TECHNISCHE UNIVERSITÄT DRESDEN

Dresden University of Technology
Faculty of Business Management and Economics

Dresden Discussion Paper Series in Economics

Verbesserung der Vergleichbarkeit von Schätzgüteergebnissen von Insolvenzprognosestudien

MARTIN BEMMANN

Dresden Discussion Paper in Economics No. 08/05

ISSN 0945-4829

Address of the author:

Martin Bemmann
Technische Universität Dresden
Lst. für Wirtschaftspolitik und -forschung
Münchner Platz 3
01069 Dresden
Deutschland

e-mail : martin.bemmann@web.de

Editors:

Faculty of Business Management and Economics, Department of Economics

Internet:

An electronic version of the paper may be downloaded from the homepage:
http://rcswww.urz.tu-dresden.de/wpeconomics/index.htm

English papers are also available from the SSRN website:
http://www.ssrn.com

Working paper coordinator:

Dominik Maltritz
e-mail: wpeconomics@mailbox.tu-dresden.de

Improving the comparability of insolvency predictions

Martin Bemmann
Working paper, 2005/06/23[1]

Abstract:

This working paper aims at improving the comparability of forecast quality measures of insolvency prediction studies.

For this purpose, in a first step commonly used accuracy measures for categorial, ordinal and cardinal insolvency predictions are presented. It will be argued, that ordinal measures are the most suitable measures for sample spanning comparisons concerning predictive power of rating models, as they are *not* affected by sample default rates. A method for transforming cardinal into ordinal accuracy measures is presented, by which comparisons of insolvency prediction results of older and present-day studies are enabled.

In the second part of the working paper an overview of influencing variables – aside from the quality of the insolvency prediction methods – is given, which affect the accuracy measures presented in the first part of the paper and thus impair sample spanning comparison of empirically obtained forecast quality results. In this context, methods for evaluating information losses that are attributable to the discretization of continuous rating scales or preselection of portfolios are developed.

Measure results of various insolvency prognosis studies are envisaged and compared with three benchmarks. First benchmark is the accuracy that can be achieved solely by taking into account legal status and industry classification of corporations. The second benchmark is the univariate prognosis accuracy of single financial ratios. As third benchmark, ALTMAN's Z-score model is examined, a multivariate insolvency prediction model, that is currently used as reference rating model in many empirical studies. It turns out, however, that the Z-score's forecast quality is so discontenting, that its application is not recommendable. Instead it is suggested to use those rating models that are cited in this discussion paper, which are fully documented and which therefore can be rebuilt and directly applied to any desired data sample. If applied to the respective target groups, their performance matches with the performance of commercial rating systems, like bureau and business scores for rather small companies, middle market rating models for SMB, or agency ratings for large public companies.

JEL classification: G33, C14

Key words: financial ratio analysis, corporate bankruptcy prediction, forecast validation, accuracy ratio, information entropy, sample selection, rating granularity

[1] Assistant at the chair of Economic Policy and Economic Research at Technische Universität Dresden, Email: bemmann@wipo.wiwi.tu-dresden.de or martin.bemmann@web.de. Updated German and English versions of this working paper are available at www.ssrn.com.

Table of contents

1 Why to measure accuracy of insolvency predictions?

The ability to provide powerful predictions of corporate insolvencies is of great importance both from an individual and a collective point of view. From an individual perspective, alongside owners, employees, customers, suppliers, certified accountants, and other counterparties of companies, in particular banks are interested in accurate insolvency predictions, as in case of bankruptcies of their costumers they have to reckon considerable losses.[2,3] From a bank's perspective powerful insolvency predictions are a fundamental prerequisite for enabling risk commensurate credit charges and/or the embodiment of non-financial credit terms (limits, required collaterals), for improving the cost-efficiency of its credit processes (by identifying those critical cases that require a more elaborate supervision and support by credit experts), for improving the liquidity of the bank's assets and for increasing the controllability of credit risks via securitization of individual credits or entire portfolios, and for determining and controlling economic and regulatory capital demands.[4]

From a national economy perspective, the employment of powerful insolvency prediction models by creditors is an important precondition for guaranteeing the soundness and stability of the banking system and for the implementation of risk-sensitive loan terms which provide a motivation for incentive compatible, risk sensitive behavior of debtors.[5,6]

By measuring the *accuracy* of insolvency predictions an important attribute of methods for insolvency predictions shall be evaluated, that also possesses a major role in the current discussion accompanying the introduction of the Basel II regulatory rules.[7]

Other output-oriented attributes of rating models, which shall *not* be considered in this working paper, are the temporal stability of prognoses[8], the additional information utility of rating models when used in combination with other rating models[9], the ability of the models to *theo-*

---

[2] See also DIMITRAS, ZANAKIS, ZOPOUNIDIS (1996, p. 488) and BALCAEN, OOGHE (2004, p. 4) for an analysis of stakeholders in insolvency predictions for business companies.

[3] In an empirical study FRANKS, SERVIGNY, DAVYDENKO (2004) identify various influencing variables on loss given default rates of bank loans and find, that on average German banks can redeem only 60% of the loans outstanding, for instance by realizing collateral (ibid., p. 4). See on this BASEL COMMITTEE (2000b, p. 27f.), BASEL COMMITTEE (2000c, p.7f), S&P (2003a, p. 64, p. 66) or MOODY'S (2004a, p. 13). For further empirical results and over 60 references upon loss given default see also GUPTON and STEIN (2002).

[4] On the theoretical and empirical relevance of insolvency predictions for banks see for instance ENGLISH, NELSON (1998, p. 11f.), TREACY, CAREY (2000/1998, p. 897), BASEL COMMITTEE (2000b, p. 33) and ESCOTT, GLORMANN, KOCAGIL (2001, p. 3).

[5] see for instance DEUTSCHE BUNDESBANK (2001, p. 15), BASEL COMMITTEE (2004, §4), OeNB (2004b, p. 33)

[6] Excessive credit risks were the major cause for the about 100 bank insolvencies in Germany since the 1960ies, see FISCHER (2004, p. 13) and the literature thee cited. See also http://www.the-exit.net/plaza/-bankinsolvenzen.de (22.6.2005) for a survey on all bank insolvencies in Germany since 1945 (without specification of insolvency causes, though).

[7] see DEUTSCHE BUNDESBANK (2003), ÖSTERREICHISCHE NATIONALBANK (2004), BASEL COMMITTEE (2005)

[8] FONS (2002), CANTOR, MANN (2003), LÖFFLER (2003, 2004 a,b), ALTMAN, RIJKEN (2004), HAMILTON (2004)

[9] See SHANNON, (2001/1948, p. 13f.), KEENAN, SOBEHART (1999, p.11f), SOBEHART, KEENAN, STEIN (2000, p. 15) and STEIN ET AL (2003, p. 38). See as well LEHMANN (2003) and GRUNERT, NORDEN, WEBER (2005) who showed, that the inclusion of "soft factor" ratings into the rating systems of a German banks (statistically) significantly increased the accuracy of the bank intern rating systems.

retically (*causally*, *structurally*) explicate insolvencies[10] or the models' eligibility to derive recommendations aimed at influencing individual insolvency likelihood.[11,12]

It might be desirable to directly state the accuracy of rating models in monetary terms, e.g. as *expected profitability in basis points per volume of credit*.[13] Such a measure would not only facilitate textual interpretability, it could also be directly related to the utility of the decision maker. However, the data requirements that have to be fulfilled are prohibitive for real world applications.[14]

Accuracy measures for insolvency predictions that are subsequently presented do not make such high demands, but they can only be heuristically justified, which means, that only *"as a general rule"* – but unfortunately not always – a prediction model that is superior to another prediction model according to a certain measure will indeed increase the utility of a specific decision maker. On the other hand, these measures are *intersubjective*, as they are not based on individual cost-benefit-relations and they are not only suited for evaluating *insolvency predictions*. They can be used for evaluating *predictions* and *diagnoses* of other areas of application as well, given the events that have to be predicted (or diagnosed) are of a categorial (yes/no) type.[15] The multipurpose character of the measures and methods that are presented in the following becomes also apparent by the fact, that central terms and concepts for insolvency prediction measures have actually been adopted from other scientific domains that are also concerned with predicting (or diagnosing) uncertain events, such as signaling theory[16], meteorology [17] or medical sciences[18].

