be found in the right portion of the map. Before making any investment decision they would like to receive advice from their financial advisor at the bank. They also want to be kept up-to-date on their portfolio and on the current events of the market.

The independence variable fits reasonably well to the formed clusters. The customers with a dependent investment style have almost never used the Internet or Rabo Orderlijn, as was to be expected. For the somewhat less independent customers we see that some have used the Internet and some have not. For the independent customers we would expect this to be somewhat higher, there are apparently some customers who say to act independently but never really do (when not using the Internet or Rabo Orderlijn an advisor always comes into play).

The size variable shows that the richest customers are in general dependent. Logically they are also the older customers, as can be seen from the age component plane. The random distribution of the number of transactions component reveals that no direct relation with the found investment profiles can be made. Independent investors do not necessarily perform more transactions. The relationship length is also randomly distributed over the map, but shows at some points a correlation with the independence variable: Customers that have had longer relationships with the Rabobank have never used the Internet or the Rabo Orderlijn.

Conclusions

The SOM gives an attractive overview of the Rabobank customer sample. Based on the answers from the survey, the customers can roughly be divided into three groups, each having distinct investment preferences. Although easy to measure, we can unfortunately not use the number of transactions to determine the group membership of a customer. We can be relatively sure that dependent investors do not use the Internet or the Rabo Orderlijn.

Survey questions

1. I actively manage my investments.

2. I need a financial advisor to make the best decisions.

3. I'm willing to take risks with my investments.

4. I want to be kept informed about new savings and investment opportunities.

5. There are so many investment possibilities that it is hard to keep track of them all.

6. You have to be knowledgeable before starting to invest.

7. Investing is fun.

8. It is feasible to invest on a short-term basis.

9. I feel the need to consult with a specialist about my asset management.

10. Most of the times I follow the advice of my consultant.

11. I independently make investment decisions based on acquired information.

12. I use my financial advisor at the bank as a sparring partner.

13. I expect my bank to provide regular news updates or investment recommendations.

14. After something important happens on the exchanges that concerns me I want my financial advisor to immediately contact me.

15. In times of large fluctuations of the exchange index I like to receive information from the bank commenting on the situation.

16. I want to annually review my investment portfolio with my financial advisor.

17. I'm a long-term investor.

18. I want to regularly receive buy or sell recommendations.

19. I want to consult one and the same advisor for al my investment decisions.

20. I'm a valued investment customer for my bank.

21. Every month I want to receive several updates on the yield of my investments.

22. I want to receive an annual report about the developments in my portfolio.

23. I'm advising my friends to start investing with the Rabobank.

24. If another bank approaches me with a tempting offer, I will consider transferring my assets.

V   Statistical measures and tests

Median standard deviation

The median standard deviation is calculated by first taking the median of the distances to the median (analogous to taking the mean of the distances to the mean for the standard deviation). This measure then needs to be rescaled to a measure comparable with the normal standard deviation.
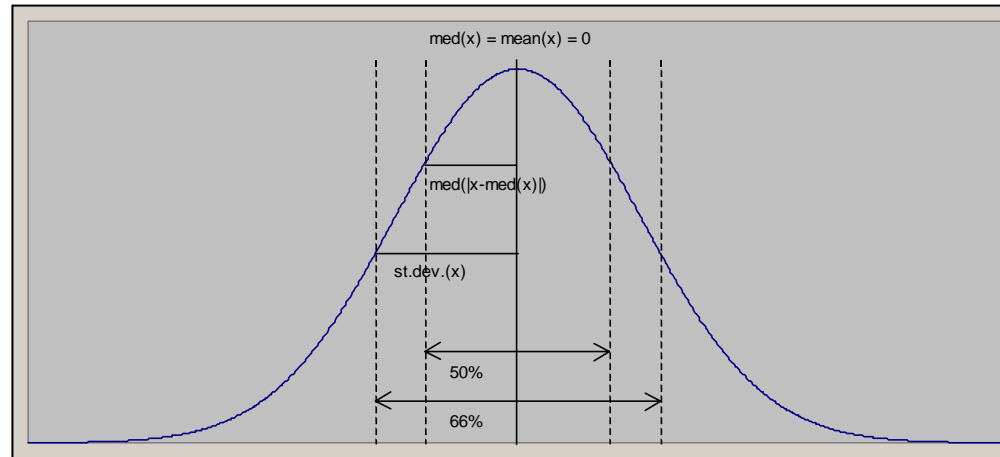
Figure A-12 The median standard deviation in relation to the standard deviation

On a normal distributed variable the area within plus or minus one standard deviation encompasses 2/3 of the distribution. The boundary of the area encompassing 1/2 of the distribution lies at plus or minus one median (according to the definition of the median). The ratio of this median to the standard deviation (on a normal distributed variable) is 0.6745. To convert the previously found measure to one comparable with a standard deviation we multiply it with 1/0.6745. Thus

$$\sigma_m = 1.483 \cdot med(|x - med(x)|)$$

Skewness

This measures the asymmetry of a distribution. It is defined as

$$S = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^3}{\sigma^3}$$

where T is the number of observations in the sample. For symmetric distributions the skewness is 0.

Kurtosis

The kurtosis measures the thickness of the tails of the distribution. It is defined as

$$K = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^4}{\sigma^4} \; .$$

A normal distributed variable has a kurtosis of 3. The kurtosis we calculated is the excess kurtosis, this is the kurtosis - 3. Values greater than 10 give rise to suspicions of non-normality.

Jarque-Bera

The final test on normality is the Jarque-Bera test. The statistic is given by

$$\frac{T}{6}[S^2 + \frac{1}{4}K^2]$$

where S is the skewness and K is the *excess* kurtosis. We can say a variable is normal distributed with 95% confidence when the result of the statistic ≤ 5.99 (X^2 distribution with 2 degrees of freedom).

VI   Descriptive analysis

Table A-1 Summary statistics per variable before cut-off, fourth quarter 1998 (continued on next page)

variable	EBIT interest coverage	EBITDA interest coverage	EBIT / total debt	debt ratio	debt-equity ratio 1	debt-equity ratio 2	net gearing	return on equity	return on assets
mean	5.07	6.73	0.20	0.61	0.79	0.56	1.76	-0.09	0.03
median	2.39	3.51	0.05	0.56	0.97	0.50	2.02	0.03	0.03
stdev	8.94	10.27	1.84	0.91	12.73	0.35	19.45	1.69	0.04
medstdev	2.71	3.10	0.06	0.26	1.01	0.27	1.44	0.04	0.02
minimum	-4.65	-4.38	-0.29	-10.59	-164.48	0.01	-237.55	-19.62	-0.16
maximum	86.50	98.50	31.30	6.02	75.34	2.52	148.33	8.23	0.25
count	282	282	288	229	289	268	292	286	289
#NA	12	12	6	65	5	26	2	8	5
> |3stdev|	0.014	0.011	0.00	0.01	0.01	0.01	0.01	0.01	0.01
> |3mstdev|	0.12	0.12	0.12	0.05	0.15	0.04	0.18	0.13	0.05
skewness	4.87	4.83	16.82	-7.10	-7.37	1.98	-5.60	-9.07	1.335
kurtosis	33.81	33.13	284.45	105.62	105.99	6.80	93.45	104.07	12.01
J.-Bera	4461.47	4887.81	106999.38	2851.11	35949.60	6.51	54740.28	3828.38	1.57

Table A-2 Summary statistics per variable after cut-off, fourth quarter 1998 (continued on next page)

mean	4.58	6.19	0.09	0.63	1.16	0.55	2.26	0.02	0.03
median	2.39	3.51	0.05	0.56	0.97	0.50	2.02	0.03	0.03
stdev	6.15	7.13	0.13	0.40	5.61	0.31	8.24	0.23	0.03
medstdev	2.71	3.10	0.06	0.26	1.01	0.27	1.44	0.04	0.02
minimum	-4.65	-4.38	-0.29	-1.03	-24.35	0.01	-33.86	-1.01	-0.05
maximum	26.79	31.39	0.61	2.14	26.29	1.59	37.89	1.07	0.10
count	282	282	288	229	289	268	292	286	289
#NA	12	12	6	65	5	26	2	8	5
> |3stdev|	0.04	0.04	0.03	0.03	0.04	0.03	0.04	0.05	0
> |3mstdev|	0.12	0.12	0.12	0.05	0.15	0.04	0.18	0.13	0.05
skewness	2.03	1.98	2.12	1.29	-0.17	1.16	-0.20	-0.35	0.35
kurtosis	4.48	3.93	6.21	4.91	12.89	1.78	10.61	13.61	1.33
J.-Bera	40.87	40.75	1.44	2.75	182.06	0.57	177.96	8.27	0.01

variable	op inc / sales	net profit margin	total assets	market value	cov net income	cov total assets	cov forecasts	beta	eps
mean	0.16	-0.06	4283.61	3985.04	0.74	0.33	-0.01	1.21	0.23
median	0.13	0.00	1377.87	782.85	0.78	0.25	0.02	1.16	0.24
stdev	0.15	0.67	20717.12	14354.96	7.29	0.27	0.64	0.59	1.36
medstdev	0.10	0.01	1455.80	991.75	1.13	0.20	0.02	0.51	0.48
minimum	-0.55	-10.39	69.27	0.71	-33.18	0.02	-5.17	-0.58	-16.16
maximum	0.86	0.77	257389.00	191264.00	34.74	1.25	3.67	2.93	6.32
count	268	294	294	266	294	294	220	264	274
#NA	26	0	0	28	0	0	74	30	20
>|3stdev|	0.01	0.01	0.01	0.01	0.03	0.02	0.02	0	0.01
>|3mstdev|	0.05	0.24	0.11	0.22	0.20	0.07	0.18	0.01	0.06
skewness	0.90	-13.10	11.23	9.82	-0.25	1.52	-4.28	0.15	-6.57
kurtosis	4.92	197.25	130.68	115.98	12.52	1.92	42.57	0.50	79.65
J.-Bera	0.99	1272.23	1.89E+08	1.1E+08	227.11	0.67	288.06	0.04	1716.30

mean	0.16	-0.01	2476.56	2705.75	0.74	0.33	0.02	1.21	0.29
median	0.13	0.00	1377.87	782.85	0.78	0.25	0.02	1.16	0.24
stdev	0.15	0.17	3352.74	4515.35	7.29	0.27	0.28	0.58	0.73
medstdev	0.10	0.01	1455.80	991.75	1.13	0.20	0.02	0.51	0.48
minimum	-0.38	-0.83	69.27	0.71	-33.18	0.02	-1.34	-0.36	-2.17
maximum	0.64	0.77	15935.84	20617.87	34.74	1.25	1.38	2.68	2.65
count	268	294	294	266	294	294	220	264	274
#NA	26	0	0	28	0	0	74	30	20
>|3stdev|	0.02	0.03	0.04	0.04	0.03	0.02	0.04	0	0.02
>|3mstdev|	0.04	0.24	0.11	0.22	0.20	0.07	0.18	0	0.06
skewness	0.78	-2.38	2.57	2.63	-0.25	1.52	-0.81	0.16	0.08
kurtosis	2.67	16.50	6.71	6.81	12.52	1.92	16.52	0.29	2.77
J.-Bera	0.26	9.48	47511.06	63562.72	227.11	0.67	15.85	0.021	0.85

