BANK DEFAULT PREDICTION MODELS

A COMPARISON AND AN APPLICATION TO CREDIT RATING TRANSITIONS

Stefan van der Ploeg*

7 January, 2010

ABSTRACT

This study examines and compares the predictive performance of multiple default prediction models and assesses the ability of these models to correctly predict credit rating transitions. The set of default prediction models that is examined comprise the probit model, logit model, hazards model and neural networks model. The focus of this study is on US banks and the examined time period starts at 1987 and ends at 2008. The out-of-sample results show that the performances of the models are not very divergent and that all models perform very adequate in the prediction of defaults. Credit rating transitions, notwithstanding the fact that various financial ratios prove to contain valuable information regarding a bank's financial condition, are demonstrated to be more difficult to predict, since out-of-sample predictive performances deteriorated relative to the prediction of defaults. Argumentation for this result can be found in the presence of subjectivity in the credit rating process and the fact that the specification of the two credit events differs to some extent.

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Department of Finance

SUPERVISOR
Prof. Dr. W. F. C. Verschoor

ERASMUS UNIVERSITEIT ROTTERDAM

ERNST & YOUNG
Financial Services Organization
Financial Services Risk Management

SUPERVISOR
Drs. J. Menken

ERNST & YOUNG

* MA-student at Erasmus University Rotterdam, Erasmus School of Economics, Department of Finance, and student intern at Ernst & Young, Financial Services Organization, Financial Services Risk Management.

Contents

I. INTRODUCTION

The ongoing credit crisis and the subsequently unsettled economic conditions during the past two years have demonstrated that a stable financial system is an important prerequisite for economic growth and stability. Consequently, and as is common in times of financial turmoil, a discussion on bank failures and on ways of precluding these failures did again rise. This preclusion of bank failures and the assessment of banks' financial condition is one of the fundamental objectives of supervisors and regulators. The assessment of a bank's financial condition is based on both on-site examinations as well as on off-site systems of bank monitoring, the latter being mainly based on regulatory reporting and other financial statement data. On-site examinations, of which the center of attention can be on all types of a bank's operations, including organizational structures, internal controls and internal risk monitoring, provide a comprehensive representation of a bank's overall strength. However, on-site examinations are usually very costly and require significant amounts of time. Hence, supervisors and regulators cannot perform these on-site examinations on a high frequency base. In order to still be able to monitor individual banks and the complete banking system in general, supervisors and regulators examine banks' financial condition on an off-site base. By way of these off-site examinations, as mentioned usually based on individual financial statement data and various models, e.g. statistical models and credit risk models, supervisors and regulators are able to monitor banks with a higher frequency. Additionally, the off-site examinations allow supervisors and regulators to examine a vast amount of banks all at once. Hence, the results of off-site examinations, usually probabilities of default or classifications into particular rating classes, provide support to the prioritization of the allocation of costly and scarce on-site examination resources. Summarizing, the off-site examination of banks' financial statement data enables supervisors and regulators to frequently monitor individual banks, to determine its financial condition and to provide a ground on which decisions are made regarding the allocation of scarce on-site examination resources.

Determining the financial condition, and accordingly the probability of default, of a bank is not only of importance in the practices of supervisory and regulation. It also plays an imperative role in numerous other financial economic practices, e.g. supporting credit analysis, calculating counterparty credit risk, extending the standard financial engineering techniques for the valuation of credit derivatives or other credit sensitive instruments. Hence, it can be concluded that the determination of current and the estimation of future financial conditions of banks is of central relevance in a wide range of financial economic practices.

However, in the light of the current credit crisis, in which a substantial amount of banks experienced massive losses as a result of credit rating transitions, widening credit spreads and loss of liquidity rather than actual defaults, the ability of merely estimating default probabilities would not appear to be entirely sufficient. This finding is supported by the fact that, as a result of banks' massive incurred losses, the Basel Committee on Banking Supervision, in June 2008, expanded the scope of the incremental risk capital charge from solely default risk to both default risk and credit rating transition risk. This credit rating transition risk comprises the probability of both direct as well as indirect losses that may arise as a result of an internal or external credit rating downgrade or upgrade.

Based on the fact that both failure events as well as credit rating transition events play an imperative role in modern risk control techniques, models that consistently perform well in the estimation of the risk of both credit events are of utmost importance. From a theoretical perspective, based on the fact that both events are credit events of equivalent nature, as a credit rating merely reflects a probability of default in a notational form, it can be expected that models achieving a high degree of predictive accuracy in default prediction, may do consistently well in the prediction of credit rating transitions.

Nonetheless, comparative literature investigating the predictive accuracy of different models in estimating defaults and in estimating credit rating transitions is very scarce. Generally, studies focus on merely one of the two credit events. The literature on bank failure prediction is broad and includes a large amount of failure prediction models, e.g. statistical models and credit risk models. The greater part of this class of literature focuses on improving predictive accuracy by appropriately determining the predictive variables, most often financial statement data. Because of conflicting empirical results however, no unified theory has been generally accepted regarding the significance of specific predictors of failure. The fact that there is no consensus regarding the significance of specific predictors emphasizes the importance of model-inherent structures, assumptions regarding the properties of the data and its relation to the probability of default. Hence, comparative studies, assessing the predictive accuracy of multiple default prediction models, embody an important class of failure prediction studies.

Examination of credit rating transition is, just as failure prediction, an established topic in financial empirical literature. However, approaches regarding the estimation of credit rating transition probabilities appear to be nearly all based on the construction of one or more credit rating transition matrices. These credit rating transition matrices approach the event of a credit rating transition from a general perspective, assessing the credit rating transition spectrum as a whole. Based on a database of historical credit rating transitions of numerous companies, a credit rating transition matrix is constructed, representing the transition probabilities of all possible transitions from all possible initial credit ratings. Hence, no firm-specific information, e.g. financial statement data, is directly applied in the determination of the credit rating transition probabilities. Yet, based on the fact that a credit rating downgrade effectively represents a deterioration of a firm's credit quality, it is straightforward to notice that firm-specific information might be able to reflect this deteriorating credit quality. Empirical literature in which credit rating transition probabilities are estimated by means of firm-specific information is extremely scarce.

The way in which this study adds to current literature is based on the fact that it examines and compares the predictive accuracy of multiple default models and assesses the ability of these models in predicting credit rating transitions using firm-specific financial statement data. Hence, the main objectives of this study are to examine what models perform best in failure prediction and to determine whether or not credit rating transitions can also be predicted using the same models and firm-specific predictive variables.

The remainder of this study is structured as follows. Section II describes the development of the models in empirical literature and presents various results of the most relevant studies. Section III provides a thorough explanation on the models employed in this study. Section IV describes the predictive variables that are employed as the predictors of failures and credit rating transitions. Section V and section VI present the results of the default prediction estimations and credit rating transition estimations, respectively. Section VII finally provides some concluding remarks and suggestions for further research.

II. LITERATURE REVIEW

Ever since the seminal work of Beaver (1966) and Altman (1968), the prediction of bankruptcy has been actively studied by academics, practitioners, and regulators. In his pioneering work, Beaver (1966) presented an univariate approach of discriminant analysis in order to assess the individual relationships between financial statement data, i.e. predictive variables, and subsequent failure events. Altman (1968) expanded the univariate approach to a multivariate discriminant analysis, allowing one to assess the relationship between failure and a set of financial characteristics. The work of Beaver (1966) and Altman (1968) initiated a intensification of studies assessing the predictability of bank failures.[1] Until the end of the 1970s, discriminant analysis remained the dominant method in the prediction of failure. In spite of this, restrictive assumptions of discriminant analysis, e.g. discriminant analysis assumes the financial statement data to be normally distributed and assumes the variance-covariance matrices of failed and non-failed banks to be equal, were proven to be violated frequently by multiple subsequent studies. As a result, it was the seminal work of Martin (1977) that introduced the first method of failure prediction that did not make any restrictive assumptions regarding the distributional properties of the predictive variables. The logistic regression, often referred to as the logit model or the logit analysis, until recently has been the most employed statistical method for the purpose of failure prediction. In the examination of the performance of the logit model, Martin (1977) examined all commercial banks that were member of the Federal Reserve System as of 1974. This group of banks contained 23 failed banks and 5,575 non-failed banks. The set of predictive variables comprised 25 ratios of financial statement data that could be classified into four groups, being asset quality, liquidity, capital adequacy, and earnings. The final preferred model included only four of the predictive variables, representing asset quality, capital adequacy and earnings.

Consistent with the results of Martin (1977) are the conclusions of Avery & Hanweck (1984). Avery & Hanweck (1984), using the logit model, examined 100 failed banks and 1,190 non-failed banks during an estimation period starting at December 1978 and ending at June 1983. The examination of these banks was performed employing an initial set of only nine predictive variables that were selected as they proved to be significant in previous studies. Five of these predictive variables proved to again be significant and displayed the correct a priori expected signs. Consistent with the findings of Martin (1977), the significant variables could be classified into asset quality, capital adequacy and earnings.

Barth et al. (1985), in their study on failures of thrift institutions between December 1981 and June 1984, again confirm the relevance of asset quality, capital adequacy and earnings. In addition to these three risk factors, Barth et al. (1985), employing the logit model, find liquidity to be an important factor in relation to subsequent failures. Liquidity is proxied by the ratio of liquid assets and total assets as well as by the natural logarithm of the total assets. The initial set of predictive variables of the study of Barth et al. (1985) only comprised 12 ratios of financial statement data that, consistent with Avery & Hanweck (1984), were selected as they proved to be significant in previous studies.

[1] Examples include Stuhr & Van Wicklen (1974), Sinkey (1975) and Altman (1977).

Thomson (1991), in a study on FDIC-insured commercial banks, examines the predictive accuracy of the logit model employing predictive variables that proxy for asset quality, capital adequacy, earnings, liquidity and management quality. Argumentation behind the inclusion of the risk factor management quality was based on an earlier study of Graham & Horner (1988) that illustrated the importance of an adequate management. The results of Thomson (1991), based on failures between 1984 and 1989, demonstrated that the probability of bank failure is a function of variables proxying for all five risk factors mentioned above. The logit model, including variables of all five risk factors, demonstrated very good classification accuracies both in-sample as well as out-of-sample.

Estrella et al. (2000), employing a logit model, examine and compare the effectiveness of simple and more complex risk-weighted capital ratios, representing the risk factor capital adequacy. They conclude that simple capital ratios predict bank failures as well as the more complex risk-weighted capital ratios and that therefore the risk factor capital adequacy can without problems be proxied by a number of simple capital ratios. Estrella et al. (2000) also examine the performance of credit ratings as a predictor of default. However, evidence in favor of credit ratings being important predictors of defaults is somewhat mixed.

In a recent study by Andersen (2008), a logit model is used to determine the most relevant predictors of defaults of Norwegian banks. Out of an initial set of 23 predictive variables, Andersen (2008) found six predictors to be most relevant. These six predictors could, consistent with numerous previous studies, be categorized into the general risk factors capital adequacy, asset quality, earnings, and liquidity.

Another model that is examined in this study is the Cox proportional hazards model. This model, first developed by Cox (1972) in a biomedical framework, differs from the probit model and the logit model, because of the fact that the dependent variables denotes the time to failure instead of solely failure or non-failure. Although early economic applications of the hazards model were in labor economics, see Kiefer (1988), since the late 1980s the hazards model is employed to the prediction of bank failures as well. One of the first studies that employed the hazards model is Lane et al. (1986). Lane et al. (1986) compared the Cox proportional hazards model to the discriminant analysis and showed that both models performed about equivalent in correctly classifying banks into default and non-default categories and that the performances of both models proved to be adequate. When the forecasting horizon was extended from one year to two years, the hazards model slightly outperformed the discriminant analysis. Whalen (1991) also examines the usefulness of the hazards model as a bank failure prediction method. Consistent with the findings of Lane et al. (1986), Whalen (1991) concludes that, with a relative small number of predictive variables constructed from only publicly available data, the hazards model could be an effective early warning tool. The overall classification accuracy of the hazards models was again very adequate, as both type I and type II errors were relatively low. Wheelock & Wilson (2000) employed the hazards models to examine the predictive ability of this model as well as to determine the most relevant predictors of failures and acquisitions. The risk factors that were found to be relevant in predicting failures yet again consisted of capital adequacy, asset quality, management quality, earnings, and liquidity. Arena (2008) employs both the hazards model as well as the logit model in order to examine and compare bank failures in East Asia and Latin America. Arena (2008) concluded that bank failures can be explained by the individual financial statement data and that systemic shocks, that are different for the two regions and not for banks within a region, primarily destabilized the already inadequate banks in East Asia. The latter conclusion signifies a regional asymmetry regarding the resilience of the banking sector to systemic shocks and hence illustrates the heterogeneity among the banking sectors of different regions. With respect to the individual financial conditions of the banks, represented in the financial statement data, Arena (2008) concluded various financial ratios that proxied for capital adequacy, asset quality, and liquidity to be most informative about future bank failures.

Empirical literature studying the Neural Networks (NN), another model that is examined in this study, for the prediction of bankruptcies, started in the early 1990s and is still vigorous at present. One of the first studies employing NN in the framework of bankruptcy prediction was Odom & Sharda (1990). Odom & Sharda (1990) used the predictive variables that were found to be significant by Altman (1968) as the input to the NN. After multiple experiments, in which the proportions of healthy firms and bankrupt firms were varied, the performance of the NN was compared to the performance of the multivariate discriminant analysis. Based on type I and type II errors, Odom & Sharda (1990) concluded that the NN outperformed the multivariate discriminant analysis employing the same set of predictive variables. Whereas Odom & Sharda (1990) applied the NN to corporate failure prediction, Tam (1991) and Tam & Kiang (1992) were the first in applying the NN to the prediction of bank failures. Both studies compared the performance of NN to that of other default prediction models and concluded that the NN outperformed other models using a one-year forecast horizon and that the logit model outperformed other models using a two-year forecasting horizon. Salchenberger et al. (1992) also compared the performance of NN to that of the logit model. Based on predictions on thrift failures, Salchenberger et al. (1992) concluded that the NN outperformed the logit model using an 18-month forecasting horizon.

All studies that are described above appear to be able to achieve adequate performances regarding the prediction of defaults. Empirical evidence on the models' predictive performances relative to each other is somewhat mixed. Concerning the risk factors that determine the financial condition of a bank, there appears to be a consensus that identifies capital adequacy, asset quality, earnings and liquidity as being most important. A fifth risk factor, management quality, is more subjective and is found to be hard to proxy for in an objective way. In an attempt to further elucidate these conclusions, table I presents an overview of the findings, regarding the importance of individual predictive variables, of the various studies.

