Wissenschaftliche Beiträge aus dem Tectum Verlag
Reihe: Wirtschaftswissenschaften
Band 74

Bernd Galler

# Scarce Data based Credit Risk Assessment

Evaluating Missing Data Methods
in PD-estimation by Logistic Regression


Tectum

**WISSENSCHAFTLICHE BEITRÄGE AUS DEM TECTUM VERLAG**

Reihe Wirtschaftswissenschaften

WISSENSCHAFTLICHE BEITRÄGE
AUS DEM TECTUM VERLAG

Reihe Wirtschaftswissenschaften

Band 74

Bernd Galler

# Scarce Data based Credit Risk Assessment

Evaluating Missing Data Methods in PD-estimation by Logistic Regression

Tectum Verlag

D6

Bernd Galler

Scarce Data based Credit Risk Assessment.
Evaluating Missing Data Methods in PD-estimation
by Logistic Regression

Wissenschaftliche Beiträge aus dem Tectum Verlag
Reihe: Wirtschaftswissenschaften; Bd. 74

Zugl. Diss. Westfälische Wilhelms-Universität Münster 2013



ISBN 978-3-8288-6159-6

(Dieser Titel ist zugleich als gedrucktes Buch
unter der ISBN 978-3-8288-3432-3 im Tectum Verlag erschienen.)

Besuchen Sie uns im Internet
www.tectum-verlag.de
www.facebook.com/tectum.verlag

**Bibliografische Informationen der Deutschen Nationalbibliothek**
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Angaben sind im Internet über http://dnb.ddb.de abrufbar.

*Preface*

Appropriately assessing and effectively managing credit risk have challenged banks for a long time. The scale of this challenge was highlighted by the financial crisis of 2008 and has become even more crucial due to the increasing introduction of restrictive banking supervisory regulations. In the meanwhile, common quantitative methods addressing this challenge have become an industry standard.

Even though finance commands sophisticated quantitative tools in order to asses credit risk, these instruments can only be as beneficial as the quality of the data they rely on. Therefore, ample and reliable data are an important precondition for the effective use of these methods. However, these conditions are rarely met in reality. Particularly, missing data are a common cause of tainting data quality known to many fields beyond credit risk assessment.

Besides transferring established means of properly dealing with missing data from statistics literature to the field of credit risk assessment, Galler evaluates various methodological procedures and, thus, contributes to the current research on credit risk models.

His analysis is based on a large database of high quality, comprising balance sheets that differ in terms of their proportion of missing data. The present study includes various methodological procedures for handling and evaluating databases with missing data, ranging from basic procedures to advanced techniques. Each of the resulting credit risk models is evaluated in terms of quality and output, while corresponding benchmarks derive from completely observed data.

Galler's study reveals that the choice of missing data method influences the output of credit risk models. More specifically, it demonstrates that the success of mitigating the observed data quality problem depends on the applied missing data method. Moreover, the study shows that the examined statistical procedures are differently sensitive to the existing proportion of missing data. Taken together, Galler's study provides clear evidence that the alleged minor problem of missing data deserves

an increased amount of consideration when building quantitative credit risk models.

*Prof. Dr. Jens Leker, May 2014*

***Acknowledgment***

*It is with immense gratitude that I acknowledge the help and support of Professor Dr. Leker. I do not only wish to thank him for providing me with the opportunity to spend satisfying years of work in my field of interest, but also for his input, encouragement and sharing his knowdledge. Working at his institute was as much a pleasure as a valuable experience.*

*Also, I would like to thank the additional members of the thesis committee, Professor Dr. Langer and Professor Dr. Vossen. I highly appreciate their support and their assistance.*

*I want to thank Dr. Kehrel for his numerous times of assistance and his cooperation over multiple years.*

*Further, I am deeply indebted to all my colleagues at the Institute of Business Administration for all their input and multitudinous discussions which helped to advance my research. My unrestricted thanks go especially to my fellow team members and friends Jan Wosnitza and Gerrit Böhm for numerous times of assistance as well as for providing well-founded positions and expertise. Also, I want to thank Elena Vakhtina for her great, highly competent and fast support whenever needed.*