Table A-3 Correlations matrix after cut-off

	EBIT interest coverage	EBITDA interest coverage	EBIT / total debt	debt ratio	debt-equity ratio 1	debt-equity ratio 2	net gearing	return on equity	return on assets	op inc / sales	net profit margin	total assets	market value	CoV net income	CoV total assets	CoV forecasts	beta	eps
EBIT int. cov.	1.00	0.98	0.92	-0.34	-0.05	-0.40	-0.01	0.21	0.67	0.15	0.18	0.25	0.41	-0.08	-0.20	0.01	0.10	0.48
EBITDA int. cov.		1.00	0.88	-0.38	-0.06	-0.44	-0.02	0.19	0.60	0.14	0.16	0.28	0.43	-0.07	-0.19	0.02	0.10	0.44
EBIT / total debt			1.00	-0.32	-0.05	-0.34	0.00	0.26	0.76	0.16	0.22	0.18	0.42	-0.06	-0.19	0.05	0.10	0.49
debt ratio				1.00	-0.10	0.81	-0.15	-0.07	-0.09	0.07	-0.14	-0.18	-0.23	-0.11	-0.06	-0.04	-0.20	-0.30
debt-equity ratio 1					1.00	-0.02	0.97	0.15	-0.05	0.00	0.11	0.01	-0.05	0.01	0.05	0.12	0.09	0.01
debt-equity ratio 3						1.00	-0.06	-0.15	0.02	0.12	-0.11	-0.27	-0.32	-0.02	0.03	0.03	-0.23	-0.14
net gearing							1.00	0.14	-0.02	0.01	0.12	0.07	0.01	0.02	0.02	0.18	0.10	0.09
return on equity								1.00	0.28	0.07	0.15	0.08	0.06	0.00	-0.06	-0.08	0.19	0.21
return on assets									1.00	0.38	0.39	0.09	0.25	-0.03	-0.20	0.10	0.05	0.68
op inc / sales										1.00	0.53	0.14	0.09	0.00	0.09	-0.05	0.06	0.31
net profit margin											1.00	0.05	0.02	0.04	0.01	0.21	0.08	0.41
total assets												1.00	0.78	0.06	-0.08	0.00	0.00	0.21
market value													1.00	-0.01	-0.06	-0.02	0.06	0.23
CoV net income														1.00	0.14	0.19	-0.03	0.04
CoV total assets															1.00	0.01	0.18	-0.17
CoV forecasts																1.00	-0.06	0.05
beta																	1.00	0.07
eps																		1.00

Table A-4 Summary statistics per cluster for final map (continued on next page)

	C 1	C 2	C 3	C 4	C 5	C 6	C 8	C 7
Matching records	82	59	48	47	29	13	8	8
Matching records (%)	27.89	20.07	16.33	15.99	9.86	4.42	2.72	2.72
Average quant. error	0.01	0.03	0.04	0.02	0.04	0.07	0.08	0.07
SP rating								
Mean	10.64	14.72	10.54	14.68	9.15	7.55	9	9
Minimum	1	9	8	10	7	1	1	7
Maximum	18	20	14	18	13	12	14	12
Std. deviation	2.74	2.88	1.58	2.02	1.38	2.68	3.43	1.5
EBITDA interest coverage								
Mean	3.76	17.64	3.72	5.44	2.39	-1.04	6.59	0.53
Minimum	-1.28	4.99	0.89	1.29	0.01	-3.56	1.92	-4.38
Maximum	13.83	31.39	17.20	11.58	9.33	0.43	24.96	5.10
Std. deviation	3.03	8.20	2.72	2.74	1.81	1.31	7.24	3.39
debt-equity ratio 2								
Mean	0.50	0.32	0.64	0.45	1.16	0.54	0.61	0.47
Minimum	0.06	0.03	0.27	0.09	0.76	0.02	0.42	0.01
Maximum	1.33	0.80	1.56	0.87	1.59	0.98	0.96	0.71
Std. deviation	0.22	0.15	0.21	0.19	0.27	0.28	0.19	0.26
return on assets								
Mean	0.02	0.06	0.03	0.02	0.04	-0.02	0.03	-0.01
Minimum	-0.02	0.03	-0.01	0.00	-0.02	-0.05	0.01	-0.05
Maximum	0.05	0.10	0.10	0.06	0.10	-0.01	0.07	0.02
Std. deviation	0.02	0.02	0.02	0.01	0.03	0.02	0.02	0.03
op inc / sales								
Mean	0.09	0.17	0.27	0.25	0.16	-0.06	0.11	-0.02
Minimum	-0.12	0.04	0.06	0.02	0.00	-0.28	0.05	-0.38
Maximum	0.25	0.38	0.61	0.64	0.37	0.21	0.24	0.17
Std. deviation	0.06	0.07	0.13	0.19	0.09	0.12	0.06	0.17

	C 1	C 2	C 3	C 4	C 5	C 6	C 8	C 7
log total assets								
Mean	2.88	3.39	3.08	3.65	2.52	2.59	2.73	2.99
Minimum	2.12	2.40	2.16	2.60	1.84	1.87	2.19	2.41
Maximum	3.48	4.20	4.20	4.20	3.71	3.36	3.59	4.00
Std. deviation	0.33	0.43	0.39	0.33	0.40	0.41	0.47	0.54
cov net income								
Mean	1.25	0.81	1.98	0.66	0.82	-3.37	-25.92	21.14
Minimum	-9.62	-17.86	-15.77	-8.19	-8.02	-15.45	-33.18	-3.04
Maximum	17.79	5.38	19.02	4.27	13.93	2.49	-12.65	34.74
Std. deviation	3.69	2.91	4.33	1.77	5.06	4.82	8.67	15.48
cov total assets								
Mean	0.24	0.20	0.72	0.23	0.19	0.42	0.20	0.91
Minimum	0.02	0.05	0.13	0.02	0.02	0.11	0.04	0.36
Maximum	0.75	0.50	1.25	0.62	0.57	0.83	0.53	1.25
Std. deviation	0.15	0.11	0.27	0.13	0.15	0.24	0.16	0.30
cov forecasts								
Mean	0.07	0.02	0.04	0.01	0.32	-1.20	0.12	0.29
Minimum	-0.29	0.00	-0.55	-0.40	0.01	-1.34	0.01	-0.04
Maximum	1.33	0.11	0.50	0.09	1.38	-0.85	0.40	1.38
Std. deviation	0.20	0.02	0.15	0.07	0.49	0.20	0.16	0.46

appendix

Table A-5 Correlations matrix for principal components analysis

	ebit interest coverage	EBITDA interest coverage	EBIT / total debt	debt ratio	debt-equity ratio 1	debt-equity ratio 2	net gearing	return on equity	return on assets	op inc / sales	net profit margin	log total assets	log market value	cov net income	cov total assets	cov forecasts	beta	eps
EBIT int. cov.	1.00	0.98	0.93	-0.35	-0.09	-0.34	-0.04	0.24	0.78	0.23	0.13	0.26	0.44	-0.12	-0.25	0.01	-0.02	0.59
EBITDA int. cov.		1.00	0.89	-0.39	-0.10	-0.38	-0.05	0.22	0.72	0.20	0.10	0.30	0.46	-0.11	-0.24	0.03	-0.01	0.52
EBIT / total debt			1.00	-0.30	-0.09	-0.28	-0.03	0.28	0.86	0.22	0.13	0.18	0.35	-0.06	-0.27	0.02	0.01	0.62
debt ratio				1.00	0.02	0.96	-0.04	0.11	-0.07	0.11	0.02	-0.11	-0.24	-0.18	0.03	-0.09	-0.05	-0.29
debt-equity ratio 1					1.00	0.11	0.96	-0.24	-0.16	-0.12	-0.03	-0.19	-0.24	-0.04	0.03	0.16	0.14	-0.09
debt-equity ratio 2						1.00	0.07	-0.01	-0.03	0.14	0.02	-0.18	-0.29	-0.17	0.03	-0.02	-0.05	-0.24
net gearing							1.00	-0.24	-0.10	-0.12	-0.03	-0.14	-0.17	-0.03	-0.05	0.23	0.13	0.03
return on equity								1.00	0.27	0.08	0.10	0.18	0.14	-0.10	-0.03	-0.12	0.14	0.14
return on assets									1.00	0.44	0.30	0.13	0.30	-0.07	-0.29	0.03	-0.03	0.74
op inc / sales										1.00	0.65	0.14	0.26	-0.03	0.04	-0.01	0.03	0.27
net profit margin											1.00	-0.03	0.06	-0.05	-0.16	0.19	-0.02	0.26
log total assets												1.00	0.86	-0.01	-0.10	-0.08	-0.07	0.16
log market value													1.00	-0.03	-0.04	-0.10	-0.09	0.24
cov net income														1.00	0.13	0.08	-0.08	0.01
cov total assets															1.00	-0.01	0.23	-0.30
cov forecasts																1.00	-0.06	0.06
beta																	1.00	-0.07
eps																		1.00