TABLE I
OVERVIEW OF PREDICTORS THAT ARE FOUND TO BE RELEVANT RELATION TO DEFAULT IN PREVIOUS STUDIES

Study	Model	Ratio[a]	General Risk Factors: Capital Adequacy	Asset Quality	Management Quality	Earnings	Liquidity
Martin (1977)	Logit	25/4	*GCARA*	*GCONI* *CI2LN*		*NITA*	
Avery & Hanweck (1984)	Logit	9/7	*KTA* *LNTA*	*NLTA* *CILNNL*		*NITA*	
Barth et al. (1985)	Logit	12/5	*NWTA*	*ISFTF*		*NITA*	*LATA* *LNTA*
Thomson (1991)	Logit	16/7	*NCAPTA*	*NCOTA* *NLTA*	*OVRHDTA* *INSLNTA*	*ROA*	*LIQ*
Andersen (2008)	Logit	27/6	*CAR*	*RMGL* *ELOSS* *CONS*		*ROA*	*NBLI*
Lane et al. (1986)	Hazard	21/7	*TETA*			*OEOI* *NITA*	*TLTD* *TLTA*
Whalen (1991)	Hazard	11/6	*CDR* *NPCR*	*TLTA*	*OHR*	*ROA*	
Wheelock & Wilson (2000)	Hazard	15/8	*TETA*	*TLTA* *OROTA* *NPLTA*	*CSTIN*	*NITA*	*LIQ*
Arena (2008)	Hazard	9/8	*TETA* *TETL*	*TLTA* *LNTA*		*ROA*	*LATL*

[a] The ratio of the total number of predictive variables screened to significant predictive variables.
GCARA = Gross Capital / Adjusted Risk Assets, *GCONI* = Charge-Offs / (Net Operating Income + Loss Provision), *CI2LN* = (Commercial and Industrial Loans + Loans to REITs and Mortgage Bankers + Construction Loans + Commercial Real Estate Loans) / Total Assets, *NITA* = Net Income / Total Assets, *KTA* = (Equity Capital + Loan Loss Reserve Allowances) / Total Assets, *LNTA* = Natural Logarithm of Total Assets, *NLTA* = Net Loans / Total Assets, *CILNNL* = Commercial and Industrial Loans / Net Loans, *NWTA* = Net Worth / Total Assets, *ISFTF* = Interest Sensitive Funds / Total Funds, *LATA* = Liquid Assets / Total Assets, *NCAPTA* = (Book Equity + Loan Loss Reserves – Loans 90 Days Past Due - Non-Accruing Loans) / Total Assets, *NCOTA* = Net Charge-Offs / Total Assets, *OVRHDTA* = Overhead / Total Assets, *INSLNTA* = Loans to Insiders / Total Assets, *ROA* = Return On Assets, *LIQ* = (Federal Funds Purchased – Federal Funds Sold) / Total Assets, *CAR* = Capital Adequacy Ratio, *RMGL* = Residential Mortgages / Gross Lending, *ELOSS* = Expected Loss based on PD / Gross Lending, *CONS* = Herfindahl Index for Loan Portfolio, *NBLI* = Norges Bank's Liquidity Indicator, *OEOI* = Operating Expense / Operating Income, *TLTD* = Total Loans / Total Deposits, *CDR* = Total Domestic Time Deposits / Total Assets, *NPCR* = (Primary Capital – Nonperforming Loans) / Average Total Assets, *TLTA* = Total Loans / Total Assets, *OHR* = Operating Expenses / Average Total Assets, *TETA* = Total Equity / Total Assets, *OROTA* = Other Real Estate Owned / Total Assets, *NPLTA* = Nonperforming Loans / Total Assets, *CSTIN* = Cost Inefficiency, *TETL* = Total Equity / Total Liabilities, *LATL* = Liquid Assets / Total Liabilities.

III. PREDICTION MODELS

As discussed in section I, one of the primary objectives of this study is to examine and to compare the predictive accuracy of multiple default prediction techniques, i.e. the probit model, the logit model, the proportional hazards model and the neural networks model. However, before examining the empirical value of the different models, it is important to thoroughly elucidate the underlying constructions and assumptions. This section provides an explanation on the multiple models.

III.A LOGIT MODEL

Since the seminal work of Martin (1977), the logit model has become one of the most commonly applied parametric failure prediction models in both the academic literature as well as in the banking regulation and supervision. The logit model is a form of a Generalized Linear Model (GLM) and is based on a binomial regression in which the dependent variable is dichotomous. The covariates can be continuous, categorical or a combination of both. This section describes the construction of the logit model utilized in this study.

In describing the construction of the logit model, let the probability of failure π_i first be defined as a linear function of a vector of covariates $X_i^{'}$ and a vector of regression coefficients β:

$$\pi_i = X_i^{'}\beta + \varepsilon \,. \tag{1.1}$$

A very common way of calculating this linear function is by way of Ordinary Least Squares (OLS). However, using this linear function, predicted values of the dependent variable π_i do not necessarily take on values between 0 and 1. Obviously, this is because of the fact that both $X_i^{'}$ as well as β can take on any real value. The first step in resolving this problem is to replace the probability of failure π_i by the odds ratio, which is defined as the ratio of π_i and its complement $1-\pi_i$:

$$odds_i = \frac{\pi_i}{1-\pi_i} \,. \tag{1.2}$$

Contrary to the real value of the probability of failure, the real value of the odds ratio can only take on any positive value, removing the upper bound of 1. The second step is to take the logarithm of the odds ratio. This results in the logit of the probability of failure, also referred to as the log-odds:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i) \,. \tag{1.3}$$

Taking the logarithm of the odds ratio removes the lower bound of 0. To clarify this, let the probability of failure π_i approach 0. This results in the odds ratio approaching 0 and the logit approaching $-\infty$. On the other hand, when π_i approaches 1, the odds ratio and the logit approach $+\infty$. Hence, taking the logarithm of the odds ratio transforms the (0,1) range into an unbounded range of $(-\infty,+\infty)$.

Based on the conclusions derived above, the logistic regression model can be derived. However, before deriving this regression model, it should be noted that the logit of the failure probability instead of the failure probability π_i itself is assumed to follow a linear model. Furthermore, it should be noted that the estimation of the logistic regression model is, as is common in GLM, based on maximum likelihood estimator.[2]

In the logistic regression model used in this study, the dependent variable Y_i takes on a value of 1 $(Y_i = 1)$ when a failure occurred in a pre-defined period following the date at which the financial statement data are determined. In other words, covariates at time $t = 0$ are related to possible failures between time $t = 0$ and time $t = T$. Obviously, Y_i takes on a value of 0 $(Y_i = 0)$ when no failure occurred between time $t = 0$ and time $t = T$. Consistent with an extensive amount of empirical literature on failure prediction, in this study T, which is the horizon of the failure prediction, equals 12 months.

---

2 Detailed information on the maximum likelihood estimator can be found in McFadden (1974).

Recall that the logit of the probability of failure follows a linear model, meaning that $logit(\pi_i)$ is a linear function of the covariates and the estimated coefficients:

$$logit(\pi_i) = X_i^{'}\beta \, , \tag{1.4}$$

where $X_i^{'}$ represents a vector of firm-specific covariates and vector β represents the corresponding estimated regression coefficients. The regression coefficients can be interpreted in a way consistent with a linear model. However, it should again be noted that the dependent variable is the logit of the probability of failure instead of the probability of failure itself. Hence, coefficient β_j represents the change in $logit(\pi_i)$ instead of the change in π_i as a result of a one-unit change in covariate x_j, ceteris paribus. Exponentiating equation 1.4 results in the following equation:

$$\frac{\pi_i}{1-\pi_i} = \exp(X_i^{'}\beta) \, . \tag{1.5}$$

Contrary to equation 1.4, this model presents the effect of changes in the covariates on the odds ratio. For example, when, ceteris paribus, x_j changes by one unit, the odds ratio changes by $\exp(\beta_j)$. Effects on the odds ratio are already more intuitively appealing than effect on the logit of the probability of failure. Therefore, exponentiating equation 1.4 into equation 1.5 is helpful in understanding the intuition behind this relatively simple model.

The final step in the construction of the logit model is to transform the dependent variable into the probability of failure π_i:

$$\pi_i = \frac{\exp(X_i^{'}\beta)}{1+\exp(X_i^{'}\beta)} = \Lambda(X_i^{'}\beta) \, . \tag{1.6}$$

In this equation, the dependent variable consist of the desired probability of failure, which is $\pi_i = P(Y_i = 1)$. This probability of failure π_i is based on a non-linear function of covariates and estimated coefficients. Equation 1.6 also demonstrates that this non-linear function, and with this the entire construction of the logit model, can be captured by the Cumulative Distribution Function (CDF) of the standard logistic distribution. When one finally wants to estimate the probability of bank q failing at any time between now and 12 months into the future, one should simply plug the firm-specific covariates X_q into equation 1.6. Accordingly, the out-of-sample estimations in section V and section VI are calculated by means of equation 1.6.

III.B Probit Model

The probit model is closely related to the logit model described above. Firstly, consistent with the logit model, the probit model is a form of a GLM and is based on a binomial regression in which the dependent variable is dichotomous. Secondly, the covariates can again be continuous, categorical or both. Thirdly, the estimation of the probit model is again based on maximum likelihood estimator. The main difference between the probit model and the logit model is that the construction of the logit model can be captured by the CDF of the standard logistic distribution and the construction of the probit model can be captured by the CDF of the standard normal distribution. This is represented in the following equation:

$$\pi_i = \int_{-\infty}^{X_i^{'}\beta} \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{(X_i^{'}\beta)^2}{2} \right) dX_i^{'}\beta = \Phi(X_i^{'}\beta)\,, \tag{2.1}$$

where $\Phi(.)$ represents the CDF of the standard normal distribution. The function that converts the score resulting from $(X_i^{'}\beta)$ into the probability of default represents the major difference between the logit model and the probit model.

The standard logistic distribution, assumed by the logit model, exhibits fatter tails than does the standard normal distribution, assumed by the probit model. Consequently, differing results of the logit model and the probit model basically reflect the relative strength of the assumptions regarding the transformation function. A second difference between the logit model and the probit model is based on the interpretability of the regression coefficients. Since the probit model does not have an equivalent to the odds ratio of the logit model,[3] the probit model regression coefficients are less intuitively appealing.

III.C Cox Proportional Hazards Model

The logit model and probit model described above, all result in predicted probabilities of failure at some unspecified time t over a predetermined interval, which is 12 months in this study. Consistent with these models, the Cox proportional hazards model, first developed by Cox (1972) and popularized by Kiefer (1988), can generate probabilities of bank failure within a predetermined interval. However, the proportional hazards model also generates estimates of probable time to failure. Thus, whereas other models only generate probabilities of failure within a certain interval, the proportional hazards model is able to provide an insight on when the failure will occur during this interval, allowing for a finer measurement of the effect of different variables on bank failure. Additionally, and consistent with the other models, the proportional hazards model does not make any assumption regarding the distributional properties of the predictive variables. This section describes the proportional hazards model utilized in this study.

In describing the construction of the proportional hazards model, let k denote the total number of banks $(i = 1, 2, ..., k)$ and let τ_i denote the failure time for each bank.[4] In defining the point process for the time to failure τ_i, let $Y_{i,t} = 1$ if $\tau_i \le T$ and $Y_{i,t} = 0$ otherwise. To clarify this, let $t = 0, 1, ..., T$ be the discrete points in time at which n banks are observed, starting at the beginning of the sample period $(t = 0)$ and ending at the end of the sample period $(t = T)$. Individual banks are started to be observed at some starting time t_i and are continued to be observed until failure time $\tau_i \le T$ or until censoring time $T_i \le T$. Here, censoring refers to the right-censoring of observations whenever a bank i is observed at time T_i, but not at time $T_i + 1$. In this study, as is common in bank failure hazards models, T_i usually equals the end of the sample period T. However, as banks can experience mergers, acquisitions and other reasons for a bank to be withdrawn from the data set, T_i

3 See equation 1.2 for the odds ratio of the logit model.

4 The failure time τ_i can both be discrete or continuous. In this study, τ_i is continuous and is measured in days.

does not necessarily equal T. In these cases, the final observation is at T_i and the censoring is assumed to be non-informative.

The dependent variable in the proportional hazards model is the time until failure τ_i, which implies that the survivor function, representing the probability of surviving longer than t periods, is of the following general form:

$$S_t = P(\tau_i > t) = 1 - F_t, \tag{3.1}$$

where F_t represents the CDF of the random variable t, which is the number of periods survived. The Probability Density Function (PDF) of t can be written as f_t and is equal to $-S_t^{'}$. Consistent with the literature on hazard models,[5] the hazard rate or intensity rate λ_i for the time to failure τ_i is defined as follows:

$$\lambda_i = \lim_{dt \to 0} \frac{P(t \le \tau_i < t + dt \mid t \le \tau_i)}{dt}, \text{ which equals } \frac{-S_t^{'}}{S_t}. \tag{3.2}$$

The hazard rate function in equation 3.2 represents the instantaneous probability of failure conditional on non-failure up to time t. Based on equation 3.2, multiple types of hazard models can be deducted. In a traditional proportional hazards model, the hazard function of equation 3.2 is assumed to be represented by the following:

$$\lambda_i(t \mid X_i^{'}\beta) = \lambda_{0,t}\, g(X_i^{'}\beta). \tag{3.3}$$

In equation 3.3, $X_i^{'}$ represents the set of predictive variables and β represents the vector of corresponding coefficients that are to be estimated by the model. The coefficients β illustrate the effect of the predictive variables $X_i^{'}$ on the likelihood of failure. Besides the parametric function $g(X_i^{'}\beta)$, there is also the nonparametric function $\lambda_{0,t}$, which is called the baseline hazard rate. The baseline hazard rate $\lambda_{0,t}$ comprises all factors that are not captured by the predictive variables, but do affect the likelihood of failure. Hence, $\lambda_{0,t}$ can be considered a time-effect. In constructing a probability of failure in a traditional hazard model, $\lambda_{0,t}$ is shifted by a factor determined by the parametric function $g(X_i^{'}\beta)$. This study assumes $g(X_i^{'}\beta)$ to be of an exponential form $\exp(X_i^{'}\beta)$. The resulting hazard function, known as the Cox proportional hazards model and used in a vast amount of empirical literature as well as in this study, is as follows:

$$\lambda_i(t \mid X_i^{'}\beta) = \lambda_{0,t} \exp(X_i^{'}\beta). \tag{3.4}$$

The corresponding survival function can be written as $S_i(t \mid X_i^{'}\beta) = S_{0,t}\exp(X_i^{'}\beta)$. Consistent with the hazard function and its baseline hazard rate $\lambda_{0,t}$, $S_{0,t}$ represents the baseline survival rate and only depends on time. In constructing a probability of survival, $S_{0,t}$ is shifted exponentially by $\exp(X_i^{'}\beta)$, which is determined by the set of predictive variables and the corresponding coefficients.

Models, like the proportional hazards model, in which the dependent variable comprises the time to default, or time to censoring when a bank does not have a default event in the sample period, have several advantages compared to models in which the dependent variable is merely a binary variable indicating a default or non-

[5] See for example Kiefer (1988) and Chava & Jarrow (2004).

default (Shumway, 2001). The most important reason to prefer the proportional hazards model is based on the fact that this model, as mentioned before, incorporates multiple observations for each covariate in estimating the relationship between the predictive variables and bank failure. When the condition of a bank deteriorates before the bank fails, the proportional hazards model is able to reveal this deteriorating condition by way of deteriorating covariates, i.e. financial statement data. Static models, on the contrary, only incorporate the financial statement data of one year prior to default in order to explain the default observation of the bank and hence do not allow the financial ratios to reveal the deteriorating condition of the bank. In other words, static models assume that the financial condition of a bank, in two or more years prior to a default event, is always healthy, due to the simple fact that no default event has occurred in the year after the relevant financial statement data. Hazard models, on the contrary, do not simply define the dependent variable to be a default or a non-default, but, by incorporating the time to default, allow a default event to be related to financial statement data multiple years prior to the default event.

