Kredit und Kapital, 38. Jahrgang, Heft 3
Seiten 401–433

Does Pooling of Financial Statements and Default Data across Specialized Banks Improve Internal Credit Rating Systems?

By Hergen Frerichs, Frankfurt/M.*

I. Introduction

Under the new Basel capital accord (Basel II)[1], banks will have the opportunity to estimate borrowers' default probabilities for regulatory capital calculation. Bank regulators will have to decide on the eligibility of these internal credit rating systems for the accord's internal ratings based approach (IRB approach).

The European Union plans to oblige all banks in its jurisdiction to abide by Basel II. For many European banks, it will also be worthwhile to implement the IRB approach. Under the standard approach most European corporate customers would obtain a risk weight of 100% as they are not externally rated. For many of these customers, risk weights might be reduced if banks derive default probabilities internally. Yet, particularly for small banks, it is costly to implement the IRB approach. In addition, small banks are not likely to be able to calibrate an internal credit rating system of high quality due to their small databases. For these banks it might make sense to share their databases with other banks to save costs and to improve system quality.

In Germany, both savings banks and cooperative banks have started respective projects. Coordination in these bank groups is particularly easy as banks generally do not compete with each other within a group. Each bank of the group is largely restricted to doing business with borrowers located in a prespecified region. Coordination becomes more difficult if banks compete with each other because the content of a bank's credit database is at least partly proprietary and constitutes part of a bank's competitive advantage. Yet, the improvement in system quality

* This paper is a result of a research cooperation with Deutsche Bundesbank.

1 *Basel Committee* (2003).

due to data pooling might outweigh the costs of coordination. Especially, when thinking about small banks, there ought to exist sufficiently many small banks in an economy that are not direct competitors. Moreover, the problem of releasing proprietary data to competitors can be handled by creating a neutral third party that collects data such that confidentiality is assured.

In a very general sense, pooling always dominates non-pooling if non-pooling is understood to be a subset of pooling. If the result of a pooling project is that it is optimal for each bank to stick to its own internal credit rating systems, and if this is regarded as a pooling solution, then pooling will always be better than non-pooling.[2] In this paper, pooling is defined in a narrower sense. A pooled credit rating system is defined to be a system that is derived from a pooled data set disregarding certain regional and sectoral attributes of this data set. This definition more closely captures the spirit of current pooling projects. For the segment of medium-sized companies, the group of German savings banks and that of cooperative banks are likely to design just one credit scoring function for the entire group. While there might be a further segmentation into producing and trading companies, it is quite unlikely that there will be any further regional or sectoral segmentations. Banks will be interested to keep the overall number of rating systems (e.g. for retail customers, for residential and commercial real estate, for small businesses etc.) small due to large administration, backtesting and updating expenses.

Internal credit rating systems will be submitted to the bank regulator who has to decide on their admission to the IRB approach. As these systems will be based on a vast amount of data, they can be expected to be of high quality. Yet, when analyzing these systems, regulators do not only have to evaluate the overall quality of a system, but also its quality with respect to each participating bank of the group. It might well be possible that the savings banks' system is appropriate for the sector as a whole, but it might be of inadequate quality for savings banks in east Germany, for example, if regional differences are not considered by the system. The same reasoning can be applied to a large commercial bank such as Deutsche Bank. If Deutsche Bank calibrates a system for all German corporate borrowers, it might not be adequate for every regional or sectoral subportfolio.

Therefore, the research question posed in the title is important for both banks and regulators. As it is very difficult to obtain access to credit

[2] Thanks to an anonymous referee for raising this point.

rating data for a large number of banks, a simulation approach is applied based on financial statements and default data from Deutsche Bundesbank.[3]

In each simulation step, a set of financial statements and default data is randomly drawn from the Deutsche Bundesbank database, which forms the credit portfolio of one particular bank. This bank may either use its own data, or it may pool its data with other banks to calibrate its internal credit rating system.[4] Based on the bank's own portfolio, rating system quality is evaluated for both options using either the area-under-curve (AUC) or the Brier score.[5] Over many simulation steps, it is investigated which of the two options is superior. Simulations are done for different bank sizes and different forms of specialization (regional, sectoral, sectoral-regional).

The study is based on Deutsche Bundesbank's annual accounts database, which is the most comprehensive collection of annual accounts of German non-financial companies. While the major strength of the database certainly is its size, some weaknesses are:

1. The database is somewhat biased towards large public limited west German manufacturing companies due to its rediscount business origin.[6]

2. Companies primarily submit annual accounts based on tax law to Deutsche Bundesbank, which are characterized by compilation periods of up to one year decreasing the database's value for default prediction.

3. Default is defined as the formal initiation of insolvency proceedings, which is a narrower definition than the Basel II definition[7] causing relatively low default rates in the database.

---

3 The general approach is taken from *Frerichs* and *Wahrenburg* (2003). See there for other related literature. *Frerichs* and *Wahrenburg* (2003) treat the general question of evaluating internal credit rating systems, but do not at all address the effects of specialization.

4 Internal credit rating systems are defined to consist of eight rating classes of equal size. Borrowers are classified into rating classes according to their logistic regression credit score. Pooled systems are calibrated based on the complete Deutsche Bundesbank database or on large regional or sectoral subsamples.

5 The AUC is used in *Sobehart* et al. (2000), *Blochwitz* et al. (2000), *Engelmann* et al. (2003). The Brier score is used in *Diebold* and *Rudebusch* (1989), *Winkler* (1994), *Lopez* (1999, 2001).

6 *Deutsche Bundesbank* (1998).

7 *Basel Committee* (2004), § 452.

We use data of medium-sized and large customers according to German corporate law[8] from 1994–1999, forming a 1994–1998 training sample and a 1999 validation sample. The training sample consists of 98,910 observations of 29,607 companies with an average default rate of 0.58%. The validation sample consists of 18,671 observations (= companies) with an average default rate of 0.74%.

When discussing the question whether pooling is worthwhile or not, portfolio size and particularly a larger number of defaults is the most important argument in favour of pooling.[9] If banks are not specialized in any way, pooling would generally be expected to improve system quality. Yet, as banks specialize in certain fields of business, the information contained in the overall economy might become less valuable. Therefore, specialization effects might lead to a situation in which pooling is not beneficial any more.

The major result of the study is that out-of-time pooling is beneficial for small and medium-sized credit portfolios (375 and 750 borrowers out-of-time) in most cases. This result holds for regionally and sectorally specialized banks, and also for banks that specialize both in a sector and a region. There are quite a few examples, where an increasing degree of specialization does not lead to a smaller benefit from pooling, but to a larger one. This means that specialization sometimes does not cause a specialization advantage, but a disadvantage. There seems to be information in other sectors, in which a bank does not specialize, that helps predicting credit risks in the sector, the bank is specialized in. Overall, differences in system quality are not large, which means, that in most cases we are not able to state rigorously that pooling significantly outperforms non-pooling. Yet, as expected, we observe that the benefits of pooling increase with decreasing credit portfolio size.

There are two points that might partly explain the out-of-time dominance of pooling for small and medium-sized portfolios. The first point is, that the pooled credit scoring functions are derived either on the complete Deutsche Bundesbank dataset or on some large subportfolio. This reflects the effort of German savings and cooperative banks that actually pool their data on this scale. If only two or three banks decide to pool

[8] § 267 HGB (Handelsgesetzbuch/German Corporate Law, defining the size criterion for public limited companies. Companies have to satisfy at least two of the following criteria: 1. total assets larger than 3,438 million Euros, 2. revenues larger than 6,875 million Euros, 3. a yearly average of more than 50 employees.

[9] *Frerichs* and *Wahrenburg* (2003).

their data, then the benefits from pooling are likely to be smaller. The second point concerns data restrictions. Due to the low default rate in the Deutsche Bundesbank database, randomly drawn portfolios often contain the same defaulters. On the one hand, this reflects reality in so far that if a borrower defaults, then in most cases more than one bank specialized in the borrower's sector is affected. On the other hand, although the Deutsche Bundesbank database is quite large, it does not nearly cover all medium-sized and large German companies. Therefore, the small number of defaults in the database reduces the randomness of samples, which might reduce the average quality of systems that are based on banks' own data.