---

[10] A major advantage of structural and inductive models, for instance scoring models or expert systems, compared with statistical models lies herein. For a survey of the various methods that are used for predicting insolvency see e.g. GÜNTHER, GRÜNING (2000, p. 41) or ÖSTERREICHISCHE NATIONALBANK (2004, p. 32).

[11] Herein lies a major shortcoming of statistical methods. Although they identify what combinations of variables empirically correlate well with defaults, they do not reveal causalities – and thus are potentially manipulable by economically nonsensical conducts. And this is one of the reasons, besides trying to protect their intellectual property (see e.g. BLÖCHLINGER, LEIPPOLD (2005/2004, p. 20)) and besides trying to prevent "self-fulfilling prophecies" (see e.g. KÜTING, WEBER (2004/1993, p. 350f.), why banks are reluctant to make their own insolvency prediction models transparent to their customers.
Opposed to that, inductive methods – such as scoring models or expert systems – do in principle allow theoretical explanations and derivations of recommendations aimed at influencing individual insolvency probability. Limits to their usefulness have to be recognized as they are not based on complete and consistent models.
The possibility to theoretically explicate insolvencies is one of the major strengths of structural models. They are based on complete and consistent – and ideally also realistic – models. In principle they even allow the derivation of quantitative predictions of insolvency probabilities without having to use historical insolvency data as input. Whether they are suited for deriving concrete recommendations depends on the "deepness" of their modelings. If a structural model, as e.g. the KMV Public Firm Model (see KEALHOFER (2003)) is essentially exclusively based on the *level* and *variability* of a firms' stock price, without revealing the influencing variables of these parameters, the model cannot be used for *managing* individual probabilities of default.

[12] For further, predominantly technical and organizational requirements for the implementation of rating models see KRAHNEN, WEBER (2001), CROUHY, GALAI, MARK (2001) or BASEL COMMITTEE (2000c, 2004).

[13] This approach is actually pursued in a simulation study by JORDÁO, STEIN (2003).

[14] See e.g. the model of JORDÁO, STEIN (2003). It does not only require the knowledge of all rating relevant data of all potential customers. Also the competing banks' rating models and price policies would have to be known in detail in order to establish own monetary consequences of the application of a certain rating model.

[15] For explanatory variables with more than two possible parameter values more efficient quality measures are generally available. Nevertheless, an application of the quality measures that are presented subsequently is technically feasible, if the explanatory variable is transformed to a two-state variable (for instance by assigning a value of "1".for above average [or above median] parameter values and "0" for below average [median] values, see for instance GUPTON, STEIN (2005, p. 26ff.))

[16] see *Receiver-Operating-Characteristic (ROC)* for describing error performance in data transmissions, see chapter 2.3.1, or *(information) entropy,* see chapter 2.4.

[17] see evaluation of accuracy of predictions of precipitation via BRIER-scores in BRIER (1950), WINKLER (1994)

In the following chapter three groups of accuracy measures for judging insolvency predictions are presented: within the groups a number of measures will be presented that are partly axiomatically founded, that allow good direct interpretability, or that exhibit analogies to other measures, which are widely used in their original fields of applications.

In the proximate chapter feasibilities of sample or portfolio spanning comparisons of predictive accuracy measures will be discussed and empirical results of insolvency prediction studies will be given.

For obtaining not only a relative ordering of the various models but for obtaining also a first notion of "absolute" quality of the models, three different benchmarks are consulted, that are meant to form a lower level of just about acceptable predictive quality. Surprisingly, some models that are still in commercial use today can only marginally surmount these hurdles, while other commercial models even fail at them.

[18] see evaluation of predictions of therapy results in LEE (1999) or see SWETS (1973, p. 997ff.), SWETS (1988, p. 1287ff.), or SWETS, DAWES, MONAHAN (2000, p. 4ff.) for various other medical applications.

2 Accuracy measures for judging performance of insolvency predictions

2.1 Operationalizing predictive accuracy

When referring to the *accuracy* of a method or model for insolvency prediction, in the following *the degree of accordance between the insolvency predictions and the realized insolvency events* is meant.
Mathematically comprehensible specifications of *accuracy* have to account for, whether insolvency predictions that have to be evaluated are of *categorial*, *ordinal* or *cardinal* nature.

- **Categorial insolvency predictions** are insolvency predictions that only use the two extreme predictions: "corporate A will default [within a specified time horizon (usually one year)[19]]" vs. "corporate B will *not* default".
- In case of **ordinal insolvency predictions** evaluations of the *relative probabilities of defaults* of the corporations under consideration are given: "corporate B will default with a greater likelihood then corporate A, but with a smaller likelihood than corporate C". Although ordinal insolvency predictions could in principle be arbitrarily differentiated, in practice ordinal rating systems preponderate that transform their results on a discrete 7- or 17-ary scale[20] and use a notation that was adopted from S&P.[21,22]

---

[19] If not stated otherwise, in the following accuracy of insolvency predictions will refer to a prediction horizon of *one* year. More important than the decision upon *which* horizon to choose, is the decision to choose *always the same* horizon when comparing different prediction systems. As the majority of insolvency studies report their results only for horizons of *one* year, this horizon had to be used as base of comparison.
Besides data availability, textual reasons for choosing just this horizon may be offered, too.
- A model that correctly classifies many defaults that occur within one year automatically captures with that a lot of the defaults that occur in an n-year horizon.
- At least some parts of the defaults of corporations can be referred to attributes of corporations, which cannot be remedied on short notice, e.g. low endowments with equity capital or dependencies from few key customers. Corporations that have above/below average probabilities of default within the next year for these reasons tend to have above/below average probabilities of default in subsequent years as well.
- Many rating models utilize annually updated financial statements as their major - and often exclusive - source of input. As comparison horizon, therefore, a period of at least one year – or a whole-numbered multiple of it - should be chosen.
- According to the Basel II requirements banks are obliged to estimate one-year-probabilities of default for their loans (BASEL COMMITTEE (2004, § 285, 331)), they have to update their ratings once a year (ibid. § 425) and they have to check and validate their rating systems (at least) once a year (ibid., §§ 443, 449).
On empirical findings concerning bank practice see also BASEL COMMITTEE (2001, p. 12): "The 'time horizon' over which a rating is expected to be valid (i.e. the forecast horizon of the rating) is mostly described by banks to be one year, [..]. The decision for a one-year horizon is mostly based on annual financial reporting cycles (bank and borrower), frequency of internal review of the rating, and in some cases the uncertainties of projected performance beyond one year."

[20] 17-ary scale: 1="AAA", 2="AA+", 3="AA", 4 ="AA-",5="A+", 6="A", 7="A-", 8="BBB+", 9="BBB", 10="BBB-", 11="BB+", 12="BB", 13="BB-", 14="B+", 15="B", 16 = "B-", 17 = "CCC/C",
7-ary scale: "AAA", "AA", "A", "BBB", "BB", "B", "C".

[21] In BASEL COMMITTEE (2000c, p. 23f) rating symbols of 30 rating systems of various international rating agencies are investigated. 22 agencies use letter combinations for expressing their ratings (ca. 75%), 6 agencies communicate their ratings in the form of [numerical] marks, and only 2 agencies render their ratings in terms of probabilities of default. From the 22 letter-ratings 14 exactly match the detailed ("modified") 17-ary S&P notation and 2 ratings match the abbreviated ("whole-letter") 7-ary S&P notation.
Banks, however, predominantly use *numerical* rating class denotations (ca. 85%) instead of letter combinations (ca. 15%), see ENGLISH, NELSON (1998, p. 4).

- **Cardinal insolvency predictions** assign *probabilities of default* to each corporate.

The respective methods are *downwardly compatible*: by any weakly monotone transformation (cardinal) probabilities of defaults can be converted to (ordinal) score values, which can also be merged to a finite number of rating classes. By further merging neighboring rating classes, so long until only two classes remain, ordinal insolvency predictions can be transformed to cardinal insolvency predictions.
The reverse can be achieved by using empirical default data. Ordinal predictions can be converted to cardinal predictions by utilizing historical, i.e. realized rating class specific *default rates* as *default probabilities*.[23,24,25] This approach is fraught with a few problems and was therefore extended in various ways. In order to improve individual probability forecasts some of the extended approaches for instance incorporate current macro economic indicators or other individual characteristics besides individual ratings.[26]

The type in that insolvency predictions *are* available (categorial vs. ordinal vs. cardinal) is conditional on the estimation method that is used.[27] Categorial predictions are for instance delivered by *discriminant analyses* or *neural networks,* though their output is sometimes interpreted in an ordinal fashion.[28] Ordinal insolvency predictions result for instance from applying s*ubjectively parameterized scoring or ratio models* while *logit or probit regression models* can provide cardinal predictions.