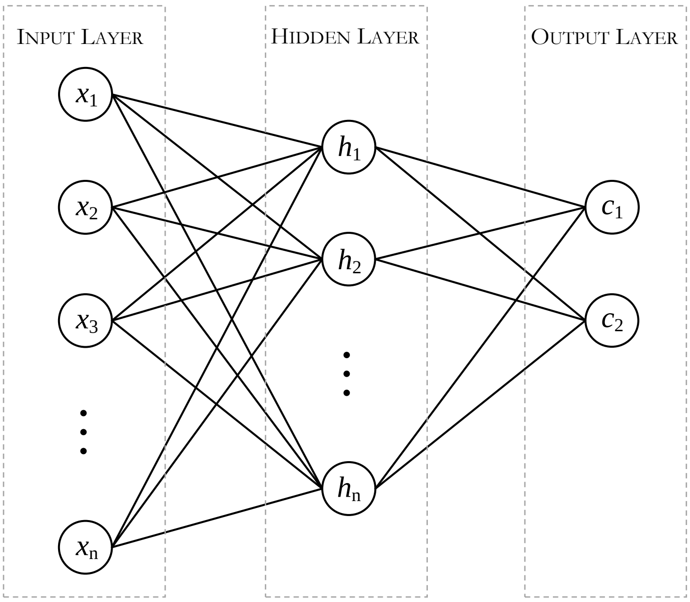
However, inherent to explicitly accounting for time to default, hazard models also have a drawback. Besides the right-censoring of observations, which is already mentioned in the description of the model, a bank can be at risk of failure before the start of the sample period $(t=0)$. This implies that the actual duration or time to failure is unknown, as bank i is already in existence, i.e. at risk of failure, before the financial statement data is available or before the start of the sample period. In this case the data is left-censored. Attempts to correct for this left-censored data are rare in empirical literature (D'Addio & Rosholm, 2002). The scarcity of corrections is based on two reasons. The first reason is because corrections for left-censored data are, often due to the non-availability of data, very complex and possible corrections result in either very restrictive assumptions or a disposure of all left-censored observations. The second reason is based on a general perception in empirical studies that left-censored data does not contain much information that can be exploited (D'Addio & Rosholm, 2002). In this study, a disposure of all left-censored data, e.g. removing all banks that where in existence before the start of the sample period, would result in the removal of too many banks with default events and would therefore result in an insufficient number of failure observations to base the different default prediction models on. Based on this, and based on the general perception that left-censored data does not contain much information that can be exploited, this study, as is consistent with empirical literature, does not correct for left-censored data.

III.D Neural Networks Model

The Neural Network (NN) method is a nonrestrictive and nonparametric alternative to bank failure prediction. Neural networks, first proposed by McCulloch & Pitts (1943) and inspired by the way the biological nervous systems works, consist of a number of homogeneous processing units, most often referred to as nodes, which are interconnected in a network. Each node can be described by a mathematical function transforming input signals into output signals. These output signals are then directed, depending on the structure and topology of the model, as input signals of nodes in the following layer. The NN model employed in this study is the Multilayer Perceptron (MLP) model, a common and dominant model in NN methods, which is based on a feed-forward network. To clarify this MLP model and NN models in general, figure I describes the architecture of the MLP model examined in this study.

As can be observed from figure I, this specific NN model is characterized by multiple layers, i.e. an input layer, a hidden layer and an output layer.[6] The input layer $(x_1, x_2, ..., x_n)$ comprises all predictive variables, i.e. financial statement data. The hidden layer $(h_1, h_2, ..., h_n)$ contains unobservable nodes, of which the value is some mathematical function of the predictive variables in the input layer. The exact form of the mathematical functions depends on both the type of the NN model as well as on user-controllable model specifications. The output layer (c_1, c_2) comprises the responses, i.e. the dependent variable specified as default or non-default. Since the dependent variable is dichotomous, the output layer is coded as two indicator variables. Each output can be described by a function of units in the hidden layer. Consistent with the functions of the hidden layer, the functions of the output layer depends on the type of the NN model and on user-controllable model specifications. As mentioned above, the MLP is based on a feed-forward network. This implies that the connections in the network flow forward from the input layer to the hidden layer or hidden layers to the output layer and that no information flows backwards.

FIGURE I
ARCHITECTURE OF MLP MODEL



The input layer comprises all predictive variables. The hidden layer contains unobservable nodes, of which the value is some mathematical function of the predictive variables in the input layer. The output layer comprises the responses, i.e. the dependent variable specified as default or non-default.

6 MLP models allow for more than one hidden layer, in which each unit of the second hidden layer is a function of the units in the first hidden layer and each response is a function of the units in the second layer. However, the MLP of this study is based on one hidden layer. This is supported by Cybenko (1989), Funahashi (1989) and Hornik et al. (1989), who state that only one hidden layer is required to approximate any arbitrary continuous function, provided that there are sufficient hidden units in the hidden layer.

Because of their nonrestrictive characteristics, NN models provide an analytical alternative to conventional failure prediction models, that can possibly be limited by assumptions on variable independence and on the relationship between the probability of default and its predictors. Moreover, because of the nonrestrictive nature, NN models can approximate numerous statistical models without hypothesizing certain relationships between variables. The relationships between variables are instead determined during the iterative learning process of the model.

A drawback inherent to the flexibility of NN models is that the underlying process of determining the relationship between the independent and dependent variables is not easily interpretable. Consequently, as one's main purpose is to examine underlying processes that explain relationships between independent and dependent variables, NN models would not be favored. On the other hand, when the interpretation of the underlying relationships between variables is not of importance, NN models might, due to their nonrestrictive characteristics, be preferred. Notwithstanding the fact that in this study it is important that rational inferences can be made regarding the relationship between financial statement data and the potential occurrence of a default and a rating transition, the NN model is estimated. Argumentation behind the examination of the NN model is based on its value as a nonparametric counterpart benchmark model.

IV. Description of Predictive Variables

Contrary to, for example, structural models, all the examined models described above, i.e. probit model, logit model, hazard model and MLP model, are not restricted by the input of predetermined predictive variables inherent to the use the relevant model. This implies that these models can contain all variables which are on beforehand expected to be of significant relevance. Researchers typically start out with either a vast number of predictive variables that proxy all risk factors that are expected to be of significant relevance or select a small set of predictive variables that were found to be of significant relevance in earlier empirical research on bank failures.

Whereas studies like Sinkey (1975) and Halling & Hayden (2006) base their research methods on the first approach, studies like Avery & Hanweck (1984), Whalen (1991) and Berg (2007) all base their methodologies on the second approach. However, regardless of the approach by which predictive variables are selected, there appears to be a consensus regarding the general risk factors that the predictive variables are intended to proxy for. These general risk factors, already accentuated in section II, for which the proxy variables are commonly found significant are capital adequacy, asset quality, management quality, earnings and liquidity. These five risk factors are commonly abbreviated as CAMEL risk factors.

This explicit differentiation between multiple risk factors originates from the early 1970s, when US federal regulators, i.e. the Federal Reserve System, the Federal Deposit Insurance Corporation (FDIC) and the Office of the Comptroller of the Currency (OCC), developed a numerical CAMEL rating system on which they based the frequency of their on-site examinations. Whereas banks with a sound CAMEL rating were examined every 18 months, problem banks with lower CAMEL ratings were examined more frequently. Since the development of the rating system and because of the simple structure, the use of CAMEL risk factors has become widespread in empirical literature. Predominantly empirical literature on US bank failures employs financial ratios that are related to the CAMEL risk factors (Arena, 2008).

Capital adequacy (C), usually measured by various ratios of capital to assets, is a measure of a bank's financial strength, i.e. a bank's ability of withstanding future unanticipated and abnormal losses. According to the Federal Reserve Board, a bank should maintain an amount of capital that properly reflects the nature and scope of risks to that bank. Hence, measures of capital adequacy should reflect the effect of credit risk, market risk and other forms of risk. In this study, the risk factor capital adequacy is proxied for by 4 financial ratios.

C1 Equity / Net Loans

Since equity protects a bank against asset breakdown, C1 measures the amount of protection, based on investments in equity, against losses in the loan book. When this ratio is high, a bank has a large equity cushion and is therefore able to absorb losses in the loan book. Hence, C1 is expected to exhibit a negative relationship with the probability of failure.

C2 Equity / Liabilities

This ratio represents a bank's leverage and is a measure of equity funding of the balance sheet and therefore another measure of capital adequacy. Based on the same argumentation as C1, this ratio is expected to exhibit a negative relationship with the probability of failure.

C3 Capital Funds / Total Assets

Capital funds comprise equity, hybrid capital and subordinated debt and, consistent with variable C1, measures an amount of protection against losses in the loan book. This ratio is expected to exhibit a negative relation with the probability of failure.

C4 Subordinated Debt / Capital Funds

This ratio represents the percentage of the total capital funds that is provided in the form of subordinated debt. The subordinated debt is the least permanent form of capital and, hence, a lower ratio is preferred. Consequently, this ratio is expected to exhibit a positive relationship with the probability of failure.

Asset quality (A) is a measure of the quantity of existing and potential future credit risk associated with the loan portfolios, the investment portfolios, the other assets, the other real estate owned and off-balance sheet transactions. The asset quality, largely depending on the risk management system, has a direct impact on all components of a bank's financial performance. Large quantities of classified assets will have a negative impact on earnings because of lower interest income and higher provisions to the loan loss reserve. On the contrary, inferior asset quality can diminish the liquidity of the loan portfolio, which consequently has a negative impact on capital adequacy. Because of the direct impact on all components of a bank's financial performance, the ability of the management to identify, measure, monitor and control the credit risk is of great importance. In this study, the risk factor asset quality is proxied for by 8 predictive variables.

A1 Loan Loss Reserves / Gross Loans

This ratio represents the percentage of the total loan portfolio that is reserved for charge-offs. A higher ratio indicates a poorer quality of the loan portfolio. Consequently, this ratio is expected to exhibit a positive relationship with the probability of default.

A2 Loan Loss Provision / Net Interest Revenue

This ratio represents the provisions in the profit & loss account relative to the interest income. Since a higher risk in the loan portfolio should be reflected by higher interest margins and not by higher loan loss provisions, a lower ratio is preferred. Consequently, this ratio is expected to exhibit a positive relationship with the probability of default.

A3 Impaired Loans / Gross Loans

This ratio measures the percentage of the total loans that is doubtful. The lower this ratio is the better the loan portfolio, i.e. the higher the asset quality. A lower ratio is preferred and, hence, A3 is expected to exhibit a positive relationship with the probability of default.

A4 Net Charge-Offs / Average Gross Loans[7]

This ratio represents the written-off loan loss reserves less the recoveries as a percentage of gross loans. It is therefore a measure of the percentage of current loans that have finally been written off. A lower ratio is preferred and therefore expected to exhibit a positive relationship with the probability of default.

A5 Net Charge-Offs / Net Income Before Loan Loss Provision

This ratio also represents a relative amount of net charge-offs. Here, the net charge-offs are measured as a percentage of the net income before loan loss provisions and is therefore determined by net income instead of the loan portfolio. A lower ratio is preferred and, hence, A5 is expected to exhibit a positive relationship with the probability of default.

A6 Total Earning Assets / Total Assets

This ratio measures earning assets as a percentage of total assets. Earning assets comprises all assets generating interest income or yielding fee income, which reflect the major sources of a bank's net income. A high ratio indicates a high asset quality and A6 is therefore expected to exhibit a negative relationship with the probability of default.

A7 Non-Earning Assets / Total Earning Assets

This ratio measures the amount of non-earning assets relative to the total of earning assets. Earning assets, including loans and leases yielding interest income or fee income, are, as mentioned, the major source of a bank's net income. A large amount of earning assets relative to non-earning assets is preferred. Hence, the relationship between A7 and the probability of default is expected to be positive.

[7] Average indicates that the data is averaged using the arithmetic mean of the value at the end of year t and at the end of year $t-1$. In order to not lose observations, when data is only available for one year, ratios are nonetheless calculated based on the single observation. The same applies to ratios of the oldest available year of a bank's financial statement data.

A8 Ln(Total Assets)

This variable, measuring the size of the bank, is expected to exhibit a negative relationship with the probability of default. Argumentation behind this expected relationship is based on the fact that large banks, holding more assets, are better able to diversify their portfolios and are therefore better able to reduce their asset risks. An alternative argumentation behind the expected negative relationship is related to the possible reluctance of regulators to liquidate large banks.

Management quality (M) is usually very difficult to measure objectively based on financial statement data. The capability of the management to identify measure, monitor and control all the risks of a bank's activities should ideally be reflected in this general risk factor. However, based on the use of external financial statement data, it is very hard to proxy for the aspects that together determine the quality of the management. Some researchers assume the quality of the management to be implicitly reflected by the financial statement data in the other four general risk factors. However, following Altman (1968), Halling & Hayden (2006) and Godlewski (2007), among others, this study attempts to explicitly proxy for the quality of the management. Consequently, two ratios are employed.

M1 Total Operating Expenses / Total Operating Income

Ratio M1 reflects the ratio of operating expenses and operating income. High operating expenses are not by definition problematic. However, high operating expenses in combination with relatively low operating income, i.e. a high ratio of M1, are not preferred and are therefore expected to result in a higher probability of default.

M2 Personnel Expenses / Total Operating Expenses

Variable M2 is expected to exhibit a negative relationship with the probability of default. Argumentation behind this expectation is based on the fact that a bank that has high personnel costs is probably able to attract higher quality personnel and is therefore able to reduce the risk of default.

The risk factor earnings (E) does not only reflect the quantity and historical trend of the earnings. It also reflects possible factors that may deteriorate the sustainability or quality of the earnings. Both the quantity as well as the quality of earnings can be deteriorated by excessive credit risk. This excessive credit risk can result in losses on loan portfolios and possibly require additional funds to loan loss provisions. Additionally, excessive market risk, i.e. the earnings of a bank have a high exposure to the volatility of interest rates, may deteriorate both the quantity as well as the quality of a bank's earnings. A third factor which may affect the quality of earnings is based on the reliance on extraordinary gains and nonrecurring events. When bank earnings heavily rely on such one-time gains, the quality of the earnings is low, future earnings are hard to forecast and adversely affected by possible non-occurring of favorable events. Hence, for a high quality of earnings, it is essential that profits come from a solid operating base. In this study, the risk factor earnings is proxied for by 8 predictive variables.

E1 Net Interest Margin

The net interest margin measures the net interest income relative to the earning assets. A high ratio indicates low-cost funding or simply higher margins demanded. A high net interest margin is preferred as long as asset quality is maintained and, hence, the net interest margin is expected to exhibit a negative relationship with the probability of default.

E2 Return On Average Assets (ROAA)

This ratio is a measure of the efficiency and operational performance of a bank and is determined by the asset returns. A high ratio implies a high efficiency and a high operational performance and is therefore preferred, implying a negative expected relationship with the probability of default.

E3 Return On Average Equity (ROAE)

This is a measure of the returns on shareholder funds and, hence, quantifies a bank's efficiency in generating profits from every unit of shareholder funds. A high ratio implies a high profitability and is therefore expected to exhibit a negative relationship with the probability of default.

E4 Dividend Pay-Out

This is a measure of the amount of post-tax profits paid out to shareholders. In general, the higher the ratio the better, but not if it is at the cost of restricting reinvestment in the bank and its ability to grow its business. The relationship between dividend pay-out and the probability of default is discussable, but is expected to be negative a priori.

E5 Income Net of Dividend / Average Equity

This ratio measures the income, net of dividend payouts, per unit of equity. Since equity is measured as the average equity, variable E5 effectively measures by what percentage the equity has increased from internally generated funds. A high ratio is preferred and E5 is expected to exhibit a negative relationship with the probability of default.

E6 Non-Operating Items / Net Income

This ratio indicates the percentage of net income that consists of unusual items. A high ratio signifies that income does not come from a solid operating base and is therefore not preferred. Hence, the expected relationship with the probability of default is positive.

E7 Profit before Tax / Interest Expense

This ratio measures the profitability of the bank, which is scaled by the total expenses on interest. A high ratio signifies a high profitability and is therefore preferred. The relationship with the probability of default is expected to be negative.