There is one major criticism against the approach of this paper.[10] From a purely statistical point of view, the importance of explanatory variables indicating sectoral or regional specialization can be very easily investigated by including them (e.g. as dummy variables) in the pooled credit scoring function and testing for their statistical significance. As a consequence, the complex simulation setup in this study would be superfluous, and given the correct scoring function is chosen, larger portfolios always dominate smaller ones. It is true that the overall significance of regional or sectoral specialization can be tested in this way. Yet, it must be pointed out that the significance of specialization is only tested for the optimal credit scoring function on the pool level. For each particular bank, not only the specialization variable changes, but the bank's optimal credit scoring function might differ considerably from the pooled function in the number and type of explanatory variables. As this study allows each specialized bank to choose its own optimal credit scoring function, it allows for more degrees of freedom than the dummy variable approach on the pool level. With its simulation approach, this paper also takes into consideration that, as each particular bank is autonomous in choosing its rating approach, regulators are only able to dismiss a particular credit scoring function as inferior if they are able to prove this inferiority on the bank's own data. As it is technically inevitable to restrict the range of credit rating models, and as banks may specialize in many more dimensions apart from regions and sectors, the question posed in the title cannot be answered generally for all credit rating systems and all sorts of specializations. Nevertheless, the study gives first answers with respect to two obviously important dimensions for banks in Germany.

---

10 Thanks to an anonymous referee who made this criticism.

The remainder of the paper is organized as follows. In section II, the simulation set-up is described in more detail. Section III reports results, and section IV concludes.

II. Simulation Set-up

For ease of exposition, we use the term "bank" to stand for a bank's credit portfolio of medium-sized and large companies. We neglect all other businesses a bank might be engaged in. One assumption needed for our approach is that banks calibrate a specific internal credit rating system for the segment of medium-sized and large corporate borrowers, which seems to be a reasonable assumption due to peculiarities of this market segment.

Bank size is defined by the number of balance sheets in a bank's 1994–1998 training sample. We use bank sizes of 1,875, 3,750, 7,500, and 15,000 observations. As there are on average 3.34 balance sheets per company, the number of training sample companies ranges from 561 to 4,490. We do not consider smaller portfolios because it becomes difficult to derive a sensible credit scoring function for such small portfolios due to the small number of defaults. We do not consider larger portfolios because due to the limited overall size of the database, there would not be enough randomness in the composition of larger portfolios.

The 1999 out-of-time sample of each bank consists of all companies that are part of the training sample and that stay customers in 1999. In addition, we simulate new business by randomly drawing new customers such that the bank's portfolio size and portfolio default rate stays constant. In doing this, we construct the situation that the out-of-time sample reflects an average year. As a result there are between 375 and 3,000 companies in the out-of-time samples.

For each company, banks have access to the complete annual accounts history as it is available in the Deutsche Bundesbank database. For each bank size, we randomly draw five hundred credit portfolio compositions representing different banks.

Each bank has the choice whether it calibrates its rating system only on its own historical financial statements and default data, or uses an externally given credit scoring function, which is based on a larger data set. This larger data set may either be one that is based on data from the whole economy (the complete Bundesbank data set) or one that is based

on data from a large region or sector of the economy like south or north-west Germany or manufacturing or trade. The system for which banks use only their own data is termed the "stepwise" system, and the one which uses an externally given credit scoring function is termed the "pooled" system.

Both the stepwise and the pooled system are based on logistic stepwise selection procedures; only the samples on which system calibration is based differ. Stepwise selection procedures choose those financial variables from a set of forty-nine financial variables that are significant given predefined enter and stay criteria. The set of financial variables consists of forty-eight variables, which have been found to be good default indicators in the German credit risk literature, and one variable from Altman (1968) (Table 1). Forty-one variables are taken from Niehaus (1987), three ratios from Hüls (1995), and four ratios from Deutsche Bundesbank (1999). After selecting financial variables, a logistic regression is carried out to derive credit scores and default probabilities.

We employ four different kinds of pooled systems that are based either on the complete Deutsche Bundesbank database or large regional or sectoral subportfolios (south, north-west, manufacturing, and trade). We avoid overfitting by applying a conservative procedure proposed by Shumway (2001) designed to account for the lack of independence between firm-year observations. In this procedure, the significance level for the stepwise selection procedure is corrected by multiplying the value of the partial F test statistic, which is necessary to obtain a confidence level of 90 %, by the average number of firm-years per company in each training sample. These new values of the partial F test are used as thresholds to decide on the significance of a variable. This procedure is conservative as it is assumed that there is only one observation per company although on average there are three financial statements. As a result, the four pooled systems are based on four to seven financial ratios.

For the stepwise system, we use two different sets of entry and stay criteria for the stepwise selection procedures. On the one hand, we employ the Shumway (2001) procedure used for the pooled system. On the other hand, we neglect the lack of independence between firm-year observations, and simply use a significance level of 5 % for the stepwise selection procedures. Our analysis shows that, as expected, the number of variables that are selected by the stepwise selection procedures significantly increases if we implement the second alternative. Using the first

alternative, there are about 1–3 financial variables in the credit scoring systems of the smallest banks (1,875 in-sample observations), about 2–3 in those of medium-sized banks (3,750 in-sample observations, and about 3–5 in those of the large banks (7,500 in-sample observations). Using the second choice, these numbers change to 3–5, 4–7, and 6–9, respectively.

There clearly is a trade-off between system quality and overfitting. The fact that the Shumway procedure is conservative might lead to inferior system quality because important financial variables might be dropped from the credit scoring functions. To neglect the lack of independence might on the other hand lead to overfitting in-sample such that system quality appears to be higher than it actually is. Our simulation findings are that implementing the significance level of 5% considerably increases the quality of the stepwise system relative to the pooled system. This quality increase is not only observed in-sample, where it might be caused by overfitting, but also out-of-time although to a smaller extent. In the following, we report results for the significance level of 5%. If we used the Shumway procedure for the stepwise systems instead the dominance of the pooled system would be even larger.

Banks determine credit scores and default probabilities exclusively based on annual accounts information. They do not add any qualitative information. The inclusion of qualitative factors is not a prerequisite for admittance to the internal ratings based approach of Basel II.[11] Yet, many banks base their internal credit ratings on some qualitative components like management quality.[12] Since we do not have any additional qualitative information in our dataset, we might underestimate system performance, particularly for small banks.

We employ the area-under-curve statistic (AUC) and the Brier score to quantify rating system quality. The AUC measures the quality of ranking borrowers from high to low default risk. If low credit scores are defined to indicate high default probabilities, then all borrowers that actually defaulted in a learning sample should be assigned a relatively low credit score, and those that did not default a relatively high credit score. The AUC is only concerned with ranking, and does not assess the accuracy of

11 *Basel Committee* (2004), § 417.

12 *Basel Committee* (2000) surveyed large international banks. Most banks assign ratings using considerable judgmental elements. The relative importance of qualitative versus quantitative factors ranged from very minor to more than 60%. *Günther* and *Grüning* (2000) report that 72 of 146 surveyed German banks use qualitative criteria for default prediction. 38 of 49 banks state that the quality of default prediction has been improved by the inclusion of qualitative factors.

default probability estimates. It is equivalent to the independent sample Mann-Whitney non-parametric test statistic and can be interpreted as the probability that the score of a randomly chosen company from the sample of defaulted companies is (correctly) lower than the score of a randomly chosen company from the sample of solvent companies. The AUC ranges from 0 % to 100 %. A perfect AUC value of 100 % is attained if exactly those borrowers defaulting in the future receive the lowest credit scores. A value of below 50 % means that the system performs worse than a system which randomly allocates credit scores to borrowers.