The type in that insolvency predictions *have to be* available is dependent on their intended usage: if a decision maker has only two options for action available - e.g. accepting or refusing a potential debtor, positively or negatively deciding about the eligibility of some financial assets as collateral[29] - categorial predictions do suffice. For a more differentiated, qualitative

---

[22] "To provide finer rating gradations to help investors distinguish more carefully among issuers, […] Standard and Poor's in 1974, and Moody's in 1982 started attaching plus and minus symbols to their ratings." CANTOR, PACKER (1994, p. 2)

[23] See S&P (2004a, p.11): "Many practitioners utilize statistics from this default study and CreditPro® to estimate probability of default and probability of rating transition. It is important to note that Standard & Poor's ratings do not imply a specific probability of default; however, Standard & Poor's historical default rates are frequently used to estimate these characteristics."

[24] See e.g. the annually and quarterly updated default studies of STANDARD AND POOR'S and MOODY'S, cf. S&P (2004a) and MOODY'S (2004).

[25] see also STEIN (2005, p. 1218): „While it is not always the case that powerful models are calibrated accurately to probabilities of default, it is empirically feasible to calibrate a model to real-world default probabilities by performing a default study on the historical behavior of each model score. Thus assuming a bank has historical data and can perform ROC analysis, it can also calibrate a model to a similar level of accuracy using the same machinery."

[26] Cf. KEENAN (1999): growth rate of inflation adjusted industrial production index, absolute (!) number of new speculative grade issuers, share of speculative grade issuers in all issuers, yield of 10year bills, etc.; BANGIA, DIEBOLD, SCHUERMANN (2002): state of the economy (expansionary vs. recessionary); S&P (2004b, p.3): unemployment rate, slope of yield curve, aggregated business profits, distribution of outlooks of speculative grade issuers; HAMILTON (2004): individual rating outlook and rating history.

[27] For a survey of diverse insolvency prediction methods see for instance GÜNTHER, GRÜNING (2000), ÖSTERREICHISCHE NATIONALBANK (2004).

[28] See for instance ALTMAN, SAUNDERS (1998, p. 1737) for assigning discriminant analysis cut-off scores to (ordinal) rating classes.

[29] see DEUTSCHE BUNDESBANK (1999)

evaluation, e.g. as foundation for stipulating different types or amounts of collateral, (at least) ordinal insolvency predictions are required. As a basis for quantitative decisions, e.g. for pricing loans or for determining economic or regulatory capital requirements[30], cardinal predictions are indispensable.

[30] According to the new regulatory requirements that were developed by the Basel Committee for banking supervision, which come into force as from January 2006, bank internal rating systems have to be based on *cardinal* predictions of insolvency (*probabilities of defaults*), see BASEL COMMITTEE (2004, in particular §461f.).

2.2 Performance measures for categorial insolvency predictions

Categorial insolvency predictions divide corporations into two groups, "presumably insolvent" vs. "presumably not insolvent". Unfortunately, none of the prediction methods in use today come even close to yielding such selective *and* correct prognoses (for empirical findings see chapter 3.5).[31] Apart from "lucky strikes" in small samples, categorial predictions therefore inevitably will exhibit *errors*.

Two kinds of errors have to be taken into account: errors of type I – actual defaults that were forecasted as non-defaults (also called *α-error* or *false negative proportion*) and errors of type II – actual non-defaults that were forecasted as defaults (also called *β-error* or *false positive proportion*).

It is common use to state errors of type I only in relation to the number of real defaulters and errors of type II in relation to the number of real non-defaulters (see Figure 1). [32] The term *100% -error type I* is also referred to as *hit rate* and *error of type II* also as *false alarm rate,* which might be somewhat misleading.[33]

	forecasted non-defaults	**forecasted defaults**	
actual non-defaults	✓ correct forecast	✗ type II error	Σ=100%
actual defaults	✗ type I error	✓ correct forecast	Σ=100%

Figure 1: contingency table

The unweighted average[34] of both error rates or a weighted average of them could be chosen as comprehensive predictive quality measures, whereas the shares of defaulters and non-

[31] Insolvencies are also triggered by *random events,* events whose realizations can not be predicted with certainty even by state-of-the-art methods and which can only be described by probability density functions (see e.g. RiskMetrics (1997, p. 43ff.), HULL (2003)). Such events might embrace the occurrence/ non-occurrence of extraordinary damages or the concrete realization of currency exchange rates, commodity prices, and interest or inflation rates.
That *on principle* – at least based on financial statements data – no deterministic (and always true) insolvency predictions can be made, is suggested by OHLSON's (1980, p. 129) "error"-analysis: „[T]he reports of the misclassified bankrupt firms seems to lack any 'warning signals' of impeding bankruptcy. All but two of the thirteen companies reported a profit. The two losses were minor […] and these two companies had strong financial positions […]. Other ratios analyzed showed the same 'healthy patterns'. It is not surprising that these firms were misclassified, especially if one considers the profile of the nonbankrupt firms […]. None of the misclassified bankrupt firms had a 'going-concern' qualification or disclaimer of opinion. […] Some of the firms even paid dividends in the year prior to bankruptcy."

[32] See for instance SWETS (1973, p. 995), ENGELMANN, HAYDEN, TASCHE (2003, p. 13), and OENB (2004c, p. 21).

[33] See ibid. An alternative, probably more intuitive definition (which is *not* in accordance with the above mentioned definition) of *hit rate* would be: *share of really insolvent corporations in all corporations that were forecasted as insolvent*. An alternative and probably more intuitive definition of *false alarm rate* would be: share of all "false alarms" (non-insolvent corporations that were predicted as insolvent) in all alarms (predictions of insolvency). For further ratios that can be derived from contingency tables see SWETS, DAWES, MONAHAN (2000, p. 25f.)

[34] See BALCAEN, OOGHE (2004, p.12) and the literature there cited.

defaulters in the examined sample or in the basic population ("Bayesian error"[35]), or the error specific costs[36] that are involved (errors of type I: credit losses, errors of type II: foregone credit margins and "cross-selling-revenues"[37]) could be used as weighting coefficients.

A delving analysis on advantages and disadvantages of the various summary measures for *categorial predictions* shall be passed on as modern methods of insolvency prediction do no longer rest upon *categorial* but upon *ordinal* or *cardinal* predictions, for which specific measures (see chapters 2.3 and 2.4) exist.[38]

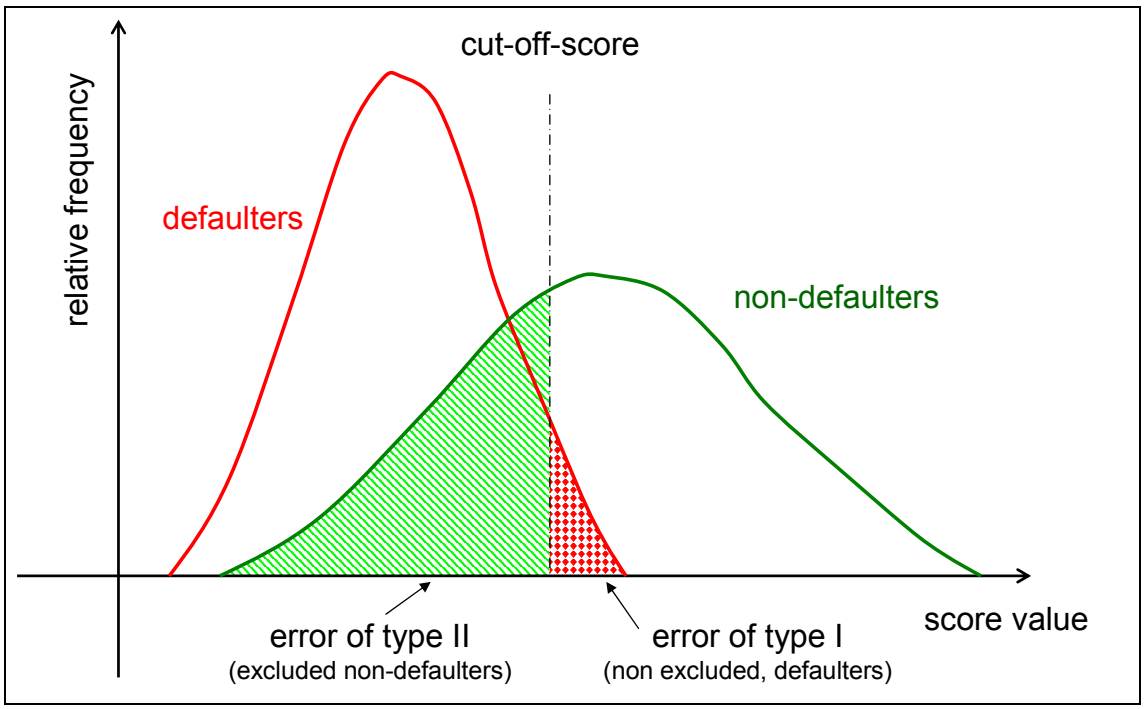

Figure 2: classification errors subject to chosen cut-off-score and rating score probability density functions for defaulters and non-defaulters[39]