*My friend Siegfried Kusper did not only offer me support and encouragement. He was also willing to share his immense knowledge founded on years of experience in the banking industry. Therefore, my sincere thanks go to him.*

*In addition, I wish to thank Jochen Klöpper and BAWAG P.S.K. They provided me generously with relevant data which did not only facilitate my research but also helped to enhance the quality of the data base of this thesis as well as of other research projects. The work of many years would not have been possible without their support.*

*I cannot find words to express my gratitude to my life companion Simone. She assisted and encouraged me patiently in addition to being an inspiration of hard work and discipline. Last but not least I would like to thank my*

*grandparents, my parents, my brother and all other members of my family for supporting me over the course of many years.*

# Contents

# List of Figures

# List of Tables

# 1 Introduction

In the last few years, we have seen dramatic developments in economic fluctuation. In the middle of the last decade, when real estate prices in the United States started to decline, mortgages became increasingly harder to repay for individuals as rising house prices were previously used to avoid defaults. This worked either by refinancing or sale of the respective property.[1] Subprime lenders, that is to say debtors with particularly low creditworthiness, were especially affected.[2] In 2007 many lenders could not meet their obligation to pay anymore and the bubble in the housing market finally burst as banks' claims resulting from their mortgage activities were rendered worthless. As mortgages were securitized in Asset Back Securities (ABS), such as Mortgage Backed Securities (MBS), formerly illiquid loans were transferred into liquid assets and could thus be easily traded between investors.[3] Risk inherent in these mortgages, which was formerly allocated to few investors, was therefore easily distributed around the globe. Therefore, the US real estate market collapse, the so called Subprime Crisis, triggered a global shock wave.[4]

On September 15, 2008 the fourth largest investment bank, Lehman Brothers, which was heavily invested in the respective securities, collapsed as a result.[5] This event amounted to the largest bankruptcy filing ever recorded for the US. AIG had issued insurance for a volume $400 billion against defaults of subprime mortgages in the form of Credit Default Swaps (CDS).[6] One day later, the breakdown of AIG followed.[7] As investors' risk awareness grew, credit spreads began to rise.[8] Banks lost their trust in each other as institutions were not able to assess how much of the contaminated assets were held by other institutions and hence whether the solvency of a potential obligor was seriously impaired. Thus, the inter-bank loan market was crippled. Therefore,

[1] Bhardwaj and Sengupta (2012), p.1504-1505.
[2] Bhardwaj and Sengupta (2012), p.1503; El Gaied et al. (2012), p.83; Reinhart and Rogoff (2008), p.340.
[3] Bhardwaj and Sengupta (2012), p.1505; El Gaied et al. (2012), p.83.
[4] Mishkin (2011), p.50-52.
[5] Mishkin (2011), p.52-53.
[6] Mishkin (2011), p.53-54.
[7] Mishkin (2011), p.52.
[8] Eichengreen et al. (2012), p.1313; Mishkin (2011), p.55-56.

the Subprime Crisis was joined by a liquidity crisis which caused central banks to intervene. As many bank institutions suffered severe losses due to the declines in value of their assets, governments and central banks assisted these institutions in order to avoid a collapse of the banking sector.[9] The world had arrived at the latest financial crisis.

The financial crisis soon began to spread to the real economy as tightened bank liquidity caused borrowing costs for companies to rise sharply, business and consumer confidence eroded and assets devalued. As a consequence, global economic activity contracted considerably.[10] The global contraction, which lasted until 2009, was labeled "the Great Recession". Due to great public efforts to moderate the consequences of the crisis, public households were strained considerably which confronted numerous countries with refinancing problems.[11] A phenomenon prior labeled highly improbable suddenly reached the attention of investors: sovereign default. As investors started to estimate the probabilities of sovereign default higher than before, sovereign bond spreads began to rise for numerous countries, especially in the Eurozone.[12] Several countries using the common currency experienced serious liquidity problems. Since then sovereign liquidity problems have been remaining and the Eurozone members have implemented several rounds of public safety nets in order to challenge impending defaults of membership countries.

In short, in the last few years we have been confronted with an increasingly uncertain environment. As uncertainty rises, so does the significance of an individual company's risk management.