The Brier score B is not only concerned with the ranking of borrowers, but also with the accuracy of default probability estimates. It is defined as

$$B = \frac{\sum_{i=1}^{n} (\hat{p}_i - I_i)^2}{n}, \tag{1}$$

where $\hat{p}_i$ is a system's default probability estimate for borrower i, $i = 1, ..., n$, and I_i is the indicator variable of default (1 if default, zero otherwise). The Brier score relies on a quadratic loss function, often used in economics, and belongs to the family of strictly proper scoring rules, meaning that banks minimize their expected score by reporting their probability estimates honestly.[13] The Brier score ranges from zero (defaulters are attached a default probability of 100 % percent and non-defaulters one of 0 %) to some maximum value (defaulters are attached a default probability of zero percent and non-defaulters one of 100 %). A system with an AUC of 100 % does not necessarily have a Brier score of zero, as default probability estimates for defaulters will usually be below 100 %, and those for non-defaulters above zero. Vice versa, a system with a Brier score of zero will also have an AUC of 100 % showing that the Brier score evaluates ranking accuracy plus the accuracy of default probability estimates. The Brier score seems to be more closely related to the bank regulator's objectives than the AUC as regulatory capital requirements directly depend on default probability estimates, and not only on the ranking of borrowers.

We perform three sets of analyses concerning regionally specialized banks, sectorally specialized banks, and banks that specialized both in a sector and a region.

[13] Cf. *Winkler* (1994).

In Germany most small banks are regionally focused because they either belong to the group of savings banks or that of cooperative banks. To a large amount, both groups are restricted to doing business in a prespecified region. While it is obvious that our analysis is of interest to these banks, it might also be of interest for large banks that do loan business all over Germany. If regional characteristics play an important role in the quality of rating systems, then it makes sense even for large banks to take these characteristics into account when calibrating rating systems.

There might not be many banks that are primarily specialized by industrial sectors. Yet, many banks estimate different credit scoring functions for manufacturing and trade. In our model set-up, this is equivalent to specializing in industries.[14] For example, Deutsche Bundesbank (1999) estimates different credit scoring functions for manufacturing, trade, and other industries.

Finally, we consider the case that both sectoral and regional specialization might influence rating system quality. We consider the case that banks specialize first in manufacturing or trade and additionally in a region. For example, savings or cooperative banks that are regionally constrained and that are of the opinion that there are structural differences in predicting defaults for manufacturing and trading companies are covered by this analysis.

We now provide for descriptive statistics of the differently specialized subportfolios introduced above. In Table 2, descriptive statistics of the regions we define are summarized. There are six regions:

1. South Germany (comprising the states of Bavaria and Baden-Wuerttemberg)
2. North-west Germany (comprising the states North Rhine-Westphalia, Lower Saxony, Schleswig-Holstein, Hamburg, and Bremen)
3. South-east Germany (comprising Bavaria, Saxony, Saxony-Anhalt, Thuringia)
4. North-east Germany (comprising Berlin, Brandenburg, Mecklenburg-Western Pomerania, Lower Saxony, Schleswig-Holstein, Hamburg, and Bremen)

[14] If we use the term bank interchangeably for credit portfolios, then a bank that estimates two different credit scoring functions for manufacturing and trade actually constitutes two banks in our analysis. This must be taken into account when addressing the issue of bank size.

5. East Germany (comprising Berlin, Brandenburg, Mecklenburg-Western Pomerania, Saxony, Saxony-Anhalt, Thuringia)
6. Big German cities (comprising Berlin (12 % share in portfolio observations), Munich (16 %), Hamburg (23 %), Cologne (19 %), and Frankfurt (30 %))

Our definition of regions is driven by presumed differences in economic structure. South Germany performs better economically than north-west Germany. East Germany performs worse than south and north-west Germany. The big cities may differ structurally from the rest of the country. In addition, we take into account how banks specialize regionally. Banks from the south are likely to expand to the south-east after the reunification, while banks from the north are likely to expand to the north-east.

In the last two columns of Table 2, we give some evidence that the defined regions actually differ structurally. For the year 1996, which lies in the middle of our training sample, we observe the share of a region's gross domestic product in Germany's total gross domestic product and do the same thing for the number of insolvencies. It can be seen, that the share of insolvencies in the south is considerably lower than the share of GDP. In the east, we observe the opposite; the share in insolvencies is much higher than the share in GDP. For the other regions, differences are not that large. Comparing these shares with those in the Deutsche Bundesbank database, we see that the share of insolvencies is always (at least weakly) larger than the share of observations in the Deutsche Bundesbank database if the same holds in the whole economy. One noteworthy difference is that east Germany is clearly underrepresented in the Deutsche Bundesbank database.

With respect to default rates, it can be said that overall default rates as recorded in the Deutsche Bundesbank database are rather low at an average of 0.58 % in the 1994–1998 training sample and 0.74 % in the 1999 validation sample. This is due to the strict default definition, but might also reflect to some extent a bias in the database towards higher quality companies caused by the database's rediscount business origin. Since rating system quality strongly depends on the number of training sample defaults, results based on extraordinarily low default rates might be misleading. For this reason, we scale the overall in-sample and out-of-time default rate up to 1.7 %. This value is taken from Carey (1998) as being a representative default rate for commercial loan portfolios of large U.S. banks.[15] Unfortunately, we do not have data on representative default rates of German credit portfolios to estimate average portfolio default rates.

To preserve differences in average default rates in different regions or sectors, we scale the default rate of each region or sector such that the ratio of a region's or a sector's in-sample default rate to the overall in-sample default rate remains the same. For example, for east Germany, the in-sample ratio equals 1.44%/0.58% = 2.4828 such that in-sample and out-of-time default rates are scaled to 1.7%*2.4828 = 4.22%. In this way, banks that operate in riskier than average regions have more default information to base their credit scoring model on than those that operate in less risky regions.

In Table 3, we summarize information on the sectoral subportfolios we define. There are eight industries: 1. Manufacturing (D), 2. Trade (G), 3. Wholesale Trade (51), 4. Automobile Trade (50), 5. Construction (F), 6. Metal Production and Production of Metal Products (DJ), 7. Mechanical Engineering (DK), and 8. Automobile Production (DH, 28, 31, 34). The definition of industrial sectors follows the industrial classification according to the classification code of the German Federal Statistical Office (WZ93). Respective classification codes are shown in brackets.

Similar to Table 2, the last two columns of Table 3 show 1996 shares of the sectors in total German GDP and insolvencies. It can be seen that the share of insolvencies in manufacturing is much lower than the share in GDP. For trade and construction, the opposite relationship holds. Unfortunately, the sectoral structure of the Deutsche Bundesbank database is not as much in accordance with the overall German industrial structure as the regional structure. First of all, absolute shares of manufacturing and trade are much larger in the Deutsche Bundesbank database than in the economy. Only the construction industry is similar. And then, only for the construction industry and automobile production differences in shares of insolvencies and GDP have the same sign in the database and in the economy. Thus, the database seems not to be entirely representative of the German economy with respect to its sectoral composition.

In Table 4, descriptive statistics of sectoral-regional subportfolios are summarized. We define eight subportfolios by dividing the data set into two sectors (manufacturing and trade) and into four regions (south, north-west, south-east, and north-east). For these subportfolios, we do not have data for the entire economy so that we are not able to compare structural characteristics.

---

15 Take the portfolio structure for commercial loan portfolios of large U.S. banks in *Carey* (1998), p. 1380, and multiply it with default probabilities given in Table III, Panel B, second column.

III. Simulation Results

1. Is Pooling Beneficial for Regionally Specialized Banks?

Tables 5 to 7 show results for regionally specialized banks. In Table 5 and 6, each randomly drawn bank calculates the AUC and the Brier score using both the stepwise and the pooled system. Resulting AUC values and Brier scores are directly compared. In Table 7, the AUC value and Brier score distributions for the stepwise and the pooled system are compared.

