BUSINESS FAILURE PREDICTION

Statistical models for non-listed companies

FORECASTING AF KONKURS

Statistiske modeller for unoterede virksomheder

Cand.merc.fir
Kandidatafhandling
Udarbejdet af Morten Nicklas Bigler Jensen
Vejledt af Jeppe Christoffersen
Afleveret: 17. maj 2016
Antal anslag: 163.112 (71,7 normalsider)
Antal sider inklusiv tabeller og figurer: 80

Executive summary

Danske unoterede virksomheder repræsenterer suverænt størstedelen af alle danske virksomheder. Alligevel har noterede virksomheder været centrum for litteraturen omkring forudsigelse af konkurs. Jeg adresserer denne asymmetri i litteraturen of drager fordel af en omfattende datatilgængelighed af regnskabsdata for unoterede virksomheder gennem Orbis databasen. Jeg observerer at tidligere studier har været inkonsistente i rapporteringen af forudsigelsesgrad. På baggrund af tidligere studier, udvikler jeg min egen metode til at måle forudsigelsesgraden for mine modeller. Dette performance mål, ΔTC, viser hvor meget en given udlåner vil spare ved at applikere mine modeller relativt til naivt at låne ud til alle lånekandidater. Dette mål tager højde for den asymmetriske omkostningsprofil for hhv. type I og type II fejl samt tager højde for konkursfrekvensen.

Jeg bestemmer tre statistiske teknikker, der med succes tidligere er blevet anvendt til at modellere forudsigelse af konkurs; multipel diskriminantanalyse, logistisk regresionsanalyse og varighedsanalyse (hazard analysis). Konkursinformation er paneldata af natur. Jeg finder at de statistiske egenskaber ved varighedsanalyse er attraktive for modellering af konkursforudsigelse. Jeg konkluderer at min varighedsmodel opnår den bedste forudsigelsesgrad, når jeg applikerer mine modeller på et sekundært datasæt.

Jeg udvikler tre modeller; to modeller bygget på logistisk regression og én model bygget på varighedsanalyse. Jeg applikerer mine tre modeller på et sekundært datasæt for at teste forudsigelsesgrad. Derudover applikerer jeg også Z''-score modellen på mit sekundære datasæt. Jeg konkluderer at mine tre modeller viser bedre forudsigelsesgrad end Z''-score modellen. Af mine tre modeller finder jeg, at min model bygget på varighedsanalyse viser stærkest forudsigelsesgrad. Ved at applikere min varighedsmodel opnår jeg ΔTC på -13,0%. Det betyder, at långivere kan opnå en besparelse på 13,0% ved at applikere min model ift. at naivt at låne ud til alle. Min varighedsmodel er drevet af fire input variable; (1) "total gæld / totale aktiver", (2) "indtjening før renter og skat / finansielle omkostninger", (3) en dummyvariabel, der tager værdien 1 hvis egenkapitalen er negativ og (4) "tid", der måler et selskabs alder.

Baseret på usande forudsætninger estimerer jeg den samlede besparelse, alle danske långivere vil opnå, ved at applikere mine modeller. Denne estimerede årlige besparelse svarer til 66% af værdien af alle danske selskaber, noteret på fondsbørsen i København. Potentialet ved en overlegen konkursmodel er af substantiel karakter.

Jeg konkluderer, konsistent med tidligere studier, at det til en vis grad er muligt at forudsige konkurs af virksomheder ud fra finansiel information.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The advantages of an accurate model for business failure prediction **(BFP)** are obvious. Business failure involves many parties and large costs (Gepp, Kumar 2008).

The use of business failure models are ever-present. Institutions that could benefit of an accurate and simply implementable BFP model include governments, banks, auditors, managers, analysts and other stakeholders (Koh 1992, Dimitras et al. 1996, Kumar, Ravi 2007). BFP models are important for two reasons; (1) BFP models are very useful for those (managers, authorities, etc.) that can take action to prevent failure (Dimitras et al. 1996) and hence reduce the loss (Meyer, Pifer 1970). (2) BFP models can help the company's lenders or investors to assess the probability of default for the company, and on this basis select which companies to lend money or invest in (Dimitras et al. 1996). Overall, accurate BFP models will contribute to stable economic growth for the benefit of all involved (Gepp, Kumar 2008).

BFP models for non-listed companies

Non-listed companies represent the vast majority of all Danish companies. Non-listed companies represent >99% of all Danish companies (Nasdaq 2016, Danish Statistics 2016)[1]. Prominent and highly cited studies, including Beaver (1966), Altman (1968), Ohlson (1980), Zmijewski (1984) and Shumway (2001) develop BFP models for listed companies. According to a recent literature review by Appiah et al. (2015) +95% of previous BFP models are based on data from listed companies. I find it hard to understand that a relatively small number of companies present the majority of previous research in BFP. As early as 1968, Altman suggested that an area for future research would be to *"extend the analysis to relatively smaller asset-sized entities, where the incidence of business failure is greater than with larger corporations"* (Altman 1968). Multiple articles, including Altman (1968) and Adnan Aziz, Dar (2006) and the recent study by Appiah et al. (2015) suggest BFP models for smaller entities. Yet, listed companies have remained in the spotlight.

Data availability

[1] Corrected for multiple share classes 148 companies are listed in Denmark. Total number of active Danish companies equals ~300 thousand

"…small and medium sized firms (SMEs) in most jurisdictions are not obliged to publish company accounts, suggesting that prior studies are limited to listed firms"(Appiah et al. 2015). The lack of financial data for non-listed companies might be an explanation for the relative small number of BFP models for non-listed companies. However, the Orbis database comprise extensive "detailed financials" for non-listed companies in several European countries, including Germany, Greece, Ireland, Portugal, Spain, Sweden and Denmark[2]. Only a handful of non-European countries possess the same data availability. The data is available to everyone with access to the Orbis database. I address this mismatch in academia.

I obtain data for non-listed limited companies from the Orbis database. I find that 66% of Danish active companies are included in my dataset[3]. This coverage of detailed company financials is economy-wide.

My raw dataset contains more than 300.000 unique CVR-numbers (Danish "company numbers", which is unique for each company), almost 2.000.000 firm years (observations) and more than 27,000 unique bankruptcies over a 10-year period, including annual reports for the period 2003-2012 and bankruptcy data for the period 2003-2014. My sample shows average annual bankruptcy frequency of 1,2% and hence my sample is well representing the bankruptcy frequency in Denmark of 1,3%[4].

Variables

The vast majority of models use financial ratios extracted from income statements and balance sheets as input variables (Adnan Aziz, Dar 2006, Balcaen, Ooghe 2006, Appiah et al. 2015). Other variables employed include market based variables (Beaver et al. 2005, Agarwal, Taffler 2008, Hoque et al. 2013) cash-flow measures (Casey, Bartczak 1985, Dambolena, Shulman 1988, Hoque et al. 2013) and industry dummies (Chava, Jarrow 2004). Previous studies employing a mix of financial ratios and market-based measures conclude that market-based ratios add incremental information to the model (Shumway 2001, Hillegeist et al. 2004, Beaver et al. 2005). However, market based variables are not available for the vast majority of Danish companies.

The employment of cash flow variables has shown a mixed evidence (Balcaen, Ooghe 2006). Proponents of cash flow measures in BFP include Gombola, Ketz (1983), Gentry et al. (1985), Gentry et al. (1987), Aziz, Lawson (1989) and Sharma, Iselin (2003). Opponents of cash flow measures in BFP include Casey, Bartczak (1984), Gentry et al. (1985), Gombola et al. (1987) and Aziz et al. (1988). Financial ratios have evidently shown predictive success in BFP (Beaver et al. 2005).

[2] For the companies mentioned the Orbis database possess "detailed financials" for +20% of all non-listed companies. See appendix
[3] See chapter 4.1.5: "Validating data"
[4] See chapter 4.1.5: "Validating data" and appendix

I develop BFP models for non-listed companies; an area that many researchers have suggested, but only a few have explored. I benefit from the extensive data availability in Denmark. I employ financial ratios derived from income statements and balance sheets, which have evidently shown predictive ability for bankruptcies.

1.1 Research process

1.1.1 Motivations

The area of BFP models for non-listed companies is neglected in the literature. Non-listed companies represent the vast majority of Danish companies and a superior BFP model specifically developed for non-listed companies is desirable. I have access to a comprehensive dataset of non-listed Danish companies, and I am able to match financial data with the undesirable event of bankruptcy. The benefits of superior BFP models are multiple and desirable by many parties. Particularly in these days where "disruptive" and "fintech" are trending buzzwords. Statistical models for BFP enables analysts to analyze a large number of companies quickly (Petersen, Plenborg 2012). Bankruptcy companies destroy value for the community by not yielding sufficient income to service their obligations. A superior BFP model may help entities in discriminating between value-adding companies and value-terminating companies. This ability to discriminate may fence value-terminating companies in obtaining financing for value-terminating projects and hence benefit the whole economy.

1.1.2 Research question

The objective of this paper is to develop a superior BFP model. The research question is formalized as;

> *"A superior statistical model for business failure prediction of non-listed companies is yet to be developed. I have access to a comprehensive dataset with financials for non-listed Danish companies. On this basis; is it possible to develop a general business failure prediction model that is implementable for non-listed companies?"*

In order to answer the overall research question, I determine several questions that will govern the road towards solving the research question;

- What is written in academia within the area of BFP, and what are the key findings?
- What techniques are used for BFP?
- How do researchers compare model predictive abilities?
- How do researchers determine the independent variables of BFP models?

1.1.3 Limitations

In order to structure this paper I set four limitations. These limitations are;

(1) I focus on statistical models for forecasting BFP: by this, I exclude theoretical models and artificial intelligence models[5].

(2) I include only accrual-based accounting measures as input variables for my statistical models: by this, I exclude all other explanatory variables, including market variables, industry dummies, qualitative measures, external economic conditions and cash flow measures. I find that accrual based measures have proved predictive ability[6].

(3) I include only non-listed, Danish companies in my analysis: This includes startup companies, SMEs and multinational companies (for example LEGO is included). I do not discriminate between the different company classes during model development, as my objective is to develop a universal model applicable for everyone. However, I provide predictive success measures for different company classes[7].

(4) I focus on predicting bankruptcy based on the latest available annual report. Albeit I find evidence that financials of bankruptcy companies are inferior up to five years prior to bankruptcy, I do only provide success rate measures of predictability, based on "latest available annual report" data[8].

1.1.4 Contributions

I develop several BFP models for non-listed companies. These are companies that (1) represent the vast majority of all companies and (2) are not well represented in BFP academia. The contributions are multifold.

The contributions include:

(1) Multiple articles mention the asymmetric cost function of type I and type II errors, but only few quantify this cost function. I develop a tool for comparing model performance across multiple statistical approaches. My approach quantifies the cost function and utilizes this information when comparing models out-of-sample[9].

(2) I provide my final models, determinants of bankruptcy and coefficient estimates. I show the superiority of my models compared to the Z''-score model developed by Altman, whom is one of the entrepreneurs within the BFP area. I provide robustness checks of the models developed, and show the impact of

[5] Justification provided in chapter 2: "Business failure prediction – a brief overview".
[6] See chapter 4.2.2: "Accrual based accounting measures".
[7] See chapter 5.2.7: "ΔTC for different accounting categories".
[8] See chapter 4.1.1: "Matching bankruptcy with annual accounts".
[9] See chapter 3.1.1: "Success rate measurement".

simulating on the cost function assumption[10]. To my knowledge, no BFP model for non-listed companies has been developed from such an extensive dataset as the one I apply.

1.1.5 Structure

Firstly, I aim to create an understanding of the BFP problem. I create a fundament for the forthcoming analysis. Secondly, I elaborate on my data employed. I argue that my dataset is extensive, elaborate on the shortfalls of my data availability and aim to create an understanding of my dataset, by providing descriptive statistics. Thirdly, I take advantage of the foundation previously set. I explain the model development process and provide results for my final models. I apply my models to a holdout sample and apply my own developed method for assessing predictive success.

This paper is divided into 7 chapters:

Chapter 1: Introduction (page 4-9). This chapter aims to justify the raison d'être of BFP models, and why BFP models for non-listed companies are desirable. This chapter also formalizes the limitations and contributions of this paper.

Chapter 2: Business failure prediction – a brief overview (page 10-12): this chapter aims to create a full picture on BFP. This chapter briefly elaborates on statistical models, artificial intelligence techniques and theoretical models. This chapter justifies my limitation to focus on only statistical models.

Chapter 3: Literature review on statistical models (page 13-34). This chapter aims to create an overview of previous studies related to statistical models for BFP. Firstly, this chapter determines and elaborates on several key terms that are necessary to understand, in order to develop BFP models. Secondly, this chapter determines pioneers within selected statistical approaches. Thirdly, this chapter aims at determining state of the art for the BFP problem, and elaborates on these approaches and previous findings. Throughout this chapter, I explain how I apply findings to my model development process.

Chapter 4: Data (page 35-59). This chapter aims to explain the datasets employed in developing my BFP models. Overall, this chapter elaborates on the road from a raw dataset to a truncated dataset that enables me to develop statistical models. This chapter includes everything related to the data, including data availability, explanatory variables and descriptive statistics. Firstly, this chapter elaborates on my two initial datasets and the merging and matching procedure employed to create a master dataset. Secondly, this chapter elaborates on data availability and justify the choice of truncating data. The procedure for truncating data is elaborated. Thirdly, the data is validated and compared with external sources. Fourthly, the chapter elaborates on explanatory variables employed in my models. Fifthly, this chapter outlines the procedure for model

[10] See chapter 4: "Analysis".

development, including backward testing and exclusion of counter-intuitive explanatory variables. Sixthly, this chapter provides descriptive statistics for my dataset.

Chapter 5: Analysis (page 60-77). This chapter is the product of chapter 2, chapter 3 and chapter 4, where I set the stage for developing BFP models. In chapter 5, I apply my findings. Firstly, I recall the model development process and I develop three models. Secondly, I aim to validate my models developed; I develop numerous unreported models and determines that my final models yield superior holdout sample predictability. Thirdly, I organize a horse race on holdout sample results of my three models developed and Altman's Z''-score. I apply two different approaches in distributing companies into (i) forecasted default and (ii) forecasted non-default, and provide results with both approaches. Fourthly, I compare these two approaches, and discuss which one to apply. Fifthly, I simulate on my underlying assumption regarding the cost function. Sixthly, I provide a comparison of in-sample and out-of-sample results. Seventhly, I put my results into perspective and estimate the impact of applying BFP models on the Danish market.

Chapter 6: Conclusion (page 78-79). This chapter provides conclusions for my research question.

Chapter 7: Perspective, future research and final words (page 80). In this chapter, I present my proposals for future research and provide some final comments. Proposals for future research include the inclusion of qualitative explanatory variables. Final comments include comments and critique on my approach to developing BFP models.

References (page 81-85).

CHAPTER 2: BUSINESS FAILURE PREDICTION – A BRIEF OVERVIEW

The BFP literature consists of a considerable body of research, including more than 150 different models for BFP (Bellovary et al. 2007), many of which have proved high predictive ability. Given the broad number of models included in research papers since the 1960s, it is clear that a literature review is necessary in order to create an overview of the "state of the art" articles that have shaped the research of BFP.

In this chapter, I provide a helicopter view of the literature on BFP. This chapter divides BFP into three categories, and briefly explain each of them. Furthermore, I elaborate on the trends over time within the BFP area. This chapter justifies my focused research area; statistical models for BFP.

2.1 Categories of BFP

Following the framework of Adnan Aziz, Dar (2006), the approach for BFP can be divided into three main categories;

Table 1: Categories of BFP

Model category	*Main features*
Statistical models	Focus on symptoms of failure Drawn mainly from company accounts Follow classical standard modelling procedures
Artificial intelligence expert system models (AIES)	Focus on symptoms of failure Drawn mainly from company accounts Heavily depend on computer technology
Theoretical models	Focus on qualitative causes of failure Drawn mainly from information that could satisfy the theoretical argument of firm failure proposed by the theory Usually employ a statistical technique to provide a quantitative support to the theoretical argument

Source: Adnan Aziz, Dar (2006)

Statistical models

The first real, published academic research paper on the BFP problem was published in 1966 by Beaver. The first model was a simple univariate model with only single input variables. Since then the research on BFP

using the statistical approach has evolved. The statistical approaches used over time include multiple discriminant analysis (MDA) (Altman 1968, Dambolena, Khoury 1980, Altman 1993, Gunasekaran et al. 2009), conditional probability models (including linear probability, logit and probit models) (Meyer, Pifer 1970, Ohlson 1980, Zmijewski 1984, Altman, Sabato 2007) and hazard models (Luoma, Laitinen 1991, Shumway 2001, Beaver et al. 2005).

AIES models

AIES systems are systems of artificially intelligence, and aims to simulate the knowledge and reasoning of humans. These methods include machine learning, which means that the system "learns" and improves its problem-solving as a function of previous learning (Adnan Aziz, Dar 2006). Close to all AIES models depend on statistical methods, hence they are to be considered as extensions/sophistications, or automated processes, of the statistical approach. Bellovary et al. (2007) conducts a literature review over time (1960s – 2007) and concludes that "Neural Networks" (NN) was the primary method used in studies during the 1990s and 2000s. Adnan Aziz, Dar (2006) concludes that AIES models perform marginally better than statistical and theoretical models. However, Adnan Aziz, Dar (2006) provide a solution for model choice in empirical application. They provide a ranking solution according to the model's adjusted standard error. The findings indicate that MDA and Logit (both statistical models) may be more reliable.

Theoretical models

Theoretical models try to evaluate the qualitative causes of business failure. These models are often case-driven, and try to go further than just predicting company failure. They theoretically explain the drivers behind a business failure. Statistical models are driven by empiricism. They seek to find correlations with company fundamentals and the event of bankruptcy. Theoretical models are products of reasoning. They include balance sheet decomposition measures, gambler's ruin theory and cash management theory (Adnan Aziz, Dar 2006).

Literature development over time

After Altman (1968) published his article employing the MDA approach, the literature on BFP has evolved rapidly. MDA models were the primary method in the 60s and 70s, but then the literature saw a shift towards logit analysis (a conditional probability approach) and neural networks (an artificially intelligence approach) in the 80s and 90s (Bellovary et al. 2007). Albeit the MDA is no more the favorite approach by researchers,

and not really applied anymore, out-of-sample applications yield high predictive results, and the original Z-score model (MDA model by Altman (1968)) is often used as baseline model when comparing newly developed models (Altman, Narayanan 1997, Balcaen, Ooghe 2006).

Table 2: Distribution of primary models applied over time

	MDA	Logit	Probit	NN	Other *
1960s	67%	0%	0%	0%	33%
1970s	79%	4%	4%	0%	14%
1980s	51%	29%	5%	2%	13%
1990s	12%	22%	4%	47%	15%
2000s**	17%	25%	0%	33%	25%
total	37%	21%	4%	23%	15%

** others include LPM, judgmental, cusp catostrophy and hazard*
*** 2000-2004*

Source: (Bellovary et al. 2007)

From the literature review by Bellovary et al. (2007) MDA shows to be the most widely applied model throughout time, with 37% of all models in their review use MDA as primary approach to BFP. One can also conclude that logit was frequently applied throughout the 1980s and 1990s, and that logit analysis has been preferred over probit analysis. NN, an extension of the classical statistical models, where the researcher take advantage of more sophisticated computer programs, has been trending since 1990s. However, Adnan Aziz, Dar (2006) conclude that statistical models may be more reliable.

The number of factors (explanatory variables) included in previous studies over time has been around 8-10 on average, but varies from one to 57 (Bellovary et al. 2007). The number of factors included in a model, and the precise combination of ratios, seems to be of minor importance with respect to the overall predictive power, because included factors are correlated (Beaver et al. 2005). Beaver (1966) yielded as high as 92% model accuracy (overall success rate) with only one variable on a paired sample (50/50 distribution of failed and non-failed firms).

2.2 Summary of business failure prediction – a brief overview

I divide the approach to BFP into three main categories; (1) statistical models (2) AIES models and (3) theoretical models. I find that models employing artificial intelligence techniques (AIES, including NN) have gained popularity during recent years. However, I find that models derived from artificial intelligence techniques are a sophistication of statistical models, and that statistical models may be more reliable.

This paper is limited to focus solely on statistical models. I find that MDA models were popular in the 60s, 70s and 80s. Logit models were popular during the 80s, 90s and 00s. Hazard models have also been applied to the BFP problem.

CHAPTER 3: LITERATURE REVIEW ON STATISTICAL MODELS

In the following, a thorough review of key terms and statistical techniques applied to the BFP problem is conducted.

My overall objective is to create an overview of the literature to date, to formalize widely used models throughout the literature, and the theoretical shortfalls and biases related to the respective models. The ultimate objective is to create fundamental understanding of the complex and extensive literature on BFP, enabling me to create my own models for BFP. Throughout the literature review, I provide information on how I specifically employ my findings in model development.

This chapter is divided into two sections; (1) "Key terms in BFP" and (2) "Review of statistical models under examination".

(1) "Key terms in BFP": In this section, I define key terms, in order to create an understanding of the fundamentals of the BFP problem. These terms include success rate measurement, definition of business failure, sampling method and validity measures. During the chapter, I address how I implement findings into my model development. In the end of this section, I provide a table summarizing my approaches, based on the findings from this section.

(2) "Review of statistical models under examination": In this section, I determine the pioneers within selected statistical approaches and uncover the "state of the art" methods for the BFP problem. Statistical models under examination include (i) "Multiple discriminant analysis" (MDA), (ii) "Conditional probability models", primarily logistic regression analysis (logit) and (iii) Hazard analysis (survival analysis). This section prepares the grounds for model development. I emphasize the methodological issues related to the respective models and approaches. This is an important step in creating an understanding of the findings in academia, and to create a critical approach to the models.

3.1 Key terms in BFP

This chapter elaborates on the fundamentals of BFP. I discuss key terms and provide information on how I incorporate my findings into my final stage of model development.

In the following I elaborate on (1) success rate measurement, including type I and type II errors, quantification of the cost distribution, and cut-off points, (2) definition on business failure, including a discussion of when the "real" business failure takes place, (3) sampling methods, including clean data criterion, matching procedures and oversampling and (4) validation, where I argue for employing a holdout sample.

3.1.1 Success rate measurement

To assess the predictive ability, researchers apply several measures. Performance measures include overall predictive rates, type I and type II errors (or type I and type II success rates) Receiver Operating Curve (ROC), trade-off function, gini-coefficient, R^2 type measures (including pseudo R^2 measures) and measures based on entropy (Balcaen, Ooghe 2006).

"Overall predictive power" is easy interpretable and enables the researcher to compare results from different models. However, this measure has shortfalls. One key shortfall is the asymmetry between type I and type II errors.

Type I vs. type II errors

Type I errors refer to misclassification of bankrupt firms as non-bankrupt. Type II errors are the reverse – non-bankrupt firms misclassified as bankrupt firms (Bellovary et al. 2007, Beaver et al. 2011). Within the literature there is a consensus that Type I errors are more costly than Type II errors (Bellovary et al. 2007). This makes sense. A Type I error implies a company going bankrupt, hence a loss of business, where a type II error implies opportunity costs from not lending, seen from a lender's point of view (Gepp, Kumar 2008).

The costs associated with type I and type II errors respectively are mainly intangible or not measureable, depending on the user of the BFP model. Users include investors, lenders and accountants (going-concern justification) (Koh 1992, Dimitras et al. 1996, Kumar, Ravi 2007).

Table 3: Examples of classification costs to different users

USER	TYPE I	TYPE II	INTUITIVELY THE LARGEST COST
INVESTOR	Loss of investment	Loss of dividends (or other indirect return)	Type I
LENDER	Loss of loan	Loss of interest rates	Type I
ACCOUNTANT	Loss of reputation, risk of lawsuits (Koh 1992)	Loss of existing and potential clients	Type I

Source: (Koh 1992), own compilation

Table 3 aims to justify costs associated with type I vs. type II errors respectively, from three users' point of view. According to table 3, it is clear that type I errors are more costly for all users mentioned, relative to type II errors. However, the quantification of the relative costs associated with type I vs. type II respectively, is hard to determine.

Quantification of the error distribution of type I vs. type II errors

Altman et al. (1977) formalize and quantify the cost-function of type I vs. type II errors. They take the stand of lenders (more specifically banks);

Type I errors: is a function of gross loan losses and gross loans recovered. They estimate this to ~70%, i.e. 70% of loans issued to "failure companies" are lost money. Type II errors: is a function of opportunity costs from not lending, i.e. a function of interest rates and opportunity costs of lending to another company with similar risk measures. They quantify this term at ~2%

Overall they conclude that type I costs are ~35 times more costly than type II errors (70% / 2%), i.e. a cost ratio of 35x.

I apply the same approach for Danish companies for the period 2008-2010, which mirrors my holdout sample. From this analysis I find average real interest rates, for newly issued loans, of 4,10% (type II costs of 4,10%)[11] and estimate a recovery rate of 26,15% (type I costs of (1-26,15%) 73,85%) over the period[12], i.e. from my analysis type I costs are ~18 times more costly than type II costs. I apply these numbers for the assessment of "success rate". I emphasize that this is a very rough estimation. The calculation of recovery rate does not include collaterals, nor interest payments before default. However, this is a "best guess" estimation, and I find it necessary for quantifying the asymmetric cost function. The numbers underlying the calculations are to be found in appendix.

Altman et al. (1977) also highlight that this is the first study to explicitly formalize and quantify the asymmetric cost function. However, one should note that (1) this is an approximation and (2) other costs than those mentioned are not evaluated. Such costs for type I errors include loss for other stakeholders, for example employees. Costs for type II include loss of value creation, given that the borrowing company did not obtain financing for positive net present value investments. Such measures are hard to quantify, and

[11] Source: statistikbanken.dk, DNRNUPI, average of real interest rates for newly issued loans to non-financial companies, for the period January 2008 – December 2010. Numbers underlying the calculation in appendix.

[12] Source: finanstilsynet.dk: "statistisk materiale" for the period 2007-2010. I estimate the recovery rate as [total recovery for the period 2008-2010 / total charge-offs for the period 2007-2009], i.e. I lag the data. I denote that this is not a perfect measure, as charge-offs also include losses on private consumers. However, this approach is the same approach as (Altman et al. 1977), and provides a fair assessment of the loan-loss recovery rate.
Numbers underlying the calculation in appendix.

support the argument that the relative costs of type I and type II errors respectively, are a subjective choice (Balcaen, Ooghe 2006).