E8 Interest Income / Interest Expense

This ratio measures the profitability of a bank from the perspective of interest income, which is one of the major sources of a bank's income. When the ratio of interest income to interest expenses is high, this signals a high profitability on interest, which is preferable and will decrease the probability of default.

The risk factor liquidity (L) is measured to determine a bank's exposure to liquidity risk. In the determination of liquidity risk, it is vital to not only focus on current sources of liquidity and funding needs, but also to focus on future sources of liquidity and future funding needs. Consequently, banks, in order to operate in a sound manner, need to maintain a level of liquidity sufficient to meet current as well as future financial obligations. This implies that a bank should be able to manage unanticipated changes in funding sources and to manage unanticipated market conditions that directly affect the liquidity of assets. Hence, liquidity risk is based on current liquidity, future liquidity and the ability to ensure that liquidity can be maintained. In this study, the risk factor liquidity is proxied for by 7 predictive variables.

L1 Net Loans / Total Assets

This liquidity ratio indicates what percentage of the assets of a bank is tied up in loans. The higher this ratio the less liquid the bank will be. Hence, the relationship with the probability of default is expected to be positive.

L2 Net Loans / Deposits & Short-Term Funding

This ratio measures the amount of loans, being not exceedingly liquid, relative to the liquid deposits & short-term funding. A high ratio signifies low liquidity and is therefore not preferred. However, a large amount of deposits & short-term funding results in a higher risk of possible deposit run-off. Hence, the relationship with the probability of default is discussable.

L3 Liquid Assets / Deposits & Short-Term Funding

This liquidity ratio measures a bank's ability of withstanding a possible deposit run-off. It measures what percentage of customer funds and short-term funds could be met if they were withdrawn unexpectedly. They higher this ratio the higher the liquidity and, hence, the relation between this ratio and the probability of default is expected to be negative.

L4 Fixed Assets / Liquid Assets

This ratio simply measures the amount fixed assets relative to liquid assets. A high ratio implies that a large part of a bank's assets is in the form of fixed assets instead of liquid assets, signifying a low liquidity. The relationship between this liquidity measure and the probability of default is expected to be positive.

L5 Fixed Assets / Total Assets

This ratio is similar to L4. Here, fixed assets are measured as a percentage of total assets instead of liquid assets. Consistent with L4, this ratio is expected to be positively related to default.

L6 Deposits & Short-Term Funding / Total Loans

This ratio measures what percentage of total loans consists of the potentially more volatile deposits and short-term funding. A high ratio denotes a higher deposit run-off risk and is therefore not preferred. The relationship with the probability of default is expected to be positive.

L7 Deposits & Short-Term Funding / Total Assets

This ratio is similar to L6. Here, deposits & short-term funding is measured as a percentage of total assets instead of total loans. Consistent with L6, the relationship with the probability of default is expected to be positive.

Before advancing to the first estimation results, it is worthwhile to note that high correlations among predictive variables can lead to multicollinearity. When the objective is to simply predict defaults or rating transitions from a set of predictive variables and when no inferences are made regarding the individual impact of predictive variables on defaults or rating transitions, multicollinearity is not a subject of matter. However, when the objective is to estimate models that are based on a set of variables on which one can make rational inferences on, multicollinearity can be problematical due to the fact that it can bias coefficients and corresponding standard errors. This study, following the second objective, employs a general-to-specific modeling procedure in order to arrive at a final profile of variables on which rational inferences can be made. Using the general-to-specific modeling procedure, which will be elaborated on in the subsequent section, this study should take into account probable presence of multicollinearity. In attempting to avoid problems caused by multicollinearity, the predictive variables that are described above originate from a set of 49 variables that is, based on individual correlation coefficients, reduced to an amount of 29 variables that do not have individual correlation coefficients higher than 0.80.[8] Furthermore, in the estimation of the models, both in default prediction as well as in rating transition prediction, effects of variables exhibiting doubtful correlation coefficients that are moderately close to 0.80 are anticipated in the general-to-specific modeling procedures. This ensures that variables with doubtful correlation coefficients moderately close to 0.80 are not present in any specific-form model.

V. PREDICTION OF DEFAULTS

As discussed in section I, this study addresses both the comparison between multiple default prediction models as well as the ability of these models to predict credit rating transitions. This section comprises the first objective, which is the determination of the most accurate model in the prediction of defaults. The estimation period starts at 12/31/1987 and ends at 12/31/2008. Furthermore, this study solely examines US banks. The reason behind focusing on US banks only is based on the fact that restricting the examination to one country automatically controls for an important source of heterogeneity; domicile (Nickell et al., 2000; Arena, 2008). Moreover, US bank default information is much better available than default information on banks in any other country.

The employed procedure in determining the predictive performance is based on general-to-specific modeling and subsequent out-of-sample estimations, based on model-generated probabilities of default. This section starts with a description of the definition of default and the resulting default sample. This is followed by a

[8] The determination of correlation coefficients is based on the sample that is used for the estimation of defaults. However, it is confirmed that also in the estimation of rating transitions exceedingly high correlation coefficients are not present in any of the estimated models.

description of the utilized procedure of general-to-specific modeling employed in order to arrive at a final profile of variables on which the model-generated probabilities are based. Thereafter, the resulting specific-form models will be discussed and the out-of-sample estimations will be examined in order to draw conclusions regarding the predictive performances of the models.

In examining the relationship between financial statement data and bank defaults and the ability of multiple models to detect these possible relationships, it is of utmost importance to clearly define the assumed default events. Empirical literature however is not harmonious regarding the definition of a default event. Whereas some studies assume only a bankruptcy filing to be a default event, other studies define a non-payment of a contractual financial obligation as one of multiple default events. This study defines the following events to be default events: (i) bankruptcy filing, (ii) becoming in receivership, (iii) Federal Deposit Insurance Company (FDIC) closing and (iv) a credit rating transition to a default rating class.
The data on (i) bankruptcy filings comprises all Chapter 7 and Chapter 11 bankruptcy filings within the sample period and is obtained from Bloomberg Professional. The sample of (ii) banks in receivership comprises all banks that, within the sample period, could only avoid a bankruptcy by reorganization with the assistance of a court-appointed trustee. The sample of banks related to the third default event, (iii) FDIC closing, comprises all banks that failed and subsequently were closed by the FDIC. The (iv) credit rating transitions to a default rating class comprise all S&P and Fitch Ratings downgrades of long-term local issuer credit and senior unsecured debt, respectively, to the D and SD/RD rating classes.[9] S&P assigns a bank the rating class D when the bank has failed to meet one or more of its contractual financial obligations and S&P believes the bank will fail to meet all or substantially all of its contractual financial obligations as they come due. S&P assigns the rating class SD when a bank has failed to meet one or more of its contractual financial obligations, but S&P believes the bank will continue to meet payment obligations on other issues or classes in a timely manner. Fitch Ratings assigns a bank the rating class D when a bank has, according to Fitch Ratings, entered into bankruptcy filings, administration, receivership, liquidation or other formal winding-up procedure, or any other procedure which has ceased business. Fitch Ratings assigns a bank the rating class RD when a bank has, according to Fitch Ratings, not been able to meet one or more contractual financial obligations, but when the bank is believed not to have entered into bankruptcy filings, administration, receivership, liquidation or other formal winding-up procedure, or any other procedure which has ceased business. Based on the default events defined above, a sample of 64 default banks is obtained, of which 40 bank defaults are used to determine the final profile of predictive variables for each model, i.e. in-sample general-to-specific modeling, and 24 bank defaults are used to examine the out-of-sample predictive performance. In addition to the sample of 40 default banks that conform to one or more of the above defined default events, 1000 banks that did not conform to the default events were randomly selected and represent the sample of healthy banks in the in-sample estimation of the models. The 1000 healthy banks result in 7819 bank years of financial statement data and together with bank years at which a default bank was still healthy, i.e. observations of an eventually

9 Moody's credit rating classes do not comprise a default rating class. Hence, Moody's is not examined.

defaulted bank prior to its default, the aggregate sample of healthy banks comprises 8112 bank years.[10] In order to generate a sample on which the out-of-sample predictions are based, a default frequency consistent with the default frequency of the initial sample is employed. This results in an out-of-sample group of 24 default banks and 600 healthy banks.

As previously mentioned this section, the out-of-sample predictive performance of the multiple models is based on specific-modeling. This implies that the initial models, containing 29 predictive variables, are altered in a way that the final profile of estimators only contains a limited number of variables. In the selection of the final set of predictive variables this study takes into account (i) the statistical significance of the variables, i.e. variables that are incorporated in the specific-form models all have *p*-values less than 0.10, (ii) the correctness of the signs of the coefficients denoting the relationship between the predictors and the probability of default, and (iii) the in-sample fit of various profiles containing only significant variables. The latter criterion implies that when including a variable that satisfies criterion (i) and (ii) leads to a decrease in the in-sample fit measure of the relevant model, this variable is not incorporated in the specific-form model. The main objective of introducing the criteria and subsequent general-to-specific modeling is to determine a set of predictive variables that consistently prove to be important in relation to defaults in-sample, on which rational inferences can be made, and of which it can be expected that it will also perform well in out-of-sample estimations of future defaults, i.e. the estimated relationship is expected to be robust.

Before starting to describe the results of the general-to-specific modeling procedures, it should be noted that the MLP model, because of its nature, is not altered in the way described above. As explained in section III, the underlying process of determining the relationships between predictive variables and dependent variables is not easily interpretable. The MLP model, due to its structure of hidden layers among other reasons, does not result in coefficients, with corresponding signs and *p*-values, that can both be related to the probability of default. Based on this, no inferences can be made regarding the relevance of predictive variables. Moreover, the MLP model cannot conform to the abovementioned criteria of general-to-specific modeling. Hence, and because of the fact that the MLP model automatically assigns a large weight to predictive variables with high relevance and low weight to predictive variables with low relevance, the MLP model is not subject to general-to-specific modeling and is not present in the results on the specific-form models. However, the MLP model is present in the out-of-sample predictions and acts as a non-parametric counterpart benchmark model.

[10] In this context, a bank is defined healthy when none of the defined default events occurred during the year subsequent to the date of the financial statement data. Typically, the date of the financial statement data is 12/31/yyyy. In some exceptional cases, in which the financial statement data is of a date before 12/31/yyyy, the final date of the horizon is simply one year past this date of reporting, maintaining a one-year horizon.

TABLE II
ESTIMATION RESULTS OF GENERAL-TO-SPECIFIC MODELING PROCEDURE

	C2 [*p*-value]	A1 [*p*-value]	A2 [*p*-value]	A3 [*p*-value]	A5 [*p*-value]	M2 [*p*-value]	E5 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L4 [*p*-value]	Int. [*p*-value]
Probit	-0.034	0.175		0.110		-0.777	-0.004	-0.316	0.013	0.183	-2.522
	[0.014]	[0.003]		[0.000]		[0.004]	[0.030]	[0.005]	[0.008]	[0.031]	[0.000]
Logit	-0.093	0.497		0.218	0.002	-1.134	-0.010	-1.204	0.049	0.410	-5.706
	[0.008]	[0.000]		[0.000]	[0.093]	[0.029]	[0.023]	[0.001]	[0.001]	[0.046]	[0.000]
Hazard	-0.049	0.430	0.006	0.147		-1.007		-1.344	0.031	0.391	
	[0.077]	[0.000]	[0.027]	[0.000]		[0.015]		[0.000]	[0.006]	[0.005]	

C2: Equity / Liabilities, A1: Loan Loss Reserves / Gross Loans, A2: Loan Loss Provision / Net Interest Revenue, A3: Impaired Loans / Gross Loans, A5: Net Charge-Offs / Net Income Before Loan Loss Provision, M2: Personnel Expenses / Total Operating Expenses, E5: Income Net of Dividend / Average Equity, E8: Interest Income / Interest Expense, L1: Net Loans / Total Assets, L4: Fixed Assets / Liquid Assets.

Table II presents the empirical results of the probit model, the logit model and the hazard model in their specific form, based on the described general-to-specific modeling criteria. Results of the MLP model are not present in this table because of the reason mentioned above. One first imperative conclusion from table II is based on the large resemblance of predictive variables present in the different specific-form models. Table II shows that out of 29 initial predictive variables, the general-to-specific modeling of the probit model and the logit model results in practically the same set of variables that are most relevant based on the general-to-specific modeling criteria. The single difference between the results of the probit model and the logit model is the presence of A5, the ratio of net charge-offs and net income, in the specific form of the logit model. This result implies that when one wants to determine what variables are most relevant in relation with defaults, the probit model and the logit model appear to be about equivalent. The final profile of estimators of the hazard model is also very similar to the results of the probit model and the logit model, i.e. the specific form of the hazard model contains 8 relevant variables of which 7 are also present in the probit model and the logit model. Based on the fact that the initial set of estimators includes 29 predictive variables, this is a very large resemblance, indicating that the three models appear to be consistent in the in-sample determination of the variables most related to the occurrence of future default. A second conclusion from table II is that all the general risk factors, i.e. capital adequacy, asset quality, management quality, earnings, and liquidity, are present in the specific forms of all three models. This result is consistent with the results of Thomson (1991) and Halling & Hayden (2006), among others, and supports the consensus regarding the relevance of these general risk factors.

A third conclusion that can be drawn based on table II is that variables C2, A3, M2, E8 and L1 and L4 consistently prove to be the most relevant variables in relation to defaults. All these variables are present in all three models and consistently prove to be significant in relation to defaults. Note that, in describing the results presented in table II, the coefficients cannot directly be related to the probability of default and are not directly comparable across models. This is because of the fact that all models determine the probabilities of default using a different method.[11] Based on the fact that coefficients are not directly related to the probability

[11] The probit model uses the probit link function defined in equation 2.1, the logit model uses the logit link function defined in equation 1.6 and the hazard function uses the survival function defined in the section below equation 3.4.

of default and are not directly comparable across models, the following discussion on the results will primarily consider the correctness of the relevant signs.

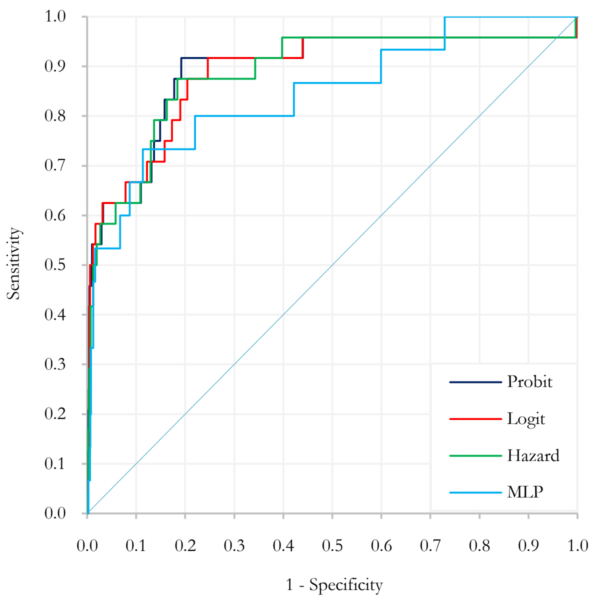
The results on variable C2, representing the ratio of equity and liabilities, imply that larger amounts of equity protects a bank against asset breakdown and an increase in C2, after applying the relevant link function, therefore results in a decrease of the probability of default. The significant relevance of the ratio of equity and liabilities is consistent with results of Arena (2008) and Godlewski (2007), among others. The fact that this simple capital ratio can contain valuable information concerning future bank failures is consistent with the results of Estrella et al. (2000), who examine the informative content of simple capital ratios relative to that of more complex risk-weighted capital ratios and conclude that the simple capital ratios perform equivalently well in the prediction of failures as more the complex capital ratios do.