In Table 5, results for the direct AUC value comparison are shown depending on region, on sample type (in-sample, out-of-time), and on sample size. A region's average default rate is also given. For each category, we report three statistics:

1. The relative frequency that a bank's stepwise performs worse than the pooled system. If this relative frequency is below 50 %, then the stepwise system performs better than the pooled system. If both systems are equally informative and assuming independence, and five hundred simulations, we would expect the relative frequency to lie between 44 % and 56 % at the 99 %-confidence level. In the Table, those entries are printed in bold that lie outside the range (25 %; 75 %).

2. The first entry in brackets states the relative frequency that the stepwise system performs significantly worse than the pooled system at the 10 %-significance level using the DeLong et al. (1988)-test for AUC values and the Bloch (1990)-test for Brier scores. Relative frequencies above 30 % are printed in bold.

3. The second entry in brackets states the opposite case that the relative frequency of the stepwise system performs significantly better than the pooled system.

The upper and the lower panel differ with respect to the pooled credit scoring function used in the simulations. In the upper panel, the pooled credit scoring function is derived on the complete Deutsche Bundesbank database. In the lower panel, it is derived on either the subportfolio of south or north-west German companies.

In-sample, system quality is generally higher for bank's that use the stepwise system than for those that use the pooled system. This seems to be at least partly caused by overfitting. The inferiority of the pooled system decreases as bank size decreases. This is intuitive as smaller

banks benefit less from their own database than large banks. These observations refer to specialized banks as well as non-specialized banks ("Overall").

With respect to regionally specialized banks, it can be seen that if the pooled system is based on the complete economy, then the results do not differ from the non-specialized banks in many cases. There are three large exceptions. For medium-sized banks (3,750 observations) from the south-east and small banks from the east (1,875 observations), pooling is not at all beneficial. In these cases, the overfitting bias seems to clearly outweigh the sample size disadvantage. The opposite holds for banks focusing on big city-companies although this result is less significant. These are indicators for structural differences in these regions relative to the economy.

If the pooled system is not based on the complete economy, but rather on a regional sub-sample, the benefit from pooling generally decreases. As this effect cannot be seen out-of-time, it seems to be due to the overfitting bias. It becomes obvious that in-sample results should be interpreted with care as it is always possible in-sample to fit a rating system perfectly to the data.

Out-of-time, pooling is mostly inferior to non-pooling for large sample sizes (7,500 and 15,000 observations) and superior for small sample sizes (1,875 and 3,500 observations) although there are only a few cases for which pooling seems to be significantly superior, e.g. for the north-east, the pooled system is often significantly better than the stepwise system.

In Table 6, the same analysis is performed for Brier scores. It is striking how similar results are although AUC values and Brier scores measure system quality in quite different ways. All results stated for the AUC analysis also hold for the Brier score analysis.

In Table 7, the relative frequency shows that the stepwise system performs worse than the pooled system calibrated on data from the complete economy or from a large region (numbers in brackets) based on the distribution of AUC values and Brier scores for the alternative systems. Statistically, for AUC values we take the 10%-quantile of the AUC value distribution of the pooled system and calculate the relative frequency that the AUC values of the stepwise system fall below this threshold. For Brier scores, we take the 90%-quantile of the Brier score distribution of the pooled system and calculate the relative frequency that the Brier

scores of the stepwise system are above the threshold. This procedure is similar to calculating the power of a test given a type-I error.

Relative frequencies below 10% indicate the inferiority of the pooled system, while values above 10% indicate its superiority. In Frerichs and Wahrenburg (2003), this kind of representation is used to identify inferior internal credit rating systems. There, values of at least 50% are taken to indicate a clear underperformance of a given system relative to the pooled system. From Table 7, it can be seen that there is just one case for which the 50%-threshold is reached (Big cities, in-sample). Otherwise results in Table 7 reinforce many of our former findings. Regional specialization seems not to play a large role for the out-of-time quality of internal credit rating systems. In most cases, quality differences between the stepwise and the pooled system are quite alike for specialized and non-specialized banks. Only for out-of-time Brier scores, pooling seems to be even more beneficial for regionally specialized banks than for non-specialized banks.

If we use the stepwise system instead of the pooled system to derive quality thresholds, then there are a number of cases, in which the quality of the pooled system is actually significantly (with a relative frequency of at least 50%) worse than the stepwise system. These are the numbers printed in bold and separated by a semi-colon. Yet, as all these cases refer to in-sample observations they indicate structural differences, but not necessarily quality differences.

To sum up this section's results, in-sample results favor non-pooling most probably due to overfitting. Out-of-time, results indicate that pooling is generally beneficial for small to medium-sized banks (1,875 and 3,750 observations) whether or not they are regionally specialized, while it is inferior for larger banks. Overall, differences in system quality are small.

2. Is Pooling Beneficial for Sectorally Specialized Banks?

Tables 8 and 9 display simulation results for sectorally specialized banks. Table 8 is equivalent to Table 5 giving direct comparisons of system quality for the stepwise and the pooled system based on AUC values. As again the respective analysis based on Brier scores gives very similar results, we do not display it. Table 9 is equivalent to Table 7 reporting results from the comparison of AUC value and Brier score distributions for the stepwise and the pooled system.

As in the former section, the quality of the pooled system increases relative to the stepwise system as bank size decreases. In-sample, the pooled system calibrated on the complete economy is mostly inferior to the stepwise system. Yet, if the pooled system is calibrated on a sectoral sub-sample (manufacturing or trade), then it is actually superior to the stepwise system for small and medium-sized samples (1,875 and 3,750 observations). This indicates that deriving pooled credit scoring functions depending on sectors might improve system quality. Out-of-time, this result is reinforced, but to a considerably smaller extent.

In-sample, there are again some industries that seem to differ structurally from the overall economy as the stepwise system is clearly superior to the pooled system (trade, 7,500 observations; wholesale trade, 3,750, and construction, 1,875). Yet, only for construction this result also holds out-of-time. As this result is particularly strong, it clearly indicates the construction industry differs considerably from other industries, and credit portfolios that are concentrated in this industry do not benefit from pooling.

Out-of-time, pooling is mostly beneficial for small and medium-sized credit portfolios (1,875 and 3,750 observations) and particularly for portfolios specialized in metal production and metal products, in mechanical engineering, and in automobile production.

Results in Table 9 reinforce those in Table 8. It is noteworthy that out-of-time banks in the sectors metal production and metal products, and automobile production using the stepwise system would be regarded as using an inadequate internal credit rating system by the standards set up in Frerichs and Wahrenburg (2003). For the construction industry the opposite holds. The pooled system clearly underperforms the stepwise system.

3. Is Pooling Beneficial for Sectorally-Regionally Specialized Banks?

Tables 10 and 11 display simulation results for banks that are specialized either in manufacturing or in trade and that are additionally specialized in one of four regions. The pooled credit scoring function is either based on manufacturing or trading companies, respectively, from the whole economy. Due to decreasing sample sizes, we are only able to simulate small and medium-sized portfolios (1,875 and 3,750 observations) in theses segments. Again, we leave out the Brier score analysis as it is very similar to the AUC value analysis.

While in-sample results are mixed – pooling dominates non-pooling in quite a few cases – out-of-time results are again clearly in favour of pooling, which is in accordance with our former results for small and medium-sized portfolios. Pooling seems to be particularly beneficial to credit portfolios specialized in trade. Results in Table 11 show the same picture. Out-of-time, the stepwise systems of banks specializing in manufacturing and the north-east, or in trading and the north- or south-east, will be identified as using inadequate rating systems based on Frerichs and Wahrenburg (2003).

Overall, it seems that any specialization advantages banks might have are considerably smaller than the added value that results from a sectoral pooled credit scoring function.