Koh (1992) provides a full article for the discussion of type I vs. type II errors. They do not quantify a specific cost ratio (like 35x in Altman et al. (1977)), but provide a formula for estimating the expected loss, and calculates total costs for different cost ratios.

$$EC = (PN)(PI)(CI) + (PG)(PII)(CII)$$

Where

- EC = expected misclassification cost of using the model
- PN = prior probability of non-going concerns (bankruptcy frequency in percentage)
- PG = prior probability of going concerns (1-(bankruptcy frequency in percentage))
- PI = (# Type I errors / number of non-going concerns)
- PII = (# type II errors / number of going concerns)
- CI = misclassification cost of a type I error
- CII = misclassification cost of a type II error

This formula quantifies the ex-ante cost-function, i.e. the expected loss.

Obviously, the objective is to minimize the cost function.

My approach of success rate measurement

On the basis on the findings of Altman et al. (1977) and Koh (1992) I develop my own success rate criteria. My approach is to quantify the total costs with the following formula

$$TC = DF * T1EF * CT1 + (1 - DF) * T2EF * CT2$$

Where

- TC = total costs, percentage
- DF = Default frequency
- T1EF = Type I errors frequency (type I errors / total defaults)
- CT1 = costs associated with type I errors
- (1-DF) = Going-concern frequency
- T2EF = type II errors frequency (type II errors / total going concerns)
- CT2 = costs associated with type II errors

My equation is an ex-post measure, and enables me to quantify the costs associated with my developed models. The TC measure is easy to interpret and easy to apply in real life.

> Example: assuming DF=1,5% (default frequency of 1,5%), T1EF=75% (type I errors of 75%), CT1=73% (i.e. 27% recovery rate, hence (1-27%) 73% type II costs), (1-DF)=98,5% (going concern companies), T2EF=5% (type I errors of 5%) and CT2=3,4% (i.e. opportunity costs of lost real interest rate of 3,4%), then TC = 0,99%. This is, that the total loss, as a percentage of all loans, equals 0,99%.

This number is easily applied in real life. The total loan loss of a given lender equals:

$$Average\ loan\ size * number\ of\ loans\ issued * 0{,}99\%$$

This interim step leads to my final success rate measurement. Earlier articles focus on overall predictive ability (Balcaen, Ooghe 2006) or type I costs (see e.g. Shumway 2001, Beaver et al. 2005). To control for the asymmetric cost function and the low bankruptcy frequency[13] I develop my own success rate criteria. My approach is intuitive and easy to understand.

My approach is quantified by:

$$\Delta TC = \frac{TC_{Developed\ model\ applied}}{TC_{Lend\ to\ all}} - 1$$

"ΔTC" has a real-world meaning. ΔTC of -15% means that by applying a given model, a lender will experience a decrease of 15% in costs, relative to the scenario, where the naïve lender lends money to all.

The implied assumptions behind this approach are that all companies in my data sample will borrow an equal amount of money. I acknowledge that this is a rough estimation, but enables me to quantify the success rate measure, taking into account the asymmetric cost distribution.

The bankruptcy frequency in my sample is only around 1,3% annually[14]. Assuming an equal cost distribution (i.e. the costs associated with type I vs. type II are equal) the overall predictive rate, for a given model, must exceed 98,7%, in order to out-perform the naïve approach of "lending to all". A quantification of the asymmetric cost distribution and applying this measure, gives flavor to the final assessment, and enables a quantification of the impact of applying my models.

Cut-off points

[13] I find that bankruptcy frequency equals 1,3% per annum. See chapter 4.1.5: "Validating data"
[14] See chapter 4.1.5: "Validating data"

The asymmetric distribution of type I and type II errors is widely recognized throughout academia. However, many researchers apply a cutoff of 0,5 and thus assume a symmetric loss-function across the two types of classification errors (Ohlson 1980, Balcaen, Ooghe 2006, Gepp, Kumar 2008).

The objective of cutoff points is to distribute companies into two groups; (1) predicted bankruptcy and (2) predicted non-bankruptcy.

For my analysis, I apply two approaches.

The first approach (*"percentile approach"*) is the approach applied by Shumway (2001), Chava, Jarrow (2004), Beaver et al. (2005) and Altman, Sabato (2007); this is, I rank and divide predicted probabilities of default into percentiles with 5 percentage points steps (5%, 10%, 15% and 20%). The percentiles will define my cut-off point. All companies with predicted probabilities in the X% percentile, will be classified as 'bankrupt'. All companies not in the X% percentile, will be classified as 'non-bankrupt'.

The second approach (*"cut-off approach"*) is the traditional approach, applied most widely throughout the literature (see e.g. Meyer, Pifer (1970), Ohlson (1980) and Zavgren (1985). This approach is simply assigning a cut-off point. All companies with a predicted probability above a given cutoff point will be classified as 'bankrupt'. All companies with a predicted probability below this cutoff point will be classified as 'non-bankrupt".

Figure 1: Distributing companies into predicted default and predicted non-default respectively and success rate measurement

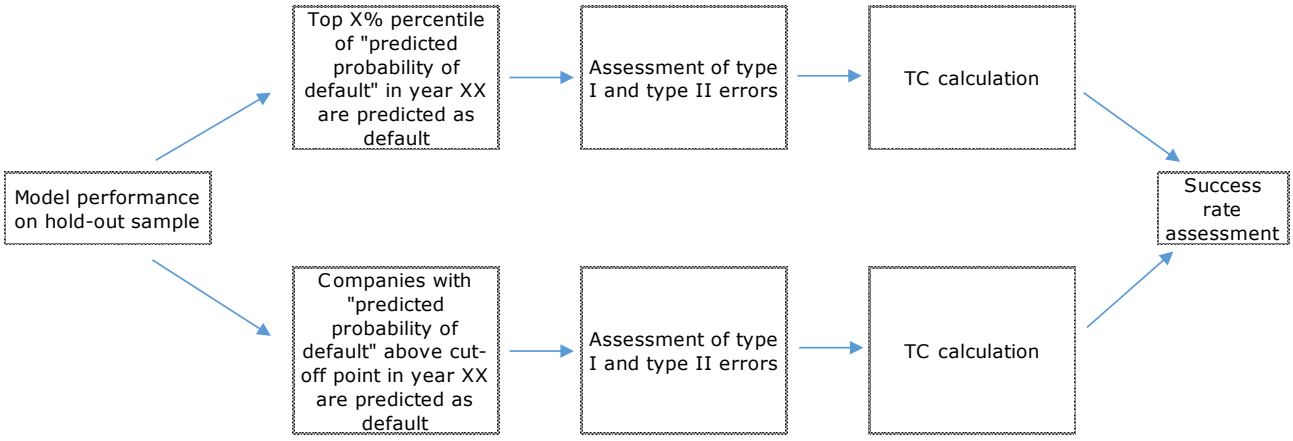


Figure 1 summarizes the two approaches I apply in distributing companies into either bankrupt or non-bankrupt.

3.1.2 Definition of business failure

The dependent variable of the statistical models is the definition of "business failure", which takes the value 1 if failed and 0 if not. This raises the question; what is the real definition of business failure and how does one determine the time of the business failure event?

84% (71% for "protection sought from creditors" and 13% for "creditors' or voluntary liquidation, appointment of receiver") of previous studies apply the legal definition of bankruptcy (Appiah et al. 2015). This definition allows an objective criterion for dating the failing firms, and easily split the sample into failed and non-failed firms (Charitou et al. 2004). This suggests that there is a general agreement on the legal definition of business failure in academia.

Other determinants of the dependent variable include; suspension of stock exchange listing, going concern qualification by the auditor, composition with the creditors, breach of debt covenants and company reconstruction (Appiah et al. 2015).

Balcaen, Ooghe (2006) criticizes the arbitrary separating of samples into either business failure or non-business failure. The business failure definition is not a clear-cut; some researchers argue that one can only separate into business failure, non-business failure and a "grey-zone" (Peel, Peel 1987, Appiah et al. 2015). The separation of samples into "failed" or "non-failed" is not a clear-cut procedure. One may argue that the use of a dichotomous dependent variable is in contrast with reality (Appiah et al. 2015). Albeit the definition of business failure is blurred, a researcher must do some simplifications, in order to formalize a statistical model for bankruptcy prediction, and the most common solution is applying the legal definition, albeit this not being a perfect measure.

The real objective of a business failure study must be to determine when a company faces challenges, and ultimately is not able to meet the condition of going-concern, which might lead to loss from customers, lenders, employees and the community. When a researcher applies the legal definition of bankruptcy as determinant, one should keep in mind the fact that the 'real' business failure might occur before filing for bankruptcy.

I apply the legal definition ("filing for bankruptcy") and match the event of bankruptcy with the latest available annual accounts. Indeed, the "real" business failure occurs at another time. However, the true point in time of business failure is unknown. I hypothesize that the latest available company accounts paint the picture that financial health of the company is deteriorating and the company is moving towards the undesirable event of bankruptcy. This is, I relate these numbers to the event of bankruptcy; the financial information in the latest available company accounts is the information that should be explanatory in BFP.[15]

The legal definition offers some important advantages. The moment of failure can be objectively dated, and is easy to implement for the researcher (Charitou et al. 2004, Balcaen, Ooghe 2006). Filing for bankruptcy is often considered as the ultimate business failure (Bellovary et al. 2007)

[15] See chapter 4.1.1: "Matching bankruptcy with annual accounts" for a thorough explanation of matching procedure applied.

In addition, one should keep in mind that the legal definition of bankruptcy varies across country boarders. Appiah et al. (2015) finds that 53% of studies originate from USA. In the US the legal definition is different from the definition that applies to Danish corporations. In the US a company can file for different parts of bankruptcy, including chapter 7, which implies that the company will be liquidated and the bankruptcy trustee will gather and sell the debtor's nonexempt assets, in order to cover creditors' claims (uscourts.gov 2016), and chapter 11, which is frequently referred to as a "reorganization" bankruptcy; this implies that the company may seek adjustments of debts, either by reducing debt or by extending the time for repayment (uscourts.gov 2016a). The primary difference between filing for chapter 7 vs. filing for chapter 11, is, that when filing for chapter 11, the company is still going concern, and liquidation may, but must not, take place. When filing for chapter 7, the objective is liquidating the company. To my knowledge previous papers on US data, apply the definitions of bankruptcy indiscriminately.

Without going into details, the Danish definition is much similar to the chapter 7 in the US; *"bankruptcy, legal means by which a debtor's assets are to be distributed among all creditors"* (Vistrup Lene 2016)[16]. Based on the findings that legal definition is the far most applied determinant in separating samples into failed vs. non-failed companies, the inconsistency in legal definitions across borders might lead to complications when comparing cross-border research; i.e. results from US studies might not be directly applicable to Danish companies.

Albeit different definitions of business failure, Hayden (2003) found that three different models developed for three different definitions of failure (bankruptcy, delay in payment and loan restructuring) have very similar structures regarding the selected variables. Adnan Aziz, Dar (2006) also hypothesize that the predictive power on an individual model is independent of the dataset being used, also across country borders, provided that the data has been drawn from reliable and dependable sources. They also emphasize that this is not a finding, but a hypothesis from what they observe, and suggest future research may well be able to test the trueness of this hypothesis.

When applying models to Danish data I use the legal definition as dependent variable. This is in line with the majority of previous studies, and allows me to objectively and easily allocate businesses into two groups; business failure=1 and non-business failure=0.

3.1.3 Sampling methods

In 1984, some years after the emerging trend of BFP began, Zmijewski (1984) published a critical article about the statistical shortfalls of previous studies. Specifically, Zmijewski (1984) mentioned two implications with the estimation techniques applied to date; (1) oversampling distressed firms and (2) complete data criterion bias.

[16] free translation

Oversampling distressed firms

For the period 2003 to 2012 the bankruptcy frequency in Denmark for all firms is 1,3% per annum (minimum 0,7% in 2006 and max 2,2% in 2010) (DST)[17]. Zmijewski (1984) highlights that previous studies use rates of 1.5% to 50%. The well-known Z-score model by Altman (1968) is conducted on a sample of 33 failing companies and 33 non-failing companies, hence a rate of 50%.

If the model is to be used in a predictive context, the samples of failing and non-failing firms should be representable for the whole population (Ooghe, Joos 1990). One might expect biased results when oversampling distressed firms. On the contrary, the Z-score model, applying a 50% rate, has performed consistently well over time, in out-of-sample tests (Altman 2000), albeit this method statistically introduces bias into the estimates.

For my data, I address the problem of over-sampling. I find that bankruptcy frequency (company bankruptcy as a percentage of total companies per in a given year) differs marginally some years. However, I conclude that I do not oversample failed companies[18].

Complete data criterion

Zmijewski (1984) also mentions the shortfalls of the "complete data criterion". One of the fundamentals of modern statistics is the assumption of random estimation samples. When including in analysis only the observations that fit the need of the researcher, a researcher breaches the assumption of random samples.

> *"When applying non-random estimate samples, the classical statistical methods are applied inappropriately and the resulting model cannot be generalized"* (Balcaen, Ooghe 2006).

Zmijewski (1984) finds that the use of non-random variables does not significantly change the overall predictive rates. Only the individual group classifications (type I and type II errors) and estimated probabilities seem to be affected by the use of non-random variables.

In my analysis, I apply a "complete data criterion", as I find it necessary for conducting the analysis and fulfilling my objective. My initial hypothesis is that I might get oversampling of larger companies, as data availability for larger companies might be higher relative to smaller companies. However, I find that after

[17] Source: DST (Statistics Denmark). Calculation: non-seasonally adjusted bankruptcies per year / total companies in year. Numbers underlying calculations in appendix.
[18] See chapter 4.1.5: "Validating data"

applying a complete data criterion, I am left with companies, where "average total assets" are ~36% smaller than before applying a complete data criterion. Total assets is my proxy for "company size". However, also other entries should be determinants of company size. These include number of employees, total equity and net earnings. After applying a complete data criterion, I find a change of +20% (number of employees), +3% (total equity) and +40% (earnings after tax)[19].

Arbitrary matching failed companies with non-failed companies

Many of the academic papers practice a matching procedure for their failed companies, in order to obtain a sample with 50% failed companies and 50% non-failed companies. This matching is performed arbitrary, and often by age, size and industry code (Balcaen, Ooghe 2006). Researchers employing this procedure include Altman (1968), Zavgren (1985) and Gentry et al. (1985). I do not apply any matching procedure, and I thus avoid this bias. I develop general models on economy-wide data, with bankruptcy frequency equal to the overall frequency in Denmark[20].

Other concerns and comments

Other concerns regarding sampling include over/under sampling of industries, size and age. A model developed on US data, might perform different when applied on Danish data, as the mix of industries is different. This shortfall might be reduced by developing specific models for e.g. (1) industries, (2) size class and (3) age of company. Albeit an appealing approach, I do not possess sufficient data on industries. Size and age are explicitly included in some models.

3.1.4 Validation

Jones (1987), Adnan Aziz, Dar (2006) and Bellovary et al. (2007), among others, argue to applicate models on a secondary sample, in interest of stronger test of predictive availability. Albeit holdout sample application yields stronger test of predictive validity (Adnan Aziz, Dar 2006), the findings of Adnan Aziz, Dar (2006) and Bellovary et al. (2007) indicate that less than half of their studies under review applied a validation sample.

I develop my models on a dataset with annual reports for the period 2003-2007 (5 years), and apply the models on a holdout sample with annual reports for the period 2008-2010 (3 years) to validate the performance of the models. I find that my models indeed show predictive success when applied on a holdout sample.

[19] See chapter 4.1.4: "From Rawdata to Cleandata"
[20] See chapter 4.1.5: "Validating data"

3.2 Summary of key terms in BFP

I address four key terms in BFP; success rate measurement, definition of business failure, sampling methods and validation.

Success rate measurement: I find that there are several ways of measuring predictive success, including overall predictability rate and type I errors vs. type II errors. I determine two procedures for distributing companies into either (i) failed or (ii) non-failed, based on predicted probability of default. These two procedures I address as "percentile approach" and "cutoff approach" respectively. I note that previous studies are inconsistent in providing success rate measurement and previous studies are thus difficult to compare. I compute a quantification of the asymmetric cost distribution, and develop a new and intuitive way of success rate measuring; ΔTC. My approach is simple, and quantifies the savings a given lender may face by applying my models compared to the naïve approach of "lend to all".

Definition of business failure: I find that researchers have previously applied several definitions on "business failure". Furthermore, I discuss when the "real" business failure takes place. I apply the most frequently applied definition; "legal bankruptcy" as my ultimate business failure definition. However, I note that the legal definition is different across country boarders, but find evidence from an article that this does not influence the BFP model.

Sampling methods: I find that previous studies apply several sampling methods when computing their samples. I address the shortages of arbitrarily matching failed companies with non-failed companies (which may lead to oversampling failed companies), and the problem with applying a complete data criterion. I find that I apply an average bankruptcy frequency that is much similar to the Danish bankruptcy frequency[21].

Validation: I simply conclude that I apply my models to a holdout sample in order to validate predictability of models developed.

[21] See chapter 4.1.5: "Validating data"

Table 4: Summary of my approaches

TERM	MY APPROACH
SUCCESS RATE MEASUREMENT	I apply my own measure "ΔTC", developed on the basis on previous articles. This measure measures the percentage change of applying my BFP models compared to the naïve approach of "lend to all". By applying this measure, I am able to include an asymmetric cost function. Furthermore, I apply an assumption of the cost ratio. I apply a cost ratio of ~18x. This is, I assume that type I errors are 18 times more costly than type II errors, from a lenders point of view.
DEFINITION OF BUSINESS FAILURE	I apply the legal definition of bankruptcy ("filing for bankruptcy") as my ultimate determinant of business failure. I match the event of "filing for bankruptcy" with the latest available annual report.
SAMPLING METHOD	I apply a clean data criterion. I avoid arbitrarily matching failed companies to non-failed companies, and aim to generate samples, that mirrors the total population.
VALIDATION	I apply a holdout sample for validating purposes

3.4 Review of statistical models under examination

The most frequently applied statistical models in academia include MDA and logit models[22]. Hazard models overcome one of the most criticized fundamental challenges of the MDA, logit and general cross-sectional approaches; the fact that the MDA and logit models do not include time-variables (Shumway 2001), and that most studies include only one observation for each company (see e.g. Altman 1968, Meyer, Pifer 1970, Ohlson 1980). Even if a study, such as Lennox (1999), include several observations for the same firm (multiple entries for the same firm for different years, hence panel data), this implies statistical shortages. A logit model with pooled data, as the one developed by Lennox (1999), breaches the assumption of independent observations, as the accrual based performance of one company in time t, will affect the performance of the same company in time t+1 (Balcaen, Ooghe 2006). Hazard models, also known as survival analysis, overcome this shortage by explicitly taking into account the time variable, and the non-random distribution of observations; hence ultimately neglect the bias produced when analyzing panel data with a logit model.

On this basis, my focus for the rest of this paper will be on MDA (as a base-line model), logit (as it has been widely applied and is well suited for a statistical problem with dichotomous dependent variables) and hazard models (as they explicitly consider time, and allow for non-random variables and ultimately enables more data input).

This chapter focuses on selected statistical models. In the following, I elaborate on (1) multiple discriminant analysis, including elaboration on several of the models developed by Altman, one of the most prominent and highly cited researchers within the area of BFP, (2) conditional probability models, including a short

[22] See chapter 2.1: "Categories of BFP"

introduction to linear probability models and probit models as well as a throughout review of logit models, and (3) survival analysis models (or hazard models), including justification of its statistical superiority to panel data problems.

3.4.1 Multiple discriminant analysis

The Multiple Discriminant Analysis (MDA) approach to BFP is one of the most used and recognized approaches in forecasting bankruptcy. Already in 1968, Altman (1968) published the first multivariate study, relying on the MDA approach. The result was the well-known and recognized Z-score model. In other comparable studies with other statistical approaches and overall objectives of developing new models for BFP, Altman's Z-score model seems to be frequently used as a 'baseline' model (Altman, Narayanan 1997, Balcaen, Ooghe 2006). Furthermore, the Z-score model is used for educational purposes (Petersen, Plenborg 2012). The Z-score model seems to be a generally accepted standard model (Balcaen, Ooghe 2006).

> *"MDA is a statistical technique used to classify an observation into one of several a priori groupings dependent upon the observation's individual characteristics. It is used primarily to classify and/or make predictions in problems where the dependent variable appears in qualitative form, e.g., male or female, bankrupt or non-bankrupt."* (Altman 1968)

In BFP, the two groups are failed vs. non-failed. Most frequently financial data is used as input to the model. The MDA attempts to derive a linear combination of these characteristics which best discriminates between the groups.

The review of the MDA approach is solely due to the findings of several literature reviews, including Dimitras et al. (1996), Balcaen, Ooghe (2006), Bellovary et al. (2007) and Appiah et al. (2015); that the MDA approach is frequently mentioned, and the model was one of the first movers in BFP. The statistical fundamentals behind the MDA approach has been extensively criticized since it was applied to a BFP problem in 1968[23]. Yet, it has proven to deliver great out-of-sample accuracy rates, over different periods (Altman 2000).

Altman's Z-score model & further developments

The fact that Altman's model has been generally accepted is justified. Altman (2000) performs an out-of-sample test of the original Z-score model in different periods; 1969-1975, 1967-1995 and 1997-1999, and finds that using a cut off score of 2.675 the predictive accuracy is 82%-94%. However, this is a truth with modifications. Altman (2000) does not explicitly address the asymmetric cost function related to type I and

[23] To come later in this chapter

type II errors respectively[24]. Furthermore he applies datasets with distribution of 50% failed and 50% non-failed, and thus not mirror the overall bankruptcy frequency in the population (see e.g. Zmijewski (1984) who addresses the arbitrary sampling method).

Table 5: Classification and prediction accuracy of the Z-score (1968) failure model

Original sample (25)	Holdout sample (33)	1969-1975 predictive sample (86)	1976-1995 predictive sample (110)	1997-1999 predictive sample (120)
95%	96%	82%	85%	94%

Source: (Altman 2000)

Table 5 shows that during different periods, the model has performed consistently well in predicting out-of-sample business failures.

The original Z-score model coefficients were as follows (Altman 1968):

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

Where

- X_1 = working capital / total assets (5)
- X_2 = retained earnings / total assets (4)
- X_3 = EBIT / total assets (1)
- X_4 = market value of equity / book value of total debt (3)
- X_5 = sales / total assets (2)

An analysis of the relative contribution is presented by the numbers in the brackets, i.e. "EBIT / total sales" is the variable in Altman's analysis, with the highest relative contribution to the whole model.

Note that the model does not have an intercept, which is due to the statistical package utilized. Other software programs have a constant term, which standardizes the cut-off score at zero if the size of the two samples are equal (Altman 2000). Altman's (1968) Z-score model uses a cut-off score of 2.675.

The original Z-score model was the result of a sample of only 66 listed companies; 33 in each group (failed vs. non-failed). The companies were all manufacturing companies. Furthermore, the original Z-score model includes market variables. In X_4 the market value of equity is a part of the ratio, which is a complication since market variables obviously are not available for non-listed companies.

Albeit the small estimation sample and the fact that the model is estimated from a sample of manufacturing companies, the Z-score model has been widely used. Since the original model was published, Altman has published several other articles regarding BFP, and several books.

[24] See chapter 3.1.1: "Success rate measurement"

Altman et al. (1977) intentioned to update the Z-score model to adapt to new developments in bankruptcies, including (1) new financial reporting standards (primarily capitalization of leases), (2) average size of companies recent years had increased significantly, (3) to generate a general model, as the first Z-score model was focused on manufacturing companies. Also other variables than the original Z-score model were used. The estimated model was named "ZETA model". In the article, the authors also conclude that although the statistical properties of the data, which indicated a quadratic structure was appropriate, the linear structure of the same model outperforms the quadratic in test of model validity. In order to construct the ZETA model, an analyst must do several adjustments to the dataset applied, including capitalization of leases, deduction of goodwill and intangibles from assets and expense research and development costs rather than capitalize them. As this information is not available for my study, I refrain from applying the ZETA analysis to my study.

Ad hoc adjustments to the market value of equity of the original Z-score model are not scientifically valid (Altman 2000), and thus Altman (1993) published model coefficients to be used for privately held entities. He updated the X_4 variable – now to include book value of equity rather than market value of equity.

The new coefficients were as follows (Altman 1993):

$$Z' = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5$$

Where

- X_1, X_2, X_3 and X_5 are the same ratios as in the original Z-score model
- X_4 is changed: the market value of equity is substituted by the book value of equity

As one of the input variables is changed, also the other coefficients are changes. For example X_1 is changed from 1.2 to 0.7. Altman (1993) did not test the new model on a secondary sample, due to limited data availability for privately held entities. The "Z**'**-score" is still developed from manufacturing companies.

Altman went even further, and developed yet another model: "Z**''**-score". The justification was, that a model with "sales / total assets" developed from data on manufacturing companies, is not applicable for other companies, as the coefficient on "sales / total assets" is normalized for manufacturing companies (Altman 1993, Altman 2000). In order to develop a universal model, Altman included retailers into the testing sample, dropped the "sales / total assets" variable and updated the estimates. The new coefficients were as follows (Altman 1993):

$$Z'' = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4 + 3.25$$

Where

- X_1 = working capital / total assets

- X_2 = retained earnings / total assets
- X_3 = EBIT / total assets
- X_4 = book value of equity / total liabilities

I observe that this model has a constant in the equation, hence the cut-off score equals zero in his model. The cutoff score is derived from median Z''-score for bankrupt US entities (Altman 2005) and thus he implicitly assume a symmetric cost function[25]. The Z''-score model was tested on both US manufacturers and US non-manufacturers, and accuracy and reliability remained high. Albeit cutoff should equal zero in a model with an intercept, I find another cutoff score to be optimal when incorporating an asymmetric cost function[26].

The estimated Z''-score model does not include market related variables and is applicable for non-manufacturing entities and is hence sufficient for my analysis of Danish companies, where the data set contains non-listed firms. In my analysis, I apply the Z''-score model to non-listed Danish companies, and test the predictability on a holdout sample.

4.4.1.1 Critique of the MDA approach

The MDA approach is based on three restrictive assumptions; (1) the independent variables included in the model are multivariate normally distributed, (2) the variance-covariance matrices are equal across the failing and non-failing group and (3) the prior probability of failure and the misclassification costs are specified. Several authors stress the possible bias from the two first assumptions, but often they do not test whether the model satisfies the assumptions. (Balcaen, Ooghe 2006)

Assumption #1; multivariate normally distributed independent variables

This assumption seems in general to be violated, which may result in a bias in the standard errors hence the significant test. One way of correcting (or transform) variables is by logarithmic transformations (Balcaen, Ooghe 2006). The procedure of transforming variables is employed by Beaver et al. (2005).

Assumption #2; equal variance-covariance matrices

A violation of this assumption may lead to misleading significance test, when testing differences in variable means between the failing and non-failing group.