Table A-6 Eigenvalues and variance of the principal components

Eigenvalues	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Value	5.097	2.274	2.176	1.453	1.407	1.298	0.942	0.859	0.795	0.612	0.508	0.235	0.142	0.092	0.050	0.032	0.019	0.010
% of variance	28.3	12.6	12.1	8.1	7.8	7.2	5.2	4.8	4.4	3.4	2.8	1.3	0.8	0.5	0.3	0.2	0.1	0.1
Cumulative %	28.3	41	53	61.1	68.9	76.1	81.4	86.1	90.6	94	96.8	98.1	98.9	99.4	99.7	99.8	99.9	100

Table A-7 Correlations between original variables and principal components

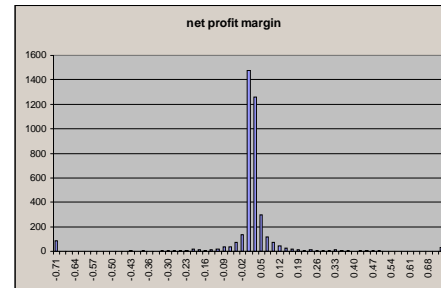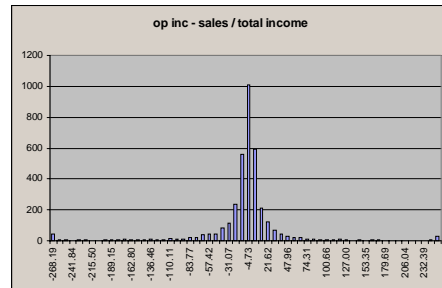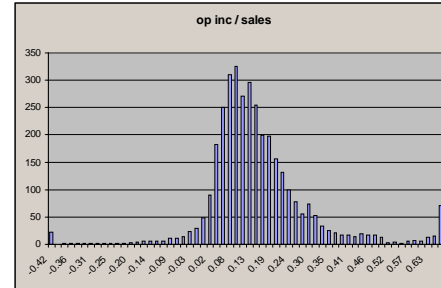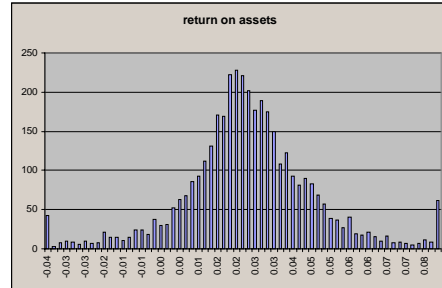| Principal components	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
EBIT int. cov.	0.92	0.15	0.05	-0.18	-0.09	0.03	0.08	-0.01	-0.16	-0.07	-0.16	-0.05	-0.10	-0.04	-0.02	0.01	0.01	0.08
EBITDA int. cov.	0.90	0.16	-0.02	-0.17	-0.06	0.02	0.08	0.02	-0.18	-0.07	-0.22	-0.05	-0.16	-0.05	-0.04	-0.02	-0.01	-0.06
EBIT / total debt	0.90	0.16	0.13	-0.17	-0.18	0.04	0.15	0.00	-0.06	-0.02	-0.10	-0.02	0.11	0.11	0.16	0.01	0.00	-0.01
debt ratio	0.84	-0.01	0.39	-0.02	-0.16	0.01	0.18	-0.06	-0.01	0.09	0.10	0.04	0.20	0.06	-0.13	-0.01	-0.01	0.00
debt-equity ratio 1	-0.17	0.83	0.28	-0.16	0.35	-0.01	0.03	-0.04	0.14	-0.09	0.04	0.02	0.02	-0.05	0.02	-0.10	-0.05	0.01
debt-equity ratio 2	-0.25	0.80	0.29	-0.19	0.34	0.04	0.03	-0.09	0.11	-0.17	-0.02	0.00	0.00	0.04	-0.03	0.11	0.04	-0.01
net gearing	-0.42	-0.32	0.72	-0.25	0.08	-0.13	0.27	0.01	-0.06	0.11	-0.02	-0.04	-0.02	-0.02	0.01	-0.06	0.09	0.00
return on equity	-0.42	-0.45	0.66	-0.29	0.08	-0.11	0.23	0.05	0.01	0.05	-0.04	-0.08	-0.04	-0.02	0.02	0.06	-0.08	0.00
return on assets	0.27	-0.09	0.50	0.63	0.16	0.12	-0.33	0.02	0.11	-0.12	-0.11	-0.30	0.02	0.02	0.00	-0.01	0.00	0.00
op inc / sales	0.43	-0.30	-0.32	-0.16	0.69	-0.17	0.03	0.12	0.13	0.12	0.04	-0.02	-0.10	0.18	-0.02	-0.02	0.00	0.01
net profit margin	0.59	-0.27	-0.32	-0.08	0.63	-0.09	0.03	0.01	-0.03	0.03	0.01	-0.05	0.16	-0.18	0.02	0.03	0.01	0.00
log total assets	-0.05	0.14	0.01	-0.19	0.05	0.81	-0.11	0.05	0.11	0.50	-0.09	-0.03	0.00	-0.01	0.00	0.00	0.00	0.00
log market value	-0.07	0.08	-0.30	0.47	-0.06	-0.02	0.66	-0.12	0.43	0.07	-0.19	-0.03	-0.02	-0.02	0.00	0.00	0.00	0.00
cov net income	0.00	0.37	0.15	0.40	0.06	-0.15	0.14	0.76	-0.21	0.15	0.01	0.07	0.00	-0.01	0.00	0.01	0.00	0.00
cov total assets	0.31	-0.36	0.09	-0.28	-0.13	0.34	-0.03	0.42	0.49	-0.37	0.06	0.07	0.00	-0.03	0.00	-0.01	0.01	0.00
cov forecasts	0.37	-0.28	0.48	0.46	0.33	0.24	-0.03	-0.23	-0.08	-0.06	-0.06	0.32	-0.06	-0.01	0.02	0.01	0.00	0.00
beta	-0.32	-0.09	-0.20	0.09	0.17	0.59	0.41	0.00	-0.40	-0.27	0.23	-0.10	0.00	0.03	0.00	-0.01	0.00	0.00
eps	0.72	0.16	0.18	0.13	-0.12	-0.08	0.05	-0.11	0.16	0.17	0.54	-0.05	-0.12	-0.05	0.03	0.02	0.01	0.00

Figure A-13 Histograms per variable after cut-off (continued on next two pages)

return on assets

op inc / sales

op inc - sales / total income

net profit margin

total assets
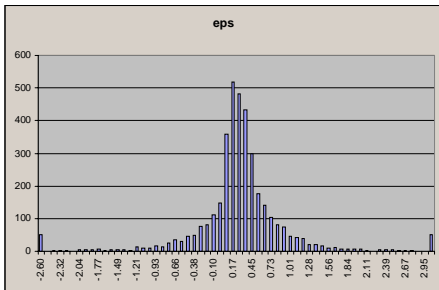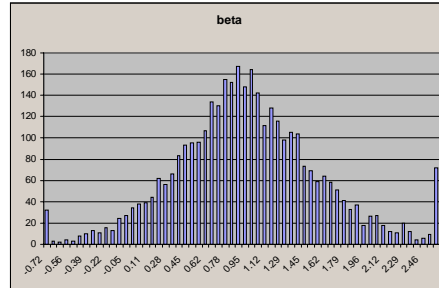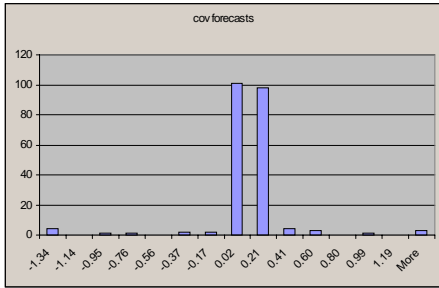
market value