IV. Conclusion

Based on a large financial statements and default database of Deutsche Bundesbank, we use a simulation approach to answer the research question whether pooling of financial statements and default data improves the quality of internal credit rating systems of regionally, sectorally, or sectorally and regionally specialized banks if these factors are disregarded in the pooling function. This research question is important as the economic success of any commercial bank, whether it is obliged to implement Basel II or not, depends to some extent on the quality of its internal credit rating system. Banks that will be obliged to implement the IRB approach of Basel-II have an additional incentive to think about pooling as they need regulatory approval for their rating system.

The study's primary result is that pooling improves the quality of internal credit rating systems, particularly of small banks, even if pooling does not take into account regional and sectoral factors although banks are regionally or sectorally specialized. For some specialized banks the results are extraordinarily strong, while for banks specialized in the construction industry pooling does not seem to be of value.

When evaluating results, it needs to be taken into account that we strengthen the point of pooling by setting up the pooling functions either based on the complete Deutsche Bundesbank database or on a large part of it. If two banks decide to pool their data, then the quality improvements are likely to be smaller. Also, the low default rate in the Deutsche Bundesbank database might affect results as discussed in the introduction.

For future research, it would be interesting to increase sample sizes by including small companies in order to mitigate data problems and allow simulating larger credit portfolios. Off-course, it would also be of considerable interest to estimate the effects of pooling on real data such as that of the savings or cooperative bank sector.

Table 1

Financial Ratios Used as Independent Variables in Credit Scorin[16]

Financial variables are taken from Niehaus (1987), Hüls (1995), Deutsche Bundesbank (1999), and Altman (1968) (cf. footnote to table). The column 'Hypothesis' indicates whether the value of the financial variable is expected to be generally lower or higher, respectively, for insolvent (I) observations than for solvent (S) observations.

Variable	Ratio	Hypothesis
V1	operating profit (before taxes) / revenues	I < S
V2	EBITDA (excl. extraordinary items) / revenues	I < S
V3	earnings before financial expenses / total assets	I < S
V4	operating profit (before taxes and financial expenses) / total assets	I < S
V5	EBITDA (excl. extraordinary items) / total assets	I < S
V6	(EBITDA (excl. extraordinary items) + financial expenses) / total assets	I < S
V7	EBITDA (incl. extraordinary items) / total assets	I < S
V8	(revenues – expenses for raw materials and supplies – amortization of fixed assets – other operating expenses) / total assets	I < S
V9	EBITDA (incl. extraordinary items) / revenues	I < S
V10	EBITDA (excl. extraordinary items) / total debt	I < S
V11	EBITDA (incl. extraordinary items) / total debt	I < S
V12	EBITDA (excl. extraordinary items) / (total debt – cash)	I < S
V13	EBITDA (incl. extraordinary items) / (total debt – cash)	I < S
V14	EBITDA (excl. extraordinary items) / (total debt – cash – securities – trade receivables)	I < S

Continue page 419

Table 1: *Continued*

Variable	Ratio	Hypothesis
V15	EBITDA (incl. extraordinary items) / (total debt – cash – securities – trade receivables)	I < S
V16	EBITDA (excl. extraordinary items) / short-term debt	I < S
V17	EBITDA (incl. extraordinary items) / short-term debt	I < S
V18	(short-term debt * 360) / revenues	I > S
V19	(trade payables + liabilities from accepted bills) * 360 / revenues	I > S
V20	(cash + securities + trade receivables) / short-term debt	I < S
V21	working assets / short-term debt	I < S
V22	(working assets – short-term debt) / total assets	I < S
V23	(working assets – short-term debt) / revenues	I < S
V24	(cash + securities + trade receivables – short-term debt) / (operating expenses – amortization of fixed assets)	I < S
V25	adjusted equity capital / total assets	I < S
V26	(equity capital + total earnings) / total assets	I < S
V27	adjusted equity capital / total debt	I < S
V28	(equity capital + total earnings) / total debt	I < S
V29	short-term debt / total assets	I > S
V30	short-term bank debt / total debt	I > S
V31	(adjusted equity capital + pension provisions + long-term debt) / long-term assets	I < S
V32	adjusted equity capital / (total assets – cash – properties)	I < S
V33	adjusted equity capital / (fixed assets – properties)	I < S
V34	revenues / total assets	I < S
V35	(debt from accepted bills + trade payables) * 12 / expenses for raw materials and supplies	I > S
V36	trade receivables * 12 / revenues	I > S

Continue page 420

Table 1: *Continued*

Variable	Ratio	Hypothesis
V37	finished goods * 12 / revenues	I > S
V38	raw materials and supplies * 12 / expenses for raw materials and supplies	I > S
V39	amortization / (fixed assets + reductions of fixed assets + amortization)	I < S
V40	investments / (fixed assets + reductions of fixed assets + amortization – investments)	I < S
V41	investments / amortization	I < S
V42	(adjusted equity capital + provisions/2) / total assets	I < S
V43	(trade payables + debt from accepted bills + bank debt) / (total debt – received advance payments)	I > S
V44	(trade receivables + inventories) / revenues	I > S
V45	(adjusted equity capital + pension provisions) / total assets	I < S
V46	earnings before taxes on income and interest paid / total assets	I < S
V47	earnings before taxes on income / adjusted equity capital	I < S
V48	net interest result / revenues	I < S
V49	retained earnings / total assets	I < S

[16] Variables V1–V41 are taken from (*Niehaus* (19879, p. 75–76). The variable 21 of (*Niehaus* (19879) is not sufficiently defined so that we do not use it. V42–44 are from *Hüls* (1995), p. 241, Table 22 (V42 = K_122, V43 = K_68A, V44 = K_85). The variable K_08EP cannot be calculated because we do not have data on the change in pension provisions, but V5 is very similar. The variable K_35 = V19, and the variable K_79 = V34. V45–48 are from *Deutsche Bundesbank* (1999), p. 55 (V45 = Equity/pension provision ratio, V46 = Return on total capital employed, V47 = Return on equity, V48 = Net interest rate). The capital recovery rate cannot be calculated because it is not sufficiently defined. The equity ratio equals V26. V49 is taken from *Altman* (1968).

Table 2

Descriptive Statistics of Regions

The Deutsche Bundesbank dataset is divided into regional subsamples based on a company's domicile. Each region collects companies in prespecified German states (Bundesländer). The table shows the number of observations, the number of defaults and the default rate for each regional subsample. It also displays the share of a regional subsample in the total sample. These shares are compared with a region's 1996 share in German total GDP and insolvencies (data from Federal Statistical Office, Germany). (South: Bavaria, Baden-Wuerttemberg; North-West: North Rhine-Westphalia, Lower Saxony, Schleswig-Holstein, Hamburg, Bremen; South-East: Bavaria, Saxony, Saxony-Anhalt, Thuringia; North-East: Berlin, Brandenburg, Mecklenburg-Western Pomerania, Lower Saxony, Schleswig-Holstein, Hamburg, Bremen; East: Berlin, Brandenburg, Mecklenburg-Western Pomerania, Saxony, Saxony-Anhalt, Thuringia; Big cities: Berlin, Munich, Hamburg, Cologne, Frankfurt)

	1994–1998 training sample					1999 validation sample					1996 share of German total	
Region	# obs	% of total	# defaults	% of total	Default rate in %	# obs	% of total	# defaults	% of total	Default rate in %	GDP	Insolvencies
Overall	98,910	100	573	100	0.58	18,671	100	138	100	0.74	100	100
South	36,219	37	201	35	0.55	6,651	36	35	25	0.53	31	23
North-West	41,850	42	178	31	0.43	8,240	44	58	42	0.70	39	34
South-East	20,560	21	196	34	0.95	4,002	21	47	34	1.17	24	30
North-East	17,725	18	102	18	0.58	3,365	18	30	22	0.89	24	27
East	7,434	8	107	19	1.44	1,411	8	37	27	2.62	16	30
Big cities	9,433	10	56	10	0.59	1,836	10	14	10	0.76	–	–