Assumption #3; prior probabilities should be determined before estimation

[25] See chapter 3.1.1: "Success rate measurement"
[26] See chapter 5.2: "Holdout sample application"

This assumption relates to the determination of the optimal cutoff point (Balcaen, Ooghe 2006). As addressed in chapter 3.1.1: "Success rate measurement" the cost function is not symmetric. Simply estimating a model without prior probabilities of default will lead to the implicit assumption of symmetric loss-function across the two types of classification rates.

Hard interpretation on coefficients

Moreover, in MDA models, the standardized coefficients cannot be interpreted like the slopes of a regression equation and hence do not indicate the relative importance of the different variables (Zavgren 1985, Altman, Sabato 2007).

3.4.2 Conditional probability models

The conditional probability models for BFP refers to models, where the dependent variable is binary and hence the dependent variable equals 0 or 1. The output of the model is directly interpreted as the conditional probability of success given x ($P(y = 1|x)$). The probability of y=1, given x, is referred to as the "probability of success". Whether y=1 is a success in its purest sense is doubtful. In my model, the success criteria (y=1) is the undesirable event of bankruptcy, which is rarely associated with success. However, this is the general accepted terminology for conditional probability models and thus this is the terminology I apply.

Conditional probability models include three statistical approaches; linear probability models (LPM), logistic regression models (logit) and probit models (probit). Throughout the history of conditional statistical approaches for BFP logit is the far most applied technique (Bellovary et al. 2007). Researchers applying at least one of the conditional probability approach to BFP include Ohlson (1980), Mensah (1983), Zmijewski (1984), Gentry et al. (1985), Zavgren (1985), Lo (1986), Dambolena, Shulman (1988), Aziz, Lawson (1989), Platt et al. (1994), Lennox (1999), Charitou et al. (2004) and Altman, Sabato (2007).

Linear probability models (LPM)

Meyer, Pifer (1970) were the first to apply the LPM approach to BFP (Dimitras et al. 1996). A LPM employs the OLS (ordinary least squares) procedure. Applying the OLS procedure for a statistical problem with a binary dependent has some statistical drawbacks:

1) The variance depends on x, thus the homoscedasticity assumption is violated (Wooldridge 2015).
2) Due to the linearity of the model, the model can predict values below zero and above one. This is undesirable, as the objective of the model is to predict probabilities, and common knowledge tells us that probabilities cannot go below zero, nor exceed one.

Albeit significant statistical shortfalls of LPM, the approach has advantages; the model is easy interpretable. The beta coefficients are easily linked to the probability, as $\Delta P(y = 1|x) = B_j \Delta x_j$. This is, a change in the independent variable, leads to a linear change in the probability of success, and albeit LPM breaching the assumption of homoscedasticity, other work has shown that t and F statistics are typically not far away from the values obtained with a valid estimator. Albeit statistical shortfalls OLS statistics are not completely meaningless (Wilke 2015).

Logit / probit models

Even Altman, the inventor and proponent of the original Z-score model (1968), turned to logit analysis (see Altman, Sabato (2007)). As to be elaborated in this chapter the statistical features of probit/logit models seems appealing to the BFP problem.

Researchers frequently favors the logit approach before the probit approach (Bellovary et al. 2007). I will follow this trend. The logit/probit models overcome the shortfalls of the LPM. The models apply a non-linear function that takes only values between zero and one. The main difference between logit and probit, is, that the logit assumes a standard logistic distribution for the error rate (e), and probit assumes a standard normal distribution for the error rate (e) (Wooldridge 2015). When applying the logit approach, no assumptions are made regarding the distribution of the independent variables (Balcaen, Ooghe 2006). Logit also allows for disproportional samples ,where the MDA assumes equal distributions. (Balcaen, Ooghe 2006).

The drawback of these models is that interpretation is harder than with the LPM approach. This is because the dependent variable, Y, is changing with the level of x (Wilke 2015). The magnitude of the coefficients itself is not useful. However, the direction of the effect (i.e. the sign of the coefficient) is similar to LPM (Wilke 2015).

The models apply the maximum likelihood estimation approach (MLE), which is a non-linear structure. When applying MLE the heteroscedasticity in $var(Y|x)$ is automatically accounted for (Wooldridge 2015)

Ohlson (1980) was the first researcher to apply the logit approach to the BFP problem. Furthermore, Ohlson's model is used for educational purposes (Petersen, Plenborg 2012). Zmijewski (1984) was the first to apply the probit approach. Since then, the logit model has been applied much more frequently than the probit model, and therefore I will keep my focus on the logit model (Bellovary et al. 2007).

The costs of type I and type II errors do not need to be accounted for, before the estimation of the model. If one apply a cut-off point of 0.5, the researcher implicitly assume a symmetric cost function. A researcher could derive an optimal cutoff point in order to minimize the total cost (see e.g. Beaver et al. (2011)). The

drawback of this approach is the fact that the cost ratio assumption is subjective, and might differ from one lender to another (Balcaen, Ooghe 2006).

With regard to statistical properties, the logit model seems to be better suited for the BFP problem, than that of MDA. However, MDA has shown great out-of-sample predictability in several previous studies. The review of Adnan Aziz, Dar (2006) show that on average, MDA models yield 85% model accuracy, while logit models yield 87%.

3.4.2.1 Critique of the logit approach

Despite the nice features of the logit approach, the procedure imply several shortfalls, which I address.

The logit approach is extremely sensitive to multi-collinearity; i.e. inclusion of highly correlated variables must be avoided (Balcaen, Ooghe 2006). Beaver et al. (2005) explicitly address the high correlation between ratios, and emphasize that due to the high correlation between explanatory variables, the precise combination of ratios used seems to be of minor importance. They used only three explanatory variables for their analysis. These findings indicate that by employing only a few, well-founded explanatory variables, the model might obtain a high accuracy and the statistical drawback of correlated variables might be reduced.

Furthermore, the logit approach is sensitive to outliers and missing values (Balcaen, Ooghe 2006). In my analysis, I overcome the problem with missing values by implementing a "complete data criterion"[27]. However, implementing a complete data criterion may also introduce sampling bias[28]. With outliers I do not want to exclude these observations. I determine the 1st and 99th percentile for all ratios generated, and instead of excluding observations outside this range, I set the observation to equal the 1st or 99th percentile respectively. This approach allows me to keep the observations, albeit the observation might seem "extreme".

> Example: If observation for company j, for the variable "Net income / total assets", is below -2,3 (1% percentile) ➔ the observation is set at- 2,3. The reasoning is that the information in the observations below -2,3 is negligible. However, I do not want to exclude these observations, as this observations clearly imply a company with significant negative return on assets. The solution I apply is to transform the observation to equal the 1% percentile, which in this case equals -2,3.

In addition, the logit/probit models lack the inclusion of time. Bankruptcy is by nature panel data. Applying a logistic analysis on panel data violates one of the basic assumptions of logit models "randomly distributed explanatory variables" – this is similar to the case of the MDA approach. By including several observations for the same company for several years, I introduce bias to the results (Shumway 2001). This crucial

[27] See chapter 4.1.4: "From Rawdata to Cleandata"
[28] See chapter 3.1.3: "Sampling methods"

statistical shortfall that evidently introduce bias to the results and standard errors (and ultimately significance tests), is solved by the implementation of the simple Hazard procedure[29].

I develop two logit models. The first model, "Logit 5y" include five years of data, with several observations for the same company. With this model, I do not correct for serial correlation of explanatory variables[30] and in theory this model is not statistical valid. The second model, "Logit 1y", include only one year of data. With Logit 1y I does not include panel data, and thus avoid breaching the assumption. However, Logit 1y only includes one year of data, hence a substantial reduction of the estimation sample.

3.4.3 Hazard models (survival analysis)

Both MDA and logit models rely on the assumption of randomly distributed variables. They are in nature cross-sectional and static models able to perform statistical analysis for an event in time=t. They lack the inclusion of the time dimension. The statistical properties of those cross-sectional models are not optimal for panel data, of which bankruptcy data indeed is classified as. Hazard models, also known as survival analysis, solve these statistical shortfalls.

The possibilities of survival analysis is multifold. Survival analysis has been applied to numerous problems, including lifetime expectancy, time between trades in financial markets, product durability and even duration of wars (Kiefer 1988). Hazard models have been applied to a variety of accounting issues (Beaver et al. 2005) including the duration of consecutive earnings increases (Beatty et al. 2002). Researchers applying the hazard approach to BFP include Lane et al. (1986), Luoma, Laitinen (1991), Laitinen, Kankaanpaa (1999), Shumway (2001), Kauffman, Wang (2001) and Kauffman, Wang (2003)

Lane et al. (1986) were the first to apply the survival analysis approach to the BFP problem (Luoma, Laitinen 1991). The review by Gepp, Kumar (2008) conclude that a number of previous studies, including Luoma, Laitinen (1991) and Laitinen, Kankaanpaa (1999) does not show significant superiority of the hazard approach in holdout sample applications. However, there seem to be a consensus that the statistical features of survival analysis are superior to logit or MDA techniques. Luoma, Laitinen (1991) indicate that a dataset of significant size would reveal the superiority of survival analysis. Shumway (2001) was the first to apply survival analysis to a dataset of significant size (Gepp, Kumar 2008)[31], and found that his hazard model is superior in holdout sample applications.

[29] See chapter 3.4.3: "Hazard models (survival analysis)"
[30] i.e. breach of the underlying assumption of randomly distributed explanatory variables
[31] (Shumway 2001) included just short of 29.000 companies in his estimation sample

Hazard models explained

Classical hazard models model BFP as a timeline, where businesses are represented by a lifetime distribution (Gepp, Kumar 2008). As static models (e.g MDA and logit) can include only one observation per company, these models include only a snapshot of the financial health of a given company. Hazard models enables the researcher to include several observations per company, i.e. allows the researcher to analyze panel data. I emphasize that bankruptcy by nature is panel data.

There are three reasons why hazard models are more appropriate for BFP modelling compared to static models (Shumway 2001).

1) Hazard models considers time at risk: Some companies may be at risk several years before filing for bankruptcy, while others may file during the first year in risk. A deterioration of financials may not lead to bankruptcy at first, but over time, deterioration financials may lead to bankruptcy. Static models fail to include this path. Hazard models adjust for it automatically.
2) Hazard models include time-varying covariates: explanatory variables change over time. If a firm's financial health is deteriorating over time, this change in covariates is included into the hazard model estimation. Static models fail to do so. Hazard models can also account for the possibility that firm age might be an important explanatory variable.
3) Hazard models are able to include much more data: Due to statistical properties, static models can include only one observation for each firm. A lack of doing so may yield invalid estimators. Hazard models allow the researcher to include unlimited years of data for the same company, i.e. they may produce more efficient out-of-sample forecasts. The hazard model can be thought of as a logit model that includes each firm year as a separate observation.

Survival analysis covers several techniques. For my analysis, I apply the hazard model technique used by (Shumway 2001). The interpretation of coefficients equals the interpretation of the logit models, but overcome the most significant shortfalls of the logit procedure. Beaver et al. (2005) and Shumway (2001) provide nice explanations of the hazard technique.

Albeit previous studies find it hard to show superiority in holdout samples, the theoretical and statistical features of the survival analysis technique are indeed appealing for BFP modelling.

3.4.3.1 Critique of Hazard models

Albeit the hazard technique shows statistical superiority to MDA and logit approaches for BFP, this method also has disadvantages. There is evidence indicating that sample construction, more specifically the proportion of failed and non-failed companies, may affect the estimation of the hazard model (Gepp, Kumar

2008). However, my bankruptcy frequency well reflect the population[32] and thus this problem should be of minor importance. Hazard models are also subject to multicollinearity problems (similar to logit models). However, I apply standard backward testing[33] and according to Gepp, Kumar (2008) researchers can easily avoid multicollinearity problems by applying this procedure.

3.5 Summary of review of statistical models under examination

I determine three statistical approaches, on which I focus, including multiple discriminant analysis, conditional probability models (with primary focus on logit models) and hazard models (survival analysis).

Multiple discriminant models: I focus on the MDA models developed by Altman, one of the most prominent and highly cited researchers in BFP. I find that Altman has modified his original Z-score. This modified model is the Z''-score, and is applicable for (i) not only manufacturing firms and (ii) non-listed companies. I decide that I applicate this model on my holdout sample, and compare the out-of-sample predictability of this model with my models.

Conditional probability models: These models include linear probability models, probit models and logit models. This section focuses on the logit models. Logit models are well suited for problems with a binary dependent variable, as the output is directly interpretable as probability of default. I address the shortfall of the undesirable combination of panel data and logit models. Several previous studies include only one observation per company. I determine that I develop two logit models; (i) the first model includes 5 years of data, and (ii) the second model includes only 1 year of data.

Hazard models: I find that hazard models overcome most of the shortfalls of the MDA and logit models. Hazard models explicitly accounts for time, and enables the researcher to include multi-period observations for the same company, i.e. panel data. Hazard models may be viewed as an extension of the logit model; *"a hazard model can be interpreted either as a logit model done by firm year, or it can be viewed as a discrete accelerated failure-time model."* (Shumway 2001).

[32] See chapter 4.1.5: "Validating data"
[33] See chapter 4.2.4: "Model development procedure"

CHAPTER 4: DATA

In the following, I elaborate on my datasets and data availability. I explain the steps I follow in preparing a final dataset that suits my needs for generating BFP models. Furthermore, I elaborate on the independent variables I employ and provide descriptive statistics. The aim of this chapter is to clarify my data used and provide an overview of my procedures for truncating data and the relationships between selected variables.

This chapter is divided into three sections; (1) "Datasets employed", (2) "Explanatory variables" and (3) "Descriptive statistics".

(1) "Datasets employed": In this section, I seek to clarify on my data used. I provide sources for my data, and explain the journey from two datasets including raw data towards a final truncated dataset fulfilling my needs for developing BFP models. This section includes (i) elaboration on the procedure applied when matching the event of bankruptcy with annual accounts, including a discussion of the time lag I observe between the latest available annual account and the event of filing for bankruptcy. (ii) Preliminary words on data availability, including a discussion of data availability for small companies. (iii) Explanation of datasets. (iv) An elaboration of the journey from raw data towards truncated data. (v) Data validation, where I compare my data with external sources, aiming for validating the reliability and extensiveness of my data.

(2) "Explanatory variables": In this section, I explain the Danish bankruptcy procedure and elaborate on the initial input variables for model development. This section includes (i) a discussion of financial ratios, where I seek to cover the financial profile of a company and (ii) a discussion of the procedure applied in model development.

(3) "Descriptive statistics": In this section, I provide descriptive statistics prior to model development. This information provides an initial idea of the determinants of bankruptcy. I show that the mean of financials are in-line with expectations for failed vs. non-failed companies respectively.

4.1 Datasets employed

In the following, I (1) describe my procedure for matching company statements with the event of bankruptcy, (2) provide preliminary words on data availability for Danish companies, (3) explain my datasets employed, (4) describe the procedure applied for truncating data into a clean dataset and (5) validate my data with external sources.

Initially I have two datasets; (1) a dataset with annual accounts for Danish companies for the period 2003-2012 and (2) a dataset with bankruptcy information for the period September 2003 – December 2014. Both datasets include unique company identifiers (CVR-numbers) that enables me to merge the two datasets.

The initially step was to merge the two datasets into one dataset.

The first dataset with accounting data is obtained from the Orbis database. The second dataset with bankruptcy data is obtained from konkurs.dk.

4.1.1 Matching bankruptcy with annual accounts

The dataset with bankruptcy data includes CVR-numbers and dates for "filing for bankruptcy". The dataset with company accounts includes CVR-numbers and a broad range of financials.

The latest available company accounts are the last information available for outsiders for determining the financial health of a given company, and ultimately the probability of bankruptcy. It is assumed that the financials of the latest available company accounts include information that should reveal the lack of financial health of a company. On this basis, I (1) apply a matching procedure that ensures that the information in the company accounts is available before bankruptcy and (2) matches the event of bankruptcy with the latest available annual accounts. Danish companies are required to file annual accounts no later than five months after fiscal year end (Erhvervsstyrelsen 2016a). I lead annual accounts by a minimum of six months.

The dependent variable takes the value 1 if two requirements are met; (1) the company filed for bankruptcy and (2) fiscal year ends at least 6 months prior to filing for bankruptcy. The dependent variable is computed as 0 otherwise.

Company accounts prior to the matched company accounts are considered non-bankruptcy, inline with (Lennox 1999, Shumway 2001). The predicted probability of default is in reality a probability of default in any future and not within a specific time frame (See e.g. Bellovary et al. (2007))[34], dependent on the financials of company statements.

[34] Bellovary et al. (2007) provide a review of previous studies, including model accuracy by "year before failure". Most prior studies predict business failure one year prior to bankruptcy, and many report predictive success up to five years prior to bankruptcy

Addressing time lag between latest available company accounts and filing for bankruptcy

Table 6: time lag between company accounts and filing for bankruptcy

Matched by time difference		
Interval lag (years)	Percent matched	Cum.
0,5-1,5	10,32%	10,32%
1,5-2,5	51,26%	61,58%
2,5-3,5	35,38%	96,96%
3,5-4,5	2,59%	99,55%
4,5-5,5	0,45%	100,00%

Estimated bankruptcies available for hold-out sample validation				
2008	2009	2010	2011	2012
6 years*	5 years*	4 years*	3 years*	2 years*
				35,95%
			79,27%	
		98,26%		
	99,78%			
100,0%				

** post years of bankruptcy data available*

Source: Rawdata

Table 6 shows the distribution of the time lag from the fiscal year end of latest available company accounts to the date of filing for bankruptcy. An "interval lag" = "0,5-1,5" of 10,32% means that 10,32% of matched bankruptcies file for bankruptcy 0,5-1,5 years after the latest available company accounts. From my data I notice that the lag between the latest available annual accounts and filing for bankruptcy is often longer than the expectation of 0,5 – 1,5 years. I am not aware of any studies observing such a considerable time lag[35]. I find that companies under reorganization proceedings may postpone filing of company accounts to one month after finalizing reorganization proceedings (Erhvervsstyrelsen 2016a). This may be explanation for the time lag I observe, between latest available company accounts and filing for bankruptcy.

The right table in table 6 summarizes the estimated bankruptcies available for holdout sample validation. I estimate that only 79% and 36% of bankruptcies are included in 2011 and 2012 respectively. Annual reports for 2012 are matched with bankruptcy filings that occur maximum two years after the fiscal year end, as the dataset including bankruptcy data includes 2014 as last year of observations.

> Example: estimated bankruptcy availability of 36% for 2012 is estimated by: $10{,}32\% + 0{,}5 * 51{,}26\%$. This is, I estimate that only 36% of bankruptcies, which should have been matched with 2012 company accounts, are computed as bankruptcy. This implies that 74% of bankruptcies related to annual reports for 2012, are not included in the estimation.

[35] (Lennox 1999) observes average time lag of 14 months,

Figure 2: Graphic illustration of bankruptcy availability

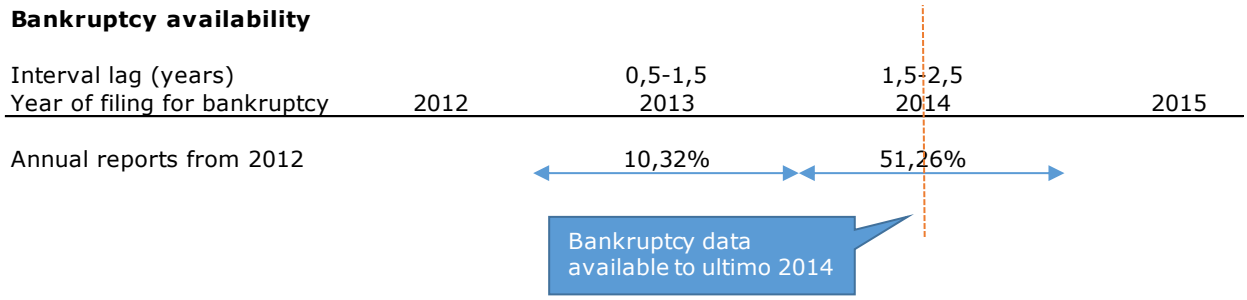


Figure 3 aims to graphically present the example.

This causes complications for the last years of the dataset. Assuming table 6 pictures the normal distribution of time lag between latest annual report and filing for bankruptcy for Danish companies, I am missing bankruptcy information for the last annual accounts of my dataset. Information of annual accounts goes to 2012. Information of bankruptcies goes to 2014. This implies maximum time lag of 2 years. Almost 40% of companies file for bankruptcy more than 2,5 years after the latest available annual accounts. This means that potentially many annual accounts from 2012 are not matched with the event of bankruptcy, if they file for bankruptcy more than two years after fiscal year end. This is, they are computed as non-bankrupt, albeit these company accounts potentially are the latest company accounts prior to filing for bankruptcy. If these years are included in the holdout sample, my models are predicting an event of which I do not have sufficient information.

Table 7: Computation of dependent variable - example with missing information

Filing for bankruptcy 01.07.2015

Fiscal year	Variable	
2007	0	
2008	0	
2009	0	
2010	0	
2011	0	
2012	0	(latest available observation)

Filing for bankruptcy 01.07.2011

Fiscal year	Variable	
2007	0	
2008	0	
2009	0	
2010	1	(latest available observation)
2011	n.a.	
2012	n.a.	

The right table in example 7 shows the matching procedure, where annual accounts prior to "latest available annual accounts before bankruptcy" are computed as zero, and the matched annual account is computed as one.

The left table in example 7 shows the complications related to the extensive time lag between "latest available annual accounts" and the event of bankruptcy. This example is hypothetical, where a company files for bankruptcy after 2014, i.e. the event of bankruptcy is not included in the dataset. If this data was to be

included into the holdout dataset, assuming the latest available annual accounts from 2012, and the company files for bankruptcy after 2014, it is still computed as a zero, as the bankruptcy filing is not known.

On this basis, I do not include the two last years of the dataset, 2011 and 2012, in holdout sample. In chapter 5.2.6: "ΔTC over time in holdout application" I show the impact on predictive success for the years 2011 and 2012.

4.1.2 Preliminary words on data availability

My initial dataset is massive, including just less than 2 million firm years (observations). Albeit the size of the dataset is substantial, the dataset includes numerous missing observations. This is because the legal requirements are heterogeneous for different firm categories.

Table 8: Accounting classes in Denmark

Accounting class	Brief explanation	Catogory distribution	Regulatory requirements
A	Personally held firms (non-limited firms)		
B*	Small, limited firms	Balance ≤ DKK 36m, revenue ≤ 72m, employees ≤ 50	Growing with company category and size
C1*	Medium sized category C, limited firms	Balance > DKK 36m, revenue > 72m, employees > 50	
C2*	Large sized category C, limited firms	Balance > DKK 143m, revenue > 286m, employees > 250	
D	Listed companies and governmental A/S		

Source: e-conomic.dk
** included in sample*

For a company to shift up to a higher company class it is required that the company has exceeded at least two of the three size restrictions for at least two consecutive years (e-conomic.dk 2016).

Table 8 shows the Danish accounting classes. The regulatory requirements are growing with the accounting category, i.e. a firm in class C2 face higher disclosure requirements than that of a firm in class A.

This affects my ability of including variables. The most significant regulatory impacts are (1) A cash flow statement is voluntary for class B firms (Elling 2008) and (2) revenue disclosure is not a requirement for class B and class C1 firms, as it is allowed to summarize revenue with other accounts, i.e. it is allowed to only disclosure gross profit (FSR 2012). This is, I cannot expect data to completely include revenue nor cash flow statements.

Companies of accounting classes B, C1 and C2 are included in my dataset.

4.1.3 Dataset explained

The merged dataset contains almost 2 million firm years (observations). The dataset is extensive. Most other studies are employing a much smaller dataset on primarily listed companies.

Gepp, Kumar (2008) suggests that it would be valuable research to apply the Cox model (a hazard technique) to a large dataset and compare it to logit and MDA. I apply another hazard technique that is similar to the model applied by Shumway (2001) to a large dataset and compare it to a logit model.

After merging my two initial datasets with company accounts and bankruptcy information, I am left with a raw dataset. This merged dataset I address as "Rawdata". This dataset contains all information on non-listed companies from the Orbis database over a 10-year period merged with bankruptcy data from konkurs.dk. This information is raw and unfiltered data. I apply several adjustments and truncations to the dataset in order to achieve a clean dataset that fulfills my requirements for variables that are necessary for deriving my models. I set up multiple criteria. After truncating Rawdata I end up with a new dataset, "Cleandata". Furthermore, Cleandata is divided into two sub-datasets; (1) "Cleandata0307" and (2) "Cleandata0810". Cleandata0307 is my estimation sample and Cleandata0810 is my holdout sample that I use for validation purposes. Variables for the years 2011 and 2012 are excluded due to lack of bankruptcy information[36].

Figure 3: Graphical presentation of Rawdata, Cleandata and two subsamples

4.1.4 From Rawdata to Cleandata

Going from Rawdata to Cleandata is a systematically journey involving several steps.

First step was to determine key variables that are crucial to include in the models. Crucial variables are to be elaborated in chapter 4.2: "Explanatory variables". I apply a complete data criterion to the dataset[37]. I apply the following rule; if at least one observation is missing, I exclude all firm years for the company.

The variables included in the complete data criterion include: current assets, total assets, total equity, current liabilities, non-current liabilities, EBIT, net income, retained earnings (approximated as difference between total equity and share capital) and financial expenditures. Listed companies are not included in Rawdata, nor

[36] See chapter 4.1.1: "Matching bankruptcy with company accounts"
[37] I include only observations that fit my needs. See chapter 3.1.3: "Sampling methods" for explanation of clean data criterion and the possible biases related.

Cleandata. After this procedure, the dataset includes non-listed companies, where variables mentioned are available for all firm years.

After applying the complete data criterion, I generate financial ratios according to chapter 4.2: "Explanatory variables".

Extreme values

I observe that some of the ratios generated show extreme values. Previous academic articles apply winsorizing procedures, and drop variables at the 1% and 99% level, respectively (see e.g. Shumway (2001) and Beaver et al. (2005)).

According to chapter 3.4.2: "Conditional probability measures" logit/probit models are sensitive to outliers.

On this basis, I determine the 1st and 99th percentile respectively for all ratios generated. I do not drop these observations, but transform them.

> Example: If "Net income to total assets" is less than 1% percentile (-2,3) then the observation is transformed to equal the 1% percentile (-2,3).

The example shows the transformation applied. It is assumed that the information in these extreme values do not include significant information for model estimation. However, I do not want to exclude these observations. The solution is the transformation approach that I apply.

Figure 4: Graphic illustration of the road from Rawdata to Cleandata

Figure 5 summarizes the journey from Rawdata to Cleandata.