cov net income

cov total assets

---
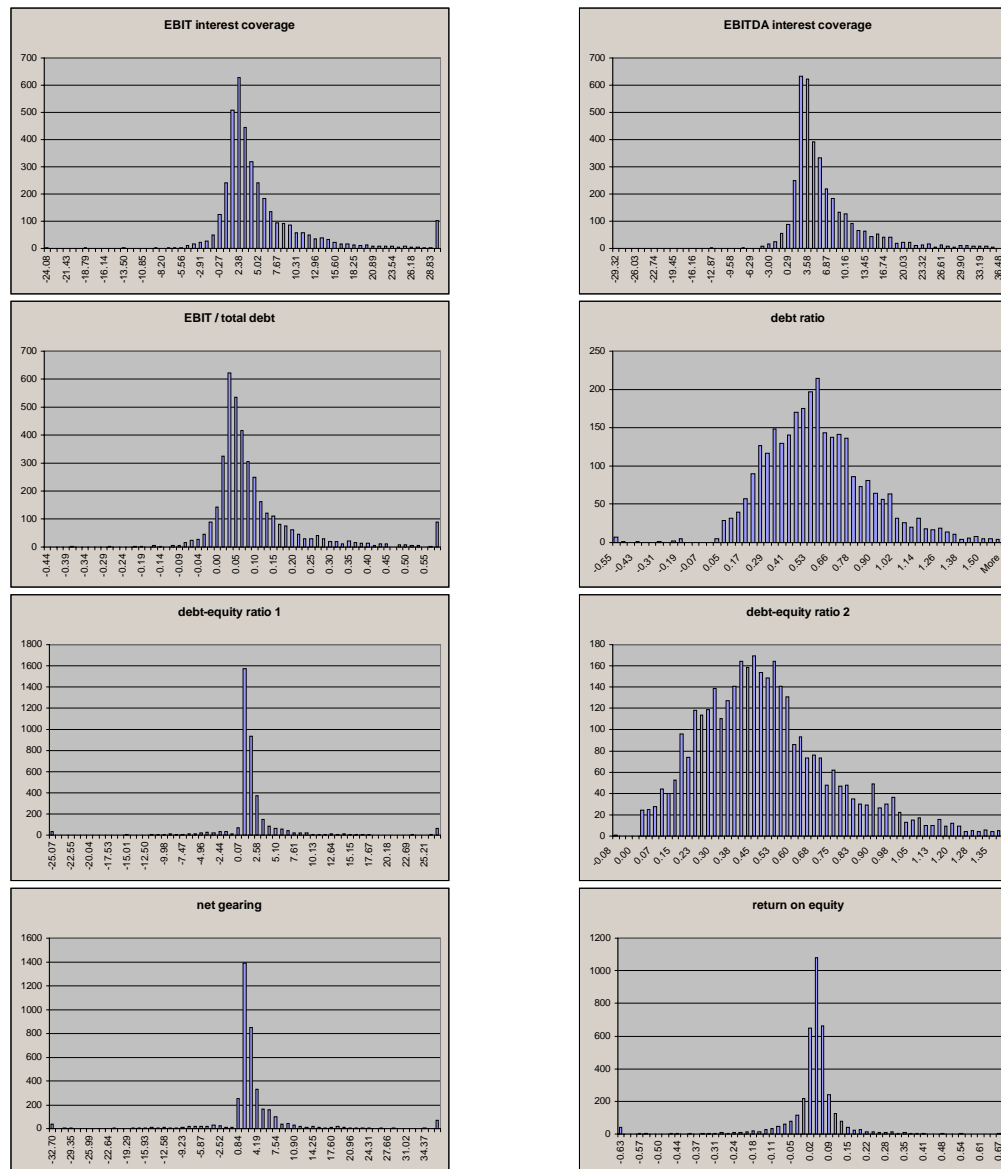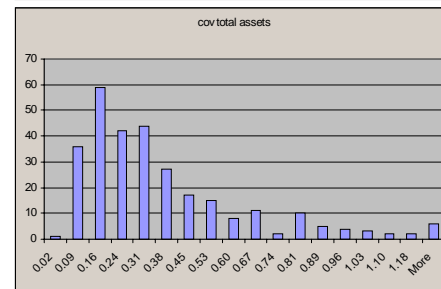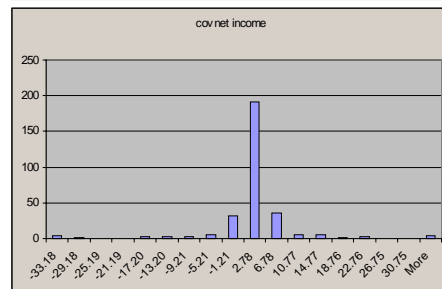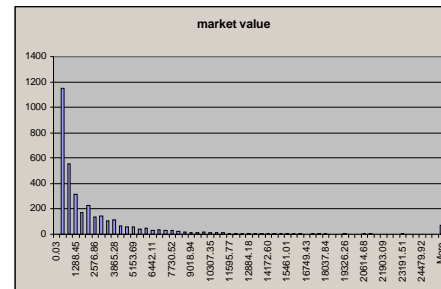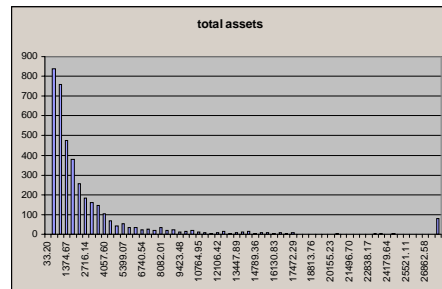
[46] The histograms for coefficient of net income, coefficient of total assets and coefficient of forecasts are slightly different because they only represent companies from the Consumer Cyclicals sector in stead of all the sectors.

cov forecasts

beta

eps
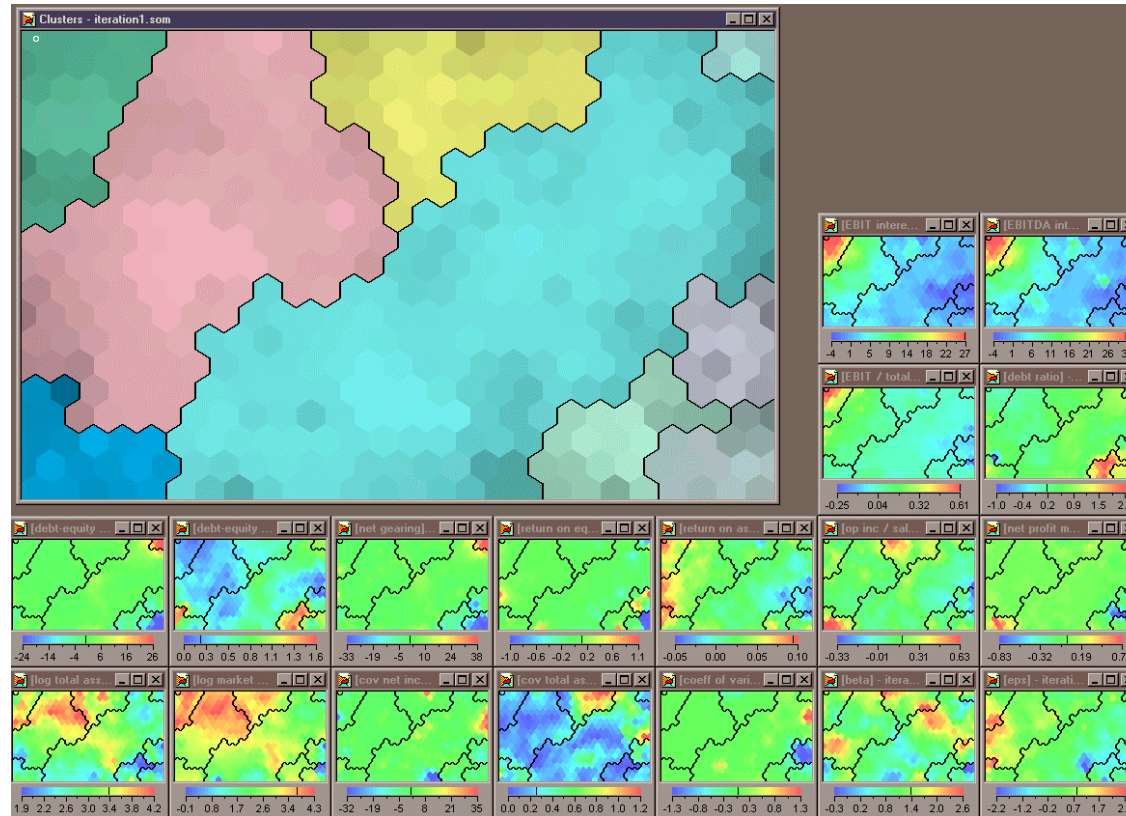
Figure A-14 Self organizing map of sector consumer cyclicals 1998 fourth quarter, iteration 1
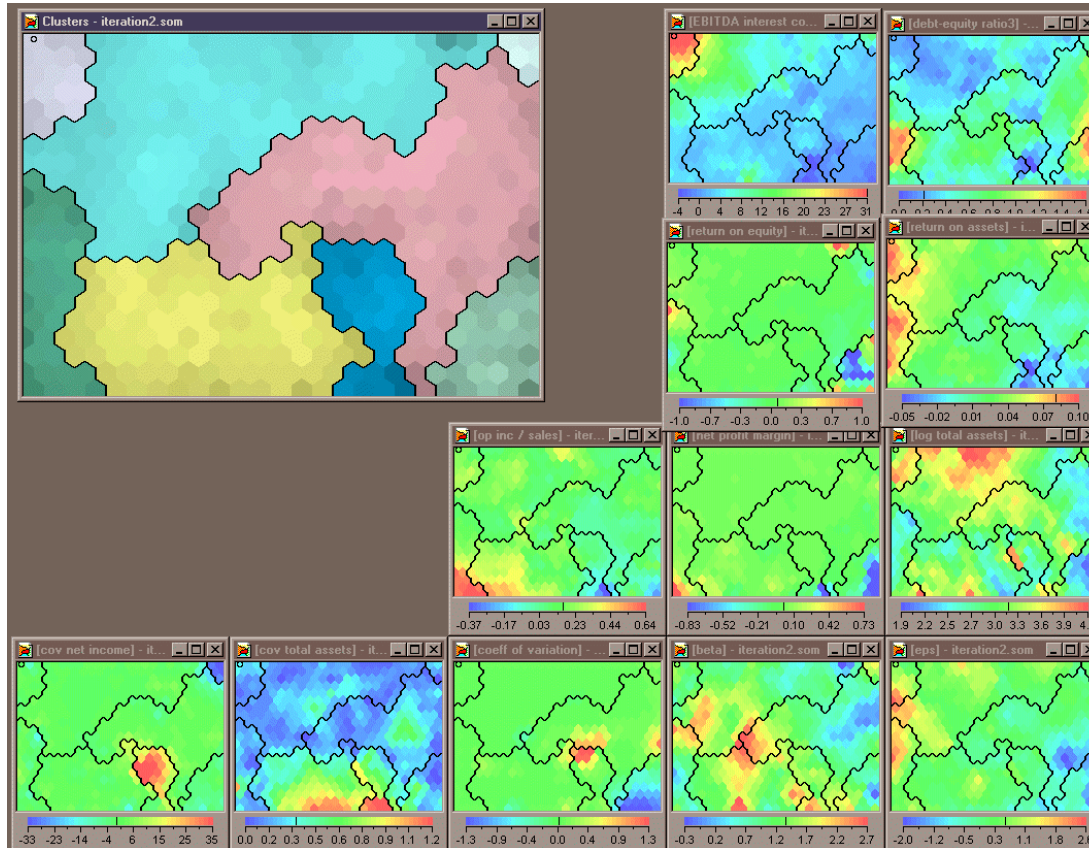
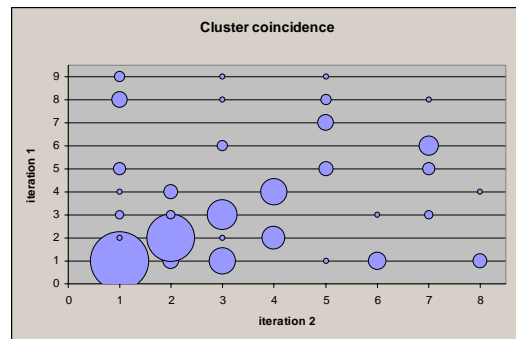Figure A-15 Self organizing map of sector consumer cyclicals 1998 fourth quarter, iteration 2