Table 3

Descriptive Statistics of Industries

The Deutsche Bundesbank dataset is divided into sectoral subsamples based on a company's industrial classification according to the classification code of the Federal Statistical Office, Germany (WZ93). The table shows the number of observations, the number of defaults and the default rate for each sectoral subsample. It also displays the share of a sectoral subsample in the total sample. These shares are compared with a sector's 1996 share in German total gross value added (GVA) and insolvencies (data from Federal Statistical Office, Germany). (Industrial classification according to WZ93: Manufacturing = D; Trade = G; Wholesale Trade = 51; Automobile Trade = 50; Construction = F; Metal Production = DJ; Mechanical Engineering = DK; Automobile Production = DH, 28, 31, 34)

	1994–1998 training sample					1999 validation sample					1996 share of German total	
Sector	# obs	% of total	# defaults	% of total	Default rate in %	# obs	% of total	# defaults	% of total	Default rate in %	GVA	Insolvencies
Overall	98,910	100	573	100	0.58	18,671	100	138	100	0.74	100	100
Manufacturing	46,539	47	292	51	0.63	8,625	46	69	50	0.80	22	12
Trade	37,467	38	158	28	0.42	7,233	39	42	30	0.58	11	19
Wholesale T.	23,665	24	102	18	0.43	4,671	25	27	20	0.58	5	8
Automobile T.	9,526	10	25	4	0.26	1,810	10	8	6	0.44	1	2
Construction	6,514	7	98	17	1.50	1,025	5	21	15	2.05	6	22
Metal Prod.	8,574	9	44	8	0.51	1,582	8	9	7	0.57	3	3
Mech. Engin.	8,062	8	77	13	0.96	1,469	8	18	13	1.23	3	2
Automob. Prod.	12,786	13	59	10	0.46	2,392	13	12	9	0.50	7	3

Table 4

Descriptive Statistics of Regions by Sector

The Deutsche Bundesbank dataset is divided into sectoral-regional subsamples based on a company's industrial classification according to the classification code of the Federal Statistical Office, Germany (WZ93) and domicile. The table shows the number of observations, the number of defaults and the default rate for each subsample. It also displays the share of a regional subsample in the total sample. (Industrial classification according to WZ93: Manufacturing = D; Trade = G; South: Bavaria, Baden-Wuerttemberg; North-West: North Rhine-Westphalia, Lower Saxony, Schleswig-Holstein, Hamburg, Bremen; South-East: Bavaria, Saxony, Saxony-Anhalt, Thuringia; North-East: Berlin, Brandenburg, Mecklenburg-Western Pomerania, Lower Saxony, Schleswig-Holstein, Hamburg, Bremen; East: Berlin, Brandenburg, Mecklenburg-Western Pomerania, Saxony, Saxony-Anhalt, Thuringia; Big cities: Berlin, Munich, Hamburg, Cologne, Frankfurt).

	1994–1998 training sample					1999 validation sample				
Sector – region	# obs	% of total	# defaults	% of total	Default rate in %	# obs	% of total	# defaults	% of total	Default rate in %
Overall	98,910	100	573	100	0.58	18,671	100	138	100	0.74
Manufacturing										
South	19,017	19	115	20	0.60	3,453	18	19	14	0.55
North-West	18,608	19	86	15	0.46	3,580	19	26	19	0.73
South-East	9,816	10	104	18	1.06	1,879	10	24	17	1.28
North-East	6,126	6	35	6	0.57	1,110	6	13	9	1.17
Trade										
South	11,573	12	53	9	0.46	2,146	11	9	7	0.42
North-West	18,045	18	60	11	0.33	3,644	20	24	17	0.66
South-East	6,813	7	41	7	0.60	1,356	7	12	9	0.88
North-East	8,877	9	44	8	0.50	1,717	9	8	6	0.47

Table 5

Is Non-Pooling Inferior to Pooling for Regionally Specialized Banks? Direct Quality Comparison (AUC)

The table shows the relative frequency (in %) that a rating system calibrated on a bank's own historical data performs worse (in brackets: significantly worse/better at 10% level) than a system calibrated on data from the complete economy or from a large region (*= south, **= north-west). For each randomly drawn bank, the AUC is directly compared. Results are shown for banks that either do not specialize or that specialize in one of six regions. They are based on 500 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%.

Region	Average Default Rate	In-sample (# obs)				Out-of-time (# obs)			
		15,000	7,500	3,750	1,875	3,000	1,500	750	375
AUC (pooled system calibrated on data from complete economy)									
Overall	1.70%	**14** (0; 26)	**23** (0; 11)	31 (2; 10)	46 (5; 6)	**25** (1; 14)	48 (6; 6)	65 (15; 2)	74 (21; 2)
South	1.63%		**17** (0; 8)	35 (0; 6)	44 (5; 5)		36 (2; 3)	62 (14; 2)	73 (20; 1)
North-West	1.25%		**23** (1; 14)	38 (4; 6)	60 (7; 2)		**24** (1; 17)	57 (8; 6)	62 (13; 4)
South-East	2.80%			**9** (0; 27)	38 (2; 6)			63 (11; 1)	**75** (21; 1)
North-East	1.69%			39 (4; 3)	63 (8; 1)			65 (16; 1)	**78** (**31**; 0)
East	4.22%				**1** (0; **62**)				50 (2; 5)
Big cities	1.74%				**87** (17; 1)				61 (15; 0)
AUC (pooled system calibrated on data from a large region)									
South*	1.63%		**2** (0; **47**)	**17** (0; 18)	34 (3; 10)		46 (2; 7)	68 (17; 3)	**81** (23; 2)
North-West**	1.25%		**5** (0; **45**)	**19** (1; 17)	44 (6; 4)		64 (13; 2)	72 (16; 1)	**75** (20; 1)
South-East*	2.80%			**0** (0; **88**)	**8** (1; **37**)			61 (5; 2)	70 (15; 3)
North-East**	1.69%			**15** (1; 22)	43 (4; 9)			61 (11; 3)	**76** (27; 1)

Table 6

Is Non-Pooling Inferior to Pooling for Regionally Specialized Banks? Direct Quality Comparison (Brier Score)

The table shows the relative frequency that a rating system calibrated on a bank's own historical data performs worse (in brackets: significantly worse/better at 10% level) than a system calibrated on data from the complete economy or from a large region (*= south, **= north-west). For each randomly drawn bank, the Brier score is directly compared. Results are shown for banks that either do not specialize or that specialize in one of six regions. They are based on 500 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%.

Region	Average Default Rate	In-sample (# obs)				Out-of-time (# obs)			
		15,000	7,500	3,750	1,875	3,000	1,500	750	375
Brier score (pooled system calibrated on data from complete economy)									
Overall	1.70%	**18** (1; **31**)	28 (1; 15)	33 (4; 12)	47 (6; 7)	28 (2; 18)	50 (7; 6)	66 (16; 4)	71 (20; 4)
South	1.63%		**21** (0; 19)	35 (1; 9)	41 (3; 8)		56 (9; 5)	69 (23; 3)	74 (21; 2)
North-West	1.25%		35 (0; 18)	41 (5; 12)	57 (5; 3)		33 (3; 15)	55 (11; 7)	61 (14; 6)
South-East	2.80%			**13** (0; **34**)	39 (2; 10)			60 (13; 4)	70 (19; 3)
North-East	1.69%			28 (3; 9)	51 (5; 3)			73 (25; 3)	**81** (**35**; 2)
East	4.22%				**1** (0; **62**)				41 (3; 12)
Big cities	1.74%				67 (10; 1)				58 (9; 4)
Brier score (pooled system calibrated on data from a large region)									
South*	1.63%		**6** (0; **48**)	**20** (1; 21)	34 (2; 12)		54 (8; 7)	69 (20; 4)	72 (25; 2)
North-West**	1.25%		**8** (0; **50**)	**21** (2; 23)	43 (4; 7)		49 (9; 9)	62 (14; 5)	73 (19; 4)
South-East*	2.80%			**0** (0; **85**)	**12** (0; **38**)			63 (12; 4)	63 (19; 4)
North-East**	1.69%			**8** (0; **42**)	32 (3; 14)			65 (14; 5)	**75** (27; 2)

Table 7

Is Non-Pooling Inferior to Pooling for Regionally Specialized Banks? Comparison of Quality Distributions

The table shows the relative frequency (in %) that a rating system calibrated on a bank's own historical data – the stepwise system – performs worse (";xx"= better) than a pooled system calibrated on data from the complete economy or from a large region (numbers in brackets: *= south, **= north-west). For the AUC (Brier scores), relative frequencies are shown that the AUC (Brier score) of the stepwise system is lower (higher) than the 10% (90%)-quantile of the AUC distribution of the pooled system. Results are shown for banks that either do not specialize or that specialize in one of six regions. They are based on 500 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%.