The general risk factor asset quality is represented by 4 variables in total. The results on A1 demonstrate that increasing loan loss reserves relative to gross loans results in an increasing probability of default. This is according to the expectations, since high loan loss reserves, as a percentage of gross loans, indicates a poor loan portfolio quality and, hence, an increase in A1 results in an increase of the probability of default. The loan loss reserves is also found to be a predictor of failure and a proxy for the risk factor asset quality in for example Espahbodi (1991). The sign of the second asset quality variable, which is A2, is also according to expectations. An increase in the loan loss provisions, relative to the interest income, signals that higher risk in the loan portfolio is reflected by higher loan loss provisions, where ideally higher risk in the loan portfolio should simply be reflected by higher interest margins. Hence, a low ratio is preferred and the positive sign of A2 in table II confirms the expected relationship. The results on A3 are apparent and easily interpretable. When doubtful loans, as a percentage of gross loans, increase, this simply indicates a decrease of the quality of the loan portfolio and an increase of the probability of default. This expected relationship is profoundly supported by the results on A3 in table II. The relevance of impaired loans as a predictor of failures has extensively be discussed and confirmed in empirical literature on bank failures. Examples of studies that confirmed the relevance of impaired loans include Gajewski (1988), Whalen (1991), Halling & Hayden (2006), and Godlewski (2007). The results of variable A5 are also according to a priori expectations. A5 indicates what percentage of net income is used for charge-offs. Low charge-offs indicate a high asset quality, which is on its turn negatively related to the probability of default, explaining the positive coefficient of A5. The ratio of charge-offs and net income is found to be an important proxy for asset quality in Martin (1977), among others.

Concerning the results on the third general risk factor, which is management quality, the personnel expenses as a percentage of total operating expenses, signified by variable M2, proves to have a negative relationship with the probability of default. This is consistent with the expectations. A probable explanation for this relationship is based on the fact that when a bank has relatively high personnel costs, it probably attracts high quality personnel, leading to a decrease in the probability of default. However, in the interpretation of the results on M2, one should be cautious. Management quality remains a risk factor that is hard to approach in an objective way and therefore, based on these results, it can only be concluded that personnel expenses, as a predictive variable on itself, plays an imperative role. Whether or not this proxy truly reflects the quality of the management still remains doubtful. The ratios E5 and E8, variables that are proxying for the general risk factor earnings, consistently display negative signs. An increase in the variable E5, signals a higher income, net of dividend, per unit of equity. Variable E5 was expected to have a negative relationship with the probability of default and this negative relationship is confirmed by the results in table II. The results on E8 are easily interpretable. When the ratio of interest income to interest expenses is high, this signals a high profitability on interest, which is preferable and will decrease the probability of default. This negative relationship between E8 and the probability of default is confirmed by the results in table II. The ratio of net loans to total assets, which is represented by L1, proves to be the most prominent proxy for the general risk factor liquidity. The sign is, as expected, positive, indicating that a bank of which a large amount of the total assets is tied up in loans is less liquid and, hence, more likely to default than a bank of which only a small amount of the total assets is tied up in loans. The importance of the ratio of net loans and total assets is confirmed by a vast amount of empirical literature. Examples of these studies include Sinkey (1975), Avery & Hanweck (1984), Gajewski (1988), Whalen (1991), Wheelock & Wilson (2000), Godlewski (2007), and Arena (2008). The second important proxy for liquidity is L4, which represents the ratio of fixed assets and liquid assets. Again the sign is, according to expectations, negative. This negative sign is easily interpretable and narrowly related to that of L1. A large amount of fixed assets relative to liquid assets indicates a low liquidity, which obviously results in a higher probability of default.

The results of the discussed specific-form models can be sensitive to the sample used to estimate these models. Based on this possible unwanted sample dependency, conclusions regarding the relative predictive performance of the different models based on in-sample estimations and in-sample model fit would possibly be incorrect. Furthermore, from the practical perspective of regulators and supervisors, it can be stated that models that only perform well in-sample and ex post are not very useful. Hence, in order to determine the relative predictive performance of the models, out-of-sample testing is performed. The ability of a model to correctly classify banks into default or non-default in an independent out-of-sample group provides a reliable estimate of the actual default predicting performance. In the out-of-sample testing, as previously mentioned, a sample containing 24 default banks and 600 non-default banks is randomly selected, while ensuring that no banks are present in both the estimation sample as well as in the holdout sample. The out-of-sample assessment of predictive performance is based on the Receiver Operating Characteristic (ROC) curve, which is widely used as a measure of out-of-sample performance. For the comparison of multiple models, the area under the ROC curve (AUC) is measured relative to the area of a unit square. The AUC value represents the probability that the examined model assigns a higher probability of default to default banks than to non-default banks. When a model has no predictive performance, i.e. the model has no ability to correctly classify banks into default or non-default categories, the AUC equals 0.50. A value of 1.00 indicates perfect discrimination between default banks and non-default banks. A main advantage of the ROC curve as the out-of-sample measure of performance is that this measure is not subject to the arbitrary choice of a certain cut-off probability, on which decisions are based regarding the classification of banks into default or non-default. Whereas assuming a high cut-off probability leads to numerous type I errors, assuming a low cut-off probability leads to numerous type II errors. The arbitrary choice of the cut-off probability and possible disturbing effects that this choice can have on an objective assessment of models' predictive performance, makes conclusions drawn on a single cut-off value incomplete. Hence, the ROC curve, employing the entire range of possible cut-off probabilities, is selected. The MLP model, while not providing estimated coefficients like the other models do, does provide estimated probabilities of default. These estimated probabilities of default are, as mentioned previously, based on the weights that are automatically assigned to all 29 predictive variables. Hence, whereas the MLP model is not present in the results on the specific-form models in table II, it is present in the out-of-sample predictions, which are based on all 29 predictive variables for the MLP model. Figure II presents the ROC curves of the probit model, the logit model, the hazard model and the MLP model and table III presents results regarding the corresponding AUC values. Before moving to the results, it should be noted that the *p*-values presented in table III, denoting the significance of the differences between the AUC values, are determined based on the procedure of Hanley & McNeil (1983). This procedure takes into account the correlation between the AUC values that is induced by the paired nature of the data, i.e. correlation that results from the fact that the different ROC curves are derived from the same holdout sample of banks.

FIGURE II
ROC CURVES OF DEFAULT PREDICTION MODELS



Sensitivity represents the probability of correctly identifying a default bank and is calculated by dividing the number of correctly identified non-default banks by the sum of the correctly identified non-default banks and the incorrectly identified default banks. Specificity represents the probability of correctly identifying a non-default bank and is calculated by dividing the number of correctly identified default banks by the sum of the correctly identified default banks and the incorrectly identified non-default banks. The 45 degree reference line with AUC value 0.50 represents a random model with no ability to correctly classify banks into default or non-default categories.

TABLE III
DEFAULT PREDICTION: RESULTS ON AUC VALUES

			H_0: AUC = 0.50	Probit	Logit	Hazard	MLP
	AUC	Std. Error[a]	p-value[b]	p-value[c]	p-value[c]	p-value[c]	p-value[c]
Probit	0.891	0.044	0.000	-			
Logit	0.887	0.044	0.000	0.781	-		
Hazard	0.885	0.043	0.000	0.847	0.949	-	
MLP	0.846	0.059	0.000	0.464	0.529	0.570	-

[a] Standard errors are based on nonparametric assumption.
[b] Denotes asymptotic p-value.
[c] Denotes Hanley & McNeil (1983) p-value.

Based on both the graphical representations in figure II as well as on the first column of table III, denoting the AUC values, it can first be concluded that the predictive performance of the different models appears to be not very divergent. The probit model, the logit model and the hazard model display ROC curves and AUC values that are practically equal, i.e. AUC values of 0.891, 0.887 and 0.885 respectively. The MLP model appears to achieve a somewhat lower predictive performance of 0.846. Based on the equivalence of the probit model and the logit model, both in-sample as well as out-of-sample, it can be concluded the transformation function, the only difference between the probit model and the logit model, does not appear to be of great relevance in this study. A second conclusion that can be drawn from figure II and table III is that, despite small discrepancies in AUC values between the models individually, all models prove to achieve a high predictive performance with AUC values ranging from 0.846 to 0.891.[12] The fact that all models are better than random in correctly classifying banks into default or non-default categories is confirmed by the results in the third column of table III. Results, under the null hypothesis that the AUC equals 0.50, are highly significant for all models, indicating an outperformance of randomly assigning banks to default or non-default categories. However, this statistic is not very appealing, since figure II already displays the fact that the AUC of all four models are evidently closer to unity than the AUC value of 0.50 corresponding to the 45 degree line representing random classification. The final, and probably most imperative conclusion is that, based on the right side of table III, there appears to be no statistically significant differences between the out-of-sample predictive performances, i.e. AUC values, of all four models. The predictive performance of the hazard model, for example, does not significantly differ from the predictive performances of all the other models. This conclusion is not consistent with the conclusion of Shumway (2001), stating that hazard models have superior forecasting abilities when compared to static models like the probit model and the logit model.

The fact that the results of the probit model and the logit model do not significantly differ from each other and from the hazard model is, based on the ROC curve and the AUC values, no unexpected result. Based on the p-values regarding the predictive performance of the MLP model, when compared to the other models, it can be concluded that this MLP model does not significantly perform worse than the other models.

[12] Tape (2006) provides a rough guide for classifying AUC values, in which AUC values ranging from 0.50 to 0.60 are classified as 'fail', from 0.60 to 0.70 as 'poor', from 0.70 to 0.80 as 'fair', from 0.80 to 0.90 as 'good', and from 0.90 to 1.00 as 'excellent'. In providing a comparison with empirical literature on bank failure or corporate failure prediction, Chava & Jarrow (2004) find AUC values ranging from 0.739 to 0.911, Porath (2004) finds AUC values ranging from 0.793 to 0.814

Summarizing the conclusions regarding the predictive performances of the models, it can first be stated that all models perform very adequate in correctly classifying banks into default or non-default categories. This conclusion implies that bank supervisors and regulators may find a useful role for the models and their relevant predictors in the construction of their supervisory and regulatory frameworks. In addition to bank supervision and regulation, the determination of a bank's financial condition plays an imperative role in numerous other financial economic practices, e.g. supporting credit analysis, calculating counterparty credit risk, extending the standard financial engineering techniques for the valuation of credit derivatives or other credit sensitive instruments. Particularly in these financial economic practices, in which complete firm-specific information is not always available due to confidentiality, models that perform very adequate employing merely financial statement data that is effortlessly available, may perhaps be very beneficial. The second conclusion is that there appears to be no statistically significant differences between the performances of individual models when compared to each other.

VI. PREDICTION OF CREDIT RATING TRANSITIONS

Based on the results of the previous section, it can be concluded that the examined models performed very adequate in the prediction of defaults and that the out-of-sample performances of the different models were approximately equal. This section addresses the second objective of this study and examines whether or not credit rating transitions, a credit event of similar nature, can also be predicted using the default prediction models.

In the examination on whether or not credit rating transitions can also be predicted, it is important to clearly describe the event to be predicted and the way the credit rating transitions relate to defaults. Hence, in order to be able to draw conclusions regarding the predictability of rating transitions, this event should be strictly specified. Since both an actual default as well as a credit rating downgrade represent a deteriorating credit quality, the events are, as mentioned above, events of similar nature. However, the two credit events are not identical. A default essentially represents a deterioration of the credit quality to or below a level that is equivalent to the credit quality of D rated banks. A credit rating downgrade also represents a deteriorating credit quality, but the rating class to which the bank will be downgraded can, and often will be, different from the credit rating class D. Hence, in the prediction of defaults, the state of the bank at the end of the considered horizon is a priori specified. The ending state of a bank after a credit rating downgrade, on the contrary, is not strictly specified and can, despite the fact that it has been downgraded or upgraded, be any rating class. Basically, the only simple condition for the dependent variable to be positive is that the credit rating at the end of the considered horizon is different from the credit rating at the beginning of the horizon. However, this broad specification allows the dependent variable to represent a multitude of different events. For example, a downgrade from AAA, in which the credit quality falls below the minimum required level of credit quality for AAA rated banks, and a downgrade from CCC, in which the credit quality falls below the minimum required level of credit quality for CCC rated banks, are assumed to present an equivalent amount of information and are assumed to be related to the financial statement data in an identical way. Since these two events are evidently not identical, the dependent variable, rating downgrade or rating upgrade, is not homogeneous. In order to create more homogeneity in the events that are represented by the dependent variable, this study groups the rating transitions on the initial ratings class, resulting in the fact that the estimations of downgrades and upgrades are performed for each initial rating class independently. The independent estimation implies that all examined downgrades (upgrades) represent a change in credit quality to or below (above) the minimum required (maximum) credit quality of the initial rating class at time t resulting in more homogeneity in the predicted events.

Besides the fact that a default represents a deterioration of the credit quality to the level D and that a credit rating downgrade can represent multiple changes in credit quality, there is another aspect relating to possible differences between actual defaults and credit rating transitions. This aspect is based on the fact that credit ratings, typically provided by expert analysts, are not solely based on quantitative information, but include the subjective judgment of analysts. The subjective credit opinion of the analyst is usually based on various qualitative and quantitative factors. As credit ratings are strongly influenced by the analysts' subjective opinions, it is probable that the entire credit rating process might not be captured by a model that merely includes financial statement data. A second source of subjectivity in the credit rating process is based on the interrelation between credit rating agencies and issuers to be rated. The fact that credit rating agencies are profit-orientated and solely operate in function of the issuers, and therefore do not operate in function of the investors, results in an agency problem. The misaligned incentives, i.e. rating agencies have a quasi-regulatory responsibility and at the same time solely operate for the issuers opting for high credit ratings, are illustrated by Borrus (2002) and Galil (2003), among others.

Based on the abovementioned sources of subjectivity in the credit rating process and based on the discrepancy between the actual specifications of the two events, it is probable that models that perform very adequate in the prediction of defaults might not actually perform that adequate in the prediction of credit rating transitions. However, since the models performed very well in the prediction of defaults and since the two credit events, despite some discrepancies, both represent a deteriorating credit quality, one would still expect that the employed financial statement data contains valuable information regarding possible future credit rating transitions.

In the determination of the default models' ability to predict credit rating transitions, the models are tested in two distinctive ways. First, the specific-form models that are the result of the general-to-specific modeling procedure in the previous section, i.e. default prediction, are tested on their predictive accuracy. In other words, the exact same set of predictive variables for each model is used to examine the ability of these resulting specific-form models when they are used to predict credit rating transitions. The second way of testing is when the models start with the initial set of 29 variables, after which general-to-specific modeling is performed, using the same criteria as in the previous section, in order to construct specific-form models concentrated on predicting credit rating downgrades or credit rating upgrades. Based on this, it can be concluded whether or not the performance in predicting downgrades or upgrades can be improved, relative to the performance of the models using the profile of default predictor variables. A second conclusion that can be drawn is whether or not the same set of variables, as in default prediction, play an imperative role in the prediction of downgrades and upgrades.

The remainder of this section is structured as follows. Firstly, the credit rating transition sample is described. Secondly, the results of both the in-sample estimations as well as the out-of-sample predictions, using the specific-form models resulting from default prediction, will be discussed. Finally, new specific-form models are constructed and the results of the new out-of-sample predictions will be presented.