In addition, all rating models that generate categorial predictions are faced with a conflict of objectives concerning type I and II error rates. Down to the model's parameterization they can achieve type I error rates of 0% and simultaneously type II error rates of 100% and vice versa – arbitrary many different trade-offs between these two extremes are usually feasible, too (for an illustration see Figure 2).[40] In the light of infinitely many alternative error-I-II-combinations it would be licentious to choose just one concrete error combination as basis for

[35] The Bayesian error rate states, which proportion of all prognoses are incorrect without differentiating between errors of types I and II. In the context of insolvency predictions this measure is poorly suited when applied to the basis populations, which are (at least for time horizons of *one year*) usually characterized by defaulter proportions of less than one or two percents. An obviously worthless rating system that always made the prediction "non-default" would cause a Bayesian error rate of just one or two percent, while an exceptionally good rating system (cf. chapter 3.5) that would cover virtually all defaulters by having to exclude only 5% of all non-defaulters would be characterized by a Bayesian error rate of 5%, cf. OeNB (2004a, p. 117ff.)

[36] see for instance NANDA, PENDHARKAR (2001, p. 155ff.)

[37] see for instance OeNB (2004b, p.33, 80)

[38] see also SWETS, PICKETS (1982, p. 24ff.) for a survey and discussion of "often used, but inadequate, [categorial] indices of accuracy"

[39] source: following DEUTSCHE BUNDESBANK (2003, p. 73), ENGELMANN, HAYDEN, TASCHE (2003, p. 5) and OeNB (2004a, p. 107)

[40] see for instance OHLSON (1980, p. 124ff.)

measuring the quality of the rating method.[41] Avoiding this sort of arbitrariness is the fundamental idea of measures for ordinal forecasts: they do evaluate a rating model's performance based on the *universe* of error-I-II-combinations that can be achieved by the respective model by applying all feasible cut-off values –not on a single, arbitrarily chosen combination.

[41] A conceivable justification for restraining the performance measurement of a rating system to one concrete error-I-II-combination would be the utilization of *optimal* (cost minimal) type I and II error combinations. However, which combination could be considered *optimal,* would not be intersubjectively identical (a bank would probably have much smaller error type II costs than a vendor of preliminary products), is dependent on conditions that may be subjectively influenceable (e.g. the concrete design of the loan's terms (interest rates, collateral, guarantees, etc.)), and may depend on non influenceable but temporally instable conditions, such as the economy wide average default rate (see on the last item BALCAEN, OOGHE (2004, p. 15)).

2.3 Accuracy measures for ordinal insolvency predictions

Although ordinal insolvency predictions are more general than categorial insolvency predictions, mere comparing statements about *relative* risks of corporations do in general *not* suffice for supporting any meaningful activity. For virtually all applications quantifications of risks are required, e.g. for assessing whether a yield premium of 1.5% p.a. is a commensurate compensation for granting a loan with a maturity of three years, whose ordinal rating implies, that *"[T]he obligor currently has the capacity to meet its financial commitment on the obligation. [But] adverse business, financial, or economic conditions will likely impair the obligor's capacity or willingness to meet its financial commitment on the obligation."* [42,43,44]

Nevertheless, a thorough analysis of accuracy measures for *ordinal predictions* is worthwhile, for the following reasons:

- Ordinal predictions correspond to that, what major rating agencies claim to deliver. [45,46]
- The appraisal of ordinal measures for insolvency predictions is meanwhile a dominant method for evaluating the quality of rating models.[47]
- There are some graphical representations of ordinal predictive quality measures available, which also facilitate the perception of quantitative measures that are derived from them.
- *Discriminative power*, which is gaged by ordinal prediction measures, is a key quality aspect of *cardinal* predictions as well – and is actually more important than any other measurable aspect of cardinal predictions, including *calibration*. Implementing *correct*

---

[42] See definition of rating grade B for long-term credit ratings according to S&P (2003b, p.7).

[43] See on this also FRERICHS, WAHRENBURG (2003, p. 13): "Under what circumstances is such a measure [*area under the ROC-curve*] useful? The ranking of borrowers is sufficient for credit risk management if banks are not able to charge different credit risk premiums for different customers in the market. In this case, banks maximize their risk-adjusted returns by not granting credit to customers with negative expected returns which is equivalent to defining a minimum credit score. Yet, this line of thought does not lead us to the AUC as a measure of system quality, but to the concept of minimized expected error costs. The AUC measures the quality of the complete ranking and not only of one threshold. Only if the threshold is difficult to define in practice, the AUC may be a sensible measure."

[44] Or, to cut a long story short: "There are no bad loans, only bad prices.", see FALKENSTEIN, BORAL, KOCAGIL (2000, p. 5).

[45] CANTOR, MANN (2003, p. 6): "MOODY'S primary objective is for its ratings to provide an accurate relative (i.e., ordinal) ranking of credit risk at each point in time, without reference to an explicit time horizon." and CANTOR, MANN (2003, p. 1, formatting added): "Moody's does *not* target specific default rates for the individual rating categories.", but also: "Moody's also tracks investment-grade default rates and the average rating of defaulting issuers prior to their defaults. These metrics measure Moody's success at meeting a secondary cardinal or absolute rating system objective, namely that ratings be useful to investors who employ simple rating 'cutoffs' in their investment eligibility guidelines.", see ibid.

[46] BASEL COMMITTEE (2000c, p.2) "Most firms report that they rate risk on a relative – rather than absolute – scale, and most indicate that they rate 'across the business cycle', suggesting that ratings should in principle not be significantly affected by purely cyclical influences." From 15 rating agencies, that provided information on this, 13 stated their ratings would measure "relative risk". Only two agencies claimed to measure "absolute risk" (KMV Corporation and Upplysningscentralen AB), see ibid. (p. 23f).

[47] See for instance MCQUOWN (1993, p.5ff), KEENAN, SOBEHART (1999, p. 5ff), STEIN (2002, p.5ff.), FAHRMEIR, HENKING, HÜLS (2002, p. 22f), ENGELMANN, HAYDEN, TASCHE (2003, p.3ff), DEUTSCHE BUNDESBANK (2003, p. 71ff.), OENB (2004a, p. 113ff.). Also the major rating agencies, cf. STANDARD AND POOR'S (2005, p.19ff.) and MOODY'S (2004c, p.3ff.), measure the quality of their rating systems by the methods and measures presented in this chapter.

(calibrated) probabilities of defaults is easier to achieve, then implementing *discriminative* probabilities of defaults.[48,49]

- Empirical comparisons of different rating models (based on the same samples) yielded largely identical orderings of the models with respect to their "accuracy", irrespective of whether ordinal or cardinal measures were used.[50] These findings imply, that quality differences between rating models are less ascribable to their differential abilities in issuing *calibrated* insolvency prognoses, which is relevant only for cardinal measures, but rather to their differential abilities in issuing *discriminative* prognoses, which is important both for ordinal and cardinal measures.