## 1.1 Problem

This is especially true for the credit risk management of banks, since they are not only faced with an uncertain environment, but also with rigorous regulations which have been intensified as a consequence of the crisis and which demand the implementation of rigorous risk management frameworks.[13] The Basel regulations especially require banks to provide

[9] European Central Bank (2009), p.22; Mishkin (2011), p.63.
[10] European Central Bank (2009), p.22; El Gaied et al. (2012), p.86; Mishkin (2011), p.57.
[11] Reinhart et al. (2012), p.73.
[12] European Central Bank (2011), p.38.
[13] Walker (2011), p.95; Basel Committee on Banking Supervision (BCBS) (2011).

their risk weighted assets with an adequately large base of regulatory capital. The youngest reforms of these regulations, known as Basel III, the implementation of which started at the January 1, 2013 and is planned to be completed by January 1, 2019, demand a higher quality and quantity of regulatory capital.[14] This puts particular pressure on banks in procuring sufficient regulatory capital. Since the Internal Ratings-Based Approach (IRB) of determining regulatory capital requirement tends to lower the regulatory capital requirements under the current Basel regulatory framework, banks have a great incentive for using this approach, rather than the Standardised Approach.[15] While the latter requires banks to rely on external assessments of their credit risk by credit rating agencies (CRA), the IRB-approach allows them to internally estimate their credit risk in order to determine their regulatory capital requirement for every credit exposure.[16] In order to estimate credit risk, banks typically rely on internal models.[17] In doing so, they consider multiple risk components, the most important of which is the parameter of *Probability of Default (PD).* It is labeled as such in this thesis as it is the only risk parameter which has to be internally estimated by banks using IRB in all cases.[18] Others might be externally provided by the supervisory body under certain conditions.[19] In addition, Probabilities of Default are an integral part of adequate credit pricing. The PD is used to determine Standard Risk-Costs and therefore influences the price for a granted loan. This component fulfills an important part in absorbing incurred losses.[20]

In light of the above, estimating PD as accurately as possible is a vitally important challenge within a bank's credit risk assessment. In order to build proper models for PD estimation, i.e. models which allow

[14] Basel Committee on Banking Supervision (BCBS) (2011), paragraph 7; Basel Committee on Banking Supervision (BCBS) (2012), p.8.
[15] Haves (2012), p.33-34.
[16] Basel Committee on Banking Supervision (BCBS) (2004), paragraph 211.
[17] Basel Committee on Banking Supervision (BCBS) (2004), paragraph 346.
[18] Florez-Lopez (2010), p.487; Haves (2012), p.34.
[19] Basel Committee on Banking Supervision (BCBS) (2004), paragraph 391.
[20] Heidorn (2012), p.384.

to deduce statistically valid inferences, banks have to rely on an adequately large database.[21] However, in banking practice, data records of obligors are often quite limited which makes modeling challenging and available obligor data precious.[22] In addition to the limited nature of obligor databases, the existing data are usually impaired by incomplete records which increases the risk of biased inferences.[23] If this condition occurs, we speak of *missing data*. The nonavailability of these records particularly amplifies the problem of scarce data. Moreover, many standard statistical procedures require complete data. This is for example the case with Logistic Regression (LR) which is widely applied by banks for PD estimation.[24] Therefore, in order to build a proper model for PD estimation which renders valid inferences, the challenge of missing data has to be appropriately tackled.

In statistics, there are numerous approaches to deal with the problem of missing data which are employed in a wide range of social sciences. These so called *missing data methods (MDMs)* vary from simple, basic procedures to statistically more elaborate methods, each one featuring individual properties. Observations containing missings can for example simply be deleted or a plausible value can be substituted for the missing value.[25] A framework for missing data handling, which is still applied, was provided by Rubin.[26] In 1977 Dempster, Laird and Rubin introduced the EM algorithm.[27] This approach makes it possible to derive *Maximum Likelihood (ML)* estimates. Instead of deleting or filling in observations, ML handles the missing data as random variables which are integrated out of the likelihood function.[28] In 1987 Rubin presented the concept of *Multiple Imputation (MI)*.[29] When applying this method, every missing value is substituted with a number of $M$ simulated values, where $M > 1$. The advancing development of computer technology by