Mean differences

The Cleandata dataset includes 691.363 firm years with 95.021 unique CVR-numbers and 10.273 bankruptcies. This is divided into two subsamples. Table 9 illustrates the difference in mean values of Rawdata and Cleandata respectively. Table 9 also shows the difference in mean values of my two subsamples Cleandata0307 and Cleandata0810 respectively.

Table 9: Differences in means: Rawdata vs. Cleandata and Cleandata0307 vs. Cleandata0810

	Mean				Mean			
Entry	**Rawdata**	**Cleandata**	**delta**	**P-value***	**Cleandata 0307**	**Cleandata 0810**	**delta**	**P-value***
Total assets (total balance sheet), DKKm	98	62	-36%	0%	61	62	2%	68%
Tangible fixed assets, DKKm	28	39	39%	0%	38	40	5%	44%
Intangible fixed assets, DKKm	3	4	48%	0%	3	5	42%	7%
Short term debt, total, DKKm	17	19	11%	29%	19	19	1%	74%
Long term debt, total, DKKm	16	19	25%	1%	19	20	3%	81%
Equity, total, DKKm	23	23	3%	51%	22	23	5%	36%
Total liabilities (total balance sheet), DKKm	98	62	-37%	0%	61	62	2%	68%
EBIT (earnings before interest and tax), DKKm	2	3	99%	0%	3	3	-18%	7%
EAT (earnings after tax), DKKm	2	2	40%	0%	3	1	-59%	0%
Number of employees	50	60	20%	7%	49	62	25%	19%
Bankruptcies	27.602	10.273	-63%		4.130	4.054		
Firm years	1.956.073	691.356	-65%		319.632	232.589		
Unique CVR numbers	302.392	95.022	-69%		84.139	81.624		
Bankruptcy frequency (bankruptcies / firm years)	**1,41%**	**1,49%**	0,07%p		**1,29%**	**1,74%**	0,45%p	

** ttest of equal means, assuming unequal variances. H0: equal means*

I emphasize that companies going bankrupt in Cleandata0810 are still included in Cleandata0307 as non-bankrupt companies[38].

I observe that by implementing a clean data criterion I significantly change the mean of the variables total assets, tangible fixed assets, intangible fixed assets, long term debt, total liabilities, EBIT and EAT. I apply a significance level of 5% (see column 5 "P-value"). After applying a clean data criterion my estimation sample has significantly changed from Rawdata. However, I argue that the dataset is still economy-wide and its coverage is superior to comparable studies. Ohlson's O-score was developed on a dataset containing industrial companies. Yet, the model is applied as a general model in other contexts than just BFP (Griffin, Lemmon 2002). Later in this chapter I provide a comparison of the size of my estimation sample vs. comparable studies.

Furthermore, I observe that the two samples Cleandata0307 and Cleandata0810 do not show significantly difference in means, on 5% level, on all entries but "Earnings after tax". Cleandata0810 include post crisis firm years and as anticipated average earnings have dropped and bankruptcy frequency has increased. Previous research including Richardson, Davidson (1984) and Barnes (1987) find that accounting ratios are unstable over time. Modern statistics, including MDA and logit models require stable relationship among variables over time, and thus *"... the relationship in future samples of companies, which are to be classified by the model, are the same as in the estimation samples of the model"* (Balcaen, Ooghe 2006). However, my analysis show that companies included in my two periods show equivalence on the majority of listed entries and thus models developed on Cleandata0307 are applicable on Cleandata0810. A holdout application of my models show predictive ability of my models developed[39].

[38] Matching procedure outlined in chapter 4.1.1: "Matching bankruptcies with annual accounts".
[39] See chapter 5.2: "Holdout sample application".

Estimation sample superiority

Cleandata0307, my estimation sample, includes 319.632 firm years with 84.139 unique CVR-numbers and 4.130 bankruptcies. The number of firm years included in my estimation sample is superior to comparable studies. According to chapter 4.1.5: "Validating data" my truncated estimation sample covers 22% of all active Danish companies. I appropriately assume that coverage is economy-wide.

Figure 5: Firm years in estimation sample relative to comparable studies

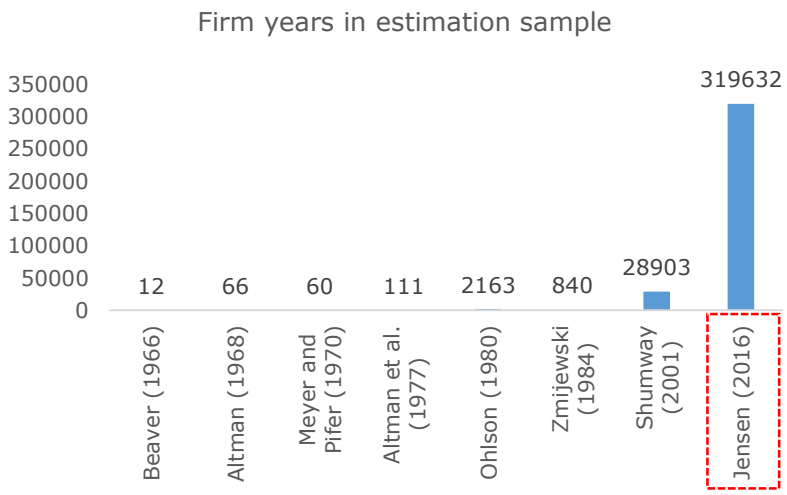


Figure 6 compares the firm years included in my estimation sample with selected prominent and highly cited studies. Even after truncating and applying a complete data criterion, my estimation sample is huge.

4.1.5 Validating data

In order to validate the extensiveness of my dataset and the reliability of bankruptcy information, I compare my dataset to data from DST (Statistics Denmark). In the following, I (1) validate the number of bankruptcies included in my dataset (obtained from konkurs.dk) with data from DST and (2) estimate the share of companies included in my sample, by comparing the number of companies included in my samples with the total number of Danish companies.

Validating bankruptcy data

Bankruptcy data is obtained from konkurs.dk. I compare data from konkurs.dk with data from DST in order to validate the reliability of the data. From konkurs.dk I am able to extract CVR-numbers and dates of filing for bankruptcy on company level. From DST I am able to extract only the total number of bankrupt companies per month. I summarize observations from konkurs.dk and compare them to total numbers extracted from DST.

Figure 6: number of bankruptcies from konkurs.dk and DST

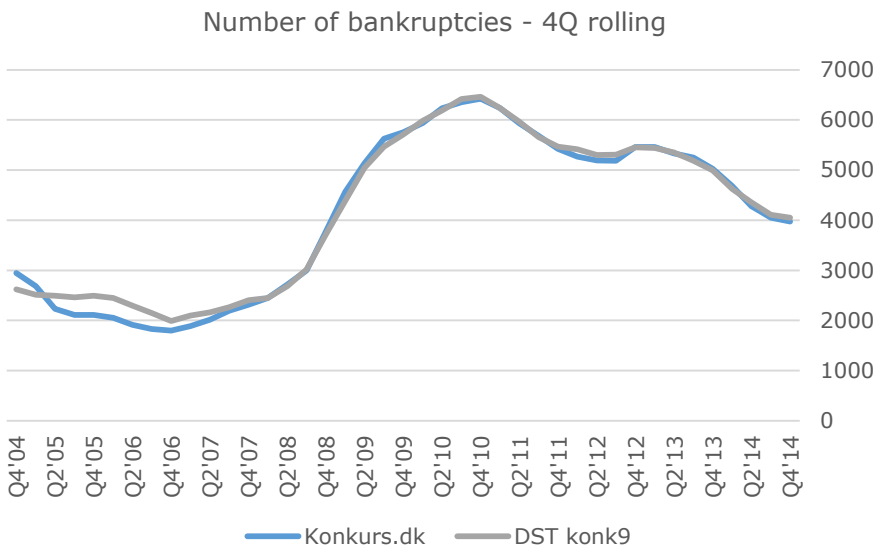


Figure 7 pictures the number of bankruptcies over time. I note that a few data points from konkurs.dk are missing in the beginning of the period. However, I conclude that the data from konkurs.dk is showing a true and fair view of the Danish bankruptcies, and hereby validate the reliability of the bankruptcy data.

Share of Danish companies included in samples

I have just validated the bankruptcy data. Next step is to determine the coverage of my datasets.

Table 10: Dataset coverage

Company coverage*	2003-2007	2008-2010	2003-2012
DST	291.220	301.890	296.398
Rawdata	157.452	238.372	194.505
Cleandata	63.926	77.530	69.135
Rawdata vs DST	54%	79%	66%
Cleandata vs. DST	22%	26%	23%

** Average companies over period*

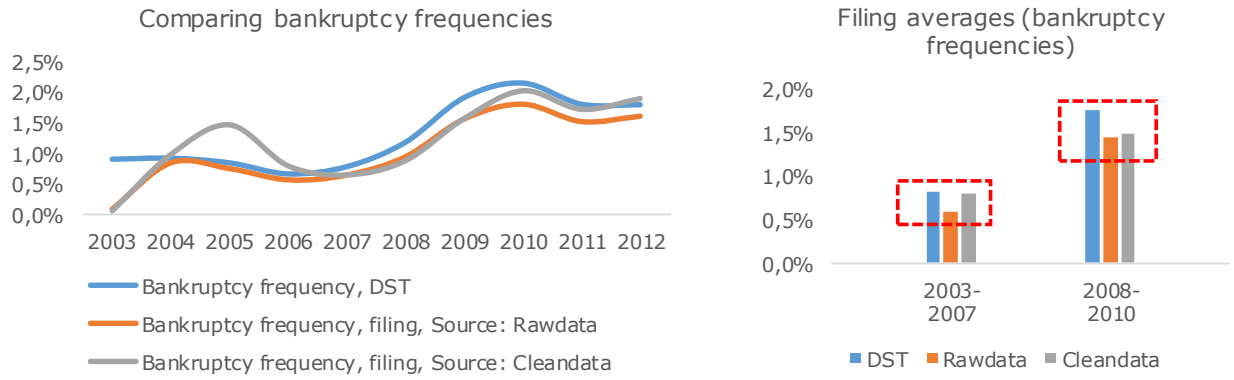
Table 10 put the extensiveness of my datasets in perspective. Rawdata and Cleandata are covering 66% and 23% of all Danish companies respectively. The scope of economy coverage is extensive. Cleandata covers more than a fifth of all active companies in Denmark. Number of companies from DST are defined as "active companies"[40]. According to chapter 4.1.4: "From Rawdata to Cleandata" I find that my estimation sample is huge relative to comparable studies. My data availability enables me to develop a model applicable for the whole economy.

[40] Defined as "activity higher than 'hobby activity' in regards to revenue generation or more than 0,5 full time employees per year" (Danish Statistics 2016), free translation

Comparing bankruptcy frequencies

From figure 8, I see that bankruptcy frequencies for DST and my datasets are not completely equal. However, they are mostly in line. On this basis, I conclude that my datasets (my samples) are well mirroring the total population, i.e. I avoid oversampling bias. I emphasize that I do not arbitrary match failed companies with non-failed companies.

Figure 7: Comparing bankruptcy frequencies over time and with external sources



4.2 Explanatory variables

This chapter seeks to uncover input variables to predict bankruptcy. In the following I (i) explain bankruptcy, including the bankruptcy process and initial thoughts on bankruptcy determinants, (ii) elaborate on input variables employed, (iii) determine initial financial ratios to be included in model development and (iv) outline the model development procedure.

4.2.1 Bankruptcy explained

In this section, I elaborate on my input variables for my models.

The bankruptcy process

In the following, I elaborate on the bankruptcy process. The objective is to create an initial understanding of bankruptcy prior to input variables generation.

The undesirable event of bankruptcy is due to the inability of paying obligations when due.

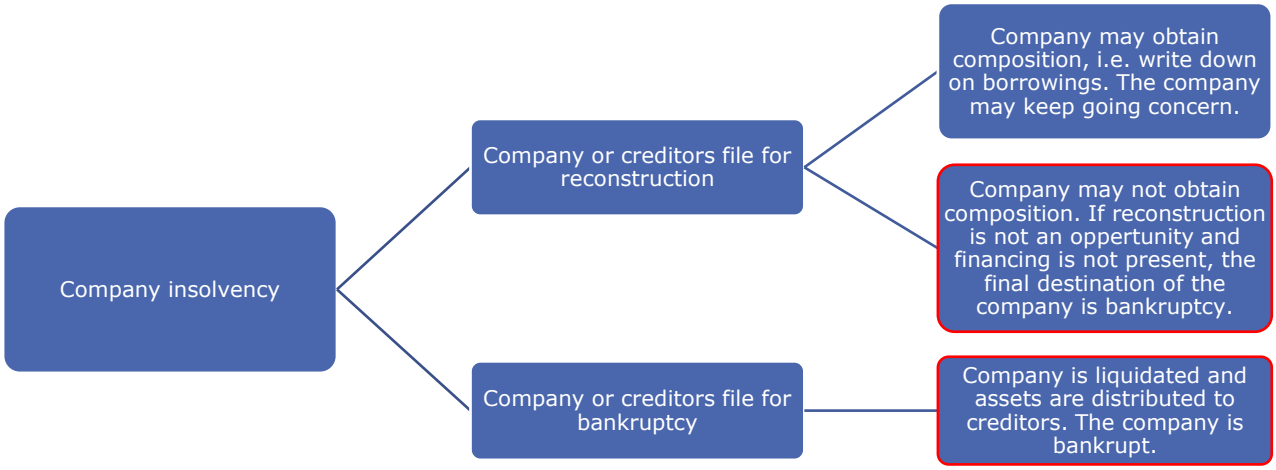
An unavoidable phrase when talking bankruptcy is insolvency. In the literature two types of insolvencies occur; (1) inability to pay obligations when due (which is closely connected to bankruptcy) and (2) when liabilities exceed total assets (i.e. negative equity). This second definition of insolvency does not necessarily

mean bankruptcy. Many technology firms or pharmaceutical companies survive for years, with liabilities exceeding total assets, and are not considered to be in lack of money of close to bankruptcy. The reason is that such firms have unrecognized intangible assets, such as the expected economic value flowing from its research and development activities. A simple measure such as insolvency is thus not sufficient for predicting bankruptcy, but it might provide an idea of the determinants of bankruptcy (Beaver et al. 2011).

The company or the creditors of the company are able to file for bankruptcy. A prerequisite of filing for bankruptcy is an event, where the company is not able to meet financial obligations when due, and that the inability of paying financial obligations when due, is not temporary (domstol.dk 2011). Bankruptcy leads to liquidation of the company, where assets are distributed to the creditors of the company.

Another related term to bankruptcy is "reconstruction". Reconstruction is also related to insolvency. The company or the creditors of the company are able to file for reconstruction. A reconstruction process may lead to either (1) obtain composition or (2) dismantling of operations (bankruptcy) (domstol.dk 2015).

Figure 8: Bankruptcy and reconstruction explained



In my dataset, the dependent variable equals one, if the company or the creditors of the company have filed for bankruptcy. Data on companies filing for reconstruction is not included into the data.

I emphasize that figure 9 is a rough simplification. The regulatory framework covering financial difficulties is extensive and out of scope of this paper.

The determinants of bankruptcy

The unpleasant journey towards bankruptcy is a process over time.

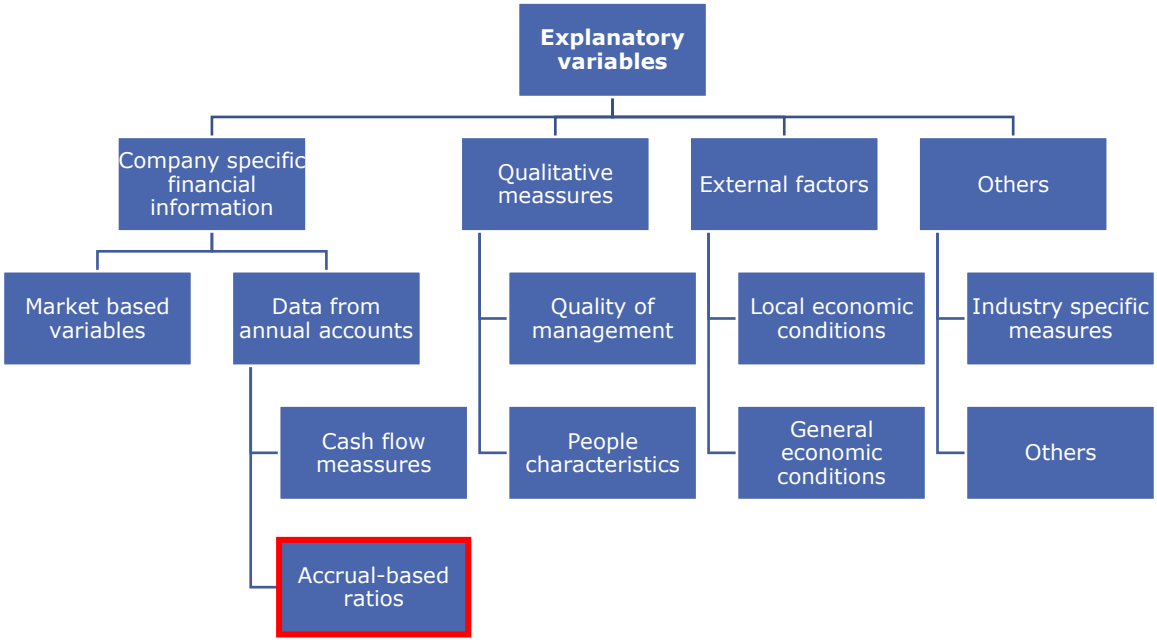
> *"Earliest symptoms of failure may be poor profitability or too fast growth compared to profitability. This leads the company to suffer from poor revenue financing which forces it to get indebted. (...) The failure process means that the firm will get involved with a vicious circle. The high indebtedness brings more financial obligations which must be paid. Poor revenue financing forces the firm to take more and more debt to pay these obligations, until they become superior."* (Luoma, Laitinen 1991)

Bankruptcy is a result of insolvency. Insolvency arises due to lack of sufficient cash. The lack of sufficient cash may arise due to lack of revenue growth financing, over-leveraging or poor profitability. Determinants of such measures include financial leverage, liquidity and profitability.

4.2.2 Accrual based accounting measures

Throughout the literature, numerous and creative inputs for statistical models are applied. Figure 10 provides an overview.

Figure 9: input variables - an overview



Source: own compilation

The vast majority of models use financial information as input for the model (Adnan Aziz, Dar 2006, Balcaen, Ooghe 2006, Appiah et al. 2015). According to Bellovary et al. (2007) 10 of the 11 most applied variables in previous studies are accrual-based measures. The prominent and highly cited studies by Altman (1968), Zmijewski (1984) and Beaver et al. (2005) apply input variables with data from annual accounts, which evidently show predictive success. In general, data from annual accounts is standard input variables

for BFP modelling. The reasons for using financial ratios are that they are (1) "hard" objective measures and (2) based on publicly available information (Balcaen, Ooghe 2006).

However, several studies apply other variables including external factors (for example BNP development and other economic indicators, including Bonfim (2009)), market based variables (see e.g. Shumway (2001), Agarwal, Taffler (2008) and Hoque et al. (2013)) or industry dummies (see e.g. Chava, Jarrow (2004)). Several researchers develop industry specific models, e.g. Altman (1968), who developed a model for manufacturing companies.

Beaver et al. (2005) conclude that the predictive ability of financial ratios has been slightly declining over time, due to increased discretion or other secular changes. They find that including market-based variables offset the slight decline in financial ratios. According to the study, market-based measures include a wide mix of information, also non-financial information. Previous studies employing a mix of financial ratios and market-based measures, conclude that market-based ratios add incremental information to the model (Shumway 2001, Hillegeist et al. 2004, Beaver et al. 2005). However, market-based measures are simply not available for non-listed companies.

The employment of cash flow variables has shown a mixed evidence (Balcaen, Ooghe 2006). Proponents of cash flow measures in BFP include Gombola, Ketz (1983), Gentry et al. (1985), Gentry et al. (1987), Aziz, Lawson (1989) and Sharma, Iselin (2003). Opponents of cash flow measures in BFP include Gentry et al. (1985), who found that *"cash flow from operations does not improve the classification results of failed and non-failed companies"*, Gombola et al. (1987), who found that cash flow from operations is not a significant predictor of bankruptcy and Aziz et al. (1988), who found that a cash flow model and two accrual-based models yield similar performances.

Albeit the findings that information from non-financial-statement might improve the predictability power, I am not in possession of such data.

A study by Bellovary et al. (2007) indicate that accrual based measures are the most applied measures in BFP, and employing only accrual based measures has evidently proved predictive ability. Additionally, *"it is well established that financial ratios do have predictive power up to at least 5 years prior to bankruptcy"* (Beaver et al. 2005)

My dataset contains data from company accounts, but lags cash flow measures. I include only non-listed companies and obviously market-based measures are not available. Accrual-based accounting data has shown predictive success for BFP. I apply accrual-based measures and find great predictive success when applied on my holdout sample.

4.2.3 Financial ratios

4.2.3.1 Which ratios to include

Several review studies including Balcaen, Ooghe (2006) and Appiah et al. (2015) discuss the problems related to the determination of input variables. Appiah et al. (2015) observe that 95% of previous studies are based on ad hoc selection of variables through statistical techniques. The criticism is that model estimation is based on empiricism due in part to the lack of real economic theory in identifying variables. Input variables are often arbitrary selected based on their popularity in literature and their predictive success in previous research (Balcaen, Ooghe 2006).

The determination of the final mix of input variables is a sport itself. The final mix of input variables are uncounted. However, the final mix of financial ratios seems to be of minor importance, as the explanatory variables are highly correlated (Beaver et al. 2005, Beaver et al. 2011). The study of Beaver et al. (2005) finds that a linear combination of ROA (return on assets), ETL (EBITDA to total liabilities) and LTA (total liabilities to total assets) capture essentially all of the explanatory power of the financial statement variables. They conclude that these three variables capture three key elements of the financial strength of a firm; profitability, cash flow generation (EBITDA as a proxy for cash flow) relative to debt levels and financial gearing.

Bellovary et al. (2007) find that the number of variables used in previous studies has been stable over time around 8-10.

Included in the literature review by Bellovary et al. (2007) is a study of variables used in previous studies. Table 11 shows the findings;

Table 11: Factors applied in previous studies

Factor	Number of studies that include	Able to include from my data
Net income / Total assets	54	x
Current ratio	51	x
Working capital / Total assets	45	x
Retained earnings / Total assets	42	x
EBIT / total assets	35	x
Sales / Total assets	32	
Quick ratio	30	
Total debt / Total assets	27	x
Current assets / Total assets	26	x
Net incom / Net worth	23	
Total liabilities / Total assets	19	x
Cash / Total assets	18	
Market value of equity / Book value of total	16	
Cash flow from operations / Total assets	15	
Cash flow from operations / Total liabilities	14	
Current liabilities / Total assets	13	x
Cash flow from operations / Total debt	12	
Quick assets / Total assets	11	
Current assets / Sales	10	
EBIT / Interest	10	x

Source: (Bellovary et al. 2007)

Column 3 of table 11 reveals the fact that my dataset is not complete and hence I am limited in which variables I am able to include into my model. Obviously, I am able to find companies where all data is available, but I do not want to truncate my datasets too much. Column 3 is a result of availability assessment. Data availability is to be explained in the following section.

4.2.3.2 Initial inputs

Accounting categories

My approach to determining financial ratios is systematic. I partly apply the approach of Altman, Sabato (2007), where I determine several accounting categories and within these categories determine financial ratios.

I determine five accounting categories; leverage, liquidity, profitability, coverage and other[41]. Financial information from these categories should generate a complete profile of a company's financial health and hence the risk of bankruptcy.

Financial ratios

From the Rawdata dataset, I have assessed the data availability.

[41] The categories leverage, liquidity, profitability and coverage are also used by Altman, Sabato (2007)

Table 12: Availability of financials from Rawdata dataset

Variable	Available observations of total observations	Key variable
Current assets	90%	x
Cash	66%	
Cash flow	66%	
Current liabilities	90%	x
Depreciation	45%	
Fixed assets	90%	
Gross profit	67%	
Financial expenses	77%	
Non-current liabilities	90%	x
EBIT	89%	x
Operational revenue	18%	
EAT	90%	x
EBT	90%	
Share capital	90%	x
Equity total	90%	x
Total assets (total balance)	90%	x
Total liabilities (total balance)	90%	x

Source: Rawdata,

Table 12 provides an overview of the financial availability. Column 3 is my assessment of financials that are key inputs for the model, i.e. are variables that I must include into my model. This assessment is based partly on data availability and partly on financial theory. I emphasize that I bring subjectivity into the model development process.

I note that the percentage in column 2 of table 12 reveals only the availability of a single variable, and not cross-variable availability. By including for example "Current assets" and "Cash" I am not left with 66% of observations, but less.

Table 13: Illustration of cross-variable availability

Observation	Variable	Availability	Variable	Availability	Total availability
1	XX	Yes	YY	Missing	Missing
2	XX	Yes	YY	Yes	Yes
3	XX	Missing	YY	Yes	Missing
4	XX	Yes	YY	Yes	Yes
5	XX	Yes	YY	Yes	Yes
Availability		80%		80%	60%

Table 13 aims to show the effect that the final model includes less total available observations, than availability of a single variable may prescribe. However, table 12 provides a preliminary overview of variable availability.

I observe that "Operational revenue" data is only available for 18% of all observations, and on this basis, the variable is excluded and hence I am not able to generate the ratio "sales to total assets", albeit this ratio is one of the most applied input variables. However, one may argue that "sales / total assets" is industry specific and might create noise when included in a model of general character (Altman 1993). A capital

heavy company, such as a manufacturing company will naturally show lower capital turnover than that of a capital light company, a consultancy company for example, where people is the true asset of the company, but is not recognized on the balance sheet.

"Depreciation" is only available for 45% of all observations, and furthermore several observations are negative for depreciation, which I am not able to explain. On this basis, I exclude depreciation; hence, I am not able to calculate EBITDA.

"Cash flow" is only available for 66% of all observations. After sorting the dataset to include key variables according to table 12, and exclude companies, where key variable observations are not available, only 61% observations include data on "Cash flow". Including cash flow has shown a mixed evidence in previous studies[42] and data availability in my dataset is low. On this basis, I do not include a cash flow measure.

From my truncated dataset, with complete data on selected variables, I generate ratios. I end up with the following preliminary input variables;

Table 14: Preliminary input variables

Accounting category	Ratios examined	Ratios explained
Leverage	tl_ta	Total liabilities / Total assets (logged)
	ebit_tl	EBIT / Total liabilities
Liquidity	nwc_ta	Net working capital / Total assets
	ca_ta	Current assets / Total assets
Profitability	re_ta	Retained earnings / Total assets
	ni_ta	Net income / Total assets
Coverage	ca_cl	Current assets / Current liabilities
	ebit_finexp	EBIT / Financial expenditures
Other	size (ta)	Total assets (logged)
	Time*	Fiscal year minus year of foundation (logged)
	ek_neg	Dummy; 1 if equity is negative, 0 if not

*** *time (age) for hazard models*

I emphasize that I have included two ratios within each accounting category.