- For those aspects of cardinal insolvency predictions (in particular for *calibration*), that cannot already be evaluated with instruments originally developed for evaluating ordinal insolvency predictions, there are currently *no powerful test procedures available!* These shortcomings are essentially caused by the empirical finding, that defaults are *correlated* between corporations. [51,52]

Although the major rating agencies do not aim at meeting any predetermined cardinal default objective with their ordinal ratings for any specified space of time,[53] they demonstrate (amongst others) the quality of their ratings, by their empirically proved ability to separate groups of corporations with clearly different, monotonically increasing default rates[54] (see Figure 3 for average rating class specific one-year-default rates according to the ratings of S&P and MOODY'S and Figure 4 for the respective multi-year default-rates).

---

[48] BLOCHWITZ, LIEBIG and NYBERG (2000, p.3): "It is usually much easier to recalibrate a more powerful model than to add statistical power to a calibrated model. For this reason, tests of power are more important in evaluating credit models than tests of calibration. This does not imply that calibration is not important, only that it is easier to carry out.", see also STEIN (2002, p. 9).

[49] For details relating to calibrating rating systems see SOBEHART ET AL (2000, p. 23f.) or STEIN (2002, p. 8ff.).

[50] See for instance KRÄMER, GÜTTLER (2003) for a comparison of predictive accuracy of S&P and Moody's ratings or SOBEHART, KEENAN, STEIN (2000, p. 14) for a similar comparison of six different rating models via ordinal and cardinal measures.

[51] see BASEL COMMITTEE (2005, p. 31f.): "These ["entropy" based] measures appear to be of limited use only for validation purposes as no generally applicable statistical tests for comparisons are available. [...] The Group [*the Validation Group is a subgroup of the Research Task Force (RTF) of the Basel Committee on Banking Supervision*] has found that the Accuracy Ratio (AR) and the ROC measure appear to be more meaningful than the other above-mentioned indices because of their statistical properties. For both summary statistics, it is possible to calculate confidence intervals in a simple way. […] However, due to the lack of statistical test procedures applicable to the Brier score, the usefulness of this metric for validation purposes is limited." and ibid, p. 34: "At present no really powerful tests of adequate calibration are currently available. Due to the correlation effects that have to be respected there even seems to be no way to develop such tests. Existing tests are rather conservative […] or will only detect the most obvious cases of miscalibration […]"

[52] When defaults are only moderately correlated, realized default rates can differ materially from their expected values – even in indefinitely diversified portfolios, see e.g. HUSCHENS, HÖSE (2003, p. 152f.).

[53] S&P (2005, p. 28, formatting added): "Many practitioners utilize statistics from this default study and CreditPro® to estimate probability of default and probability of rating transition. *It is important to note that Standard & Poor's ratings do not imply a specific probability of default*; however, Standard & Poor's historical default rates are frequently used to estimate these characteristics."

[54] see for instance MOODY'S (2005, p.7)

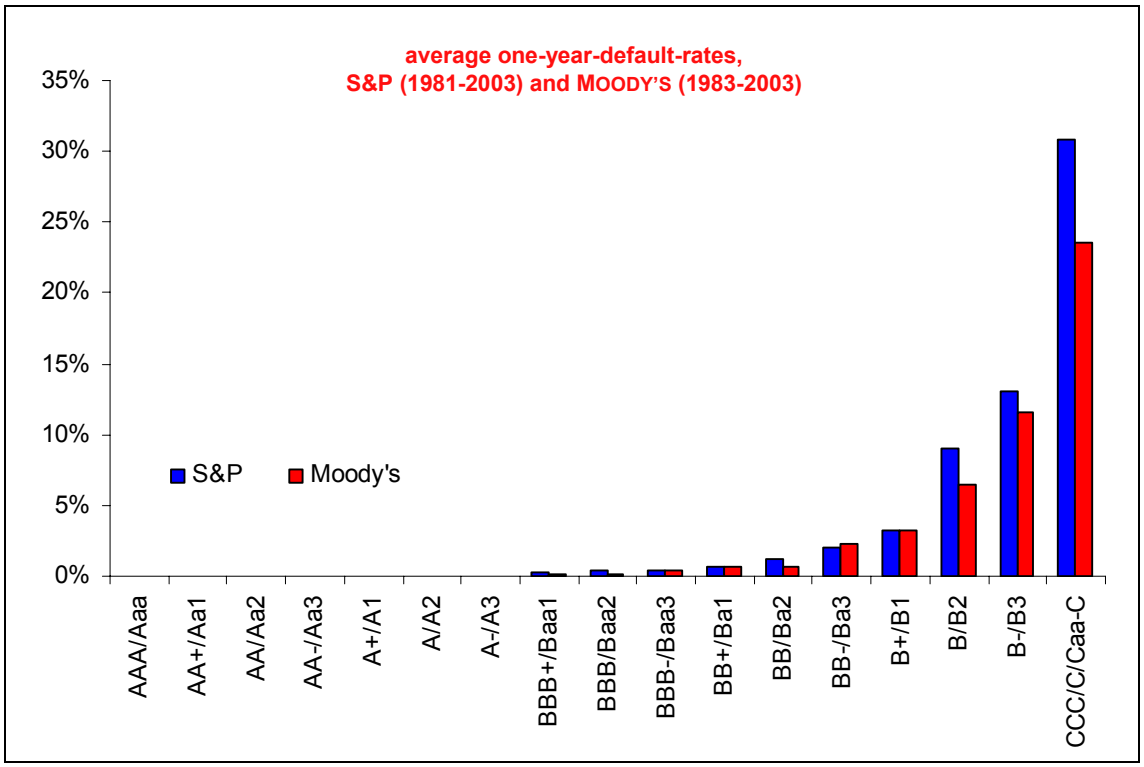

Figure 3: average historical one-year-default-rates by modified rating classes for STANDARD & POOR'S (1981-2003) and Moody's (1983-2003) ratings[55]

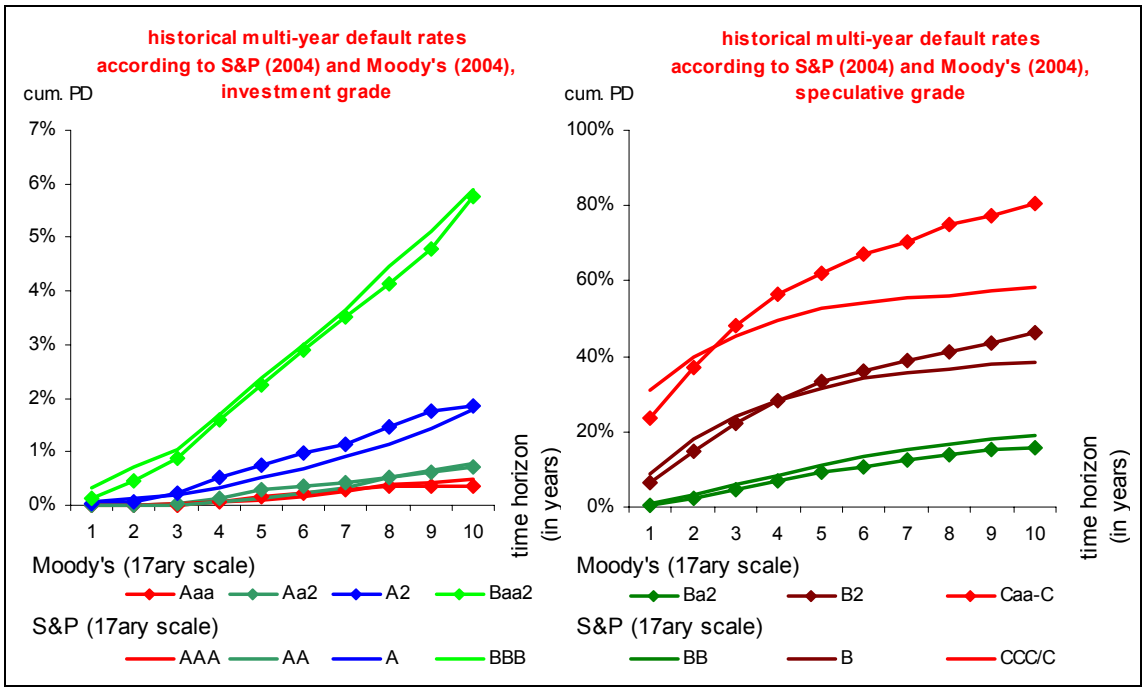

Figure 4: average historical cumulated one- to ten-year-default-rates by modified rating classes for STANDARD & POOR'S (1981-2003) and MOODY'S (1983-2003) ratings[56]

Explanatory note: Financial market participants consider ratings of S&P and MOODY'S as close substitutes concerning meaning[57,58] and quality[59,60]. Although both agencies listed

[55] see S&P (2004, p.13), MOODY'S (2004, p. 26)

[56] See S&P (2004, p.13), MOODY'S (2004, p. 26). For lack of space for the rating classes AA, A, BBB, BB and B were only displayed the middle classes (according to MOODY'S notation: Aa2, A2, Baa2, Ba2 and B2) but not the rating classes with + and - respectively *1* and *3* modifiers.

nearly identical default rates for the companies and periods they covered (S&P: 1981-2003, MOODY'S: 1983-2003) of 1.76%[61] and 1.86%[62] p.a., rating class specific one-year-default rates are smaller for all rating classes displayed in Figure 3 for MOODY'S ratings compared with the respective S&P rating classes (with a minor exception for BB-/Ba3). Wherefrom may be concluded, that MOODY'S ratings are not fully equivalent with S&P ratings but are a bit sterner.[63] However, with exception of corporations with the worst rating grades (CCC/C respective Caa/C), who account for only a very minor share of all issuers, there are only small differences concerning both agencies realized rating class specific one- and multi-year default rates.