[21] Basel Committee on Banking Supervision (BCBS) (2004), paragraph 417; Oesterreichische Nationalbank (OeNB) and Finanzmarktaufsicht (FMA) (2004a), p.64; Schewe and Leker (2000), p.171.
[22] Basel Committee on Banking Supervision (BCBS) (2005), p.2.
[23] Kaltofen et al. (2007), p.7.
[24] Basel Committee on Banking Supervision (BCBS) (2000), p.6.
[25] Schafer and Graham (2002), p.155-162.
[26] Rubin (1976), Schafer and Graham (2002), p.148.
[27] Dempster et al. (1977).
[28] Schafer and Graham (2002), p.148.
[29] Rubin (1987).

the late eighties of the last century alleviated the implementation of these techniques considerably and stimulated their application. [30] They were later joined by *Markov Chain Monte Carlo (MCMC)* techniques. In the 1990s Ibrahim presented ways of applying the EM algorithm to Generalized Linear Models (GLM) for discrete data.[31] Later in the decade, the method was expanded to continuous variables by the help of an MCMC via the *Gibbs Sampler*.[32] The application of the EM algorithm to GLMs is especially interesting for credit risk assessment since many models are based on logistic regression.

However, in spite of the ample number of dealing with missing data, statistical theory does not recommend one superior approach which is universally best suited in every situation and for every field of science. So, a proper way of missing data treatment in a certain setting outside of credit risk might yield suboptimal results when employed therein, let alone on a specific problem such as PD estimation under IRB. Rather, the best approach highly depends on the properties of the missing data, such as their underlying distribution. Credit risk literature addresses the problem of missing data only superficially as it often recommends basic procedures, but neglects to evaluate newer approaches for credit risk already known to statistical literature. However, the former are not necessarily always adequate for credit risk problems nor efficient in their use of available data. In addition, many studies in credit risk employ methods which emerged by convention of the field, rather than on the basis of theoretic considerations or empiric evidence.[33] Galler and Kehrel showed that varying missing data methods can have different influences on the distribution of financial ratios which often serve as predictor variables in credit risk models.[34] Florez-Lopez conducted an investigation of using different missing data methods in credit modeling for retail customers based on a mixture of categorical and continuous data.[35] However, the assets of an average European bank largely consist

[30] Schafer and Graham (2002), p.148.
[31] Ibrahim (1990).
[32] Ibrahim et al. (1999a).
[33] Collins et al. (2001), p.330; Florez-Lopez (2010), p.486; Oesterreichische Nationalbank (OeNB) and Finanzmarktaufsicht (FMA) (2004a), p.80-82.
[34] Galler and Kehrel (2011).
[35] Florez-Lopez (2010).

of claims on corporates, rather than retail exposures. In addition, credit risk models for corporate exposures typically rely on financial statement information, i.e., financial ratios, as variables serving as predictors of default. That is to say, when assessing credit risk for corporate obligors, we use continuous attributes in the respective model, rather than a mixture of categorical and continuous variables as in the case of models for retail exposures. Thus, we lack the knowledge of how different missing data methods perform in a credit risk modeling framework which would be applied by most banks.[36]

Therefore, the goal of this thesis is to evaluate a range of statistical methods of dealing with missing data, varying in their complexity as well as in their efficiency of handling available data, under a credit risk framework designed for estimating the PD for corporate exposures. Particularly, I am interested in how banks can on the one hand efficiently use their available customer records and on the other draw inferences, i.e., calculate the risk parameter PD, as precisely as possible in spite of missing data records. Ergo, the central research questions of this thesis are:

- RQ 1: Do missing data methods exert influence on the assessment of credit risk?
- RQ 2: Do specific missing data methods use existing data more efficiently than others?
- RQ 3a: Do specific missing data methods yield potential gains for credit risk assessment?
- RQ 3b: If RQ 3a is correct, is the performance of particular missing data methods sensitive to varying ratios of missing data?

For this investigation, I want to rely on a typical framework which could be employed by banks in order to estimate this risk parameter in a Basel environment. As Logistic Regression asserted itself as the method of choice in credit modeling in most banks, I will use it in order to construct a corporate credit rating model. Usually, credit risk models for corporate exposures are based on financial statement information which is why I

[36] Kretzschmar et al. (2010), p.2840.