"Net income to total assets" and "EBIT to total assets" show correlations of 0,77 and hence the ratio "EBIT to total assets" is not included as a profitability measure, as the information in this ratio is captured by "Net income to total assets".

The variables from table 14 will make up the explanatory variables for the initial model.

4.2.3.3 Variables explained

My aim is to create a comprehensive financial profile for each company. By distributing variables into categories, I assure that all categories of interest are covered. I assess that variables included for model estimation are sufficient, and I expect to find coherence between financial information and business failure.

[42] See chapter 4.2.2: "Accrual based accounting measures"

Leverage

My accounting ratios for leverage include "Total liabilities to total assets" and "EBIT to total liabilities". My thesis is, that a high level of leverage, provides less economic freedom during periods with deteriorating earnings generation. "Total liabilities to total assets" is a measure of financial leverage. This ratio measures the proportion of debt to be repaid relative to the assets of the firm, which are the source for repaying the debt (Beaver et al. 2005). "EBIT to total liabilities" aim to quantify the obligations of the firm relative to earnings generation.

Liquidity

My accounting ratios for liquidity include "Net working capital to total assets" and "Current assets to total assets". The aim of the liquidity measures are obvious. The event of bankruptcy is due to lack of liquidity to pay obligations when due. In a perfect world, I would have included other liquidity measures. However, accrual-based measures have shown mixed evidence for BFP[43]. "Net working capital to total assets" measure the excess short-term liquidity relative to total assets. "Current assets to total assets" measure the proportion of assets that are not fixed. I hypothesize that high proportion of current assets lead to high financial freedom.

Profitability

The reasons for including profitability measures are obvious. I include "Retained earnings to total assets" aiming for the inclusion of cumulative earnings over time. However, this measure is subject to errors. Retained earnings, calculated as the difference between total equity and share capital, are blurred by dividends and impact of fair value adjustments booked directly on the equity balance. "Net income to total assets" and "EBIT to total assets" are highly correlated and are two sides of the same coin. I choose to include only "Net income to total assets".

Coverage

The coverage measure aim for measuring the coverage ability. "Current assets to current liabilities" measure the ability of a company to meet short term obligations. "EBIT to financial expenditures" measure the

[43] See chapter 4.2.2: "Accrual based accounting measures"

interest coverage. I emphasize that a company may still be able to pay its financial expenditures albeit a ratio below one, as the company may generate cash flows greater than EBIT.

Other

Other variables included are other measures that are either (1) not from financial accounts or (2) by-products of financial statements. Size and age are included in model estimation, as *"failing firms tend to be younger and smaller"* (Balcaen, Ooghe 2006). Other studies find that smaller sized firms are more exposed to bankruptcy, including Begley et al. (1996), Beaver et al. (2005) and Bonfim (2009). "ek_neg" is computed as a dummy variable that equals one if equity is negative, and zero otherwise. This variable equals "OENEG" employed by Ohlson (1980)[44].

4.2.4 Model development procedure

4.2.4.1 General model development procedure

In previous sections, I generated financial ratios. In this section, I seek to elaborate on my approach to model development.

My approach is twofold. In previous sections, I determined accounting categories and determined several initial ratios within each category. When developing models I (1) follow the "backward selection" statistical approach (Gepp, Kumar 2008)[45] and exclude ratios that do not show significance on 5% level and (2) exclude ratios that show counter-intuitive signs. A coefficient with counter-intuitivism is a coefficient, where the sign of the coefficient (positive or negative) is not in line with economic theory.

> Example: If a coefficient suggests a positive relationship between a profitability measure and bankruptcy, this would mean that higher profitability relates to higher probability of default, ceteris paribus. This does not make sense.

In holdout application, I find that a model generated with the counter-intuitivism approach yields superior predictive ability compared to a model purely driven by empiricism[46].

[44] Ohlson (1980) computed his variable as: OENEG = 1 if total liabilities exceed total assets, zero otherwise.
[45] *"…as with traditional regression techniques, the best explanatory variables are chosen from a starting set by forward or backward selection methods"* (Gepp, Kumar 2008)
[46] See chapter 5.3: "Further topics on model development"

Figure 10: Process of determining financial ratios for the model

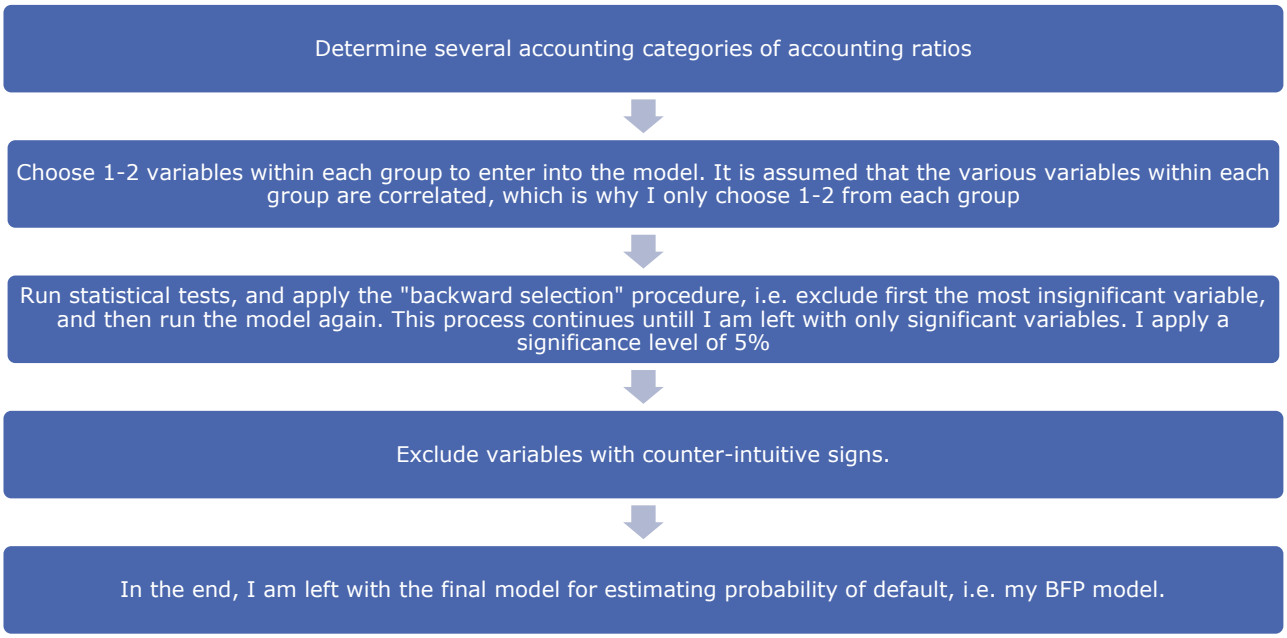


Figure 11 summarizes the model development procedure.

4.2.4.3 Transformation of ratios

Some ratios are skewed. In order to adjust for skewness, I transform into logs. Transformed ratios include "Total liabilities to total assets", "size" (total assets) and "Age" (time). Altman, Sabato (2007) use logarithmic transformations for all their variables and prove that this model is superior to a non-transformed model.

4.3 Descriptive statistics

Mean values

In the following, I picture the differences between financial ratios for bankrupt and non-bankrupt firms respectively. Descriptive statistics have the purpose of creating an understanding of the determinants of bankruptcy, prior to model development.

Table 15: Mean and median values of financial ratios for bankruptcy vs. non-bankruptcy companies

		0*		1*		
Categoty		Mean	Median	Mean	Median	P-value**
Leverage	tl_ta	0,74	0,71	1,32	0,96	0%
	ebit_tl	0,10	0,06	-0,10	-0,03	0%
Liquidity	nwc_ta	0,01	0,04	-0,39	-0,12	0%
	ca_ta	0,51	0,53	0,65	0,74	0%
Profitability	re_ta	0,09	0,17	-0,70	-0,07	0%
	ebit_ta	0,05	0,04	-0,15	-0,03	0%
	ni_ta	0,02	0,04	-0,27	-0,07	0%
Coverage	ca_cl	2,12	1,11	1,32	0,81	0%
	ebit_finexp	3,49	1,39	-0,99	-0,46	0%
Other	Size***	61	5	21	2	0%
	Age****	11,2	8,0	9,1	7,0	0%
	ek_neg	0,10	n.a.	0,43	n.a.	0%

** 1=bankrupt, 0=non-bankrupt*
*** ttest of equal means, assuming unequal variances. H0: equal means*
**** Size estimated by total assets (DKKm)*
***** time (age) for hazard models*

Bankruptcy firms have higher gearing

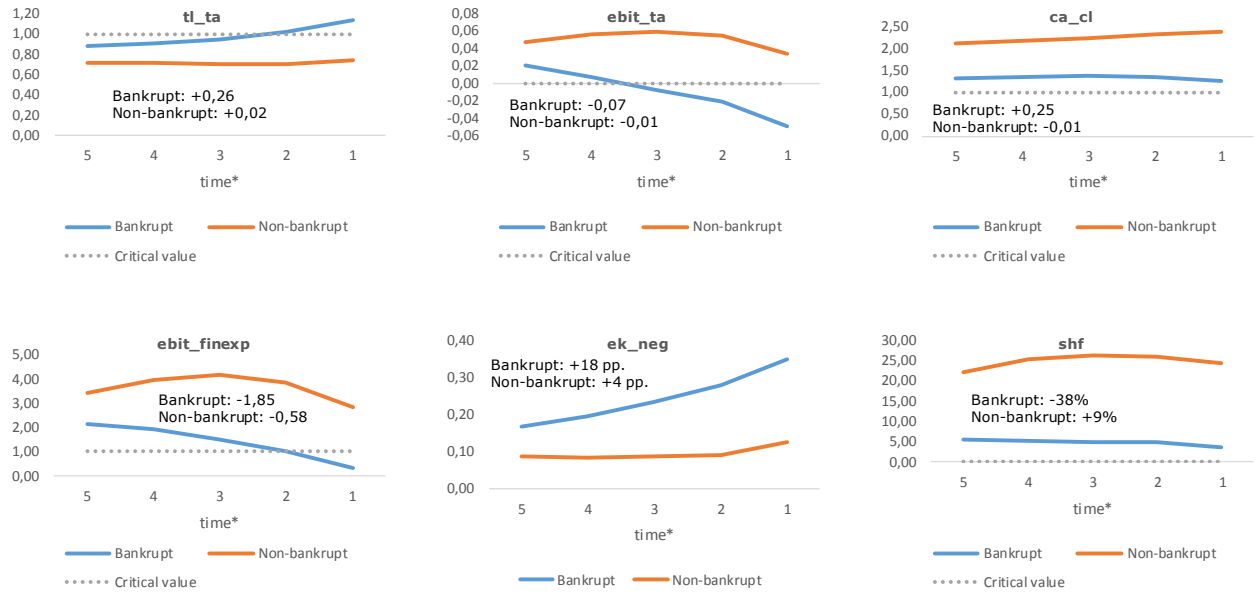
Bankruptcy firms have poorer profitability

Bankruptcy firms have poorer coverage measures

Source: Cleandata0307: estimation sample

I observe that the relationship between mean values for bankrupt and non-bankrupt companies respectively are as expected, but ca_ta. Bankruptcy companies show higher leverage, inferior liquidity (measured by nwc_ta), inferior profitability measures and inferior coverage measures. This is in-line with expectations. ek_neg of 0,43 for bankruptcy firms shows that 43% of bankruptcy firms in the estimation sample had negative equity, relative to 10% for non-bankruptcy firms. All means show significantly difference, according to the t-test (p-values in column 5).

Figure 11: Mean values of financial ratios over time



**time = [time to default] for defaulted companies and [comparable time] for non-defaulted companies*

Note: tl_ta = total liabilities to total assets, ebit_ta = EBIT to total assets, ca_cl = current assets to current liabilities, ebit_finexp = EBIT to financial expenditures, ek_neg = 1 if negative equity (the y-axis shows the percentage of companies with negative equity), shf = equity

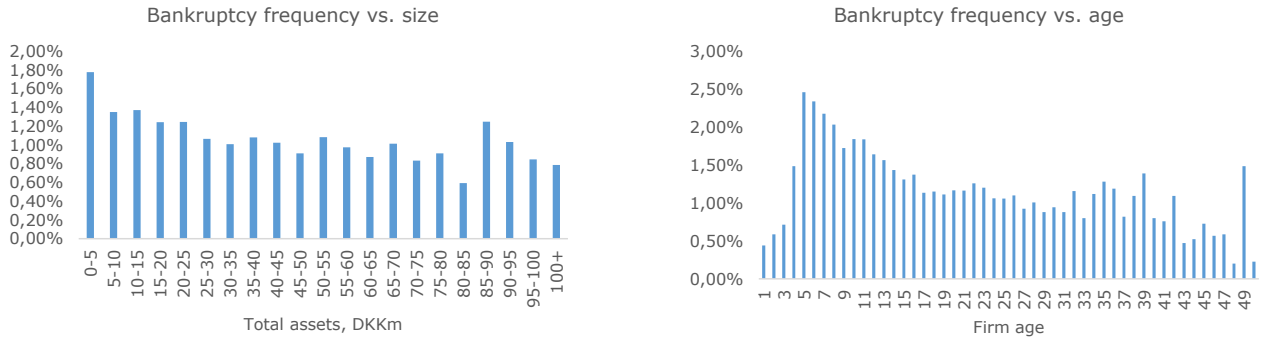
Source: Cleandata: 10 years' truncated data

Figure 12 pictures levels and trends for selected ratios of firms that filed for bankruptcy vs. non-bankruptcy firms. Additionally, the changes from 5 years before default to 1 year before default are provided. From figure 12, I observe that on average, financials are inferior as soon as five years prior to bankruptcy. This is in-line with previous findings[47]. I observe that over the period bankrupt companies have experiences a decrease in equity of 38%, while non-bankrupt companies have experienced an increase in equity of 9%. Furthermore, I observe that "EBIT to total assets" has decreased by 7 percentage points for bankrupt companies, where the decrease was only 1 percentage point for non-bankrupt companies. Overall, I find that financial health for bankrupt companies has been more deteriorating, measured on all variables, compared to non-bankrupt companies. Average fiscal account years are for the period 2004-2008 – the years just before the financial crisis in 2007.

Size and age related to bankruptcy frequency

I relate size and age to bankruptcy frequency.

Figure 12: Bankruptcy frequency vs. size and bankruptcy frequency vs. age



Source: Cleandata: 10 year truncated data

I observe that bankruptcy frequency is decreasing by company size, which equals findings in academia (Begley et al. 1996, Beaver et al. 2005, Balcaen, Ooghe 2006, Bonfim 2009).

The shape of the bankruptcy frequency vs. age looks a bit surprising. It looks that the frequency is increasing in the interval [age=1 to age=5], and then the frequency is decreasing. The development in the frequency rate for firm age 1 to 5 might be explained by the fact that when founding a limited company (ApS, A/S), the founder must put up an initial investment (Erhvervsstyrelsen 2016b)[48]. This initial investment may be sufficient to keep the company running for several years, even if the company is not profitable and does not create positive cash flows. I hypothesize that the accumulative knowledge and learning of a given company

[47] See chapter 4.2.2: "Accrual based accounting measures" and (Beaver et al. 2005)
[48] Capital requirements at registration: DKK 50t (ApS), DKK 500t (A/S)

must be positively correlated with company age. A high level of cumulative learning must negatively correlate with bankruptcy. It seems that given a company has survived for five years; then the company is starting to cumulate previous learning and hence bankruptcy frequency is declining.

Ratio correlations

Including highly correlated explanatory variables in model estimation leads to multicollinearity. Multicollinearity is the scenario when there is a high, but not perfect, correlation between two or more variables. Multicollinearity leads to increased variances of the estimated beta coefficients. This is, it is hard for the statistical program to determine the significance and the coefficient of given explanatory variable, if the variable is highly correlated with one or more other explanatory variables (Wooldridge 2015).

Table 16: Correlation matrix

Correlation matrix	konk_ ones	time*	tl_ta*	ebit_tl	nwc_ ta	ca_ta	re_ta	ebit_ ta	ni_ta	ca_cl	ebit_ finexp	size (ta)*	ek_ neg
konk_ones	1,00												
time	-0,01	1,00											
tl_ta	0,07	-0,11	1,00										
ebit_tl	-0,05	0,05	0,05	1,00									
nwc_ta	-0,08	0,13	-0,52	0,15	1,00								
ca_ta	0,04	0,09	0,01	0,09	0,32	1,00							
re_ta	-0,09	0,08	-0,48	0,14	0,71	-0,10	1,00						
ebit_ta	-0,10	0,07	-0,37	0,67	0,37	0,05	0,40	1,00					
ni_ta	-0,11	0,07	-0,27	0,44	0,48	-0,02	0,61	0,77	1,00				
ca_cl	-0,03	0,09	-0,66	-0,09	0,45	0,23	0,20	0,01	0,09	1,00			
ebit_finexp	-0,07	0,07	-0,03	0,65	0,23	0,02	0,17	0,66	0,42	0,03	1,00		
size (ta)	-0,05	0,21	-0,12	0,07	0,23	-0,14	0,38	0,17	0,25	0,06	0,09	1,00	
ek_neg	0,12	-0,09	0,37	-0,15	-0,50	0,07	-0,51	-0,36	-0,42	-0,14	-0,24	-0,24	1,00

** logged*

Source: Cleandata0307

Table 16 shows the correlation between all explanatory variables. Variables with absolute correlation above 0,5 are highlighted in red. When estimating my models and excluding variables, I keep this in mind.

From table 16 I note that several ratios share the same numerator or denominator. "EBIT to total liabilities" and "EBIT to financial expenditures" both share the same numerator and additionally one may expect an increase in financial liabilities to generate increased financial expenditures. For these ratios I observe a correlation of 0,65. "Net income to total assets" and "EBIT to total assets" both are both measures of profitability and share the same denominator. I observe a correlation of 0,77 on these ratios.

4.4 Summary of data

I elaborate on my datasets and provide a throughout description of my approach to truncating data. Furthermore, I validate the reliability of my datasets. I determine initial explanatory variables to be included in model development, and provide descriptive statistics.

Datasets employed: I find that the size of my estimation sample is huge relative to comparable studies. I validate the reliability of my data with external sources. I notice that by implementing a clean data criterion I significantly change the mean of the majority of variables. Additionally, I observe a significant time lag between latest available company accounts and filing for bankruptcy. I explain this by the fact that companies under reorganization proceedings may postpone filing of company accounts. On this basis, I address the interpretation of the predicted probability of default: The predicted probability of default is in reality a probability of default in any future and not within a specific time frame.

Explanatory variables: I elaborate on the initial inputs for model development. I use only accrual-based ratios and hence implicitly assume that all causes for the event of bankruptcy is fully explained by these measures. I find that accrual-based ratios have evidently shown predictive success in academia. I determine five accounting categories for input variables and on this basis determine the initial inputs for model development. At last, I determine the procedure for model development; backward selection and exclusion of variables with counter-intuitive signs.

Descriptive statistics: I find that annual accounts matched with the event of bankruptcy show significantly higher leverage, inferior liquidity, inferior profitability measures and inferior coverage measures. Furthermore, I find that these differences in mean values are observable as early as five years prior to bankruptcy. I also find that several input variables show high correlation, which I keep in mind during model development.

CHAPTER 5: ANALYSIS

Former chapters laid the foundation for understanding BFP models and measure predictive success. Additionally I determined statistical approaches to be applied and developed a dataset enabling me to develop my own BFP models. I am now ready to pick the fruits. This chapter is the product of all previous findings. This is the chapter where I develop and apply my BFP models.

This chapter is divided into three sections; (1) "Model development", (2) "Holdout sample application" and (3) "Further topics on model development".

(1) "Model development": In this section, I account for expected signs of coefficients and explain my three main models that I develop. I furthermore provide my coefficients and final explanatory variables for my three final models, and provide information on marginal effects.

(2) "Holdout sample application": In this section, I apply my models on a holdout sample and compare results. I provide results from my two approaches; (i) percentile approach and (ii) cutoff approach. I compare the results and also simulate on my underlying assumption of the cost distribution. Furthermore, I connect in-sample results with holdout sample results, provide information on model success over time and evaluate predictive success of my hazard model by accounting class. This chapter includes all results.

(3) "Further topics on model development": In this section, I account for further topics on model development. I show some other approaches that I have applied, and compare them to my results in (2). This section aims to validate my final models and check the robustness of my approaches applied.

5.1 Model development

The procedure for model development is explained in chapter 4.2.4: "Model development procedure". In short, I apply the backward selection procedure as a statistical approach, and leave out explanatory variables, that show counter-intuitive signs.

5.1.1 Expected sign of coefficients

Table 17: Expectations to signs

	Expected correlation to the event of bankruptcy			
	Positive	Negative	Mixed	Variable explained
tl_ta	x			Total liabilities to total assets
ebit_tl		x		EBIT to total liabilities
nwc_ta		x		Net working capital to total assets
ca_ta		x		Current assets to total assets
re_ta		x		Retained earnings to total assets
ni_ta		x		Net income to total assets
ca_cl		x		Current assets to current liabilities
ebit_finexp		x		EBIT to financial expenditures
size (ta)		x		Total assets (proxy for size)
Age			x	Years since foundation
ek_neg	x			Dummy; 1 if equity is negative, 0 if not

Table 17 conceals the expected direction of the coefficients. If the sign of the coefficients are not in-line with the expectations in table 17 they are excluded from the model.

Many of the expected correlations are obvious. Higher profitability (measured as Net income to total assets) should lead to lower probability of bankruptcy. However, some variables need explanation. Size: size on a stand-alone basis shows negative correlation to bankruptcy[49]. This is well in line with the common understanding of previous research, including Begley et al. (1996), Lennox (1999), Hayden (2003), Beaver et al. (2005), Balcaen, Ooghe (2006) and Bonfim (2009). Age: According to chapter 4.3: "Descriptive statistics", the bankruptcy frequency related to age is increasing from age=1 to age=5. From age=5 to age=50+ the bankruptcy frequency is decreasing. On this basis, I do not exclude the variable due to counter-intuitiveness.

5.1.2 Developing three models

I develop three models. The first model, "Logit 5y" is a model that includes all observations from Cleandata0307, i.e. panel data for the years 2003-2007. Including panel data into a logit model is a breach of the underlying assumptions, and I cannot trust the significance statistics. However, I develop this model in order to compare this model with two other models. The second model, "Logit 1y", is a model that include only observations for the year 2007. This model does not breach the assumption related to serial correlation. The third model, "Hazard", is a model that applies a hazard procedure similar to Shumway (2001). This model includes all observations from Cleandata0307, i.e. panel data for the years 2003-2007. This model automatically corrects for the serial correlation[50]. This allows me to include five years of data.

[49] See chapter 4.3: "Descriptive statistics"
[50] See chapter 3.4.3: "Hazard models (survival analysis)"

Table 18: Expectations to model performance

	Years included	Corrects for serial correlation	Expectations to performance*
Logit 5y	2003-2007 (5 years)	No - assumption breached	3
Logit 1y	2007 (1 year)	No (no need to)	2
Hazard	2003-2007 (5 years)	Yes	1

** where 1 = best, 2 = second best, 3 = third best (or poorest)*

Table 18 summarize the three models I develop. According to table 18, I expect Logit 1y to be superior to Logit 5y, as Logit 5y is not statistically valid. I expect Hazard to be superior to Logit 1y, as Hazard includes data for 5 years rather than just one single year. All three models are applied to a holdout sample enabling me to compare the out of sample performance of the models, according to my success rate measure ΔTC[51].

Table 19: Final input variables in three models

	Logit 5y		**Logit 1y**		**Hazard**	
Step	Variable removed	Reason	Variable removed	Reason	Variable removed	Reason
1st	ebit_tl	Statistical	nwc_ta	Statistical	ebit_tl	Statistical
2nd	ca_ta	Intuitive	re_ta	Statistical	size	Statistical
3rd	re_ta	Intuitive	ebit_tl	Statistical	nwc_ta	Statistical
4th	nwc_ta	Intuitive	ca_ta	Intuitive	ca_cl	Statistical
5th	ca_cl	Statistical	ca_cl	Statistical	ca_ta	Intuitive
6th			size	Intuitive	re_ta	Intuitive
7th			ni_ta	Statistical	ni_ta	Statistical
Final explanatory variables included in model	l_tl_ta ebit_finexp ek_neg ni_ta size		l_tl_ta ebit_finexp ek_neg		l_tl_ta ebit_finexp ek_neg l_time	

Table 19 shows the development process, and the variables excluded. The variables "Total liabilities to total assets", "EBIT to financial expenditures" and "ek_neg" (dummy for negative equity) are significant and show expected signs in all three models. I note that Logit 5y include also "Net income to total assets" and "Size" as significant variables, but these variables are excluded in the Logit 1y and Hazard models. This is in-line with the justification that the Logit 5y model over-estimate number of observations (Shumway 2001). A hazard model with the coefficients equal to Logit 5y shows that "Net income to total asssets" and "Size" are not significant at the 5% level. Contrary to the findings of Shumway (2001), I find that time show significance in my Hazard model[52].

Surprisingly, my final models include only 3-5 variables, and no variables measuring profitability are included in the models Logit 1y and Hazard. Initially I was expecting profitability to show significance in my models.

[51] See chapter 3.1.1: "Success rate measurement"
[52] However, (Shumway 2001) defines age as "time listed on stock exchange" and not firm age, as I do

"Profitability is expected to be a critical element, since prior research has shown that capital markets are concerned about the ability of the firm to repay its debts and profitability is a key source of ability to pay" (Beaver et al. 2011)

Shumway (2001) finds "Net income to total assets" to be significant with his hazard model, when applying the technique to the coefficients of Zmijewski (1984). "Current assets to current liabilities" and "Retained earnings to total assets" are excluded in all three models; in-line with the findings of Shumway (2001)[53].

The final models are as follows:

Table 20: Coefficients and p-values for final input variables

	Logit 5y		Logit 1y		Hazard	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
l_tl_ta	0,2987	0,000	0,4535	0,000	0,4612	0,000
ebit_finexp	-0,0672	0,000	-0,0702	0,000	-0,0704	0,000
ek_neg	1,0897	0,000	0,8761	0,000	1,1657	0,000
ni_ta	-0,1580	0,000				
size	-0,0438	0,000				
l_time					0,2286	0,000
cons	-4,0594	0,000	-4,0617	0,000	-5,2105	0,000

From table 20 I observe that all coefficients are highly significant, even at a 1% significance level.