Data regarding the credit rating transitions are obtained from Bloomberg Professional and consist of a full credit rating transition history of US banks that are rated by the rating agencies Standard & Poor's (S&P), Moody's or Fitch Ratings. Consistent with the examination of defaults, the sample is restricted to US banks only. S&P, Moody's and Fitch Ratings all examine multiple rating types. Regarding the rating type of S&P, this study examines the long-term local issuer credit ratings. This long-term issuer credit rating is the current opinion of S&P regarding a bank's overall financial capacity, its credit quality, to pay its financial obligations. As S&P long-term local issuer credit ratings do not take into account the nature of specific obligations, the ratings reflect, as mentioned, the overall credit quality of a bank. Regarding the rating types of Moody's and Fitch Ratings, this study examines the senior unsecured debt ratings, that, consistent with the S&P long-term local issuer credit ratings, belong to the class of long-term credit ratings. Senior unsecured debt is not secured by collateral, but is paid off before other unsecured debt in the event of default. Despite the difference between the rating types, all credit rating transitions of S&P, Moody's and Fitch Ratings are assumed to be equivalent, e.g. an S&P downgrade of the long-term local issuer credit rating is assumed to represent an equivalent amount of information regarding the change in credit quality of a bank as does a Moody's or Fitch Ratings downgrade of senior unsecured debt. This assumption is often made in empirical literature and is implemented in the Basel II capital regime.[13]

Merging multiple credit rating agencies is a necessary step in obtaining a sufficient number of credit rating transitions. A straightforward solution to a problem of insufficient observations is to extend the scope of the study to multiple countries. However, and as mentioned earlier, examining multiple countries in one sample has the undesirable effect of heterogeneity. Based on this, and based on the fact that the examination of default prediction is also based on US banks,[14] it is decided to merely focus on US banks and therefore to merge multiple rating agencies.

In order to prevent that one bank, which is rated by multiple rating agencies, is present multiple times in the estimation sample, an order of priority is created regarding the rating agencies. The rating agency with the highest priority is S&P. The rating agencies with the second and third priorities are Moody's and Fitch Ratings, respectively. This order of priority implies that banks that are rated by S&P, Moody's and Fitch Ratings are removed from the sample two times in order to ensure that the bank is only rated by S&P and is

[13] The Standardized Approach for the calculation of the capital requirements relies on external credit ratings of counterparties. These external credit ratings must be assigned by an external credit rating agency that satisfies criteria which are described in the Capital Accord. S&P, Moody's and Fitch Ratings all satisfy these criteria.

[14] The examination of default prediction is merely based on US bank because US bank default information is much better available than any other default information.

only present once in the total sample. When a bank is not rated by S&P, but it is rated by Moody's and by Fitch Ratings, the bank is removed from the sample once in order to make sure that the bank is only rated by Moody's and is only present once in the total sample.

The ratings of S&P, Moody's and Fitch Ratings, as provided by Bloomberg Professional, consist not only of a primary rating grade, but also comprise rating modifiers[15] and signs of credit watches. Although, rating modifiers and credit watches provide a finer differentiation between banks within one rating class, these additional rating scales are not taken into account in this study. The fact that the additional rating scales are not taken into account is consistent with empirical literature on credit rating transitions. The reason that these additional rating scales are not taken into account, is based on the fact that transitions within rating classes are not expected to represent a change in credit quality of a bank that is of such a magnitude that it can be explained by the different models and the employed financial statement data. Moreover, the sample size of banks per rating class, when modifiers and credit watches are included, would not be sufficient for numerous rating classes.

Subsequently, the credit rating transitions obtained from Bloomberg Professional are combined with the database containing the predictive variables, obtained from Bankscope and described in an earlier section, resulting in a final sample containing 437 US banks and a total of 4674 bank years. This total number of 4674 bank years represents the total number of observations for which the dependent variable can either take on a value of 1 or can take on a value of 0. Restricted by the Bankscope database, providing the data on the predictive variables, the sample period of the credit rating transition database now also starts at 12/31/1987 and also ends at 12/31/2008. Based on the fact that for multiple initial credit rating classes there is no sufficient number of transitions to base statistical inferences on, only the downgrades from credit rating classes AA, A and BBB and the upgrades from credit rating classes A, BBB and BB are examined. Table IV provides an overview over the examined transitions and the corresponding number of observations.

Based on table IV, it can be concluded that only for AA downgrades, A upgrades and BBB upgrades there are an adequate number of observations to divide the sample in an in-sample group and an out-of-sample group.[16]

TABLE IV

NUMBER OF TRANSITIONS PER INITIAL CREDIT RATING CLASS

	A. Downgrades		*B. Upgrades*	
	N° downgrades	N° observations	N° upgrades	N° observations
AA	52	6687	-	-
A	23	1469	73	1469
BBB	16	850	50	850
BB	-	-	19	135

N° observations indicates the total number of observations in an initial rating class and includes both transitions as well as bank years in which a bank remained in the same initial rating class.

15 Whereas the modifiers of S&P and Fitch Ratings are represented by either a '+' or a '-', the modifiers of Moody's are defined by either a '1' or a '3'. The value of '1' corresponds to '+'.

16 The 52 AA downgrades are divided into an estimation sample of 30 AA downgrades and a holdout sample of 22 AA downgrades. The 73 A upgrades are divided into an estimation sample of 40 A upgrades and a holdout sample of 33 A upgrades. The 50 BBB upgrades are divided into an estimation sample of 31 BBB upgrades and a holdout sample of 19 BBB upgrades. It is ensured that the holdout samples all have an equal transition frequency as the estimation samples.

This implies that, whereas for AA downgrades, A upgrades and BBB downgrades both in-sample estimations as well as out-of-sample tests are performed, only in-sample estimations are performed for A downgrades, BBB downgrades and BB upgrades. By estimating the models only in-sample, no inferences can be made regarding the predictive performances of the models for these initial rating classes. However, in-sample estimations do provide an insight into the relative strength of the predictive variables in the prediction of downgrades and upgrades and are therefore nonetheless functional.

Table V presents the results on the in-sample estimations of the models using the variables that are the result of default prediction. Panel A, B and C present the results on AA downgrades, A downgrades and BBB downgrades, respectively. Panel D, E and F present the results on A upgrades, BBB upgrades and BB upgrades, respectively. Note that significant coefficients with the correct sign are presented in bold.[17]

A first conclusion that can be drawn based on table V is that the number of significant variables, of which the coefficient displays the correct sign, is considerably larger in the prediction of downgrades than in the prediction of upgrades. Whereas in panel A, B and C there is a total number of 29 significant coefficients displaying the correct sign, in panel D, E and F there is a total number of 10 significant coefficients accompanied by the correct sign. It should be noted that the in-sample estimation of A upgrades results in a total of 6 significant coefficients with the correct sign, related to A3 and L1. This implies that, although this is less than for the downgrade estimations, A upgrades may be slightly better predicted than BBB upgrades and BB upgrades. A second conclusion that can be drawn based on table V is that variable A3, representing the ratio of impaired loans and gross loans, undoubtedly plays the most imperative role in describing both downgrades as well as upgrades. Furthermore, the variables E8 and L1, representing the ratio of interest income and interest expense and the ratio of net loans and total assets, respectively, prove to be important variables in the prediction of downgrades. Upgrades in general, however, do not appear to be closely related to these variables. The conclusion that variables A3, E8 and L1 display significant results in multiple initial rating classes indicates that these variables possibly play an important role in the credit rating process. Note that, variables A3, E8 and L1 were also among the variables that consistently proved to play the imperative role in relation to defaults and, particularly A3 and L1, are found to be significant in a vast amount of empirical literature. Consequently, it can be stated that variables A3, E8 and L1 are likely to be suitable indicators of the financial condition of a bank.

A third conclusion on table V is that variables C2 and M2, that consistently proved to be among the most relevant variables in relation to defaults, appear to be irrelevant in relation to credit rating transitions. This could indicate that C2 and M2, representing the ratio of equity and liabilities and the ratio of personnel expenses and operating income, respectively, are not taken into consideration in the credit rating process. Regarding the proxy for management quality, one might argue that rating agencies base their evaluation of the management efficiency on other factors, e.g. active meetings and discussions with the management of a bank, rather than the proxy employed in this study.

---

[17] Based on a significance level of 10%.

TABLE V

ESTIMATION RESULTS OF DEFAULT MODELS APPLICATION TO CREDIT RATING TRANSITIONS

	C2 [*p*-value]	A1 [*p*-value]	A2 [*p*-value]	A3 [*p*-value]	A5 [*p*-value]	M2 [*p*-value]	E5 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L4 [*p*-value]	Int. [*p*-value]
A. Downgrades from AA											
Probit	0.019	-0.195		**0.510**		**-1.319**	0.003	**-0.308**	**0.019**	-0.187	-1.685
	[0.189]	[0.098]		**[0.000]**		**[0.087]**	[0.682]	**[0.054]**	**[0.013]**	[0.740]	[0.021]
Logit	0.036	-0.396		**0.965**	0.009	-1.454	0.005	**-0.643**	**0.042**	-0.365	-3.745
	[0.180]	[0.111]		**[0.000]**	[0.263]	[0.364]	[0.710]	**[0.056]**	**[0.009]**	[0.724]	[0.016]
Hazard	0.030	-0.316	0.006	**0.668**		-1.539		**-0.600**	**0.034**	-0.501	
	[0.153]	[0.222]	[0.377]	**[0.000]**		[0.302]		**[0.059]**	**[0.017]**	[0.673]	
B. Downgrades from A											
Probit	-0.005	-0.125		**0.302**		-0.706	**-0.020**	**-0.304**	**0.023**	0.067	-2.672
	[0.805]	[0.017]		**[0.000]**		[0.358]	**[0.053]**	**[0.092]**	**[0.021]**	[0.807]	[0.003]
Logit	-0.008	-0.222		**0.562**	0.006	-0.561	**-0.049**	**-0.799**	**0.052**	0.178	-5.756
	[0.845]	[0.039]		**[0.001]**	[0.190]	[0.776]	**[0.040]**	**[0.093]**	**[0.021]**	[0.767]	[0.006]
Hazard	-0.018	-0.263	**0.028**	**0.438**		1.029		-0.704	0.033	0.106	
	[0.713]	[0.445]	**[0.036]**	**[0.000]**		[0.668]		[0.125]	[0.111]	[0.850]	
C. Downgrades from BBB											
Probit	-0.027	**0.252**		**0.343**		-1.322	-0.004	-0.146	-0.018	0.304	-0.875
	[0.561]	**[0.037]**		**[0.001]**		[0.288]	[0.517]	[0.422]	[0.161]	[0.142]	[0.439]
Logit	-0.042	**0.654**		**0.632**	-0.008	-3.843	**-0.052**	-0.482	-0.038	**0.669**	-0.788
	[0.723]	**[0.014]**		**[0.002]**	[0.575]	[0.210]	**[0.061]**	[0.315]	[0.230]	**[0.093]**	[0.763]
Hazard	-0.107	**0.581**	0.002	**0.583**		-3.433		-0.674	-0.044	**0.563**	
	[0.365]	**[0.005]**	[0.682]	**[0.000]**		[0.228]		[0.148]	[0.117]	**[0.081]**	
D. Upgrades from A											
Probit	0.005	-0.066		**-0.387**		-0.534	-0.002	-0.064	**-0.014**	0.082	0.100
	[0.244]	[0.614]		**[0.065]**		[0.449]	[0.812]	[0.468]	**[0.037]**	[0.695]	[0.879]
Logit	0.007	-0.115		**-0.813**	-0.001	-0.959	-0.004	-0.164	**-0.029**	0.175	0.723
	[0.307]	[0.663]		**[0.072]**	[0.889]	[0.508]	[0.833]	[0.390]	**[0.029]**	[0.666]	[0.576]
Hazard	0.004	-0.078	-0.005	**-0.766**		-0.997		-0.173	**-0.026**	0.127	
	[0.363]	[0.774]	[0.732]	**[0.067]**		[0.467]		[0.346]	**[0.039]**	[0.736]	
E. Upgrades from BBB											
Probit	0.030	-0.094		0.221		0.230	0.002	0.010	-0.001	-0.635	-1.617
	[0.205]	[0.516]		[0.168]		[0.856]	[0.911]	[0.779]	[0.910]	[0.146]	[0.130]
Logit	**0.078**	-0.218		0.285	0.036	4.939	0.017	0.000	0.007	-1.465	-6.185
	[0.084]	[0.546]		[0.410]	[0.007]	[0.117]	[0.604]	[0.996]	[0.781]	[0.126]	[0.024]
Hazard	0.055	-0.247	0.023	0.366		2.938		0.002	0.001	-1.373	
	[0.206]	[0.454]	[0.237]	[0.181]		[0.360]		[0.969]	[0.965]	[0.140]	

TABLE V (CONTINUED)

ESTIMATION RESULTS OF DEFAULT MODELS APPLICATION TO CREDIT RATING TRANSITIONS

	F. Upgrades from BB										
	C2	A1	A2	A3	A5	M2	E5	E8	L1	L4	Int.
	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]	[*p*-value]
Probit	-0.089	0.081		**-0.615**		0.626	-0.003	0.403	-0.022	0.347	0.090
	[0.157]	[0.463]		**[0.052]**		[0.646]	[0.860]	[0.156]	[0.133]	[0.278]	[0.939]
Logit	-0.182	0.164		**-1.071**	0.000	1.010	-0.002	0.686	-0.039	0.579	0.478
	[0.163]	[0.407]		**[0.059]**	[0.972]	[0.675]	[0.947]	[0.172]	[0.126]	[0.283]	[0.815]
Hazard	-0.084	0.397	**-0.057**	-0.776		-2.768		0.367	-0.022	1.162	
	[0.496]	[0.031]	**[0.004]**	[0.164]		[0.222]		[0.409]	[0.324]	[0.018]	

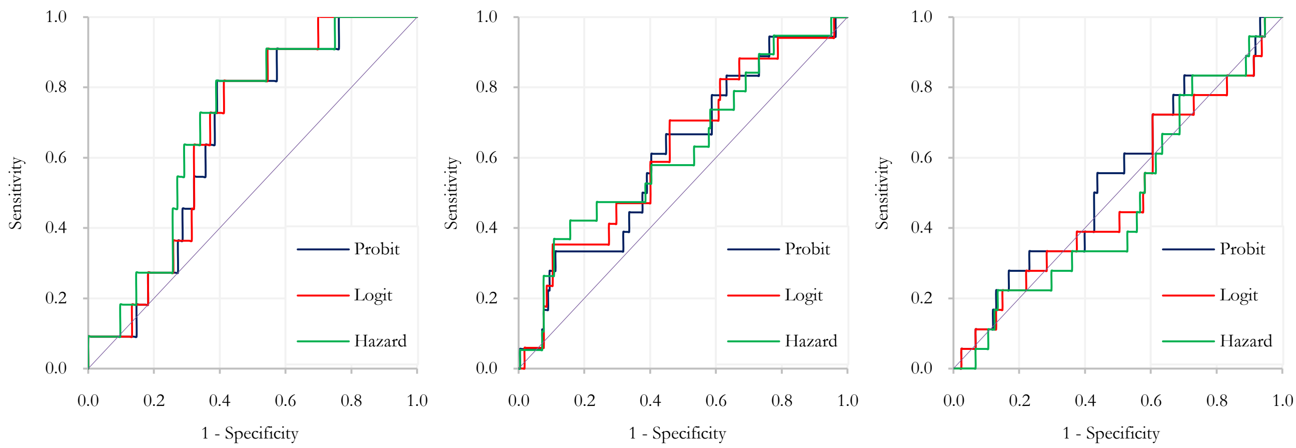
Significant coefficients with the correct sign are presented in bold.
C2: Equity / Liabilities, A1: Loan Loss Reserves / Gross Loans, A2: Loan Loss Provision / Net Interest Revenue, A3: Impaired Loans / Gross Loans, A5: Net Charge-Offs / Net Income Before Loan Loss Provision, M2: Personnel Expenses / Total Operating Expenses, E5: Income Net of Dividend / Average Equity, E8: Interest Income / Interest Expense, L1: Net Loans / Total Assets, L4: Fixed Assets / Liquid Assets.