As can be seen in Figure 3 and Figure 4, by the ratings they assigned ex-ante, both agencies were able to separate groups of companies who subsequently were characterized by clearly different realized default rates. This ability, however, is only a *necessary* but not a *sufficient* precondition for highly discriminative ratings.[64] Following two extreme examples are compatible with the figures presented above[65]:

- Example I: the rating system arranges practically all corporations to one single, middle rating class, for instance BB (Ba2) (see Figure 3), and only very few corporations to other rating classes.
- Example II: the rating system arranges practically all corporations to the two "extreme" rating classes, i.e. either AAA (Aaa) or CCC/C (Caa/C), and only a few to middle rating classes.

---

[57] see CANTOR, PACKER (1994, p. 12): „As a practical matter, however, it appears that market participants have historically viewed the Moody's and Standard and Poor's scales as roughly equivalent [...]."

[58] On a 17ary rating scale corporations that were rated both by S&P and MOODY'S receive exactly the same rating in about 40%-45% of all cases. In further 40%-45% of all cases ratings of both agencies differ by exactly one grade and in 10%-15% by exactly two grades. Differences of more then two grades appear only in 2,5%-5% of all cases, see GÜTTLER (2004, p.13 and the studies there cited). On a 7-ary scale S&P's and MOODY'S ratings coincide in 71% of all cases, in 28% they differ by one whole grade, in 1,1% by two whole grades and in 0,1% of all cases by more than two grades (own analysis basing on GÜTTLER (2004, appendix B)).

[59] see ELTON ET AL (2004, p. 2755f.) and STEINER, HEINKE (2000, p. 560f.)

[60] Based on examinations of *identical* samples of corporations, that were rated both by S&P and MOODY'S, KRÄMER, GÜTTLER (2003) and GÜTTLER (2004) find, that MOODY'S rating system is slightly superior with respect to various forecast horizons and validation measures.

[61] S&P(2004, p. 16)

[62] Special note: The value given in MOODY'S (2004, p. 26), 1.24%, is wrong, as was confirmed on inquiry at MOODY'S investors service. In MOODY'S (2005, p.17) the average all corporations default rate for the (marginally different) period 1983-*2004* is given with 1.79%.

[63] See on this GÜTTLER (2004, p.13) and the studies there cited: for corporations that were rated both by S&P and MOODY'S and where both agencies' ratings conflicted, MOODY'S was assigning worse ratings in 60% of all cases. Nonuniform results were obtained by the studies, whether the negative bias in MOODY'S ratings (or the positive bias in S&P's ratings) was attributable to negative biases mainly in the domains of investment or speculative grades.

[64] The default studies of S&P and MOODY's are based on *issuer ratings* (also *corporate credit ratings*, *implied senior-most rating*, *default ratings*, *natural ratings, estimated senior ratings*), which are intended to measure the probabilities of default of corporations – however, without reference to an explicit time horizon, see on this S&P (2003b, p. 3ff., 61ff.), CANTOR, MANN (2003, p. 6f.), MOODY'S (2004b, p. 8), and MOODY's (2005, p. 39). Ratings of specific issues, so called *issue ratings*, do additionally take into account expected losses in case of default and therefore can be considered as measures for *expected credit losses*. *Issue rating* of senior, unsecured liabilities are usually agreeing with the corporate's *issuer rating*. Depending on the order of priority in case of bankruptcy, collaterals and other influencing variables, *issue ratings* of other liabilities are derived from *issuer ratings* by "notching up or down" by usually no more than one or two points, see ibid.

[65] see on this also CANTOR, MANN (2003, p. 14)

In example I the rating system would be virtually worthless, because it hardly permits any differentiation among the various companies with respect to default probabilities (strictly speaking: with respect to default *rates*). In example II the value of information provided would be immense: the rating system would always give "extreme predictions" – i.e. predict either very low or very high chances of default - and the prognoses would be correct: as can be seen in Figure 3, hardly any AAA rated company would ever default within one year and only 0.5% would default with in ten years while about a quarter of all CCC/C rated companies would default within one year and about 80% within the next ten years.

Therefore, for determining ordinal quality of a rating model, not only default rates of the various rating classes have to be taken into account, but also the distribution of the rated companies among the various rating classes.

2.3.1 Graphical determination of predictive accuracy

The ability of a rating system to reliably discriminate between "good" and "bad" companies can for instance be visualized by ROC-[66] and CAP-curves[67] and can be quantified by various cross-convertible measures.[68]

ROC-curves were firstly used in experimental psychology in the early 1950ies.[69] Other labeling for ROC- or CAP-curves are *intelligence profile, power curve, Lorenz curve, Gini curve, lift-curve, dubbed-curve or ordinal dominance graph*.[70]

A rating system's ROC-curve is given by the universe of all combinations of *hit rates* (100% - error type I) and *false alarm rates* (error type II) that the rating system can achieve by applying different cut-off values (in case of categorial forecasts) or thresholds (in cases of ordinal or cardinal forecasts) (see Figure 5, left hand side). If the cut-off score/ threshold is chosen tight enough, all corporations would be forecasted as *insolvent* which invokes hit rates of 100% (which is desirable) but also false alarm rates of 100% (which is obviously undesirable); while when choosing a threshold too lax, hit rates and false alarm rates of 0% in each case would result. For thresholds in-between these extreme cases, there is a trade-off between bith types of errors. The quality of ordinal (and important aspects of the quality of cardinal) insolvency predictions just express itself in the nature of this trade-off. A perfect forecast system would not have to exclude a single non-defaulter in order to "hit" each defaulter (vertical course of the ROC-curve from (0%; 0%) to (0%; 100%)), tightening the threshold would only increase the false alarm rate (horizontal course of the ROC-curve from (0%; 100%) to (100%; 100%)). The ROC-curve of a rating system that assigns its ratings at random would lead along the main diagonal – each percentage point increase in hit rate would come at a cost of one percentage point increase in the false alarm rate.

ROC-curves of realistic insolvency prediction systems exhibit a concave shape (see Figure 5, see also appendix I for various empirical examples of ROC-curves). Concave shapes imply, that realized default rates continuously recede with improving credit ratings, which is also called "semi-calibration".[71]

CAP-curves (see Figure 5, right hand side) result from applying a slightly modified procedure compared with that of ROC-curves. The abscissa here does not represent *false alarm rates* (error of type I), but the share of companies – no matter whether they are defaulters or non-defaulters – that have to be excluded in order to attain a specific *hit rate*.

---

[66] "A curve […] is called a 'ROC' – sometimes short for *Receiver* Operating Characteristic, especially in the field of signal detection, and sometimes short for *Relative* Operating Characteristic, in generalized applications.", see SWETS (1988, p. 1287, formatting added).

[67] CAP … cumulative accuracy profile

[68] For alternative graphical representations of ROC-curves and for corresponding quantitative measures see SWETS, PICKET (1982, p. 31ff.), who suggest to use "binormal" ROC-curves, i.e. ROC-curves, whose abscissa and ordinate values are transformed for quantile values between 1% and 99% according to the reverse function of the Gaussian distribution. According to SWETS, PICKET (1982, p. 31f.) empirical binormal ROC-curves can be approximated quite well by linear functions – which, however, could not be confirmed in own examinations of empirical ROC-curves that are presented in Appendix I. *If* binormal ROC-curves *could* indeed be approximated by linear functions, it would be possible to inter- and extrapolate ROC-curves just on the basis of two combinations of errors of types I and II. It would also be possible to completely characterize ROC-curve with only two parameters.