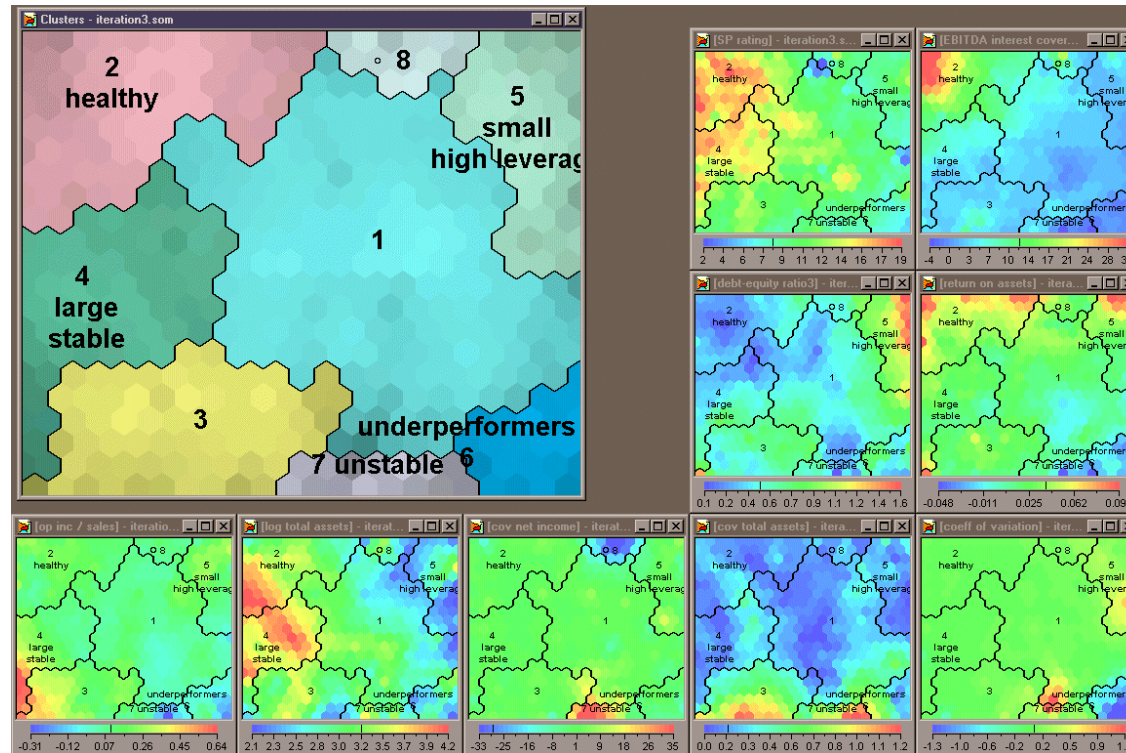Figure A-16 Cluster coincidence of iteration 1 vs iteration 2

Figure A-17 Self organizing map of sector consumer cyclicals 1998 fourth quarter, iteration 3
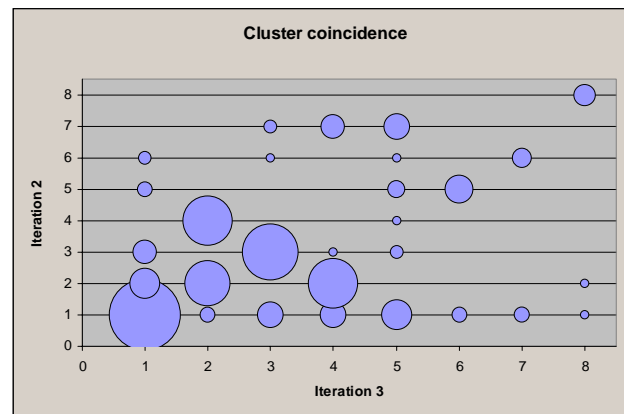

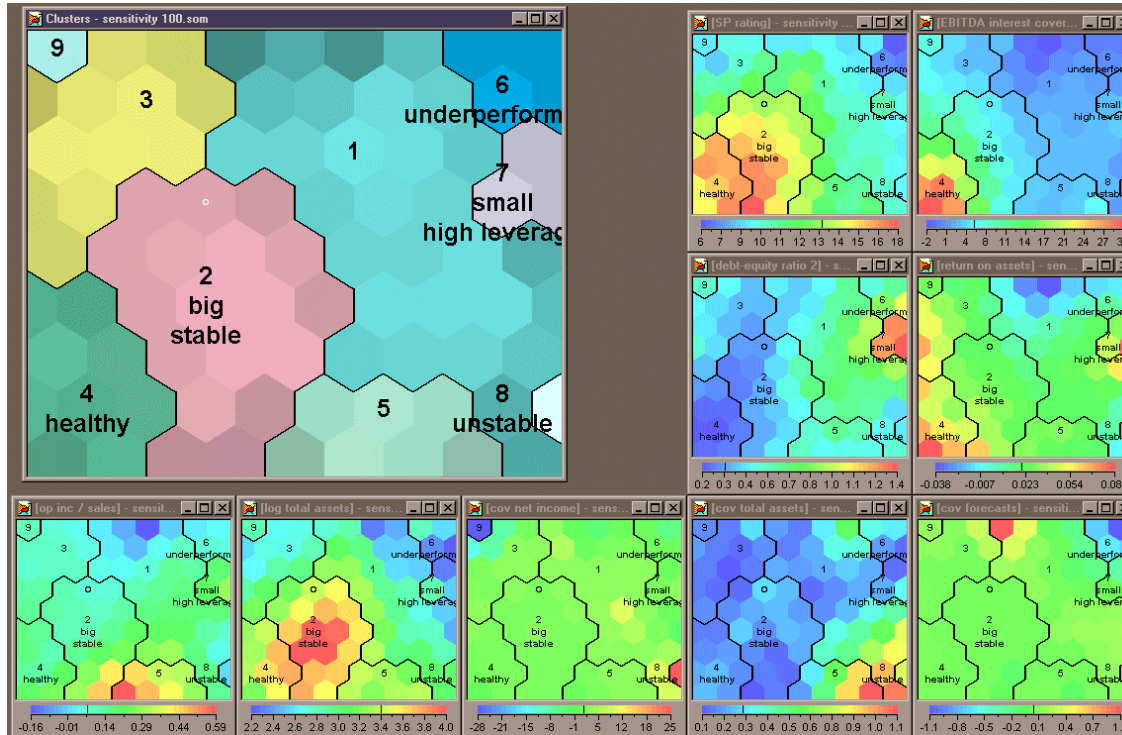Figure A-18 Cluster coincidence of iteration 2 vs iteration 3
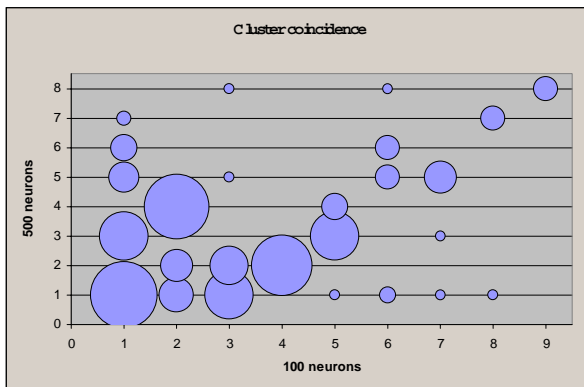
Figure A-19 Sensitivity analysis: using 100 neurons

Figure A-20 Cluster coincidence of 100 neuron som vs 500 neuron som
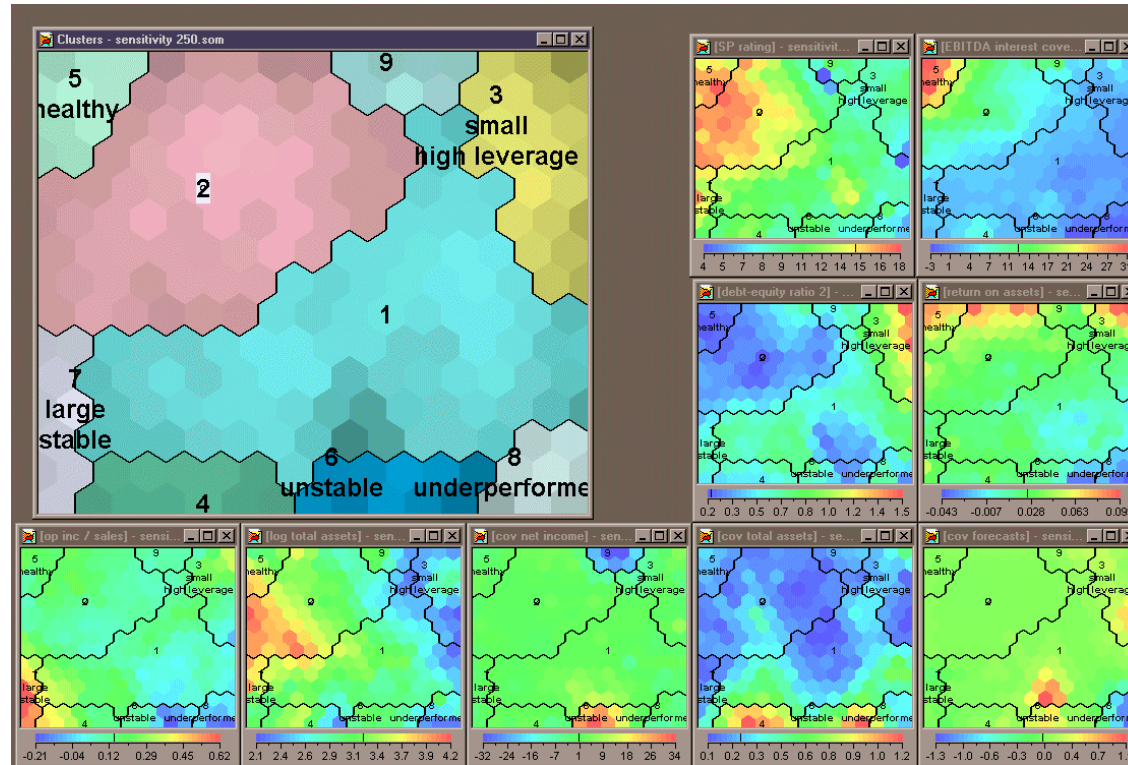
Figure A-21 Sensitivity analysis: using 250 neurons