In order to test the predictive performance of the models described in table V, out-of-sample tests are performed. Recall that these out-of-sample tests, due to a shortage of transitions from some initial rating classes, are only performed on the downgrades from AA, the upgrades from A and the upgrades from BBB. Figures III.A, III.B and III.C present the ROC curves of the out-of-sample estimations on transitions from these initial rating classes. Table VI presents the descriptive statistics regarding the corresponding AUC values.[18] Note that the out-of-sample predictive performances of the MLP model are not present in both the three figures as well as in table VI. The reason behind this is based on the fact that in the prediction of defaults the MLP model, because of the structure described earlier, did not result in a specific-form model. And since later on in this section all models are estimated a second time, in which the MLP model will again consist of all 29 variables, estimating the MLP model at this time would yield the same results as later on in this section.

Based on the figures III.A, III.B and III.C as well as on the first column of table VI it can be concluded that, consistent with the results on the prediction of defaults and already mentioned above, the out-of-sample performances of the different models are not very divergent. Whereas the hazard model appears to perform slightly better than the other models in the prediction of AA downgrades, its performance in the prediction of A upgrades appears to be equal to other models and the prediction of BBB upgrades is conceivably even worse than that of other models, resulting in the fact that again no strict inferences can be made regarding the performances of the models relative to each other.

[18] Note that table VI does not provide *p*-values regarding the differences between the models' performances relative to each other. The reason for this is based on the fact that Hanley & McNeil (1983) do not present sufficient information regarding the calculation of the correlation adjusted *p*-values when the average AUC values of two models are below 0.70, which is the case at this time. However, since the main objective of this section is to examine whether or not rating transitions can effectively be predicted and not essentially to examine which model performs best in this prediction of rating transitions, testing the performances on their differences is not an essential part of this section. Furthermore, based on the presented ROC curves and corresponding AUC values, it can already be observed that the differences between the models appear to be again very small and that tests on these differences would be highly likely to produce insignificant results. This further indicates the fact that not performing statistical tests on the differences between the models is no crucial deficiency in this section.

Figure III.A	**Figure III.B**	**Figure III.C**
ROC Curves AA Downgrades	ROC Curves A Upgrades	ROC Curves BBB Upgrades



Sensitivity represents the probability of correctly identifying a default bank and is calculated by dividing the number of correctly identified non-default banks by the sum of the correctly identified non-default banks and the incorrectly identified default banks. Specificity represents the probability of correctly identifying a non-default bank and is calculated by dividing the number of correctly identified default banks by the sum of the correctly identified default banks and the incorrectly identified non-default banks. The 45 degree reference line with AUC value 0.50 represents a random model with no ability to correctly classify banks into default or non-default categories.

Table VI
Credit Rating Transition Prediction: Descriptive Statistics on AUC Values (1)

	AUC	Std. Error[a]	H_0: AUC = 0.50 *p*-value[b]
A. Descriptive Statistics on AA Downgrades			
Probit	0.666	0.068	0.044
Logit	0.676	0.062	0.052
Hazard	0.696	0.066	0.030
B. Descriptive Statistics on A Upgrades			
Probit	0.612	0.067	0.115
Logit	0.624	0.069	0.089
Hazard	0.621	0.069	0.080
C. Descriptive Statistics on BBB Upgrades			
Probit	0.539	0.071	0.583
Logit	0.495	0.073	0.943
Hazard	0.477	0.068	0.747

[a] Standard errors are based on nonparametric assumption.
[b] Denotes asymptotic *p*-value.

A second, and more important, conclusion based on table VI is that all models achieve AUC values that appear to be significantly different from 0.50 for the AA downgrades prediction, using a significance level of 5%. Furthermore, the *p*-values on the differences between the AUC values of A upgrades and 0.50 are somewhat higher and are not all about the 5% significance level. Based on the presented *p*-values, it can however still be concluded that the models, possibly with the exception of the probit model in this particular situation, are better able to correctly classify banks into future upgrades and future non-upgrades than a random selection of banks would be. However, in discussing these results, it should be noted that AUC values below 0.70, notwithstanding the fact that they significantly differ from 0.50, are still not very persuasive.[19] The predictive performances on BBB upgrades are obviously no better than random, implying that rating upgrades from this initial rating class are not suitably captured by the models and the employed predictor variables. The inadequate resulting AUC values on BBB upgrades are, based on table V presenting the significance of the predictive variables, not very

[19] In the classification of AUC values by Tape (2006), AUC values ranging from 0.60 to 0.70 are classified as 'poor'.

surprising since there proved to be hardly any significant variables in the in-sample estimation of BBB upgrades.

The final, and most imperative, conclusion that can be drawn based on the presented figures III.A, III.B, III.C and table VI is that the predictive performances of all models decreased by a considerable amount when compared to the performances of the exact same models when used in the context of default prediction. Focusing on the difference between the prediction of defaults and of AA downgrades, which proved to be the best predicted type of rating transition, it can be noticed that whereas the models achieve AUC values of approximately 0.89 in the prediction of defaults, the same models achieve AUC values not higher than 0.70 when they are used to predict AA downgrades. Table VII presents results regarding the differences between the AUC values of default prediction and the prediction of downgrades from initial rating class AA and confirms the fact that for all models the predictive performances significantly decrease when the models are used to predict credit rating transitions.[20]

TABLE VII

DIFFERENCES BETWEEN DEFAULT PREDICTION AND AA DOWNGRADE PREDICTION

	Default Prediction		*AA Downgrade Prediction*		H_0: $AUC_{Default} = AUC_{Downgrade}$
	AUC	Std. Error[a]	AUC	Std. Error[a]	*p*-value[b]
Probit	0.891	0.043	0.666	0.068	0.004
Logit	0.887	0.044	0.676	0.062	0.006
Hazard	0.885	0.043	0.696	0.066	0.017

[a] Standard errors are based on nonparametric assumption.

[b] Denotes two-sided *p*-value.

In examining the results presented above, one should recall that the estimates on the probability of a transition are based on the set of variables that is optimized using default or non-default as the event to be predicted. Since rating downgrades and defaults are credit events that are only of similar nature and are obviously not precisely the same, it is possible that predictive variables that are highly relevant in the prediction of defaults are less relevant in the prediction of credit rating transitions. On the other hand, variables that are found to be irrelevant in relation to defaults might play a more important role in relation to credit rating transitions. In order to examine this and in order to examine whether or not the out-of-sample performance of rating transition prediction can be improved relative to the performances presented in table VI, the models are optimized one more time. In this last optimization, all models start with the initial set of 29 initial variables. The same procedure of general-to-specific modeling is used as in section V, leaving only (i) significant variables with the (ii) correct sign that collectively result in the (iii) best in-sample fit.

Table VIII presents the results on the specific-form models that are estimated concentrating on the prediction of credit rating transitions. As in section V, the MLP model is solely used in the out-of-sample comparison of performances and is therefore not present in table VIII.

[20] In these tests on the differences between default prediction and rating transition prediction, the procedure of Hanley & McNeil (1983), used to determine the correct *p*-values in the previous section, is not necessary since the ROC curves are derived from independent samples.

TABLE VIII

ESTIMATION RESULTS OF GENERAL-TO-SPECIFIC MODELING PROCEDURE

A. Downgrades from AA

	C3 [*p*-value]	A1 [*p*-value]	A3 [*p*-value]	A8 [*p*-value]	E3 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L2 [*p*-value]	L3 [*p*-value]	Int. [*p*-value]
Probit			0.369		-0.021	-0.264	0.025			-2.312
			[0.002]		[0.095]	[0.082]	[0.001]			[0.000]
Logit			0.720		-0.045	-0.604	0.054			-4.302
			[0.001]		[0.079]	[0.069]	[0.001]			[0.000]
Hazard			0.629		-0.037	-0.544	0.049			
			[0.000]		[0.088]	[0.077]	[0.001]			

B. Downgrades from A

	C3 [*p*-value]	A1 [*p*-value]	A3 [*p*-value]	A8 [*p*-value]	E3 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L2 [*p*-value]	L3 [*p*-value]	Int. [*p*-value]
Probit			0.077		-0.026	-0.351	0.026			-2.845
			[0.016]		[0.012]	[0.055]	[0.005]			[0.000]
Logit			0.152		-0.066	-1.028	0.059			-5.141
			[0.010]		[0.002]	[0.036]	[0.006]			[0.005]
Hazard			0.141		-0.061	-1.048	0.056			
			[0.006]		[0.001]	[0.028]	[0.006]			

C. Downgrades from BBB

	C3 [*p*-value]	A1 [*p*-value]	A3 [*p*-value]	A8 [*p*-value]	E3 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L2 [*p*-value]	L3 [*p*-value]	Int. [*p*-value]
Probit		0.290	0.344			-0.331				-2.202
		[0.000]	[0.000]			[0.080]				[0.000]
Logit		0.735	0.677			-1.153		0.011		-4.489
		[0.000]	[0.000]			[0.034]		[0.062]		[0.000]
Hazard		0.709	0.519			-1.205		0.010		
		[0.000]	[0.000]			[0.012]		[0.055]		

D. Upgrades from A

	C3 [*p*-value]	A1 [*p*-value]	A3 [*p*-value]	A8 [*p*-value]	E3 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L2 [*p*-value]	L3 [*p*-value]	Int. [*p*-value]
Probit	0.036		-0.606	0.170				-0.009		-2.501
	[0.004]		[0.013]	[0.002]				[0.041]		[0.000]
Logit	0.066		-1.304	0.347				-0.018		-4.529
	[0.006]		[0.014]	[0.001]				[0.044]		[0.001]
Hazard	0.043		-1.307	0.346				-0.016		
	[0.001]		[0.011]	[0.001]				[0.057]		

E. Upgrades from BBB

	C3 [*p*-value]	A1 [*p*-value]	A3 [*p*-value]	A8 [*p*-value]	E3 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L2 [*p*-value]	L3 [*p*-value]	Int. [*p*-value]
Probit				0.533					0.030	-6.726
				[0.000]					[0.038]	[0.000]
Logit				1.080					0.059	-13.245
				[0.000]					[0.036]	[0.000]
Hazard				0.988					0.055	
				[0.000]					[0.036]	

TABLE VIII (CONTINUED)
ESTIMATION RESULTS OF GENERAL-TO-SPECIFIC MODELING PROCEDURE

	F. Upgrades from BB									
	C3 [*p*-value]	A1 [*p*-value]	A3 [*p*-value]	A8 [*p*-value]	E3 [*p*-value]	E8 [*p*-value]	L1 [*p*-value]	L2 [*p*-value]	L3 [*p*-value]	Int. [*p*-value]
Probit			-0.607							-0.605
			[0.032]							[0.009]
Logit			-1.100							-0.972
			[0.036]							[0.015]
Hazard			-1.004							
			[0.038]							

C3: Capital Funds / Total Assets, A1: Loan Loss Reserves / Gross Loans, A3: Impaired Loans / Gross Loans, A8: Ln(Total Assets), E3: Return on Average Equity, E8: Interest Income / Interest Expense, L1: Net Loans / Total Assets, L2: Net Loans / Deposits & Short-Term Funding, L3: Liquid Assets / Deposits & Short-Term Funding.

The first conclusion based on table VIII is that the determination of relevant variables for every type of examined rating transition individually is equivalent across the multiple models. With the probit model in the prediction of BBB downgrades as the only exception, it can be concluded that, consistent with the results on the prediction of defaults, the models show a great uniformity regarding the in-sample determination of the variables most related to the occurrence of particular future transitions.

A second conclusion that can be drawn from table VIII is that there is a reasonable resemblance between the variables that are identified as important predictors of downgrades, i.e. AA downgrades, A downgrades and BBB downgrades. Whereas the models identify variables A3, E3, E8 and L1 as the relevant predictors of both AA downgrades and A downgrades, variables A1, A3, E8 and L2 are identified as the relevant predictors of BBB downgrades. At this point, it should be noted that variable L1 represents the ratio of net loans and total assets and variable L2 represents the ratio of net loans and deposits & short-term funding. The latter signals the importance of net loans as a predictor of future downgrades. Based on these outcomes, it can be concluded that the set of variables, containing A3, E3, E8 and L1 or L2, is most relevant in the in-sample estimations of the specific-form models on downgrades. Consistent with the results of the first estimation of credit rating transitions, the risk factors capital adequacy and management quality appear to be not very relevant. This may again indicate that these risk factors are not considered in the credit rating process or are not appropriately proxied for by the predictors employed in this study.

The third conclusion is that the variables A3, E8 and L1 are also among the variables that consistently proved to be the most relevant variables in relation to defaults and were also identified as important in the previous estimation on rating transitions. This again signals that these variables play an imperative role in both the prediction of defaults as well as in the prediction of rating downgrades and are, hence, assumed to be important determinants of the financial condition of a bank. Furthermore, it can be observed that again A3 plays an important role in relation to upgrades. This further emphasizes the relevance of this variable, representing the ratio of impaired loans and gross loans. Variable E3, representing the return on average equity and concluded to be relevant in the prediction of downgrades, however, is not present in any of the specific-form models predicting defaults. This possibly indicates that whereas credit rating agencies consider the return on equity to be important in the determination of the credit quality of a bank, this variable might not play the same important role in relation to actual defaults. The irrelevance of the return on average equity in relation to bank failures is substantiated by empirical literature, in which this predictor is not often found to be of significant importance.

The final conclusion based on table VIII is that the multiple models identify a larger number of relevant downgrade predictors than relevant upgrade predictors, with the exception of A upgrades. Whereas the models are able to identify four significant predictors of A upgrades, on the prediction of BBB upgrades they can only identify two predictors and on the prediction of BB upgrades they can only identify one relevant predictor. This finding results in the expectation that in the out-of-sample examination of the models' performances, the AA downgrade predictions and the A upgrade predictions will again be similar and that the BBB upgrade predictions will again be highly inadequate.

Summarizing the results presented by table VIII, it can be stated that again there appears to be no difference between the models regarding the determination of relevant predictors and that there is a reasonable resemblance regarding the relevant predictors of downgrades in general, i.e. downgrades from multiple initial rating classes. Furthermore it can be concluded that variables A3, E8 and L1 play the most important role in downgrade prediction and that these variables where also among the most important variables predicting defaults. The last conclusion is based on the fact that the number of relevant predictors is less in estimation of BBB upgrades and BB upgrades than for the other individual rating transitions, indicating that the out-of-sample performances of the BBB upgrades might result in inadequate outcomes.

Figure IV provides the ROC curves of the out-of-sample performances of the different specific-form models estimating the AA downgrades, A upgrades and BBB upgrades. Table IX presents the descriptive statistics on the corresponding AUC values. Note that the MLP model is present in this out-of-sample examination and that, consistent with the prediction of defaults, it is again based on the complete profile of 29 predictive variables.

Based on figure IV and table IX, it can once more be concluded that the out-of-sample performances of the models, when compared to each other, are not very divergent. The fact that the performances of the four models, based on all three individual types of rating transitions, again appear to be approximately equivalent is consistent with both the results on default prediction as well as with the earlier prediction of rating transitions, in which the models used the predictors that were found to be significant in the prediction of defaults. The MLP model, however, appears to slightly outperform the other models in the prediction of AA downgrades. Since the averages of the AUC values, when the MLP model is compared to the three other models, approaches 0.70 for these AA downgrades, the correlation adjusted *p*-values on the differences between the MLP model and the other models can be calculated. The resulting *p*-values range from 0.05 to 0.10, indicating that for this particular rating transition the MLP model appears to outperform the other models. However, the underlying process of the MLP model is very hard to interpret and, hence, no structural argumentation can be provided in order to support this finding. A possible explanation for the outperformance is based on the fact that the MLP employs a large amount of predictors, where the other models only employ a few predictors.