5.1.3 Interpretation of coefficients – marginal effects

In the following, I show marginal effects on coefficients. I emphasize that marginal effects in logit and hazard models are dependent on the value of x, where x is the explanatory variable. These models are non-linear and hence the marginal effects are non-linear as well.

Table 21: Marginal effects

				Logit 5y		Logit 1y		Hazard	
	Median	75% percentile (healthy companies)	25% percentile (non-healthy companies)	P(default) at 75% percentile	ΔP(default) 75% percentile to 25% percentile	P(default) at 75% percentile	ΔP(default) 75% percentile to 25% percentile	P(default) at 75% percentile	ΔP(default) 75% percentile to 25% percentile
tl_ta *	0,71	0,47	0,88	0,61%	0,13%p	0,82%	0,27%p	0,49%	0,16%p
ebit_finexp	1,37	5,58	-0,23	0,61%	0,29%p	0,82%	0,41%p	0,49%	0,24%p
ek_neg **	n.a.	0	1	0,61%	1,62%p	0,82%	1,90%p	0,49%	1,69%p
ni_ta	3,51%	11,0%	-1,0%	0,61%	0,01%p				
size (DKKm)	5	15	2	0,61%	0,06%p				
time (age)	8	16	4					0,49%	-0,13%p

** inverse values, i.e. 75% percentile is 25% percentile and vice versa. This is due to positive correlation to bankruptcy*
*** ek_neg values for are chosen arbitrary; 0 for 75% percentile (healthy companies) and 1 for 25% percentile (non-healthy companies)*

Source: Cleandata0307

[53] Shumway (2001) finds that "Current assets to current liabilities" is insignificant when applying the variables of Zmijewski (1984). Additionally, Shumway (2001) finds that "Retained earnings to total assets" is insignificant when applying the variables of Altman (1968).

Table 21 provides information on marginal effects. The column "75% percentile" show great financials. Companies with financials equal to this column show low financial gearing, healthy coverage, positive equity and positive profitability. The column "25% percentile" show poor financials. Companies with financials equal to this column show high financial gearing, unhealthy coverage, negative equity and negative profitability.

The columns "ΔP(default) 75% percentile to 25% percentile" (marked with blue) show the marginal change in percentage points for the financial, holding other financials constant.

I notice that a change in ek_neg from zero to one, indicating a change from positive equity to negative equity leads to a large change in predicted probability in default for all three models. Note that the change ek_neg from 0 to 1 implies tl_ta going from 0,47 to >1. This is included in the calculations. For the hazard model, holding all other financials equal, but changing ek_neg from zero to one (and tl_ta to 1,000001), yields a change in predicted probability of default of 1,69 percentage points and hence more than triples the predicted probability of default (from 0,49% to 1,69%). Substituting tl_ta to equal 1 for both "75% percentile" and "25% percentile" (in order to find the 'clean' marginal effect of ek_neg), and keep other variables constant, I observe marginal effects for ek_neg of 1,46 percentage points, 1,57 pp. and 1,49 pp. for Logit 5y, Logit 1y and Hazard model respectively. This indicates that this specific change from positive to negative equity has a great influence of probability of default.

According to the hazard model, a company going from tl_ta=0,47 to tl_ta=0,88, holding all other financials constant, yields a change in predicted probability of default of 0,16 percentage points.

It seems that a dummy variable for negative equity, ek_neg, is a strong determinant of bankruptcy. I find it hard to add variables of significance to the model, when checked for negative equity. Albeit this variable was expected to show significant, these findings are somewhat surprising. I must admit that I was expecting more variables to show significance in my final models. I find that 43% of bankrupt companies in my estimation sample show negative equity, and 10% of non-bankrupt companies show negative equity. This supports the argument that negative equity, and hence theoretical insolvency (debt is greater than assets), does not necessary lead to bankruptcy[54]. However, my analysis of marginal effects show that negative equity is a strong determinant in bankruptcy prediction.

In chapter 5.3: "Further topics on model development" I find that simply predicting companies with negative equity to go bankrupt show inferior ΔTC compared to my models.

[54] See e.g. chapter 4.2.1: "Bankruptcy explained"

5.2 Holdout sample application

My three models developed are applied on my holdout sample. This application enables me to validate the model performances in on a secondary sample; my holdout sample. As described in chapter 3.1.1: "Success rate measurement", I (1) use the measure ΔTC as my ultimate success rate and (2) I apply an asymmetric cost distribution, where type I errors are ~18 times more costly than type II errors.

I apply two approaches in distributing companies into either predicted bankruptcy or predicted non-bankruptcy. These two approaches are (1) the "Percentile approach" and (2) the "Cutoff approach"[55].

Additionally to my three models developed (Logit 5y, Logit 1y and Hazard) I also apply the coefficients of the Z''-score[56].

5.2.1 Percentile approach

In the following, I apply the percentile approach. I divide into percentiles with steps of 5 percentage points (5%, 10%, 15% and 20%) per firm year. This is, for every year in holdout sample I rank companies by predicted probability of default and distribute companies into percentiles.

Table 22: Predictive ability, percentile approach

Model	Percentile	Hold-out sample TC	Delta in TC compared to "lend to all"	Average type I success at cut-off	Average type II success at cut-off	Overall predictability rate at cut-off	Average predicted probability of default for defaulted companies *	Average predicted probability of default for non-defaulted companies *	Predictive ratio**
Z''-score***	Top 5%	1,21%	-5,7%	20,52%	95,28%	93,97%	-30	25	n.a.
	Top 10%	1,24%	-3,3%	33,33%	90,41%	89,42%	-17	27	n.a.
	Top 15%	1,30%	1,4%	43,96%	85,51%	84,79%	-12	28	n.a.
	Top 20%	1,39%	7,9%	52,89%	80,58%	80,10%	-8	30	n.a.
Logit 5y model	Top 5%	1,20%	-6,7%	21,39%	95,29%	94,00%	8,16%	1,19%	6,9
	Top 10%	1,17%	-8,9%	38,58%	90,51%	89,60%	6,27%	1,01%	6,2
	Top 15%	1,24%	-3,9%	48,94%	85,60%	84,96%	5,32%	0,87%	6,1
	Top 20%	1,34%	4,2%	56,31%	80,64%	80,22%	4,52%	0,79%	5,7
Logit 1y model	Top 5%	1,19%	-7,3%	22,03%	95,30%	94,03%	8,60%	1,50%	5,7
	Top 10%	1,16%	-10,2%	39,86%	90,53%	89,65%	6,80%	1,31%	5,2
	Top 15%	1,23%	-4,2%	49,21%	85,61%	84,97%	5,88%	1,15%	5,1
	Top 20%	1,33%	3,0%	57,50%	80,67%	80,26%	5,08%	1,05%	4,8
Hazard model	Top 5%	1,16%	-9,5%	24,12%	95,34%	94,10%	5,93%	0,85%	7,0
	Top 10%	1,12%	-13,0%	42,45%	90,58%	89,74%	4,66%	0,71%	6,6
	Top 15%	1,22%	-5,3%	50,30%	85,63%	85,01%	3,94%	0,61%	6,5
	Top 20%	1,31%	2,1%	58,34%	80,68%	80,29%	3,33%	0,55%	6,1

** Z-score for Altman's model*
*** Calculated as ([Average predicted probability of default for defaulted companies] / [Average predicted probability of non-defaulted companies])*
**** percentiles for Z''-score model are in reality bottom percentiles, as a low Z''-score indicates high probability of default*

[55] See chapter 3.1.1: "Success rate measurement"
[56] See chapter 3.4.1: "Multiple discriminant analysis".

Table 22 summarizes the holdout sample results for four models. I see that all of my three models outperform the Z''-score model at optimal percentile-cutoff (marked with red box). I find the optimal percentile cutoff to be the 10th percentile for all three models. As expected, I find that (1) the Hazard model, yielding ΔTC of -13% is superior to the Logit 1y model, yielding ΔTC of -10,2% and (2) the Logit 1y model, yielding ΔTC of -10,2% is superior to the Logit 5y model, yielding ΔTC of -8,9%.

I find the highest predictive success by applying a cutoff equal to the 10th percentile of all my models applied. However, I note that this percentile does not yield the highest overall predictability rate (column7). This is due to the asymmetric cost distribution and the assumption that type I errors are more costly relative to type II costs. By naively forecasting no companies going bankrupt, overall predictability would equal 98,26% per year (i.e. average annually bankruptcy frequency for the period = 1,74%). As argued earlier, the overall predictability rate is not a well-suited performance measure[57].

At cutoff equal to the 10th percentile I find that predictive ratio (column 10) ,which is intended to measure the relative probability of default for failed companies compared to non-failed companies, equals 6,2, 5,2 and 6,6 for Logit 5y, Logit 1y and Hazard respectively at optimal percentiles. The highest ratio is observed with the Hazard model.

If financial ratios had no predictive power, I would expect the fraction of firms (bankrupt and non-bankrupt) in each percentile to equal the percentile (Shumway 2001). With the hazard model, I observe 42% of bankruptcy firms in the 10th percentile. This finding supports the model's predictive ability. With the Z''-score, I observe 33% of bankruptcy firms in the 10th percentile. This model indeed show predictive ability, however the predictive ability is inferior to my Hazard model.

[57] See chapter 3.1.1: "Success rate measurement".

5.2.2 Cutoff approach

In the following, I show the holdout results with the cutoff approach.

Table 23: Predictive ability, cutoff approach

Model	Cut-off	Hold-out sample TC	Delta in TC compared to "lend to all"	Average type I success at cut-off	Average type II success at cut-off	Overall predictability rate at cut-off	Average predicted probability of default for defaulted companies *	Average predicted probability of default for non-defaulted companies *	Predictive ratio**
Z''-score	-4,50	1,21%	-6,2%	25,11%	93,94%	92,74%	-23	24	n.a.
	-4,00	1,21%	-6,0%	26,30%	93,52%	92,35%	-22	24	n.a.
	-3,50	1,21%	-6,0%	27,95%	92,98%	91,85%	-20	24	n.a.
	-3,00	1,22%	-5,5%	29,48%	92,35%	91,25%	-19	25	n.a.
	-2,50	1,23%	-4,7%	31,03%	91,58%	90,53%	-17	25	n.a.
	-2,00	1,24%	-3,8%	32,71%	90,77%	89,76%	-16	25	n.a.
Logit 5y model	1,8%	1,32%	2,5%	54,09%	81,91%	81,43%	4,71%	0,81%	5,8
	2,3%	1,27%	-1,7%	51,23%	84,15%	83,58%	5,08%	0,84%	6,0
	2,8%	1,25%	-3,0%	49,48%	85,16%	84,53%	5,25%	0,86%	6,1
	3,3%	1,22%	-5,6%	45,61%	87,21%	86,48%	5,58%	0,91%	6,1
	3,8%	1,17%	-8,9%	37,79%	90,77%	89,84%	6,34%	1,02%	6,2
	4,3%	1,18%	-8,7%	30,74%	92,94%	91,86%	7,03%	1,09%	6,4
Logit 1y model	2,5%	1,30%	1,0%	53,68%	82,53%	82,03%	5,17%	0,85%	6,1
	3,0%	1,27%	-1,6%	51,21%	84,15%	83,57%	5,31%	0,87%	6,1
	3,5%	1,25%	-3,0%	48,84%	85,36%	84,72%	5,84%	0,95%	6,1
	4,0%	1,21%	-6,3%	45,26%	87,55%	86,81%	6,61%	1,05%	6,3
	4,5%	1,15%	-10,4%	37,64%	91,30%	90,36%	7,26%	1,12%	6,5
	5,0%	1,17%	-9,3%	30,64%	93,18%	92,09%	7,85%	1,17%	6,7
Hazard model	1,7%	1,26%	-2,0%	51,63%	84,13%	83,56%	3,75%	0,58%	6,4
	2,2%	1,22%	-5,3%	49,61%	85,84%	85,21%	3,96%	0,61%	6,5
	2,7%	1,13%	-11,8%	45,24%	89,33%	88,56%	4,45%	0,68%	6,5
	3,2%	1,15%	-10,8%	29,38%	94,07%	92,94%	4,97%	0,75%	6,6
	3,7%	1,16%	-9,5%	23,58%	95,50%	94,25%	5,50%	0,81%	6,8
	4,2%	1,18%	-8,2%	19,26%	96,45%	95,10%	5,99%	0,86%	7,0

Underlined cut-off = midpoint of "average predicted probability" for bankruptcy vs. non-bankruptcy companies respectively, in-sample
** Z-score for Altman's model*
*** Calculated as ([Average predicted probability of default for defaulted companies] / [Average predicted probability of non-defaulted companies])*

Table 23 summarizes the holdout sample results for four models. Similar to the percentile approach I find all my three models outperform the Z''-score at optimal cutoff points. The results of applying the cutoff approach are much similar to applying the percentile approach. The Hazard model still yields the best performance.

5.2.3 Comparison: percentile approach vs. cutoff approach

In the following, I compare the two approaches (percentile and cutoff approach) and the holdout performances.

Figure 13: ΔTC at optimal percentile / cutoff point

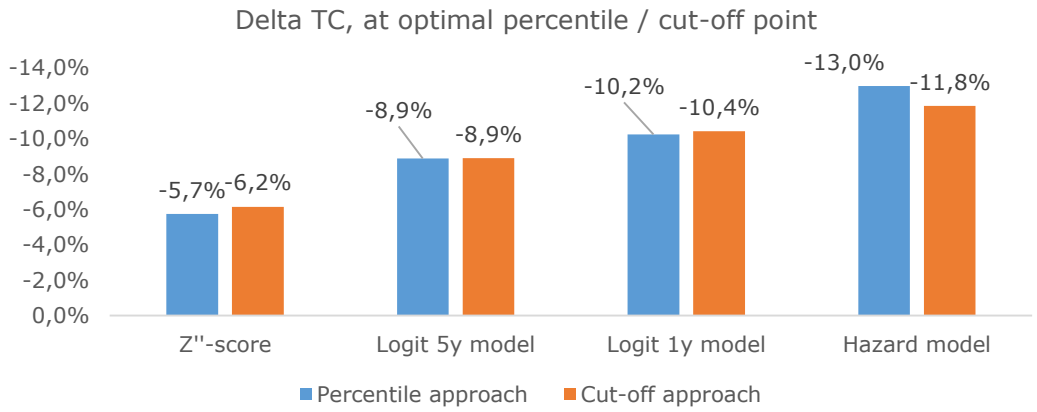


Figure 14 compares ΔTC of the two approaches. I find that the percentile approach performs marginally better relative to the cutoff approach with the hazard approach. By implementing continuous steps (i.e. not steps of 5 percentile points but indefinite small steps and not 0,5 percentage points steps but indefinite small steps for the percentile approach and the cutoff approach respectively) the predictive success would almost equal for both approaches. The percentile approach implies a constant percentile applied for all years in holdout sample, but the cutoff point in predicted probability of default, related to the percentile, is not necessarily constant over time.

I argue for the superiority of the percentile approach. Applying the percentile approach does not force the researcher to determine a cutoff prior to estimation, but allows for varying cutoff points over time. I note that recent studies, including Shumway (2001), Chava, Jarrow (2004) and Altman, Sabato (2007) apply the percentile approach in favor of the cutoff approach. Additionally the researcher gets a feeling of the predictive abilities of models. I find that at the 10th percentile I capture 42% of bankrupt companies.

5.2.4 Simulation on relative costs related to type I and type II errors

In my analysis above, I apply a cost ratio of ~18x. This is, I assume that type I errors are ~18 times more costly than type II errors[58]. In the following, I simulate on the cost ratio of type I errors vs. type II errors, and show the effects on optimal percentile/cutoff and related predictive success.

[58] As estimated in chapter 3.1.1: "Success rate measurement"

Table 24: Simulating on cost function assumptions

			Relative costs (type I / type II)							
Model	**Approach**	**Simulation**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
Z''-score	Percentile	Optimal percentile	5%	5%	10%	15%	20%	30%	30%	30%
		Delta TC	-3%	-7%	-12%	-17%	-22%	-26%	-31%	-35%
	Cut-off	Optimal cut-off	-4,5	-3,5	-3	0,5	1,5	1,5	1,5	1,5
		Delta TC	-2%	-8%	-12%	-17%	-22%	-25%	-28%	-31%
Logit 5y model	Percentile	Optimal percentile	5%	10%	10%	15%	15%	25%	25%	30%
		Delta TC	-4%	-12%	-17%	-22%	-26%	-30%	-34%	-37%
	Cut-off	Optimal cut-off	4,3%	3,8%	3,8%	2,8%	2,3%	1,3%	1,3%	1,3%
		Delta TC	-4%	-12%	-17%	-22%	-26%	-29%	-33%	-36%
Logit 1y model	Percentile	Optimal percentile	5%	10%	10%	15%	25%	25%	30%	30%
		Delta TC	-4%	-13%	-19%	-22%	-27%	-32%	-36%	-40%
	Cut-off	Optimal cut-off	3,2%	2,7%	2,7%	2,7%	2,7%	1,2%	1,2%	1,2%
		Delta TC	-7%	-15%	-21%	-25%	-28%	-31%	-34%	-36%
Hazard model	Percentile	Optimal percentile	10%	10%	10%	10%	10%	25%	30%	30%
		Delta TC	-7%	-16%	-21%	-25%	-27%	-31%	-35%	-39%
	Cut-off	Optimal cut-off	3,2%	2,7%	2,7%	2,7%	2,7%	1,2%	1,2%	1,2%
		Delta TC	-7%	-15%	-21%	-25%	-28%	-31%	-34%	-36%

Relative cost relationship ↑ → Absolute value of "Delta TC" ↑

Source: Cleandata0810: holdout sample

Table 24 summarizes the impact of simulating on the cost distribution assumption. A cost ratio of 20x is highlighted, as this is close to the applied cost ratio of 18x.

From table 24 I find that higher cost ratio (type I costs / type II costs) leads to lower (more negative, i.e. more cost reduction) ΔTC. This is in alignment with expectations. This support the findings that my models are able to discriminate between bankruptcy and non-bankruptcy, which indeed is the major objective of this paper.

Assuming the cost ratio goes towards infinity, would lead to a scenario where costs associated with type II errors go towards zero and costs associated with type I errors go towards infinity. This will lead to higher percentile (lower cutoff). A higher percentile (lower cutoff) will lead to an increase in type II errors, but since the costs associated with type II errors is going towards zero, this does not influence ΔTC calculations. On the contrary, I would observe a decrease in type I errors, which have become very costly. The conclusion is lower ΔTC, i.e. more cost reduction.

This implies that a more asymmetric cost function (higher cost ratio assumption) leads to larger savings of applying my models.

Table 24 allows researchers or practitioners to apply their own assumptions for the cost ratio, and the optimal percentile/cutoff related to this assumption.

5.2.5 Comparison of in-sample and holdout sample results

In the following, I link in-sample results to holdout sample findings.

Table 25: Linking in-sample findings to holdout sample results

Model	In-sample results			Holdout sample results		
	Pseudo R^2	Optimal percentile	Delta TC	Delta TC at percentile equal optimal in-sample	Optimal percentile in hold-out sample	Delta TC at optimal percentile in hold-out sample
Logit 5y model	0,0919	5%	-7,0%	-6,7%	10,0%	-8,9%
Logit 1y model	0,0804	5%	-7,1%	-7,3%	10,0%	-10,2%
Hazard model	0,0913	5%	-7,1%	-9,5%	10,0%	-13,0%

Table 25 summarizes the comparison of in-sample results and holdout sample results. I compare only the percentile approach. I find that the optimal percentile in-sample equals the 5th percentile. In holdout application, I find that the optimal percentile equals the 10th percentile. This is explained primary by different bankruptcy frequencies for the two samples. In-sample results are derived from the dataset Clean0307 (annual reports for the period 2003-2007) where the holdout sample results are derived from the dataset Clean0810 (annual reports for the period 2008-2010). I note that the holdout sample includes annual reports for the post crisis years and hence the bankruptcy frequency is higher, as expected. I find that the bankruptcy frequency is 1,29% and 1,74% for in-sample and holdout sample data respectively.

The ΔTC of holdout sample at optimal cutoff equal to the 10th percentile (column 7), is superior to ΔTC in-sample at optimal percentile=5% (column 4). This is explained by the differences in bankruptcy frequencies.

I recall the calculation of ΔTC:

$$\Delta TC = \frac{TC_{Developed\ model\ applied}}{TC_{Lend\ to\ all}} - 1$$

The calculation denominator is not fixed for the two samples.

Table 26: Explaining holdout ΔTC superiority, Hazard model

	In-sample	holdout sample	Change, %
Average annually bankryptcy frequency	1,29%	1,74%	
Naive approach, lending to all, TC (denominator)	0,95%	1,29%	35%
Applying model, optimal percentile, TC (numerator)	0,89%	1,12%	26%
Delta TC	**-7,1%**	**-13,0%**	

Table 26 shows the $TC_{Lend\ to\ all}$ for in-sample and holdout sample respectively. I note that $TC_{Lend\ to\ all}$ for the holdout sample is higher compared to in-sample $TC_{Lend\ to\ all}$ (denominator).

I observe that $TC_{\text{Lend to all}}$ goes up by 35%, and that $TC_{\text{Developed model applied}}$ goes up by 26%. This means that the increase in the denominator is higher relative to the increase in numerator, which leads to a more negative number, and ultimately a more negative ΔTC. I note that $TC_{\text{Developed model applied}}$ goes up.

This leads to the scenario where ΔTC for the holdout sample, at optimal percentile, shows superiority compared to ΔTC for in-sample results. The superiority is explained by the change in bankruptcy frequency.

5.2.6 ΔTC over time in holdout application

I have previously argued that the years 2011 and 2012 do not include sufficient bankruptcy data. In this section, I show the deteriorating success rate over time, due to lack of data.

Figure 14: "Predictive success" of excluded years

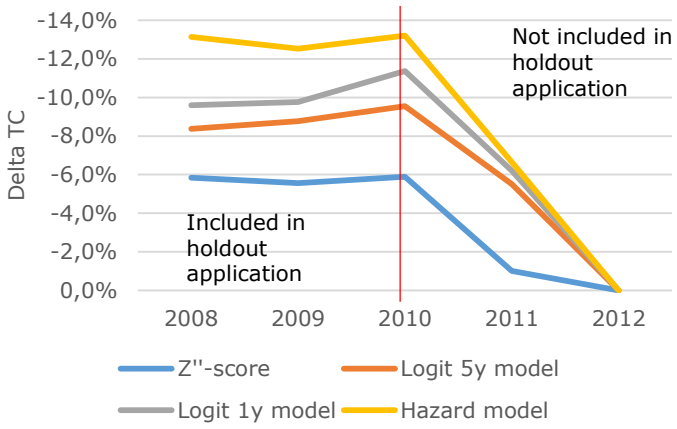


Figure 15 shows ΔTC over time at optimal percentile. I observe that the Hazard model is yielding consistent successful holdout sample predictability for the years 2008-2010. I note that ΔTC, and hence predictability, is deteriorating in the years 2011 and 2012. This is due to the time lag between the annual report and the filing for bankruptcy. It looks as my models have close to no predictive power in 2012, but I emphasize that this is due to lack of bankruptcy information[59].

On this basis, I note that ΔTC calculations for 2011 and 2012 are biased and thus excluded from the results previously presented in this chapter.

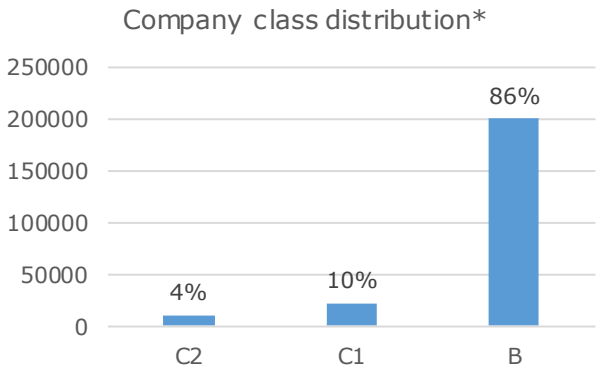
5.2.7 ΔTC for different accounting categories

In the following, I distribute my holdout sample into three subsamples, to mirror the Danish accounting classes. According to table 8: "Accounting classes in Denmark"[60] three financials determine the accounting classes; total balance (total assets), revenue and number of employees. Only the size of the balance sheet is

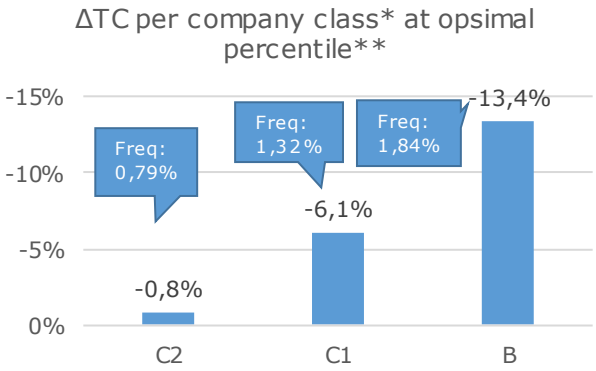
[59] See chapter 4.1.1: "Matching bankruptcy with annual accounts"
[60] In chapter 4.1.2: "Preliminary words on data availability"

available for all firm years in my sample. I approximate the Danish accounting classes by distributing companies by total assets.

Figure 15: Predictive success per proxy company class



** based on balance sheet size*
B: total assets ≤ DKK 36m, C1: total assets ≤ DKK 143m, C2: total assets > DKK 143m



** based on balance sheet size*
B: total assets ≤ DKK 36m, C1: total assets ≤ DKK 143m, C2: total assets > DKK 143m
*** optimal percentiles: C2: 5%, C1: 5%, B: 10%*

Figure 16 pictures the proxy company class distribution and the predictive success of my hazard model, measured by ΔTC. I employ the percentile approach. I observe that the majority of companies in my holdout sample (86%) are small companies with total assets ≤ DKK 36m. Applying my developed hazard model on the respective proxy company classes show inferior predictability measures for the classes C2 (-0,8%) and C1 (-6,1%) compared to predictive success of B (-13,4%) and general predictive success of all companies (-13,0%). Companies in accounting class C1 and C2 are medium and large sized companies. Furthermore, I observe that bankruptcy frequencies are high for class B companies (1,84%) and relatively low for class C1 (0,79%) and C2 (1,23%). Lower bankruptcy frequencies imply lower percentile. I hypothesize that by including fragmented cutoffs of e.g. 1 percentile points steps I would get closer to the real optimal percentile for class C1 and C2 companies respectively[61]. However, I undeniably admit that my models show superior predictive ability for class B companies.