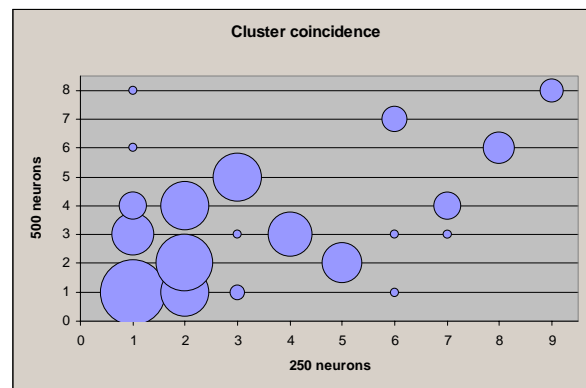Figure A-22 Cluster coincidence of 250 neuron som vs 500 neuron som
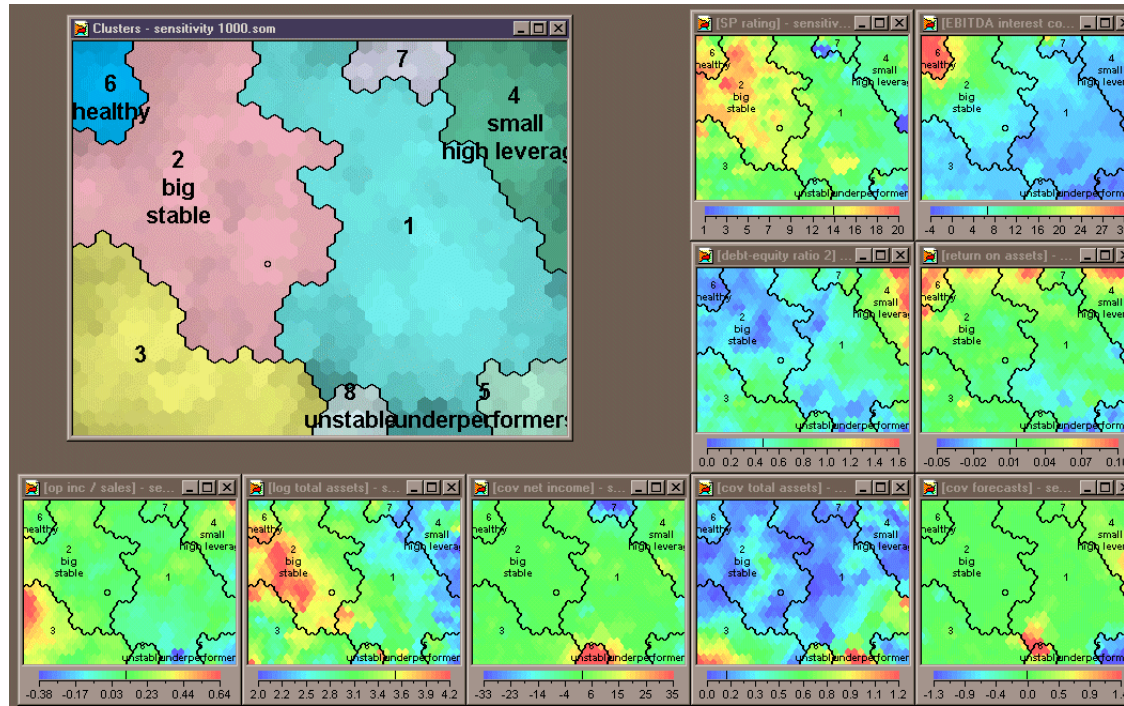
Figure A-23 Sensitivity analysis: using 1000 neurons
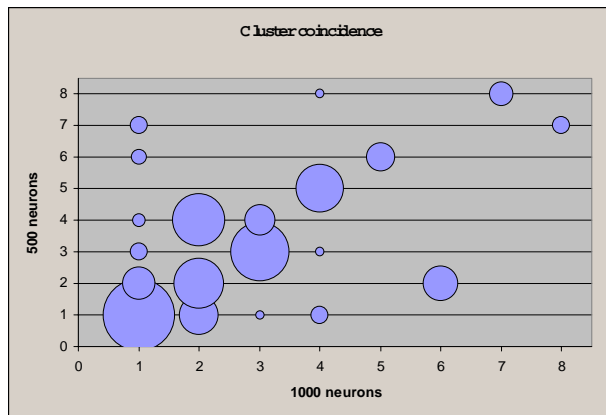
Figure A-24 Cluster coincidence of 1000 neuron som vs 500 neuron som
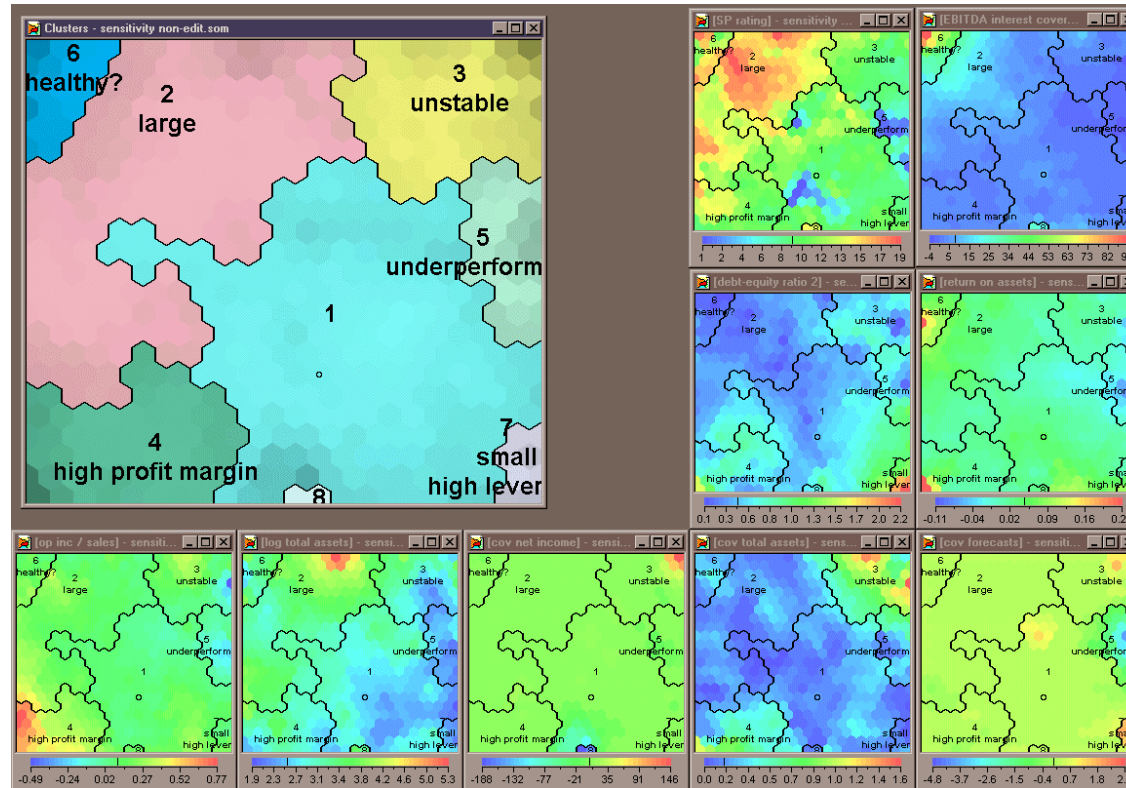
Figure A-25 Sensitivity analysis: using non-edited data
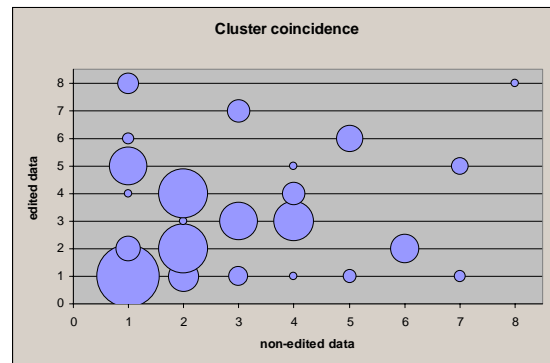