FIGURE IV.A
ROC CURVES AA DOWNGRADES

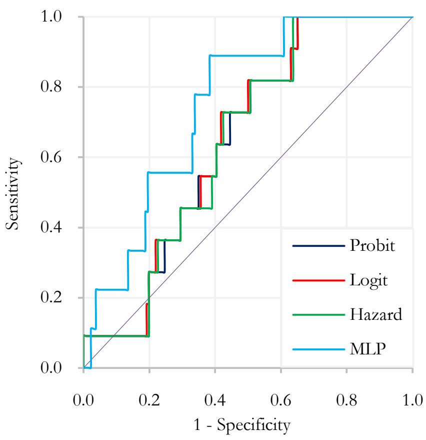


FIGURE IV.B
ROC CURVES A UPGRADES

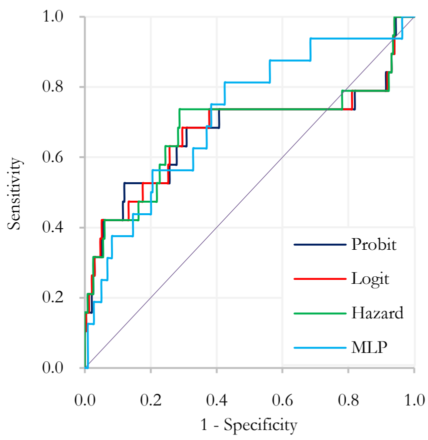
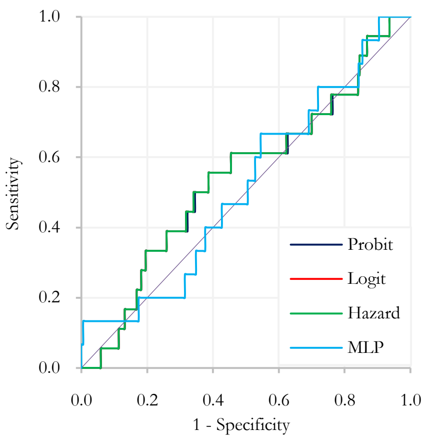


FIGURE IV.C
ROC CURVES BBB UPGRADES



Sensitivity represents the probability of correctly identifying a default bank and is calculated by dividing the number of correctly identified non-default banks by the sum of the correctly identified non-default banks and the incorrectly identified default banks. Specificity represents the probability of correctly identifying a non-default bank and is calculated by dividing the number of correctly identified default banks by the sum of the correctly identified default banks and the incorrectly identified non-default banks. The 45 degree reference line with AUC value 0.50 represents a random model with no ability to correctly classify banks into default or non-default categories.

Whereas in other out-of-sample estimations, i.e. defaults and rating transitions from other initial rating classes, the use of more predictive variables does not lead to a higher out-of-sample performance, conversely, it is possible that in the prediction of AA downgrades it does lead to an outperformance of other models that employ less predictors.

The second conclusion that can be drawn based on figure IV and table IX is that AA downgrades and A upgrades are better predicted than BBB upgrades by all four models. This conclusion is according to the expectations, which were based on the in-sample estimations that resulted in the specific-form model containing only one significant predictor. The conclusion that BBB upgrades are less well-predicted is also consistent with the findings in the previous out-of-sample prediction of rating transitions, when the models used the default predictors as predictive variables. In fact, focusing on the differences between the results of this previous estimation of rating transitions and the results presented in figure IV and table IX, the predictive performances appear to have remained approximately equal. This finding signifies that general-to-specific modeling, i.e. focusing on the accuracy of the in-sample estimation, does not result in a better out-of-sample prediction of rating transitions.

TABLE IX
CREDIT RATING TRANSITION PREDICTION: DESCRIPTIVE STATISTICS ON AUC VALUES (2)

	AUC	Std. Error[a]	H_0: AUC = 0.50 p-value[b]
A. Descriptive Statistics on AA Downgrades			
Probit	0.643	0.064	0.114
Logit	0.649	0.064	0.100
Hazard	0.644	0.064	0.112
MLP	0.751	0.065	0.012
B. Descriptive Statistics on A Upgrades			
Probit	0.672	0.085	0.013
Logit	0.674	0.084	0.012
Hazard	0.678	0.084	0.010
MLP	0.718	0.069	0.004
C. Descriptive Statistics on BBB Upgrades			
Probit	0.544	0.072	0.531
Logit	0.545	0.072	0.522
Hazard	0.545	0.072	0.522
MLP	0.518	0.076	0.821

[a] Standard errors are based on nonparametric assumption.
[b] Denotes asymptotic p-value.

This, on its turn, implies that observed in-sample relationships between financial statement data and actual transitions do not necessarily hold in the out-of-sample estimation of these credit rating transitions.

However, in the interpretation of the results, one should recall that a number of variables, i.e. A3, E8 and L1, consistently proved to be significant predictors of defaults and downgrades. Variable A3 even proved to play an imperative role in the prediction of upgrades. Based on this, it can be concluded that these variables, proxying for the risk factors asset quality, earnings and liquidity, are definitely important indicators of a bank's credit quality. However, based on the fact that the out-of-sample predictions on rating transitions are not extremely convincing, i.e. the AUC values are typically below 0.70, and that this out-of-sample performance of rating transition prediction is again worse than the out-of-sample performance of default prediction, it can be stated that rating transitions prove to be more difficult to predict than defaults are. The most plausible argument for this appears to be the subjectivity in the credit rating process. This subjectivity, as mentioned earlier this section, originates from the fact the credit rating process depends on an analyst's subjective judgments on both quantitative as well as qualitative factors. Another source of subjectivity in the credit rating process is related to misaligned incentives. Whereas credit rating agencies have a quasi-regulatory role, they simultaneously only operate in function of the issuers to be rated. A second possible argument for the fact that rating transitions prove to be harder to predict than defaults are, is based on the specification of the particular events. Whereas both events represent a change in credit quality, it is not unlikely that financial statement data relates to the two events in dissimilar ways. Summarizing the results described above, it can be concluded that, notwithstanding the fact that reliable indicators do provide some insight into the financial condition of a bank, predicting credit ratings that essentially should reflect this financial condition in an objective way proves to be a challenging subject of matter.

VII. CONCLUSION

The main objectives of this study were to examine the predictive performances of multiple default prediction models and to examine whether or not these models are also able to correctly predict credit rating transitions. Regarding the prediction of defaults and the examination of the most accurate default prediction model, the results of this study provide evidence for two central conclusions. The first conclusion is that the predictive performances of the probit model, the logit model, the hazard model, and the MLP model are not very divergent. The second conclusion is that, despite the small discrepancies across models, the predictive performances, based on ROC curves, of all four models were very adequate. These conclusions indicate that the default prediction models examined in this study, employing only firm-specific financial statement data, are valuable tools that are informative in the determination of current and the estimation of future financial conditions of a bank. Hence, the use of these models as off-site monitoring systems are supported by the results of this study. However, as the predictive performances of the models were approximately equal, the actual selection of the most suitable model seems of minor importance.

Regarding the examination of the ability of the models to correctly predict credit rating transitions, the results of this study provide evidence for three central conclusions. The first conclusion is that credit rating transitions are more complicated to predict than defaults are. This conclusion is based on the considerable decrease in the out-of-sample predictive performances of the models predicting defaults and predicting rating transitions. The second conclusion is that the impaired loans to gross loans ratio, the interest income to interest expense ratio and the net loans to total assets ratio play a central role in the prediction of credit rating transitions, particularly downgrades. These ratios were also among the ratios that consistently proved to be most relevant in the prediction of defaults. Consequently, notwithstanding the fact that rating transitions proved to be difficult to predict, it can be concluded that these ratios contain valuable information regarding the financial condition of a bank.

The third conclusion is that, consistent with the prediction of defaults, the differences between the individual models are minor. This conclusion is confirmed in both methods of rating transition prediction performed in this study.

The fact that rating transitions prove to be more difficult to predict than defaults, can be explained by the fact that the credit rating process is exposed to two sources of subjectivity and by the fact that the two credit events do not necessarily reflect the same change in credit quality. The subjectivity originates from the fact the credit rating process depends on an analyst's subjective judgments on both quantitative as well as qualitative factors and from the fact that credit rating agencies can be exposed to misaligned incentives. The fact that rating transitions and default do not necessarily reflect the same change in credit quality is related to the minimum required level of credit quality that determined the dependent variable. Whereas a rating downgrade (upgrade) is initiated by passing the minimum required level of credit quality of the initial rating class (current rating class), actual defaults by definition pass the level of credit quality of D rated banks.

A suggestion for improvement of this study would be that, in addition to bank failures, corporate failures are examined. Examining corporate failures as well, enables one to test the consistency of the performances of the models. Additionally, it can be tested whether or not corporate credit rating transitions can be predicted. A second suggestion for improvement would be the inclusion of market data and macro economic data as predictors of default or credit rating transitions. This enables one to examine whether or not these data include an informational content that is not reflected by the various CAMEL ratios. A third and last suggestion would be to examine the predictive performances of the default prediction models employed in this study relative to the predictive performances of the individual credit ratings provided by the credit rating agencies.

REFERENCES

Altman, E. I., 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23, 589-609.

Altman, E. I., 1977. Predicting Performance in the Savings and Loan Association Industry. Journal of Monetary Economics, 3, 443-466.

Andersen, H., 2008. Failure Prediction of Norwegian Banks; A Logit Approach. Working Paper, Norges Bank.

Arena, M., 2008. Bank Failures and Bank Fundamentals: A Comparative Analysis of Latin America and East Asia during the Nineties using Bank Level Data. Journal of Banking and Finance, 32, 299-310.

Avery, R. B., and Hanweck, G. A., 1984. A Dynamic Analysis of Bank Failures. Research Papers in Banking and Financial Economics, 74, Board of Governors of the Federal Reserve System.

Barth, J. R., Brumbaugh, R. D., Sauerhaft, D., and Wang, G. H. K., 1985. Thrift Institution Failures: Causes and Policy Issues. Proceedings, Federal Reserve Bank of Chicago, 184-216.

Beaver, W. H., 1966. Financial Ratios as Predictors of Failure. Journal of Accounting Research, 4, 71-102.

Berg, D., 2007. Bankruptcy Prediction by Generalized Additive Models. Applied Stochastic Models in Business and Industry, 23, 129-143.

Borrus, A., 2002. The Credit Raters: How They Work and How They Might Work Better. Businessweek.

Chava, S., and Jarrow, R., 2004. Bankruptcy Prediction with Industry Effects. Review of Finance, 8, 537-569.

Cox, D. R., 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society, 34, 187-220.

Cybenko, G., 1989. Approximation by Superpositions of a Sigmoidial Function. Mathematical Control Signals System, 2, 303-314.

D'Addio, A. C., and Rosholm, M., 2002. Left-Censoring in Duration Data: Theory and Applications. Economics Working Papers, School of Economics and Management, University of Aarhus.

Demirgüç-Kunt, A., 1989. Deposit-Institution Failures: A Review of Empirical Literature. Economic Review, Federal Reserve Bank of Cleveland, 2-18.

Espahbodi, P., 1991. Identification of Problem Banks and Binary Choice Models. Journal of Banking and Finance, 15, 53-71.

Estrella, A., Park, S., and Peristiani, S., 2000. Capital Ratios as Predictors of Bank Failure. Economic Policy Review, Federal Reserve Bank of New York, 6.

Funahashi, K., 1989. On the Approximate Realization of Continuous Mapping by Neural Networks. Neural Networks, 2, 183-192.

Gajewski, G. R., 1988. Bank Risk, Regulator Behavior, and Bank Closure in the Mid-1980s: A Two-Step Logit Model. Ph. D. Dissertation, The George Washington University.

Galil, K., 2003. The Quality of Corporate Credit Rating: An Empirical Investigation. European Financial Management Association, Helsinki Meetings.

Godlewski, C., 2007. Are Ratings Consistent with Default Probabilities? Empirical Evidence on Banks in Emerging Market Economies. Emerging Markets Finance and Trade, 43.

Graham, F. C., and Horner, J. E., 1988. Bank Failure: An Evaluation of the Factors Contributing to the Failure of National Banks. Proceedings, Federal Reserve Bank of Chicago, 405-435.

Halling, M., and Hayden, E., 2006. Bank Failure Prediction: A Two-Step Survival Time Approach. Proceedings of the International Statistical Institute's 56th Session.

Hanley, J. A., and Mcneil, B. J., 1983. A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases. Radiology, 148, 839-843.

Hornik, K., Stinchcombe, M., and White, H., 1989. Multilayer Feedforward Networks are Universal Approximators. Neural Networks, 2, 359-366.

Kiefer, N. M., 1988. Economic Duration Data and Hazard Functions. Journal of Economic Literature, 26, 646-679.

Lane, W. R., Looney, S. W., and Wansley, J. W., 1986. An Application of the Cox Proportional Hazards Model to Bank Failure. Journal of Banking and Finance, 511-531.

Martin, D., 1977. Early Warning of Bank Failure: A Logit Regression Approach. Journal of Banking & Finance, 1, 249-276.

McCulloch, W. S., and Pitts, W., 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115-133.

McFadden, D., 1974. The Reviewed Preferences of a Government Bureaucracy: Empirical Evidence. Bell Frontiers of Econometrics, Academic Press, New York.

Nickell, P., Perraudin, W., and Varotto, S., 2000. Stability of Ratings Transitions. Bank of England Working Papers 133, Bank of England.

Odom, M. D., and Sharda, R., 1990. A Neural Network Model for Bankruptcy Prediction. International Joint Conference on Neural Networks, 2, 163-167

Porath, D., 2004. Estimating Probabilities of Default for German Savings Banks and Credit Cooperatives. Discussion Paper Series 2: Banking and Financial Studies, Deutsche Bundesbank, Research Centre, 06.

Salchenberger, L., Cinar, E., and Lash, N., 1992. Neural Networks: A New Tool for Predicting Thrift Failures. Decision Sciences, 23, 899-916.

Shumway, T., 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. Journal of Business, 74, 101-124.

Sinkey, J., 1975. A Multivariate Statistical Analysis of the Characteristics of Problem Banks. Journal of Finance, 30, 21-36.

Stuhr, D. P., and Van Wicklen, R., 1974. Rating the Financial Condition of Banks: A Statistical Approach to Aid Bank Supervision. Federal Reserve Bank of New York, Monthly Review, 233-238.

Tam, K., 1991. Neural Network Models and the Prediction of Bank Bankruptcy. Omega, 19, 429, 445.

Tam, K., and Kiang, M., 1992. Managerial Applications of the Neural Networks: The Case of Bank Failure Predictions. Management Science, 38, 416-430.

Tape, T. P., 2006. Interpreting Diagnostic Tests. University of Nebraska Medical Center.

Thomson, J. B., 1991. Predicting Bank Failures in the 1980s. Economic Review, Federal Reserve Bank of Cleveland, 9-20.

Wheelock, D. C., and Wilson, P. W., 2000. Why do Banks Disappear? The Determinants of U.S. Bank Failures and Acquisitions. The Review of Economics and Statistics, MIT Press, 82, 127-138.

Whalen, G., 1991. A Proportional Hazards Model of Bank Failure: An Examination of Its Usefulness as an Early Warning Tool. Economic Review, Federal Reserve Bank of Cleveland, 21-31.