Region	Average Default Rate	In-sample (# obs)				Out-of-time (# obs)			
		15,000	7,500	3,750	1,875	3,000	1,500	750	375
Rejection frequencies using critical AUC thresholds (in %)									
Overall	1.70%	4	5	8	13	6	9	16	21
South*	1.63%		4 (1)	11 (5)	11 (10)		8 (12)	15 (22)	22 (29)
North-West**	1.25%		7 (1; **51**)	11 (6)	18 (12)		5 (22)	15 (27)	17 (27)
South-East*	2.80%			3 (0; **88**)	10 (2)			17 (16)	27 (23)
North-East**	1.69%			16 (6)	26 (14)			20 (18)	31 (28)
East	4.22%				0; **71**				15
Big cities	1.74%				**50**				18
Rejection frequencies using critical Brier score thresholds (in %)									
Overall	1.70%	5	7	9	15	7	11	16	24
South*	1.63%		5 (1)	8 (5)	10 (7)		22 (15)	30 (26)	27 (30)
North-West**	1.25%		11 (1)	13 (6)	20 (11)		6 (14)	15 (20)	16 (32)
South-East*	2.80%			3 (0; **91**)	10 (2)			20 (26)	20 (27)
North-East**	1.69%			10 (3; **52**)	17 (8)			32 (29)	39 (39)
East	4.22%				1; **68**				13
Big cities	1.74%				41				16

Table 8

Is Non-Pooling Inferior to Pooling for Sectorally Specialized Banks? Direct Quality Comparison (AUC)

The table shows the relative frequency (in %) that a rating system calibrated on a bank's own historical data performs worse (in brackets: significantly worse/better at 10% level) than a system calibrated on data from the complete economy or from a large sector (*= manufacturing, **= trade). For each randomly drawn bank, the AUC is directly compared. Results are shown for banks that either do not specialize or that specialize in one of eight sectors. They are based on 1,000 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%.

Sector	Av. DR	In-sample (# obs)				Out-of-time (# obs)			
		15,000	7,500	3,750	1,875	3,000	1,500	750	375
AUC (pooled system calibrated on data from complete economy)									
Overall	1.70%	**14** (0; 26)	**23** (0; 11)	31 (2; 10)	46 (5; 6)	**25** (1; 14)	48 (6; 6)	65 (15; 2)	74 (21; 2)
Manufacturing	1.84%		**7** (0; 29)	34 (2; 9)	50 (6; 5)		70 (12; 1)	73 (17; 1)	72 (20; 1)
Trade	1.24%		**1** (0; **48**)	**14** (0; 18)	27 (1; 14)		31 (1; 7)	51 (6; 5)	58 (13; 3)
Wholesale T.	1.26%			**4** (0; **39**)	**15** (0; 17)			30 (3; 5)	48 (6; 3)
Automobile T.	0.77%				45 (3; 5)				58 (9; 3)
Construction	4.41%				**0** (0; **71**)				2 (0; **52**)
Metal Prod.	1.51%				61 (7; 2)				**87** (25; 0)
Mech. Engin.	2.80%				47 (1; 2)				**83** (20; 0)
Automob. Prod.	1.35%				49 (4; 4)				**87** (**32**; 0)
AUC (pooled system calibrated on data from a large region)									
Manufacturing*	1.84%		39 (3; 6)	51 (5; 2)	53 (6; 3)		52 (9; 4)	65 (13; 2)	67 (16; 2)
Trade**	1.24%		36 (2; 9)	51 (7; 6)	49 (5; 6)		**81** (23; 2)	**82** (20; 0)	**81** (22; 2)
Wholesale T.**	1.26%			**79** (19; 1)	67 (11; 3)			70 (12; 2)	73 (14; 2)
Automobile T.**	0.77%				48 (4; 5)				65 (14; 5)
Metal Prod.*	1.51%				**85** (26; 1)				**88** (26; 0)
Mech. Engin.*	2.80%				58 (8; 3)				**85** (18; 0)
Automob. Prod.*	1.35%				64 (14; 1)				**87** (26; 1)

Table 9

Is Non-Pooling Inferior to Pooling for Sectorally Specialized Banks? Comparison of Quality Distributions

The table shows the relative frequency (in %) that a rating system calibrated on a bank's own historical data – the stepwise system – performs worse (";xx"= better) than a pooled system calibrated on data from the complete economy or from a large sector (numbers in brackets: *= manufacturing, **= trade). For the AUC (Brier scores), relative frequencies are shown that the AUC (Brier score) of the stepwise system is lower (higher) than the 10% (90%)-quantile of the AUC distribution of the pooled system. Results are based on 500 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%. (Av. DR = Average default rate)

Sector	Av. DR	In-sample (# obs)				Out-of-time (# obs)			
		15,000	7,500	3,750	1,875	3,000	1,500	750	375
Rejection frequencies using critical AUC thresholds (in %)									
Overall	1.70%	4	5	8	13	6	9	16	21
Manufacturing*	1.84%		3 (9)	9 (14)	14 (12)		20 (14)	22 (18)	25 (21)
Trade**	1.24%		0; **76** (8)	3 (17)	4 (14)		3 (32)	8 (29)	12 (35)
Wholesale T.**	1.26%			0; **59** (28)	2 (21)			1 (25)	4 (29)
Auto T.**	0.77%				15 (13)				15 (43)
Construction	4.41%				0; **92**				0; **68**
Metal Prod.*	1.51%				29 (53)				**70 (67)**
Mech. Eng.*	2.80%				16 (15)				46 (47)
Auto Prod.*	1.35%				12 (20)				**61 (63)**
Rejection frequencies using critical Brier score thresholds (in %)									
Overall	1.70%	5	7	9	15	7	11	16	24
Manufacturing*	1.84%		3 (16)	9 (18)	12 (17)		22 (8)	22 (14)	24 (13)
Trade**	1.24%		0; **63** (5)	4 (11)	5 (12)		9 (23)	14 (27)	20 (23)
Wholesale T.**	1.26%			1; **50** (21)	2 (20)			8 (14)	19 (22)
Auto T.**	0.77%				11 (10)				23 (20)
Construction	4.41%				0; **87**				0; **56**
Metal Prod.*	1.51%				24 (52)				**54 (52)**
Mech. Eng.*	2.80%				15 (14)				35 (29)
Auto Prod.*	1.35%				17 (34)				**50 (51)**

Table 10

Is Non-Pooling Inferior to Pooling for Sectorally-Regionally Specialized Banks? Direct Quality Comparison (AUC)

The table shows the relative frequency (in %) that a rating system calibrated on a bank's own historical data performs worse (in brackets: significantly worse at 10% level) than a system calibrated on data from either manufacturing or trade. For each randomly drawn bank, the AUC is directly compared. Results are shown for banks that are specialized in one of two sectors and additionally specialize in one of four regions. They are based on 500 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%. (Av. DR = Average default rate; Manuf. = Manufacturing)