Figure A-26 Cluster coincidence of non-edited data som vs edited data som
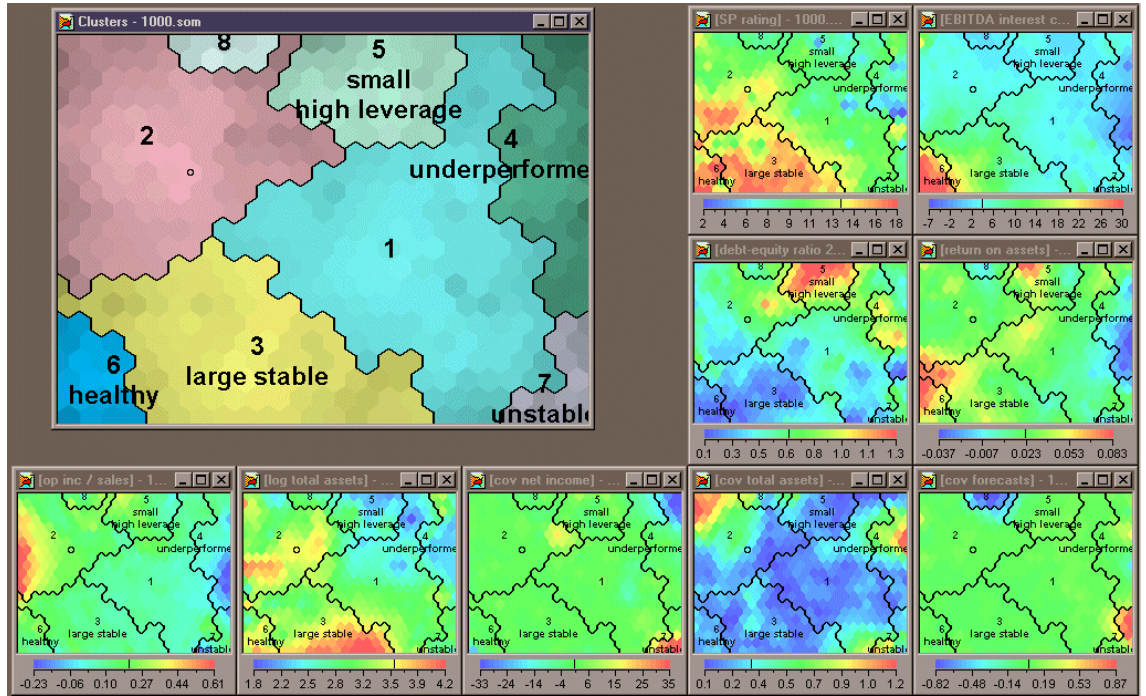
Figure A-27 Sensitivity analysis: merging all four cross-sections of 1998
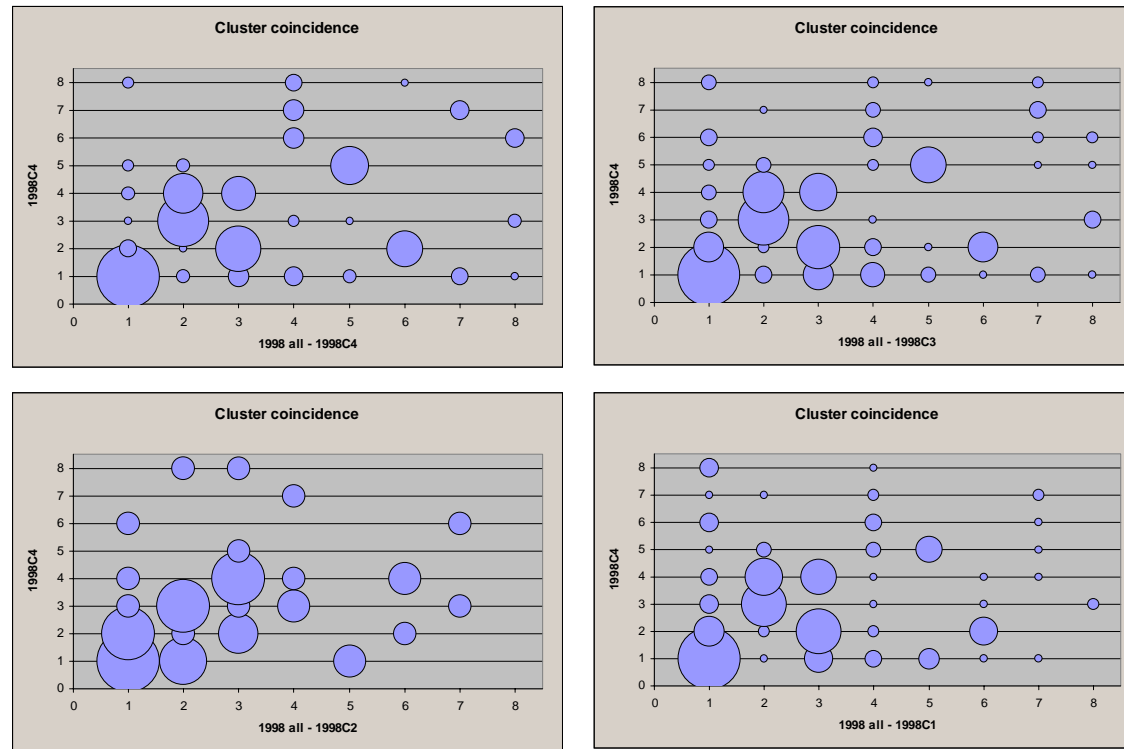
Figure A-28 Cluster coincidence of the separate quarters in the merged som of 1998 vs the final map.
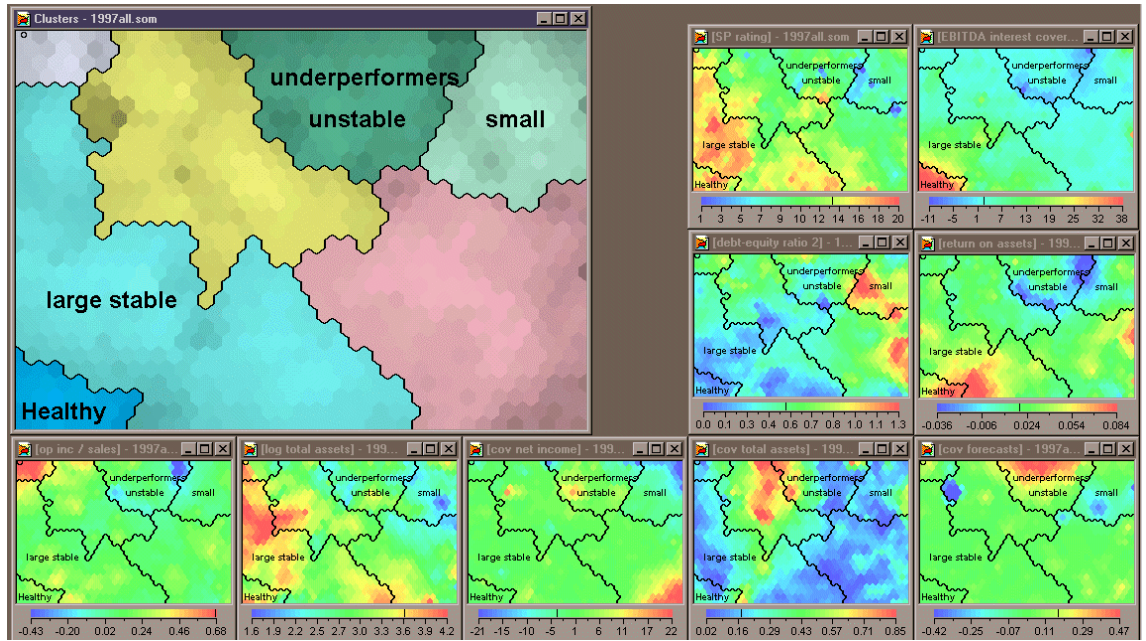
Figure A-29 Sensitivity analysis: using 1997 data
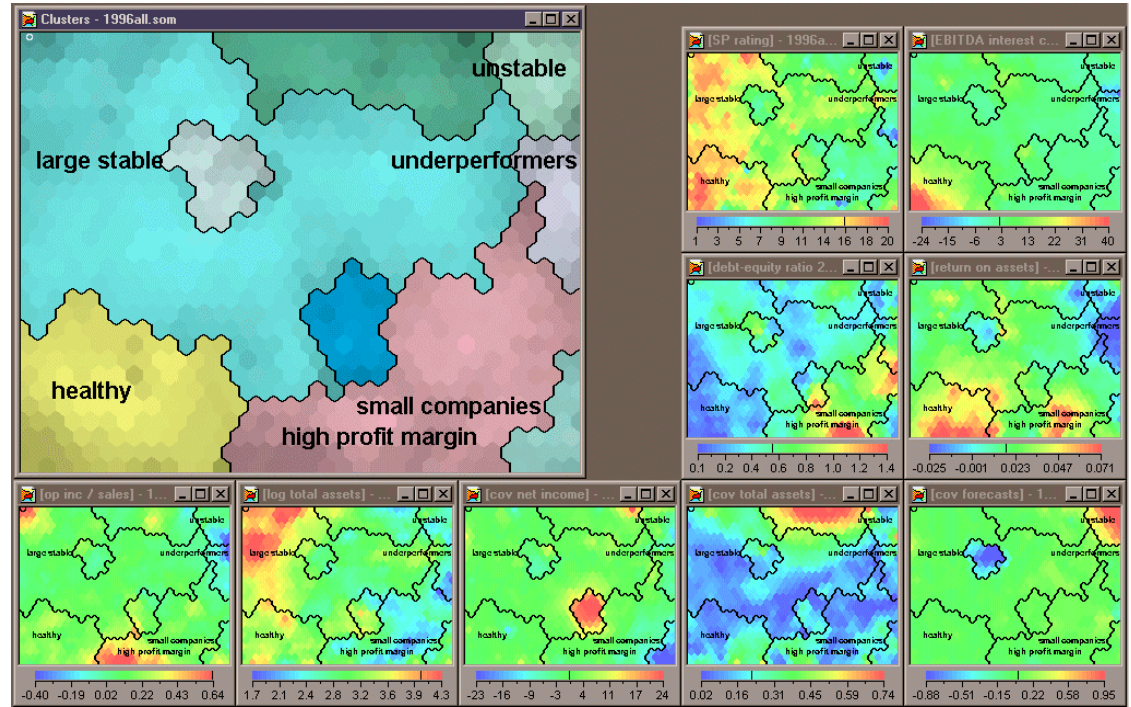
Figure A-30 Sensitivity analysis: using 1996 data