One of my objectives from the beginning was to develop a general model applicable for all Danish companies. However, based on the findings above, I emphasize that my models should be applied to companies in accounting class C1 and C2 (medium and large sized companies) with caution.

[61] The calculations of (Beaver et al. 2011, p. 111) show positive relationship between bankruptcy frequency and cutoff

5.3 Further topics on model development

Until now, I have only presented my final models. This section seeks to check the robustness of my models developed. In this chapter I present selected results from the numerous and uncounted models that I have developed in order to validate my final models developed above. I emphasize that I have not been able to develop a model that yielded higher holdout sample predictability than my final Hazard model.

Running numerous models including other variables

I was surprised that only three variables ended up significant and with the expected sign in my final models. I was hoping for more variables to be included into the models. I am especially surprised that when checked for "Total liabilities to total assets", "EBIT to financial expenditures" and negative equity, then "Net income to total assets" is not significant in the Logit 1y and Hazard models. I also substituted "Net income to total assets" with "EBIT to total assets", but found similar results. The "Retained earnings" variable was not included in any of the three models, hence no metrics measuring profitability were included in the final Logit 1y and Hazard models. I find that "Net income to total assets" and "EBIT to financial expenditures" have a correlation of 0,42, which might explain why "Net income to total assets" is not significant. If I run a hazard model where "EBIT to financial expenditures" is substituted with "Net income to total assets", I find that "Net income to total assets" is significant and has the expected sign.

I have run countless models. I have added previously excluded variables to the final models, without success. I have included variables that aim to measure changes over time. These variables are the ones applied by Ohlson (1980); CHIN[62] and INTWO[63]. By including these two variables I reduce my estimation sample to only include three years of data; 2005-2007. With the hazard approach, CHIN turned out to be significant. INTWO was not significant. However, including CHIN into the model yielded poorer predictability on my holdout sample. I find holdout sample ΔTC of -12.6% for a hazard model including [explanatory variables equal to my final hazard model[64] + CHIN] only slightly inferior compared to ΔTC of -13,0% of my final hazard model.

Model development on non-fitted data

[62] CHIN: a variable that is intended to measure change in net income: $(NI_t - NI_{t-1})/(|NI_t| + |NI_{t-1}|)$

[63] INTWO: A dummy variable: One if net income was negative for the last two years, zero otherwise

[64] Variables included: "l_tl_ta" (logarithm of total liabilities to total assets), "ebit_finexp" (EBIT to financial expenditures), "ek_neg" (dummy equal to 1 if equity is negative) and "l_time" (logarithm to time, where time equals company age)

Furthermore, I generated a model on non-fitted data. This model is derived from a dataset, where I do not transform extreme values. I applied the hazard procedure with [explanatory variables equal to my final hazard model]. I found lower pseudo R^2 of 0,076 compared to my final hazard model, yielding pseudo R^2 of 0,091. Applied on my holdout sample, this model on non-fitted data yielded inferior ΔTC of -12,9%, slightly inferior to my final hazard model, yielding ΔTC of -13,0%. Furthermore, I found that the variable ebit_finexp was insignificant. Furthermore this model showed decreasing ΔTC from 2008 to 2010. My final Hazard model showed consistent ΔTC over the period. I found that this model on non-fitted data yielded superior ΔTC in 2008, but slightly inferior ΔTC in 2009 and inferior ΔTC in 2010 compared to my final Hazard model.

Model misspecification: Testing for quadratics and interaction terms

I applied a RESET test (Wooldridge 2015) to check for functional form misspecification bias and found evidence that quadratics and/or interaction terms should be included into the model. However, to keep the model simple I do not include quadratics or interaction terms. Including quadratics and/or interaction terms might improve predictive ability of my models. Including quadratics or interaction terms have yielded mixed evidence in previous studies. Altman et al. (1977) find that a quadratic structure for their model is appropriate, but the linear structure of the same model outperforms the quadratic in tests of model validity. Lennox (1999) find that including non-linearity improves the model's explanatory power. However, Beaver et al. (2005) find that a linear combination of their three variables capture essentially all of the explanatory power of the financial statement variables used in their three models.

Running a model purely driven by empiricism

I also developed a hazard model purely driven by empiricism. This is, excluding only insignificant variables and not excluding counter-intuitive variables. This model included seven explanatory variables[65] compared to four in my final Hazard model. This model showed coefficients with counter-intuitive signs; e.g. I found that "Retained earnings" should be positively correlated with bankruptcy. This model, purely driven by empiricism, showed inferior ΔTC of -11,7% when applied to my holdout sample, compared to my final hazard model ΔTC of -13,0%. This finding supports my approach of excluding counter-intuitive variables.

Hazard model excluding age

[65] Variables included: tl_ta, ca_ta, re_ta, ni_ta, ebit_finexp, ek_neg and time

I developed a hazard model excluding age as explanatory variable[66] and found inferior ΔTC of -10,3%, when applied to my holdout sample, compared to ΔTC of -13,0%. My final Hazard model indicates positive relationship of age and the event of bankruptcy (the sign of time (age) is positive). The inclusion of this variable yields superior holdout sample predictability.

Simply predict all companies with negative equity to go bankrupt

According to chapter 5.1.3: "Interpretation of coefficients – marginal effects", I find that 43% of bankruptcy firms in my estimation sample showed negative equity, and that this is a strong determinant in BFP. I applied this finding to my holdout sample in order to validate the superiority of my models developed. Simply predicting all companies with negative equity to go bankrupt yielded ΔTC of only -2,5%, which indeed is inferior to ΔTC of -13,0% of my Hazard model. This finding supports the superior predictability of my models. It also supports my findings of including more variables than only ek_neg; a dummy for negative equity.

With this approach I find type I errors of 49,58% and type II errors of 15,32%.

5.4 Results in perspective

This chapter seeks to put my results in perspective. In the following, I quantify the total savings in absolute terms. This is, in cash – how much the entire lending industry, which is defined by "entities that lend money to Danish companies", could save.

[66] Age excluded from final Hazard model, i.e. variables included were: tl_ta, ebit_finexp and ek_neg

Table 27: Naively quantifying annual savings

	Year	2008	2009	2010	Total	Average savings per year
a	total liabilities outstanding (DKKm), holdout sample	3.228.404	2.893.274	2.872.081		
b	Holdout sample coverage*	25,5%	26,4%	25,1%		
c	Total liabilities outstanding (DKKm) total population, estimated **	12.645.777	10.959.372	11.430.533		
d	ΔTC, hazard model***	-13,1%	-12,5%	-13,2%		
e	Total savings (DKKm) by implementing my hazard model, estimated ****	1.661.604	1.372.981	1.508.982	4.543.568	1.514.523

** as estimated by chapter 4.1.5: "Validating data": Cleandata coverage for the year*
*** calculated as a/b*
**** at percentile = 10%*
***** calculated as c*d*

Table 27 show the quantification. In row a I provide information on total liabilities from my holdout sample per year. In row b I show the estimated coverage of my holdout sample. These numbers are estimated in chapter 4.1.5: "Validating data". In row c I estimate the total liabilities outstanding for all Danish companies. This is calculated as a/b. I implicitly assume that all companies not included in my holdout sample have liability profiles equal to my holdout sample. In row d I provide ΔTC for my hazard model, applying the 10th percentile. In row e I quantify the total savings per year. This is calculated as c*d. I implicitly assume that lenders have not applied any model when lending, and that lending has been conducted by the naïve "lend to all" approach. This is not true. The true savings of employing my hazard model, compared to the current employed models by the industry is unknown.

Albeit calculations are conducted on non-true assumptions, I find the quantification appealing. By this, I am able to naively quantify the savings in DKK by implementing my models. I also get an estimate of the potential of a superior BFP model.

From table 27 I see that the average savings per year equals DKK 1.514.523m. This annual saving equals ~66% of the total market value of all Danish listed companies as of May 2016[67]. The savings are enormous. Great BFP models are indeed desirable by lenders.

[67] Source: Factset, total market value of "Copenhagen all share" of DKK 2.277.399m (as of 09.05.2016).

5.5 Summary of analysis

I develop three final models for BFP. I find that the Hazard model yields superior success rate in holdout sample application, measured with my previously developed ΔTC approach. I find ΔTC of -13,0% with the Hazard model. Furthermore, I find that my three final models all yield superior predictability compared to the Z''-score model, albeit the Z''-score indeed show predictive ability. My final Hazard model includes only four explanatory variables, and surprisingly no profitability measures are included in the model. I find that one specific variable; a dummy variable that takes the value 1 if negative equity, shows high marginal effect. Albeit I initially expected more variables to show significance, the models prove successful predictability when applied on a holdout sample. I simulate on my underlying assumption regarding the cost distribution and finds that higher cost ratio of type I vs. type II errors respectively leads to improved ΔTC and higher percentile cutoff (lower absolute cutoff). For my holdout sample application, I find my Hazard model yields consistent predictive ability, where ΔTC is roughly constant over the period. I find that my hazard model yields best predictive success when applied on small companies. Furthermore, I provide key findings on some of the numerous and unreported models that I have developed. These models include (i) models, where I add previously excluded variables to my final models (ii) a model including two new variables measuring changes over time, (iii) a model developed on non-fitted data, (iv) a model purely driven by empiricism, (v) a hazard model excluding age, (vi) a model simply predicting all companies with negative equity to go bankrupt. I find that none of these models yield superior predictive success, compared to my final hazard model.

CHAPTER 6: CONCLUSION

The literature on business failure prediction consists of massive body of research. I find that non-listed companies represent the vast majority of companies in Denmark. Yet, listed companies have been in the spotlight for business failure prediction, where +95% of previous papers are developed for listed companies. I address this mismatch in academia and develop business failure models for non-listed companies.

I find that financial ratios from company accounts have evidently proved predictive ability for business failure prediction. I employ accrual-based measures in my models. Previous studies are inconsistent in providing success rate measures. I find that previous studies focus on overall predictive rate, type I errors (or type I correctly predicted) and type II errors (or type II correctly predicted). I find that several researchers acknowledge an asymmetric cost function, but only a few quantifies this ratio. I quantify this ratio at 18x and on this basis, I generate my own measure, ΔTC, which takes into account (1) the asymmetric cost function and (2) the low bankruptcy frequency. The point is that the rare event of bankruptcy is costly. ΔTC is to be interpreted as the savings a given lender may face by implementing my models compared to the naïve approach of "lend to all".

I find that previous studies are employing datasets that do not mirror the total population. Previous studies apply bankruptcy frequency rates of 1,5% to 50%. From my analysis I estimate the average annual bankruptcy rate in Denmark at 1,3%. My dataset employed show average annual bankruptcy rate of 1,2% and is thus well mirroring the overall bankruptcy frequency of Danish companies. I estimate that my truncated dataset covers 23% of all active Danish companies and hence I appropriately assume that the data coverage is economy-wide. Additionally I find that my dataset is huge relative to comparable studies.

I determine three key statistical techniques that have been applied to the business failure prediction problem with success; multiple discriminant analysis, logistic regression analysis and hazard analysis (or survival analysis). Bankruptcy information is by nature panel data. I find that hazard analysis has appealing statistical features for analyzing bankruptcy. In holdout sample application, I confirm the superiority of the hazard model.

I observe that an uncounted mix of final input variables are included in previous studies. However, I find evidence that the final mix of input variables is of minor importance, as input variables are highly correlated. I apply the backward elimination procedure and additionally exclude variables with counterintuitive signs. I find that a model developed with this technique outperforms a model purely driven by empiricism.

From descriptive statistics, I find that mean values for financials included in the analysis are significantly inferior for bankrupt companies relative to financials for non-bankrupt companies. Additionally, I find that mean values for bankrupt companies are inferior up to at least five years prior to bankruptcy. The descriptive

analysis show that financials, on average, for bankrupt companies indeed are different from financials for non-bankrupt companies. My statistical models show predictive abilities and hence support this initial finding.

I develop three models; one logit model based on 5 years of data, one logit model based on 1 year of data and one hazard model based on 5 years of data. As expected, I find that my Hazard model yields superior holdout sample predictability. Furthermore, I compare my models' holdout predictability to the Z''-score model and find that all three models developed yield superior holdout sample predictability. My Hazard model is yielding ΔTC of -13%, based on four input variables; (1) "Total liabilities to total assets", (2) "EBIT to financial expenditures", (3) a dummy variable that takes the value 1 if the company has negative equity and (4) "time", which is a measure of firm age. I find that my hazard models performs best when applied to small companies. I obtain ΔTC of -13% from 42% correctly predicted bankruptcy companies and 91% correctly predicted non-bankruptcies and overall predictability rate of 90%. I emphasize that overall predictability rate is not a well suited success measure, as the cost function is not symmetric. Additionally I find that my Hazard model show predictive superiority for all holdout sample years, compared to my two logit models and the Z''-score model. My Hazard model yields consistent predictive ability over time.

I report predictive abilities for some of the uncounted models that I have developed. None of these models yield superior predictive abilities compared to my final Hazard model.

Additionally I simulate on my underlying assumption to the cost ratio. In all my analyses, I assume a cost ratio of 18x. I provide cutoff points and ΔTC measures for different cost ratio assumptions. I observe that a higher cost ratio assumption leads to higher ΔTC, as type I errors become more costly. Furthermore, I find that my models yield best predictive ability for small companies.

Based on non-true assumptions I naively estimate the total savings the Danish lending industry may face by employing my models, and find that the average annual savings are equal to ~66% of the total market value of all companies listed on OMX Copenhagen. The potential of superior business failure prediction models is indeed appealing for lenders.

I confirm previous findings and provide evidence that financial ratios show predictive abilities for the event of bankruptcy. Indeed, it is possible to develop universal business failure prediction models for non-listed companies.

CHAPTER 7: PERSPECTIVE, FUTURE RESEARCH AND FINAL WORDS

Perspective

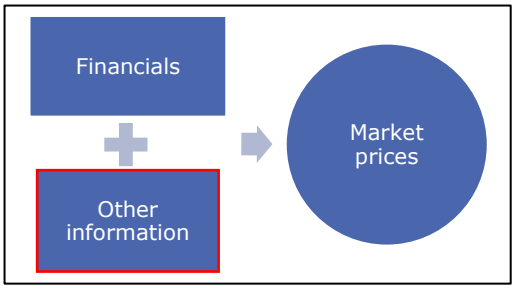
Standing at the end of the road, I can look back on a process that has been frustrating, yet highly rewarding.

In retrospect, I would have put more emphasis to the determination of input variables and put more effort on understanding the bankruptcy process. The determination of input variables has been somewhat arbitrary, and the majority of variables are selected based on frequently used variables in previous studies. Looking back, I would have spent more time on reasoning determinants of BFP before choosing final input variables. However, I am pleased with my final models, and I prove that they indeed show predictive abilities.

Future research

Albeit BFP consists of a considerable body of research, there is room for improvements. The majority of articles focus on listed companies, albeit these companies represent the minority.

By including only accrual based measures I implicitly assume that all information that influence the probability of default is reflected in the company accounts (Balcaen, Ooghe 2006). Market-based variables have evidently added incremental information to accrual-based models (Shumway 2001, Hillegeist et al. 2004, Beaver et al. 2005). However, market-based measures are obviously not available for non-listed companies. According to the efficient market hypothesis (Malkiel, Fama 1970) market prices reflect all currently available information. On this basis, it might seem reasonable to include variables that mitigate the incremental information included in market-based measures, but are not to be found in financials from company accounts. Additionally to financials, market-based variables include 'soft data' including non-company specific conditions, news flow stream and management capabilities[68]. Previous studies suggest to include qualitative measures such as quality of management or people characteristics[69], which might specially be appropriate for the study of small companies (Balcaen, Ooghe 2006)[70], where non-financial information is not simply obtained from market-based variables.



[68] According author
[69] For more information on qualitative variables, see e.g. Altman (2007): mentions that several recent studies indicating that predictive ability improves when applying qualitative variables, such as number of employees, legal form of business, region of operations and main industry.

References

Adnan Aziz, M. & Dar, H.A. 2006, "Predicting corporate bankruptcy: where we stand?", *Corporate Governance: The international journal of business in society,* vol. 6, no. 1, pp. 18-33.

Agarwal, V. & Taffler, R. 2008, "Comparing the performance of market-based and accounting-based bankruptcy prediction models", *Journal of Banking & Finance,* vol. 32, no. 8, pp. 1541-1551.

Altman, E. 1993, *Corporate Financial Distress and Bankruptcy: A Complete Guide to Predicting & Avoiding Distress and Profiting from Bankruptcy,* 2nd edition edn, Wiley Finance.

Altman, E.I. 2005, "An emerging market credit scoring system for corporate bonds", *Emerging Markets Review,* vol. 6, no. 4, pp. 311-323.

Altman, E.I. 2000, "Predicting financial distress of companies: revisiting the Z-score and ZETA models", *Stern School of Business, New York University,* , pp. 9-12.

Altman, E.I. 1968, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The journal of finance,* vol. 23, no. 4, pp. 589-609.

Altman, E.I., Haldeman, R.G. & Narayanan, P. 1977, "ZETA TM analysis A new model to identify bankruptcy risk of corporations", *Journal of banking & finance,* vol. 1, no. 1, pp. 29-54.

Altman, E.I. & Narayanan, P. 1997, "An international survey of business failure classification models", *Financial Markets, Institutions & Instruments,* vol. 6, no. 2, pp. 1-57.

Altman, E.I. & Sabato, G. 2007, "Modelling credit risk for SMEs: Evidence from the US market", *Abacus,* vol. 43, no. 3, pp. 332-357.

Appiah, K.O., Chizema, A. & Arthur, J. 2015, "Predicting corporate failure: a systematic literature review of methodological issues", *International Journal of Law and Management,* vol. 57, no. 5, pp. 461-485.

Aziz, A., Emanuel, D.C. & Lawson, G.H. 1988, "Bankruptcy prediction - an investigation of cash flow based models", *Journal of Management Studies,* vol. 25, no. 5, pp. 419-437.

Aziz, A. & Lawson, G.H. 1989, "Cash flow reporting and financial distress models: Testing of hypotheses", *Financial Management,* , pp. 55-63.

Balcaen, S. & Ooghe, H. 2006, "35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems", *The British Accounting Review,* vol. 38, no. 1, pp. 63-93.

Barnes, P. 1987, "The analysis and use of financial ratios: A review article", *Journal of Business Finance & Accounting,* vol. 14, no. 4, pp. 449-461.

Beatty, A.L., Ke, B. & Petroni, K.R. 2002, "Earnings management to avoid earnings declines across publicly and privately held banks", *The Accounting Review,* vol. 77, no. 3, pp. 547-570.

Beaver, W.H. 1966, "Financial ratios as predictors of failure", *Journal of accounting research,* , pp. 71-111.

[70] Mentions 11 studies, which advice using non-financial or qualitative predictors

Beaver, W.H., Correia, M. & McNichols, M. 2011, *Financial statement analysis and the prediction of financial distress*, Now Publishers Inc.

Beaver, W.H., McNichols, M.F. & Rhie, J. 2005, "Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy", *Review of Accounting Studies*, vol. 10, no. 1, pp. 93-122.

Begley, J., Ming, J. & Watts, S. 1996, "Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models", *Review of Accounting Studies*, vol. 1, no. 4, pp. 267-284.

Bellovary, J.L., Giacomino, D.E. & Akers, M.D. 2007, "A review of bankruptcy prediction studies: 1930 to present", *Journal of Financial education*, , pp. 1-42.

Bonfim, D. 2009, "Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics", *Journal of Banking & Finance*, vol. 33, no. 2, pp. 281-299.

Casey, C.J. & Bartczak, N.J. 1984, "Cash flow: It's not the bottom line", *Harvard business review*, vol. 62, no. 4, pp. 60-66.

Casey, C. & Bartczak, N. 1985, "Using Operating Cash Flow Data to Predict Financial Distress: Some Extensions", *Journal of Accounting Research*, vol. 23, no. 1, pp. 384-401.

Charitou, A., Neophytou, E. & Charalambous, C. 2004, "Predicting corporate failure: empirical evidence for the UK", *European Accounting Review*, vol. 13, no. 3, pp. 465-497.

Chava, S. & Jarrow, R.A. 2004, "Bankruptcy prediction with industry effects", *Review of Finance*, vol. 8, no. 4, pp. 537-569.

Dambolena, I.G. & Khoury, S.J. 1980, "Ratio stability and corporate failure", *The Journal of Finance*, vol. 35, no. 4, pp. 1017-1026.

Dambolena, I.G. & Shulman, J.M. 1988, "A primary rule for detecting bankruptcy: Watch the cash", *Financial Analysts Journal*, vol. 44, no. 5, pp. 74-78.

Danish Statistics, (. 2016, , *Firmastatistik*. Available: http://www.dst.dk/da/Statistik/emner/virksomheder-generelt/firmastatistik [2016, May 12].

Dimitras, A.I., Zanakis, S.H. & Zopounidis, C. 1996, "A survey of business failures with an emphasis on prediction methods and industrial applications", *European Journal of Operational Research*, vol. 90, no. 3, pp. 487-513.

domstol.dk 2015, , *Rekonstruktion*. Available: http://www.domstol.dk/SAADANGOERDU/ERHVERV/Pages/Rekonstruktion.aspx [2016, April 17].

domstol.dk 2011, , *Konkurs*. Available: http://www.domstol.dk/SAADANGOERDU/ERHVERV/KONKURS/Pages/default.aspx [2016, April 17].

e-conomic.dk 2016, , *Regnskabsklasse*. Available: https://www.e-conomic.dk/regnskabsprogram/ordbog/regnskabsklasser [2016, May 15].

Elling, J.O. 2008, *Finansiel rapportering*, Gad.

Erhvervsstyrelsen 2016a, , *Indsendelse af årsrapport*. Available: https://erhvervsstyrelsen.dk/indsendelse-af-aarsrapport [2016, April 17].

Erhvervsstyrelsen 2016b, , *Stiftelse og registrering af selskaber*. Available: https://erhvervsstyrelsen.dk/stiftelse-registrering-selskaber [2016, May 5].

FSR, d.r. 2012, , *Regnskabsvejledning for klasse B og C 2012*. Available: http://www.fsr.dk/Faglige_informationer/Regnskaber/Standarder%20og%20vejledninger/Danske%20regnskabsvejledninger/~/media/Files/FSR/Faglige_informationer/Regnskaber/Standarder%20og%20vejledninger/Danske%20regnskabsvejledninger/ForeloebigtUdkast%20Regnskabsvejledning%20for%20B%20og%20C.ashx [2016, April 17].

Gentry, J.A., Newbold, P. & Whitford, D.T. 1987, "Funds flow components, financial ratios, and bankruptcy", *Journal of business finance & accounting*, vol. 14, no. 4, pp. 595-606.

Gentry, J.A., Newbold, P. & Whitford, D.T. 1985, "Classifying bankrupt firms with funds flow components", *Journal of Accounting research*, , pp. 146-160.

Gepp, A. & Kumar, K. 2008, "The role of survival analysis in financial distress prediction", *International research journal of finance and economics*, , no. 16, pp. 13-34.

Gombola, M.J., Haskins, M.E., Ketz, J.E. & Williams, D.D. 1987, "Cash flow in bankruptcy prediction", *Financial Management*, , pp. 55-65.

Gombola, M.J. & Ketz, J.E. 1983, "A note on cash flow and classification patterns of financial ratios", *Accounting Review*, , pp. 105-114.

Griffin, J.M. & Lemmon, M.L. 2002, "Book–to–market equity, distress risk, and stock returns", *The Journal of Finance*, vol. 57, no. 5, pp. 2317-2336.

Gunasekaran, A., Steven White, D., Opoku Appiah, K. & Abor, J. 2009, "Predicting corporate failure: some empirical evidence from the UK", *Benchmarking: An International Journal*, vol. 16, no. 3, pp. 432-444.

Hayden, E. 2003, "Are credit scoring models sensitive with respect to default definitions? evidence from the austrian market", *Evidence from the Austrian Market (April 2003).EFMA*, .

Hillegeist, S.A., Keating, E.K., Cram, D.P. & Lundstedt, K.G. 2004, "Assessing the probability of bankruptcy", *Review of Accounting Studies*, vol. 9, no. 1, pp. 5-34.

Hoque, M., Bhandari, S.B. & Iyer, R. 2013, "Predicting business failure using cash flow statement based measures", *Managerial Finance*, vol. 39, no. 7, pp. 667-676.

Jones, F.L. 1987, "Current techniques in bankruptcy prediction", *Journal of accounting Literature*, vol. 6, no. 1, pp. 131-164.

Kauffman, R. & Wang, B. 2003, "Duration in the digital economy: Empirical bases for the survival of internet firms", *36th Hawaii International Conference on System Sciences (HICSS), Hawaii*.

Kauffman, R.J. & Wang, B. 2001, "The success and failure of dotcoms: A multi-method survival analysis", *proceedings of the 6th INFORMS Conference on Information Systems and Technology (CIST), Miami, FL, USA*Citeseer, .

Kiefer, N.M. 1988, "Economic duration data and hazard functions", *Journal of economic literature*, vol. 26, no. 2, pp. 646-679.

Koh, H.C. 1992, "The sensitivity of optimal cutoff points to misclassification costs of type I and type II errors in the going-concern prediction context", *Journal of Business Finance & Accounting*, vol. 19, no. 2, pp. 187-197.

Kumar, P.R. & Ravi, V. 2007, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review", *European Journal of Operational Research*, vol. 180, no. 1, pp. 1-28.

Laitinen, T. & Kankaanpaa, M. 1999, "Comparative analysis of failure prediction methods: the Finnish case", *European Accounting Review*, vol. 8, no. 1, pp. 67-92.

Lane, W.R., Looney, S.W. & Wansley, J.W. 1986, "An application of the Cox proportional hazards model to bank failure", *Journal of Banking & Finance*, vol. 10, no. 4, pp. 511-531.

Lennox, C. 1999, "Identifying failing companies: a re-evaluation of the logit, probit and DA approaches", *Journal of economics and business*, vol. 51, no. 4, pp. 347-364.

Lo, A.W. 1986, "Logit versus discriminant analysis: A specification test and application to corporate bankruptcies", *Journal of Econometrics*, vol. 31, no. 2, pp. 151-178.

Luoma, M. & Laitinen, E.K. 1991, "Survival analysis as a tool for company failure prediction", *Omega*, vol. 19, no. 6, pp. 673-678.

Malkiel, B.G. & Fama, E.F. 1970, "Efficient capital markets: A review of theory and empirical work", *The journal of Finance*, vol. 25, no. 2, pp. 383-417.

Mensah, Y.M. 1983, "The differential bankruptcy predictive ability of specific price level adjustments: some empirical evidence", *Accounting Review*, , pp. 228-246.