[69] see SWETS (1988, p. 1287)

[70] see BLOCHWITZ, LIEBIG, NYBERG (2000, p. 33): power curve, lift-curve, dubbed-curve, receiver-operator curve; FALKENSTEIN, BORAL, KOCAGIL (2000, p.25): Gini curve, Lorenz curve, ordinal dominance graph; SCHWAIGER (2002, p. 27): intelligence profile (Aufklärungsprofil)

[71] see KRÄMER (2003, p. 403)

Starting in (0%; 0%) the CAP-curve of a perfect rating system leads "steeply" up – but not orthogonally – because at least PD% of all corporations have to be excluded in order to cover all defaulters (see the bold streak-point-streak-line in Figure 5, right hand side).

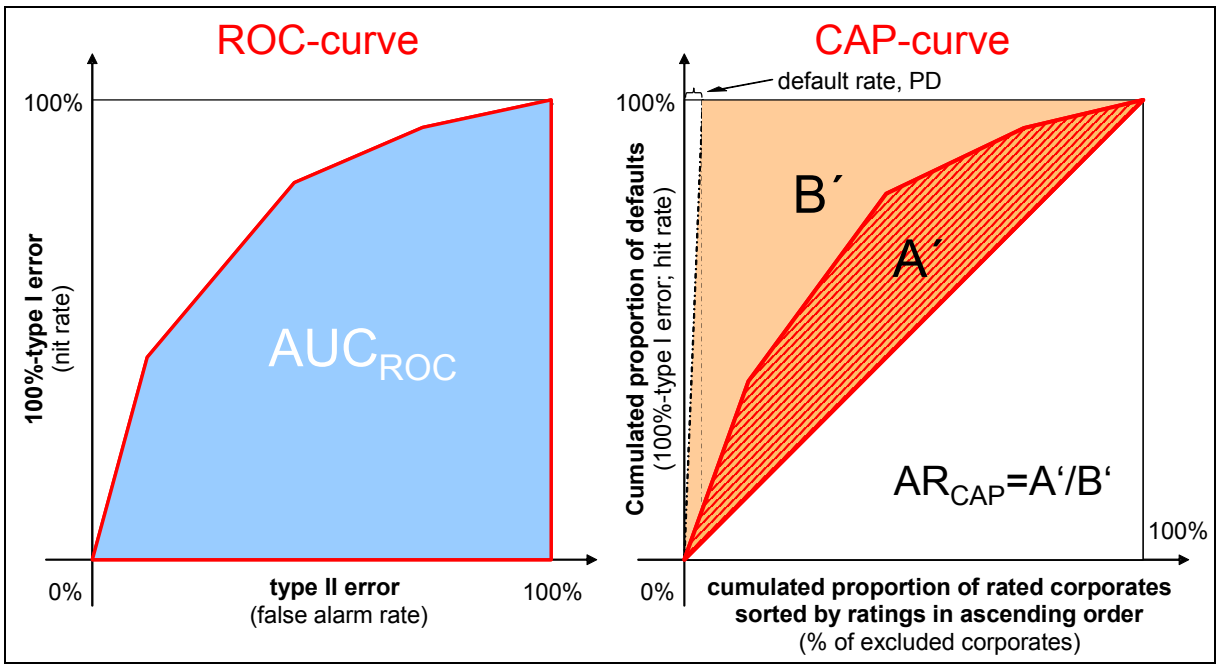

Figure 5: ROC- and CAP-curve

If - when applied to the same sample of companies - the ROC-/CAP-curve of a rating system A_1 is positioned entirely top left of the ROC-/CAP-curve of a rating system A_2, this means that rating system A_1 gives strictly better prediction for every possible threshold than rating system A_2 (cf. Figure 6, left hand side).

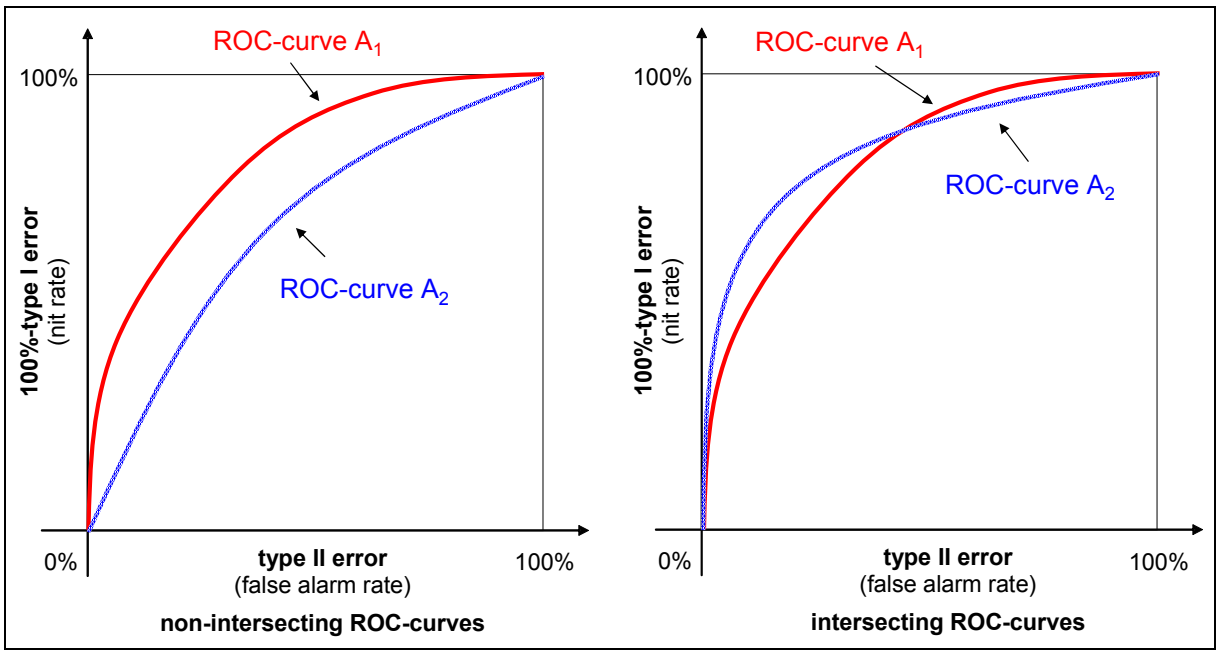

Figure 6: non-intersecting (left hand side) and intersecting ROC-curves (right hand side)

If the relative advantageousness of a rating system was solely dependent on its precision[72], system A_1 would be *objectively* superior to system A_2– irrespective of individual error costs.[73]

[72] Besides the aspects mentioned in the introductory chapter another relevant criterion for the advantageousness of rating systems might be the costs they evoke themselves, in particular the costs for obtaining and process-

If, however, the curves of both rating systems intersected at some point, no such far-ranging statements could be made.
In the example chosen (see Figure 6, right hand side) rating system B_1 is better than B_2 in differentiating among "good corporations", while system B_2 is more discriminative among "bad corporations" (for identical, low *false alert rates* it always achieves better *hit rates* than B_1).[74]

If the decision maker is *not* forced to opt exclusively for either the one ore the other rating system, he might probably be able to combine both systems to a new system B_3 that is superior to both B_1 and B_2. The same holds for a situation like that depicted in Figure 6, left hand side: although A_1 is strictly inferior to A_2, it *could* be possible to generate a system A_3 such that it is not only superior to A_2 (which is trivial) but also to A_1.

In empirical studies, it was shown, that the predictive quality of ratio based bank rating systems could be improved by including "soft factor" ratings, although the predictive quality of the ratio based ratings alone was strictly better than that of the "soft factor" ratings alone.[75]

For practical purposes in addition to graphical representations, comprehensive quantitative measures of rating accuracy are essential. Minimum requirements for such measures are that the measure shall be defined for any ROC-curve and when the ROC-curve of one rating system A_1 is top left of the ROC-curve of another rating system A_2, the measure associated with A_1 should unequivocally indicate the superiority of A_1 over A_2.[76] In the following chapter two cross-related measures are presented, that fulfill this minimum requirement.