Figure A-31 Sensitivity analysis: using 1995 data

Figure A-32 Sensitivity analysis: using 1994 data

VII   Classification model

Table A-8 Model results for all tried variable combinations. The alternating colours reflect the different variable classes, the figures in red are model performances for the selected variable in the class (continued on next page).

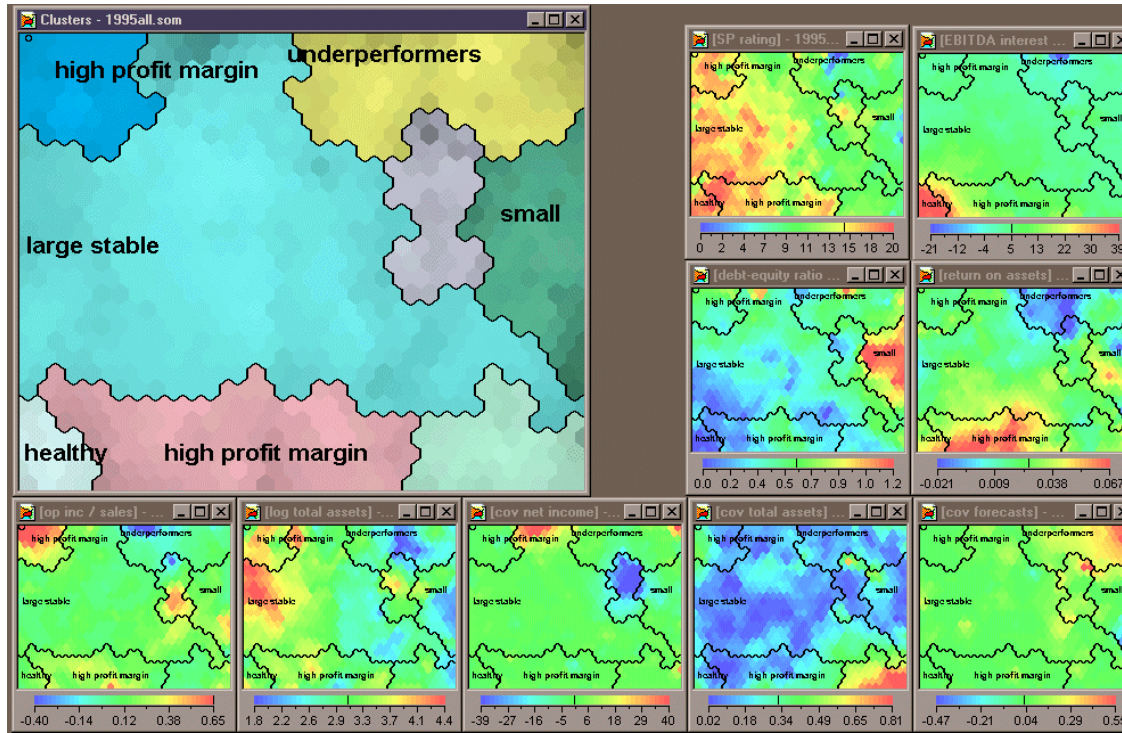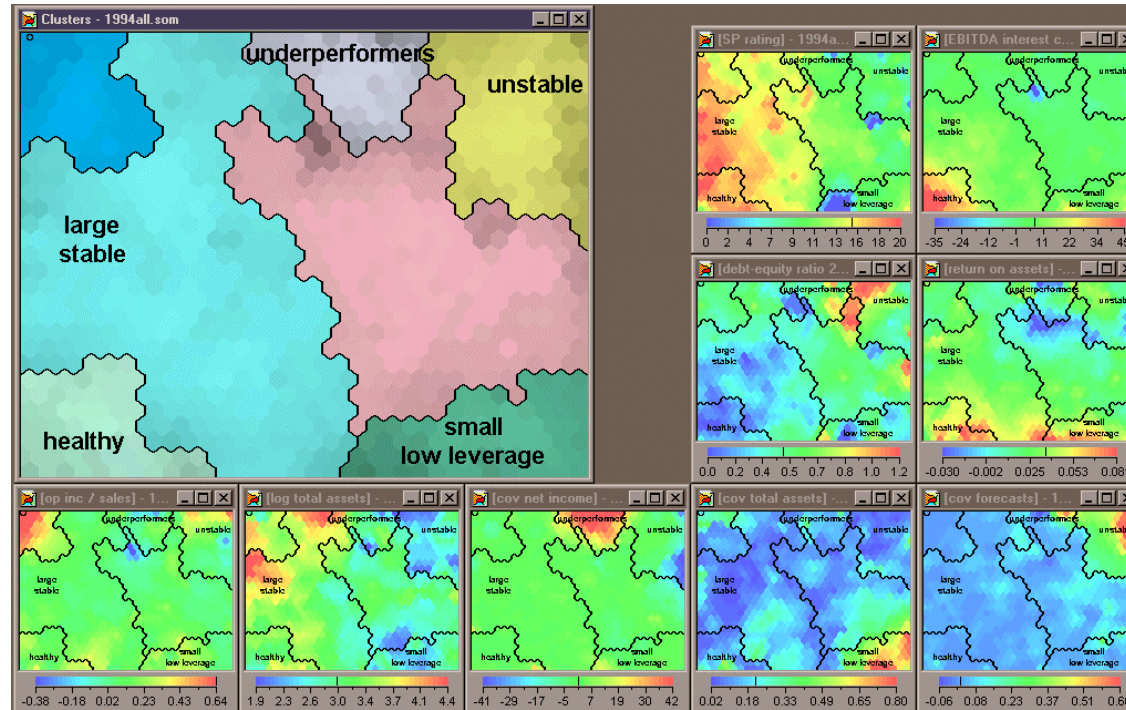| Variables																		Model results				
EBIT interest coverage	EBITDA interest coverage	EBIT / total debt	Debt ratio	Debt-equity ratio 1	Debt-equity ratio 2	Net gearing	Return on equity	Return on assets	Op. Inc. / sales	Net profit margin	Log total assets	Log market value	CoV net income	CoV total assets	CoV forecasts	Beta	Eps	MAD	success ratio max 0 notches (%)	success ratio max 1 notch (%)	success ratio max 2 notches (%)	R^2
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.59	0.23	0.57	0.76	0.66
+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.64	0.23	0.56	0.77	0.63
	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.40	0.29	0.62	0.84	0.72
		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.59	0.23	0.56	0.79	0.65
+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.65	0.19	0.55	0.80	0.65
	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.45	0.27	0.62	0.80	0.70
+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.49	0.29	0.62	0.84	0.70
+	+	+	+				+	+	+	+	+	+	+	+	+	+	+	1.47	0.30	0.57	0.80	0.69
+	+	+		+			+	+	+	+	+	+	+	+	+	+	+	1.59	0.25	0.54	0.80	0.67
+	+	+			+		+	+	+	+	+	+	+	+	+	+	+	1.49	0.29	0.62	0.84	0.69
+	+	+				+	+	+	+	+	+	+	+	+	+	+	+	1.64	0.24	0.54	0.77	0.64
+	+	+	+	+			+	+	+	+	+	+	+	+	+	+	+	1.61	0.22	0.57	0.80	0.65
+	+	+	+		+		+	+	+	+	+	+	+	+	+	+	+	1.52	0.24	0.57	0.80	0.69
+	+	+	+			+	+	+	+	+	+	+	+	+	+	+	+	1.48	0.29	0.62	0.84	0.68
+	+	+		+	+		+	+	+	+	+	+	+	+	+	+	+	1.41	0.25	0.63	0.83	0.72
+	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	1.56	0.22	0.56	0.79	0.69
+	+	+	+	+	+	+	+				+	+	+	+	+	+	+	1.51	0.27	0.62	0.79	0.67
+	+	+	+	+	+	+		+			+	+	+	+	+	+	+	1.53	0.23	0.60	0.80	0.68
+	+	+	+	+	+	+			+		+	+	+	+	+	+	+	1.49	0.25	0.61	0.82	0.68
+	+	+	+	+	+	+				+	+	+	+	+	+	+	+	1.66	0.26	0.57	0.77	0.62
+	+	+	+	+	+	+	+		+		+	+	+	+	+	+	+	1.48	0.23	0.61	0.79	0.70
+	+	+	+	+	+	+		+	+		+	+	+	+	+	+	+	1.53	0.25	0.57	0.80	0.67
+	+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	1.51	0.24	0.56	0.82	0.70
+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	1.60	0.26	0.58	0.77	0.64
+	+	+	+	+	+	+			+		+	+	+	+	+	+	+	1.49	0.26	0.61	0.79	0.68
+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	1.49	0.25	0.60	0.80	0.69
+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	1.52	0.27	0.58	0.82	0.65
+	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	1.48	0.28	0.58	0.82	0.68
+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	1.56	0.24	0.59	0.78	0.67

EBIT interest coverage	EBITDA interest coverage	EBIT / total debt	Debt ratio	Debt-equity ratio 1	Debt-equity ratio 2	Net gearing	Return on equity	Return on assets	Op. Inc. / sales	Net profit margin	Log total assets	Log market value	CoV net income	CoV total assets	CoV forecasts	Beta	Eps	MAD	success ratio max 0 notches (%)	success ratio max 1 notch (%)	success ratio max 2 notches (%)	R^2
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			1.50	0.24	0.59	0.80	0.70
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		+		1.46	0.27	0.58	0.82	0.72
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	1.53	0.25	0.56	0.80	0.68
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		1.55	0.21	0.56	0.81	0.69
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.52	0.20	0.63	0.81	0.69
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	1.55	0.25	0.55	0.78	0.68
	+		+				+				+			+	+			1.74	0.24	0.56	0.73	0.59
			+				+				+			+	+			1.90	0.21	0.53	0.71	0.54
	+						+				+			+	+			1.75	0.22	0.51	0.74	0.62
	+		+								+			+	+			1.73	0.22	0.51	0.74	0.60
	+		+				+							+	+			2.13	0.19	0.49	0.67	0.41
	+		+				+				+			+				1.68	0.23	0.56	0.74	0.64
+			+				+				+			+				1.66	0.26	0.59	0.74	0.61

Table A-9 Estimated tresholds for ordered logit classes

treshold	class
-8.11	1
-7.59	3
-7.46	4
-6.65	5
-6.34	6
-6.06	7
-4.60	8
-2.70	9
-1.62	10
-0.83	11
-0.32	12
0.39	13
1.13	14
1.97	15
2.48	16
3.41	17
4.39	18
4.86	19
	20