Sector-Region	Av. DR	In-sample (# obs.)		Out-of-time (# obs.)	
		3,750	1,875	750	375
		AUC			
Overall	1.70%	31 (2)	46 (5)	65 (15)	74 (21)
Manuf. – South	1.77%	70 (9)	64 (11)	31 (1)	60 (11)
Manuf. – North-West	1.36%	**20** (0)	53 (9)	60 (13)	69 (14)
Manuf. – South-East	3.11%		26 (1)		55 (10)
Manuf. – North-East	1.68%		**89** (4)		**93** (28)
Trade – South	1.34%		34 (2)		72 (13)
Trade – North-West	0.98%	73 (14)	67 (13)	64 (10)	**76** (17)
Trade – South-East	1.77%		40 (0)		**94** (**35**)
Trade – North-East	1.45%		63 (4)		**99** (**51**)

Table 11

Is Non-Pooling Inferior to Pooling for Sectorally-Regionally Specialized Banks? Comparison of Quality Distributions

The table shows the relative frequency (in %) that a rating system calibrated on a bank's own historical data – the stepwise system – performs worse than a pooled system calibrated on data from either manufacturing or trade. For the AUC, relative frequencies are shown that the AUC of the stepwise system is lower than the 10%-quantile of the AUC distribution of the pooled system. Results are based on 500 simulations assuming that each bank's portfolio experiences a portfolio default rate of 1.7%. (Av. DR = Average default rate; Manuf. = Manufacturing)

Sector-Region	Av. DR	In-sample (# obs.)		Out-of-time (# obs.)	
		3,750	1,875	750	375
AUC					
Overall	1.70%	8	13	16	21
Manuf. – South	1.77%	25	24	5	16
Manuf. – North-West	1.36%	4	12	26	25
Manuf. – South-East	3.11%		7		14
Manuf. – North-East	1.68%		**64**		**50**
Trade – South	1.34%		9		25
Trade – North-West	0.98%	31	28	22	36
Trade – South-East	1.77%		13		**59**
Trade – North-East	1.45%		28		**90**

References

Altman, Edward I.: "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", Journal of Finance 23 (September 1968), 568–609. – Basel Committee on Banking Supervision: "International conversion of capital measurement and capital standards: A revised framework" (June 2004), Basel. – Basel Committee on Banking Supervision: "Range of practice in banks' internal ratings systems" (January 2000), Basel. – *Bloch,* Daniel A.: "Evaluating predictions of events with binary outcomes: an appraisal of the Brier score and some of its close relatives", Technical Report No. 135 (May 1990), Stanford University, Division of Biostatistics. – *Blochwitz,* Stefan/*Liebig,* Thilo/*Nyberg,* Mikael: "Benchmarking Deutsche Bundesbank's default risk model, the KMV Private Firm Model and common financial ratios for German corporations", Working paper Deutsche Bundesbank, KMV (November 2000). – *Carey,* Mark: "Credit risk in private debt portfolios", Journal of Finance 53, 4 (August 1998), 1363–1387. – *DeLong,* Elizabeth R./*DeLong,* David M./*Clarke-Pearson,* Daniel L.: "Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach", Biometrics 44 (September 1988), 837–845. – Deutsche Bundesbank: "The methodological basis of the Deutsche Bundesbank's corporate balance sheet statistics", Deutsche Bundesbank Monthly Report (October 1998). – Deutsche Bundesbank: "The Bundesbank's method of assessing the creditworthiness of business enterprises", Deutsche Bundesbank Monthly Report (January 1999). – *Diebold,* Francis X./*Rudebusch,* Glenn D.: "Scoring the leading indicators", Journal of Business 62 (July 1989), 369–391. – *Engelmann,* Bernd/*Hayden,* Evelyn/*Tasche,* Dirk: "Testing rating accuracy", Risk (January 2003), 82–86. – *Frerichs,* Hergen/*Wahrenburg,* Mark: "Evaluating internal credit rating systems depending on bank size", Working Paper Series: Finance and Accounting, No. 115, University of Frankfurt. – *Günther,* Thomas/*Grüning,* Michael: "Einsatz von Insolvenzprognoseverfahren bei der Kreditwürdigkeitsprüfung im Firmenkundenbereich", Die Betriebswirtschaft 1 (2000), 39–59 (Use of processes for default prediction for corporate borrowers, in German with English abstract). – *Hüls,* Dagmar: Früherkennung insolvenzgefährdeter Unternehmen. Düsseldorf: IDW-Verlag, 1995 (Early identification of companies with high default risk, in German). – *Lopez,* Jose A.: "Evaluating the predictive accuracy of volatility models", Journal of Forecasting 20 (2001), 87–109. – *Lopez,* Jose A.: "Regulatory evaluation of value-at-risk models", Journal of Risk 1 (June 1999), 37–64. – *Niehaus,* Hans-J.: Früherkennung von Unternehmenskrisen. Düsseldorf: IDW-Verlag. 1987 (Early identification of company crises, in German). – *Shumway,* Tyler: "Forecasting bankruptcy more accurately: a simple hazard model", Journal of Business 74 (2001), 101–124. – *Sobehart,* Jorge R./*Keenan,* Sean C./*Stein,* Roger M.: "Benchmarking quantitative default risk models: a validation methodology", Moody's Investors Service, Global Credit Research (March 2000). – *Winkler,* Robert L.: "Evaluating probabilities: asymmetric scoring rules", Management Science 40 (November 1994), 1395–1405.

Summary

Does Pooling of Financial Statements and Default Data across Specialized Banks Improve Internal Credit Rating Systems?

Under the new Basel capital accord, banks will have the opportunity to estimate default probabilities for regulatory capital calculation. In the European Union, most banks will implement the internal ratings based approach, as most corporate customers are not externally rated. Based on data from Deutsche Bundesbank and using a simulation approach, this paper addresses the issue whether pooling of data improves rating system quality even if participating banks are regionally or sectorally specialized and the pooling does not take these factors into account. The primary result is that even under these circumstances pooling is beneficial for most small and medium-sized banks independent of their specialization. (JEL G2, G21, G28, C52)

Zusammenfassung

Verbessert die Zusammenführung von Bilanz- und Ausfalldaten spezialisierter Banken die Qualität interner Kreditratingsysteme?

Mit der Inkraftsetzung der Basel-II-Regeln werden Banken die Möglichkeit haben, Ausfallwahrscheinlichkeiten für die Berechnung des regulatorischen Kapitals selbst zu schätzen. In der Europäischen Union werden die meisten Banken den auf internen Ratings basierten Ansatz umsetzen, da die meisten Firmenkunden keine externen Ratings besitzen. Basierend auf Daten der Deutschen Bundesbank und mithilfe eines Simulationsansatzes adressiert dieser Beitrag die Fragestellung, ob die Zusammenführung von Daten mehrerer Banken die Qualität der Ratingsysteme einzelner Banken verbessert, selbst wenn diese regional oder sektoral spezialisiert sind und die Zusammenführung unter Nichtberücksichtigung regionaler und sektoraler Faktoren erfolgt. Das Hauptergebnis ist, dass die Zusammenführung von Daten selbst unter diesen Bedingungen vorteilhaft für die meisten kleinen und mittelgroßen Banken ist, und zwar unabhängig von ihrer Spezialisierung.

Résumé

La mise en commun de données de bilan et de données de pertes de banques spécialisées améliore-t-elle la qualité des systèmes internes de rating de crédit ?

L'entrée en vigueur de la règlementation Bâle II permettra aux banques d'évaluer elles-mêmes les probabilités de pertes pour calculer le capital régulatoire. Dans l'Union Européenne, la plupart des banques utiliseront les ratings internes car la plupart des entreprises clientes ne possèdent pas de ratings externes. Sur base de données de la Deutsche Bundesbank et à l'aide d'une simulation, l'auteur de cet article examine si la mise en commun de données de plusieurs ban-

ques améliore la qualité des systèmes de rating des différentes banques, même si celles-ci sont spécialisées au niveau régional ou sectoriel et même si cette mise en commun de données se réalise au-delà des limites régionales et sectorielles. Il en conclut principalement que la mise en commun de données est avantageuse pour la plupart des petites banques et des banques moyennes, indépendamment de leur spécialisation.