Meyer, P.A. & Pifer, H.W. 1970, "Prediction of bank failures", *Journal of Finance*, , pp. 853-868.

Nasdaq, O. 2016, , *Shares - share prices for all companies listed on nasdaq nordic*. Available: http://www.nasdaqomxnordic.com/shares [2016, January 12].

Ohlson, J.A. 1980, "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of accounting research*, , pp. 109-131.

Ooghe, H. & Joos, P. 1990, "Failure prediction, explanation of misclassifications and incorporation of other relevant variables: result of empirical research in Belgium.", *Working paper*, .

Peel, M.J. & Peel, D.A. 1987, "Some further empirical evidence on predicting private company failure", *Accounting and Business Research*, vol. 18, no. 69, pp. 57-66.

Petersen, C.V. & Plenborg, T. 2012, *Financial statement analysis*, Prentice-Hall.

Platt, H.D., Platt, M.B. & Pedersen, J.G. 1994, "Bankruptcy discrimination with real variables", *Journal of Business Finance & Accounting*, vol. 21, no. 4, pp. 491-510.

Richardson, F.M. & Davidson, L.F. 1984, "On linear discrimination with accounting ratios", *Journal of Business Finance & Accounting*, vol. 11, no. 4, pp. 511-525.

Sharma, D.S. & Iselin, E.R. 2003, "The decision usefulness of reported cash flow and accrual information in a behavioural field experiment", *Accounting and Business Research*, vol. 33, no. 2, pp. 123-135.

Shumway, T. 2001, "Forecasting bankruptcy more accurately: A simple hazard model*", *The Journal of Business*, vol. 74, no. 1, pp. 101-124.

uscourts.gov 2016a, , *Chapter 11 - Bankruptcy Basics*. Available: http://www.uscourts.gov/services-forms/bankruptcy/bankruptcy-basics/chapter-11-bankruptcy-basics [2016, .

uscourts.gov 2016b, , *Chapter 7 - Bankruptcy Basics*. Available: http://www.uscourts.gov/services-forms/bankruptcy/bankruptcy-basics/chapter-7-bankruptcy-basics [2016, May 3].

Vistrup Lene 2016, , *Konkurs*. Available: http://denstoredanske.dk/Samfund,_jura_og_politik/Jura/Retspleje_og_domstole/konkurs [2016, May 3].

Wilke, R. 2015, *Limited Dependent Variable Models*, Copenhagen Business School.

Wooldridge, J. 2015, *Introductory econometrics: A modern approach*, Nelson Education.

Zavgren, C.V. 1985, "Assessing the vulnerability to failure of American industrial firms: a logistic analysis", *Journal of Business Finance & Accounting*, vol. 12, no. 1, pp. 19-45.

Zmijewski, M.E. 1984, "Methodological issues related to the estimation of financial distress prediction models", *Journal of Accounting research*, , pp. 59-82.

Appendix

#1: data material for type I vs. type II costs

Table: recovery rate calculation

Pengeinstitutter: statistisk materiale
Individuelle nedskrivninger 8.9

	2007	2008	2009	2010	2011	2012	2013	2014
Impairments	6611858	20070488	50048794	38115096	34242552	36621288	26925680	20495730
Recovery	5983507	4166003	3238246	12663369	8275152	11219284	10230864	13341026
% recovery		63%	16%	25%	22%	33%	28%	

Total 26,15% (sum of recovery (year=2008, 2009, 2010) / sum of impairments (year=2007, 2008, 2009))
Type II cost 73,85%

Table: average interest rates for newly issued loans

2008M01	2008M02	2008M03	2008M04	2008M05	2008M06	2008M07	2008M08	2008M09	2008M10	2008M11	2008M12
5,546	5,307	5,835	5,731	5,757	5,913	5,894	5,603	6,036	6,329	6,372	5,828
2009M01	2009M02	2009M03	2009M04	2009M05	2009M06	2009M07	2009M08	2009M09	2009M10	2009M11	2009M12
4,801	4,844	3,86	3,381	3,396	3,535	4,114	3,814	3,609	3,478	3,555	2,734
2010M01	2010M02	2010M03	2010M04	2010M05	2010M06	2010M07	2010M08	2010M09	2010M10	2010M11	2010M12
3,084	2,645	2,884	2,296	2,463	2,827	2,601	2,142	2,709	2,88	2,828	2,913
2011M01	2011M02	2011M03	2011M04	2011M05	2011M06	2011M07	2011M08	2011M09	2011M10	2011M11	2011M12
2,49	2,302	2,592	2,708	2,253	2,111	2,751	2,356	2,836	2,726	2,717	3,102
2012M01	2012M02	2012M03	2012M04	2012M05	2012M06	2012M07	2012M08	2012M09	2012M10	2012M11	2012M12
3,214	2,398	2,88	2,103	1,866	2,22	1,709	2,007	2,494	2,279	2,374	2,742

Average interest rate 4,10%

Table: average interest rates for newly issued loans

Type I costs	73,85%
Type II costs	4,10%
Ratio	18,02

#2: Bankruptcies from DST, Rawdata and Cleandata

Source: DST: "Erklærede konkurser (historisk sammendrag) efter sæsonkorrigering og tid"

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	
Bankruptcies	1771	2329	2469	2506	2620	2495	1987	2401	3709	5710	6461	5468	5456	avg
TOT Compar Firmaer	284446	284166	281653	275712	282968	293885	298214	305319	311518	296072	298081	300733	301481	
Bankruptcy frequency,	**0,6%**	**0,8%**	**0,9%**	**0,9%**	**0,9%**	**0,8%**	**0,7%**	**0,8%**	**1,2%**	**1,9%**	**2,2%**	**1,8%**	**1,8%**	**1,3%**
Source: Rawdata	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	
Number of companies				127532	139791	155333	173195	191410	237599	237599	239919	227611	215062	avg.
Filing for bankruptcy				135	1194	1181	1001	1256	2262	3716	4326	3463	3458	
Bankruptcies, matched with year of annual report				2107	1310	1185	1709	3825	4245	3586	3495	3300	2854	
Bankruptcy frequency, filing, Source: Rawdata				**0,1%**	**0,9%**	**0,8%**	**0,6%**	**0,7%**	**1,0%**	**1,6%**	**1,8%**	**1,5%**	**1,6%**	**1,0%**
Number of companies, coverage				46%	49%	53%	58%	63%	76%	80%	80%	76%	71%	66%
Source: Cleandata	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	
Number of companies				57657	59005	62166	67141	73663	79529	78163	74897	71376	67756	
Filing for bankruptcy				35	582	915	536	479	704	1232	1522	1233	1290	
Bankruptcies, matched with year of annual report				1266	775	461	463	1165	1462	1264	1329	1202	887	
Bankruptcy frequency, filing, Source: Cleandata				**0,1%**	**1,0%**	**1,5%**	**0,8%**	**0,7%**	**0,9%**	**1,6%**	**2,0%**	**1,7%**	**1,9%**	**1,2%**
Number of companies, coverage				21%	21%	21%	23%	24%	26%	26%	25%	24%	22%	23%

#3: Financial availability per country (Orbis database)

Orbis coverage

Breakdown of companies (including branches) according to their world regions/countries versus availability of financial data

Last data update: 11/05/2016

World regions/countries	Availability of financial data							My calculations		
	Companies with detailed financials	Companies with limited financials	Companies with no recent financials	Companies without financials	Total	of which publicly listed companies	of which branches	non-listed companies with detailed financials	non-listed companies with detailed financials, as share of total companies	non-listed companies with detailed financials, as share of total non-listed companies in database
North America	35437	9364082	562843	14677871	24640233	14049	3400359			
Canada (CA)	3964	842238	2351	857993	1706546	3655	147154	309	0%	0%
United States (US)	31473	8521844	560491	13819861	22933669	10394	3253205	21079	0%	0%
Western Europe	10424981	10382996	9114669	27082364	57005010	10145	6592476	10414836	18%	18%
Andorra (AD)	7	2	9	706	724	0	2	7	1%	1%
Austria (AT)	150726	100275	196049	470556	917606	92	153412	150634	16%	16%
Belgium (BE)	453775	103354	197491	2456483	3211103	163	1233361	453612	14%	14%
Cyprus (CY)	1090	10144	66396	355290	432920	122	3	968	0%	0%
Denmark (DK)	257003	177	96301	837396	1190877	155	126333	256848	22%	22%
Finland (FI)	196639	339935	55788	752782	1345144	143	86665	196496	15%	15%
France (FR)	1320631	2246090	1889886	9674193	15130800	865	1788345	1319766	9%	9%
Germany (DE)	743803	649313	859327	1098821	3351264	846	367240	742957	22%	22%
Gibraltar (GI)	82	42	32	5351	5507	6	1	76	1%	1%
Greece (GR)	32598	7	25518	83346	141469	224	21601	32374	23%	23%
Iceland (IS)	31620	2	12403	4577	48602	18	303	31602	65%	65%
Ireland (IE)	161118	5196	102157	324925	593396	98	51	161020	27%	27%
Italy (IT)	1119863	2506637	494404	550640	4671544	322	172	1119541	24%	24%
Liechtenstein (LI)	39	4785	445	41928	47197	3	262	36	0%	0%
Luxembourg (LU)	16246	5450	12552	118487	152735	85	2908	16161	11%	11%
Malta (MT)	11583	20	11638	59172	82413	33	4	11550	14%	14%
Monaco (MC)	12	538	191	11688	12429	2	335	10	0%	0%
Netherlands (NL)	785400	1747726	1405181	668418	4606725	199	226911	785201	17%	17%
Norway (NO)	317985	30	137709	1463260	1918984	211	63109	317774	17%	17%
Portugal (PT)	395100	717	181028	105340	682185	61	69248	395039	58%	58%
San Marino (SM)	4	0	0	480	484	0	0	4	1%	1%
Spain (ES)	892612	162	716512	2755929	4365215	3214	1338712	889398	20%	20%
Sweden (SE)	445805	904315	357134	200365	1907619	612	103583	445193	23%	23%
Switzerland (CH)	1252	713766	62718	60332	838068	285	24560	967	0%	0%
Turkey (TR)	34920	128820	168876	824414	1157030	429	61008	34491	3%	3%
United Kingdom (GB)	3055068	915493	2064923	4157475	10192959	1957	924346	3053111	30%	30%
Eastern Europe	5442903	6567532	4429257	14174095	30613787	7767	1413464	5435136	18%	18%
Albania (AL)	210	27351	1676	101628	130865	0	385	210	0%	0%
Belarus (BY)	32	48493	8127	123626	180278	1	22	31	0%	0%
Bosnia and Herzegovina (BA)	31228	0	9901	2149	43278	750	362	30478	70%	72%
Bulgaria (BG)	346527	284052	293308	637878	1561765	366	11416	346161	22%	22%
Croatia (HR)	114014	7	41853	45196	201070	184	522	113830	57%	57%
Czech Republic (CZ)	226736	1304970	302175	646894	2480775	18	600796	226718	9%	9%
Estonia (EE)	121635	411	38646	84690	245382	18	237	121617	50%	50%
Hungary (HU)	482073	1621	173103	1115511	1772308	43	201230	482030	27%	27%
Kosovo (KV)	20	44494	0	7010	51524	0	119	20	0%	0%
Latvia (LV)	131238	14	69970	163863	365085	26	2428	131212	36%	36%
Lithuania (LT)	16016	98364	39806	7770	161956	32	775	15984	10%	10%
Macedonia (FYROM) (MK)	15145	44767	24787	53459	138158	389	109	14756	11%	11%
Moldova Republic of (MD)	10916	2645	619	184439	198619	690	2607	10226	5%	5%
Montenegro (ME)	306	0	2914	2789	6009	267	171	39	1%	1%
Poland (PL)	159788	1106564	145516	155841	1567709	873	24136	158915	10%	10%
Romania (RO)	750759	52	431366	1464952	2647129	833	8548	749926	28%	28%
Russian Federation (RU)	2215658	3294263	2289939	6294508	14094368	1106	239062	2214552	16%	16%
Serbia (RS)	79032	0	74413	218001	371446	865	8764	78167	21%	21%
Slovakia (SK)	201408	286427	76918	238427	803180	111	225333	201297	25%	25%
Slovenia (SI)	156061	19189	24872	187084	387206	47	309	156014	40%	40%
Ukraine (UA)	384101	3848	379348	2438380	3205677	1148	86133	382953	12%	12%
Middle East	4329	298664	346951	1099566	1749510	1875	91809	2454	0%	0%
Bahrain (BH)	77	697	11007	60161	71942	45	30497	32	0%	0%
Iran Islamic Republic of (IR)	293	169	126	3041	3629	264	333	29	1%	1%
Iraq (IQ)	130	322	88	2648	3188	99	324	31	1%	1%
Israel (IL)	2629	169462	227728	440252	840071	514	305	2115	0%	0%
Jordan (JO)	245	774	627	86163	87809	227	860	18	0%	0%
Kuwait (KW)	233	1318	11605	96805	109961	192	3302	41	0%	0%
Lebanon (LB)	73	15535	35567	66270	117445	10	2718	63	0%	0%
Oman (OM)	147	336	140	130804	131427	126	436	21	0%	0%
Palestinian Territory (PS)	52	177	51	2482	2762	45	239	7	0%	0%
Qatar (QA)	64	15802	195	4345	20406	43	848	21	0%	0%
Saudi Arabia (SA)	193	5107	1543	32375	39218	174	14204	19	0%	0%
Syrian Arab Republic (SY)	20	86	99	1768	1973	18	227	2	0%	0%
United Arab Emirates (AE)	169	88822	58142	171588	318721	118	37376	51	0%	0%
Yemen (YE)	4	57	33	864	958	0	140	4	0%	0%
Far East and Central Asia	1891544	4166912	1224507	26625316	33908279	25204	3792777	1866340	6%	6%
Afghanistan (AF)	6	1	1	11659	11667	0	3	6	0%	0%
Armenia (AM)	26	0	1	558	585	13	8	13	2%	2%
Azerbaijan (AZ)	23	1	2	761	787	10	10	13	2%	2%
Bangladesh (BD)	271	0	39	1156	1466	306	220	-35	-2%	-3%
Bhutan (BT)	15	0	0	18	33	14	0	1	3%	5%
Brunei Darussalam (BN)	2	0	4	10543	10549	0	8	2	0%	0%
Cambodia (KH)	54	1363	7	1111	2535	3	19	51	2%	2%
China (CN)	317422	38	313629	17067649	17698738	5776	1611305	311646	2%	2%
Georgia (GE)	17	133581	3	508018	641619	61	2	-44	0%	0%
Hong Kong (HK)	1902	51395	1773	1632530	1687600	260	1173	1642	0%	0%
India (IN)	30047	1072788	20718	176139	1299692	5837	175	24210	2%	2%
Indonesia (ID)	830	2	517	75296	76645	528	6574	302	0%	0%
Japan (JP)	519633	915601	476237	4003698	5915169	3673	1778525	515960	9%	9%
Kazakhstan (KZ)	3297	338418	13685	121389	476789	78	43585	3219	1%	1%
Korea Democratic People's Republic of (KP)	0	1	0	853	854	0	1	0	0%	0%
Korea Republic of (KR)	249080	16	315687	2058201	2622984	2036	337282	247044	9%	9%
Kyrgyzstan (KG)	16	0	1	594	611	16	4	0	0%	0%
Lao People's Democratic Republic (LA)	9	41733	1	292	42035	5	1	4	0%	0%
Macao (MO)	14	1	1	1521	1537	0	11	14	1%	1%
Malaysia (MY)	245046	6327	37251	38725	327349	919	53	244127	75%	75%
Maldives (MV)	1	0	0	131	132	0	2	1	1%	1%
Mongolia (MN)	217	0	73	2581	2871	230	0	-13	0%	0%
Myanmar (MM)	1	0	0	55891	55892	1	11	0	0%	0%
Nepal (NP)	242	6	8	607	863	155	383	87	10%	12%
Pakistan (PK)	654	1785	251	73656	76346	650	298	4	0%	0%
Philippines (PH)	27503	4	7387	4029	38923	250	46	27253	70%	70%
Singapore (SG)	2375	16	4319	31693	38403	665	122	1710	4%	5%
Sri Lanka (LK)	283	6448	37	1643	8411	284	33	-1	0%	0%
Taiwan (TW)	3238	1588351	3292	152204	1747085	1811	12810	1427	0%	0%
Tajikistan (TJ)	1	0	0	251	252	0	3	1	0%	0%
Thailand (TH)	482707	318	24428	10461	517914	687	53	482020	93%	93%
Turkmenistan (TM)	1	0	0	141	142	0	8	1	1%	1%
Uzbekistan (UZ)	196	8714	4038	557365	570313	1	5	195	0%	0%
Viet Nam (VN)	6415	4	1117	23952	31488	935	44	5480	17%	18%
South and Central America	1078181	21273706	430361	6708508	29490756	3629	1321239	1074552	4%	4%

Orbis coverage

Breakdown of companies (including branches) according to their world regions/countries versus availability of

Last data update: 11/05/2016

World regions/countries	Availability of financial data: Companies with detailed financials	Companies with limited financials	Companies with no recent financials	Companies without financials	Total	of which publicly listed companies	of which branches	**My calculations: non-listed companies with detailed financials**	**non-listed companies with detailed financials, as share of total companies**	**non-listed companies with detailed financials, as share of total non-listed companies in database**
Anguilla (AI)	3	1	2	500	506	2	2	1	0%	0%
Antigua and Barbuda (AG)	4	0	3	264	271	1	2	3	1%	1%
Argentina (AR)	490	320379	137	390445	711451	98	333	392	0%	0%
Aruba (AW)	2	2	35	344	383	0	3	2	1%	1%
Bahamas (BS)	37	1	40	237814	237892	19	5	18	0%	0%
Barbados (BB)	44	0	17	1798	1859	15	4	29	2%	2%
Belize (BZ)	17	0	2	6261	6280	2	1	15	0%	0%
Bermuda (BM)	1013	208	226	42581	44028	760	607	253	1%	1%
Bolivia (BO)	66	1	19	273261	273347	34	6	32	0%	0%
Brazil (BR)	13383	19760605	478	922339	20696805	416	1300977	12967	0%	0%
Cayman Islands (KY)	1478	2	411	37519	39410	1191	21	287	1%	1%
Chile (CL)	555	262943	294	360254	624046	220	23	335	0%	0%
Colombia (CO)	1051464	23	417044	2109049	3577580	76	143	1051388	29%	29%
Costa Rica (CR)	82	1	24	593885	593992	7	34	75	0%	0%
Cuba (CU)	7	1959	12	9758	11736	0	2598	7	0%	0%
Curaçao (CW)	237	21	508	6874	7640	5	1	232	3%	3%
Dominica (DM)	2	0	1	639	642	1	1	1	0%	0%
Dominican Republic (DO)	123	44515	14	109465	154117	1	17	122	0%	0%
Ecuador (EC)	373	176302	531	1000	178206	46	1471	327	0%	0%
El Salvador (SV)	70	0	18	15320	15408	29	39	41	0%	0%
Grenada (GD)	4	0	1	91	96	1	1	3	3%	3%
Guatemala (GT)	79	1	18	50554	50652	4	4	75	0%	0%
Guyana (GY)	10	0	2	303	315	9	10	1	0%	0%
Haiti (HT)	8	0	1	86	95	0	2	8	8%	8%
Honduras (HN)	44	0	5	575	624	9	5	35	6%	6%
Jamaica (JM)	65	0	16	677	758	48	5	17	2%	2%
Mexico (MX)	4155	575363	4359	50772	634649	143	13231	4012	1%	1%
Nicaragua (NI)	23	2	9	453	487	2	3	21	4%	4%
Panama (PA)	226	5	58	866303	866592	31	55	195	0%	0%
Paraguay (PY)	168	0	27	60078	60273	45	235	123	0%	0%
Peru (PE)	1687	109207	5855	433661	550410	208	1351	1479	0%	0%
Saint Kitts and Nevis (KN)	9	1	3	1437	1450	5	7	4	0%	0%
Saint Lucia (LC)	7	0	0	289	296	3	1	4	1%	1%
Saint Vincent and the Grenadines (VC)	0	0	1	606	607	0	2	0	0%	0%
Sint Maarten (SX)	2	0	12	64	78	0	0	2	3%	3%
Suriname (SR)	6	1	16	185	208	2	5	4	2%	2%
Trinidad and Tobago (TT)	37	0	8	552	597	25	13	12	2%	2%
Uruguay (UY)	1969	22163	19	5041	29192	8	8	1961	7%	7%
Venezuela (VE)	104	0	71	62768	62943	37	13	67	0%	0%
Virgin Islands (British) (VG)	128	0	64	54642	54834	126	0	2	0%	0%
Africa	166521	196578	75266	2161939	2600304	1355	67622	165166	6%	6%
Algeria (DZ)	19909	1	26474	193312	239696	4	47370	19905	8%	8%
Angola (AO)	15	0	1	12810	12826	0	36	15	0%	0%
Benin (BJ)	3	202	14	371	590	1	3	2	0%	0%
Botswana (BW)	57	1182	320	2598	4157	23	536	34	1%	1%
Burkina Faso (BF)	9	102	5	582	698	2	1	7	1%	1%
Burundi (BI)	3	62	0	152	217	0	0	3	1%	1%
Cameroon (CM)	4	19003	21	47264	66292	4	6	0	0%	0%
Cape Verde (CV)	15	0	1	2118	2134	4	11	11	1%	1%
Central African Republic (CF)	0	5	4	140	149	0	0	0	0%	0%
Chad (TD)	1	36	3	202	242	0	3	1	0%	0%
Comoros (KM)	0	15	3	64	82	0	0	0	0%	0%
Congo (CG)	0	212	17	736	965	0	20	0	0%	0%
Congo Democratic Republic of (CD)	7	210	7	594	818	0	5	7	1%	1%
Côte d'Ivoire (CI)	36	886	29	10838	11789	32	185	4	0%	0%
Djibouti (DJ)	1	148	0	184	333	0	2	1	0%	0%
Egypt (EG)	324	3770	1595	48785	54474	249	6601	75	0%	0%
Equatorial Guinea (GQ)	1	43	0	270	314	0	7	1	0%	0%
Eritrea (ER)	0	9	0	102	111	0	0	0	0%	0%
Ethiopia (ET)	10	719	2	4542	5273	0	5	10	0%	0%
Gabon (GA)	5	156	8	571	740	1	11	4	1%	1%
Gambia (GM)	1	51	2	237	291	1	5	0	0%	0%
Ghana (GH)	48	67309	7	1102	68466	34	1	14	0%	0%
Guinea (GN)	1	274	5	380	660	0	5	1	0%	0%
Guinea Bissau (GW)	0	16	0	91	107	0	0	0	0%	0%
Kenya (KE)	94	2973	18	3484	6569	62	12	32	0%	0%
Lesotho (LS)	3	10029	2	290	10324	0	0	3	0%	0%
Liberia (LR)	23	262	13	2347	2645	2	71	21	1%	1%
Libya (LY)	7	192	3	948	1150	0	18	7	1%	1%
Madagascar (MG)	3	19958	24	37753	57738	0	3	3	0%	0%
Malawi (MW)	21	372	3	2375	2771	13	2	8	0%	0%
Mali (ML)	6	190	9	448	653	0	1	6	1%	1%
Mauritania (MR)	0	113	2	349	464	0	5	0	0%	0%
Mauritius (MU)	240	365	78	99516	100199	114	8	126	0%	0%
Morocco (MA)	142999	5	15092	304976	463072	74	11652	142925	31%	31%
Mozambique (MZ)	15	0	3	13076	13094	3	6	12	0%	0%
Namibia (NA)	34	635	220	103177	104066	10	311	24	0%	0%
Niger (NE)	0	1	3	8833	8837	1	26	-1	0%	0%
Nigeria (NG)	181	38194	46	4583	43004	184	1	-3	0%	0%
Rwanda (RW)	9	79	1	50698	50787	3	0	6	0%	0%
Sao Tome and Principe (ST)	1	17	0	78	96	0	0	1	1%	1%
Senegal (SN)	3	7659	97	1190	8949	6	30	-3	0%	0%
Seychelles (SC)	16	193	7	10989	11205	1	465	15	0%	0%
Sierra Leone (SL)	2	57	2	408	469	0	4	2	0%	0%
Somalia (SO)	0	94	0	72	166	0	2	0	0%	0%
South Africa (ZA)	2154	18197	31008	771202	822561	320	58	1834	0%	0%
South Sudan (SS)	0	76	0	29	105	0	0	0	0%	0%
Sudan (SD)	2	237	1	774	1014	14	27	-12	-1%	-1%
Swaziland (SZ)	11	416	19	528	974	7	1	4	0%	0%
Tanzania United Republic of (TZ)	41	331	9	99185	99566	16	6	25	0%	0%
Togo (TG)	6	113	7	365	491	1	1	5	1%	1%
Tunisia (TN)	74	3	54	101105	101236	77	29	-3	0%	0%
Uganda (UG)	14	0	2	212386	212402	8	32	6	0%	0%
Zambia (ZM)	33	743	2	1022	1800	22	25	11	1%	1%
Zimbabwe (ZW)	79	663	23	1708	2473	62	13	17	1%	1%
Oceania	18881	4219942	18206	11195268	15452297	2236	4004530	16645	0%	0%
Australia (AU)	14356	3609798	14487	10983384	14622025	1981	4001631	12375	0%	0%
East Timor (TL)	0	0	0	47	47	0	1	0	0%	0%
Fiji (FJ)	29	0	3	281	313	17	5	12	4%	4%
Kiribati (KI)	0	0	0	237	237	0	0	0	0%	0%
Marshall Islands (MH)	51	0	5	3206	3262	52	0	-1	0%	0%
Micronesia Federated States of (FM)	1	0	0	17	18	0	0	1	6%	6%
Nauru (NR)	0	0	0	62	62	0	0	0	0%	0%
New Zealand (NZ)	4425	610144	3707	107592	725868	173	2887	4252	1%	1%
Palau (PW)	0	0	0	5	5	0	0	0	0%	0%
Papua New Guinea (PG)	16	0	3	98903	98922	12	5	4	0%	0%
Samoa (WS)	1	0	0	1132	1133	1	0	0	0%	0%
Solomon Islands (SB)	0	0	0	66	66	0	1	0	0%	0%
Tonga (TO)	0	0	0	35	35	0	0	0	0%	0%
Tuvalu (TV)	0	0	0	5	5	0	0	0	0%	0%
Vanuatu (VU)	2	0	1	294	297	0	0	2	1%	1%
Supranational	17	0	0	24	41	0	0	17	41%	41%
No country specified	298	328	4228	3560425	3565279	0	235	298	0%	0%
Total	19063092	56470740	16206288	1,07E+08	1,99E+08	66260	20684511	18996832	10%	10%