It is possible to directly infer *relative* default rates (in relation to the average sample default rate) from the ROC- and CAP-curves' slopes: the steeper the curves are at some sections, the higher is the default rate of the companies that are covered by those sections.[77] If the average sample default rate is known, also *absolute* default rates can be calculated:

Given a total number of *def* defaulters and *ndef* non-defaulters, by additionally excluding the worst remaining $\Delta x*(def+ndef)$ companies, the *total number* of correctly classified defaulters rises by $\Delta CAP(x)*def$ and thus

---

ing the data required. Banks, for instance, often use highly automated rating systems (that tend to be cheap but less discriminative) for evaluating small loans and more labor-intense and thus expansive but more discriminative systems for larger loans, see TREACY, CAREY (2000/1998, p. 905) and BASEL COMMITTEE (2000b, p.18f.)

[73] see for instance BLOCHWITZ, LIEBIG, NYBERG (2000, p.7f)

[74] According to KRÄMER (2003, p. 402, translation), when comparing real life rating systems, ROC- or CAP-curves usually intersect. "Insofar the concept of default-dominance is not helpful in many applications […]. The concept of default dominance is primarily recommendable for sorting out substandard systems." Default-dominated prognoses can be derived from dominating prognoses, see KRÄMER (2003, p. 397f.)

[75] See on this LEHMANN (2003, p.21). See as well GRUNERT, NORDEN, WEBER (2005), who only state aggregated measures for the different models (financial ratios rating, soft factor rating, combined rating) – but no CAP- or ROC-curves. In their study, the predictive ability of the soft factor rating actually was *better* than that of the financial ratios rating (see ibid, p. 519). For the LEHMANN (2003)-study see also the survey in chapter 3.5. The GRUNERT, NORDEN, WEBER (2005)-study was not incorporated into that survey due to its minor sample size (340 non-defaulters, 69 defaulters).

[76] The converse of the minimum requirement ("If $A_1 > A_2$ then is ROC_1 top left of ROC_2") does not hold in general. It can only be excluded, that when $A_1>A_2$, the ROC-curve belonging to A_1 is down to the right of the ROC-curve belonging to A_2. However, if $A_1 > A_2$ it still might be possible that both ROC-curves intersect. See on this CANTOR, MANN (2003, p. 12): "Although the accuracy ratio is a good summary measure, not every increase in the accuracy ratio implies an unambiguous improvement in accuracy."

[77] see e.g. FALKENSTEIN, BORAL, CARTY (2003/2000, p. 30f)

F 1) $$PD(x) = \frac{\Delta CAP(x) \cdot def}{\Delta x \cdot (def + ndef)}$$

F 2) $$PD(x) = \frac{\partial CAP(x)}{\partial x} \cdot PD$$

with $\partial CAP(x)/\partial x$ = tangent of the slope of the CAP-curve at point x and

PD…average sample default rate

PD(x) local default rate at point x

Proceeding analogously for ROC-curves yields:

F 3) $$PD(x) = \frac{\Delta ROC(x) \cdot def}{\Delta x \cdot ndef + \Delta ROC(x) \cdot def}$$

F 4) $$PD(x) = \frac{\Delta ROC(x) \cdot PD}{\Delta x \cdot (1 - PD) + \Delta ROC(x) \cdot PD}$$

F 5) $$PD(x) = \frac{PD}{\frac{\Delta x \cdot (1 - PD)}{\Delta ROC(x)} + PD}$$

F 6) $$PD(x) = \frac{PD}{(1 - PD) \cdot \frac{1}{\left(\frac{\partial ROC(x)}{\partial x}\right)} + PD}$$

with $\partial ROC(x)/\partial x$ = tangent of the slope of the ROC-curve at point x

If a ROC- or CAP-curve is made up of linear fragments, as in Figure 5, this means that there is no differentiation of risk possible *within* the respective fragments, i.e. a user of the rating system is faced with the same error type I and II trade-off – and thus with the same probability of default - for all corporations that are covered by that section. One possible cause could be, that the ROC- or CAP-curve under consideration is not based on a continuous score but on discrete rating classes.[78]

[78] See on this e.g. S&P (2005, p.19ff.). In the CAP-curves there given, for all linear fragments the respective rating grades are denoted.

2.3.2 Quantitative determination of predictive accuracy

The measure most often used in context with ROC-curves is the *area under the ROC curve* AUC_{ROC}:

F 7) $AUC_{ROC} = \int_0^1 CAP(x)dx$ [79,80]

As can be shown, the area under the ROC-curves equals the probability that for two randomly chosen individuals, one chosen from all defaulters and one chosen from all non-defaulters, the non-defaulter was (correctly) assigned a better score than the defaulter.[81]

A perfect rating system would receive an AUC_{ROC} score of 100%, while a purely random ("naïve") rating system would achieve a score of only 50% (on average). [82,83]

In the field of insolvency prediction it is common to use a linearly transformed version of AUC_{ROC}, with co-domains of [-100%; +100%] instead of [0%; 100%]. This transformation also ensures that "naïve prognoses" receive scores of 0% rather than 50%. The variable thus created is called "accuracy ratio" AR_{ROC}[84], with:

F 8) $AR_{ROC} = 2 \cdot (AUC_{ROC} - 0{,}5).$

It can be shown, that the Accuracy Ratio is just a special case of other, costumary ordinal measures.[85]

In contrast to ROC-curve based measures it is *not* conventional to quote AUC-measures for CAP-curves (probably because they do not enable similar probabilistic interpretations like in case of ROC_{AUC}– at least not without further adjustments). It is conventional, however, to use the CAP-Accuracy Ratio, which is determined by a slightly modified calculation:

F 9) $AR_{CAP} = A'/B'$ (cf. Figure 5, right hand side), whereby A´ stands for the area between the CAP-curve and the main diagonal and B´ for the area between the main diagonal and the CAP-curve a perfect rating could achieve given the sample default rate.

It can be shown, that AUC_{ROC} and AUC_{CAP} are identical.[86] Thus:

---

[79] see e.g. DEUTSCHE BUNDESBANK (2003, p.71ff.)

[80] For further, well interpretable measure for ROC-curves see LEE (1999).

[81] see LEE (1999, p. 455)

[82] The lowest possible value, 0%, would be achieved by a rating system whose forecasts were *always* wrong. By simply inverting the systems forecasts, the system could easily be converted to a perfect rating system.

[83] An alternative notation for AUC_{ROC} is CoC … Coefficient of Concordance, cf. LEHMANN (2003, p. 12).

[84] Further notations for Accuracy Ratio are: *GINI index* or *GINI* coefficient (see BLOCHWITZ, LIEBIG, NYBERG (2000)), *LORENZ-MÜNZER concentration measure* (see DVFA (2004, p. 599), or *power statistic* (see FAHRMEIR, HENKING, HÜLS (2000, p. 27))

[85] See on this HAMERLE, RAUHMEIER, RÖSCH (2003, p. 21f.), who show, that the accuracy ratio is a special case of the more general measure SOMER'S D, which is also defined if the variable to be explained can adopt more than only two distinct parameter values (default vs. non-default). See also SOMER (1962, p. 804f.) for a depiction of SOMER'S D relations to other ordinal measures such as KENDALL'S tau or GOODMAN and KRUSKAL'S gamma.

[86] see ENGELMANN, HAYDEN, TASCHE (2003, p. 23)

F 10) $$AR_{CAP} = \frac{\int_0^1 CAP(x)dx - \frac{1}{2}}{\int_0^1 CAP_{perfekt}(x)dx - \frac{1}{2}}$$ with

F 11) $$\int_0^1 CAP_{perfekt}(x)dx = 1 - \frac{PD}{2}$$, it follows that

F 12) $$AR_{CAP} = \frac{\int_0^1 CAP(x)dx - \frac{1}{2}}{1 - \frac{PD}{2} - \frac{1}{2}}$$

F 13) $$AR_{CAP} = \frac{2 \cdot \int_0^1 CAP(x)dx - 1}{1 - PD}$$

For a rating system with g separate classes the integral term ∫CAP(x)dx can be decomposed into g triangular and g rectangular shaped subareas with known side lengths, see Figure 7. In this figure formulas for single triangular and rectangular surface areas are given.

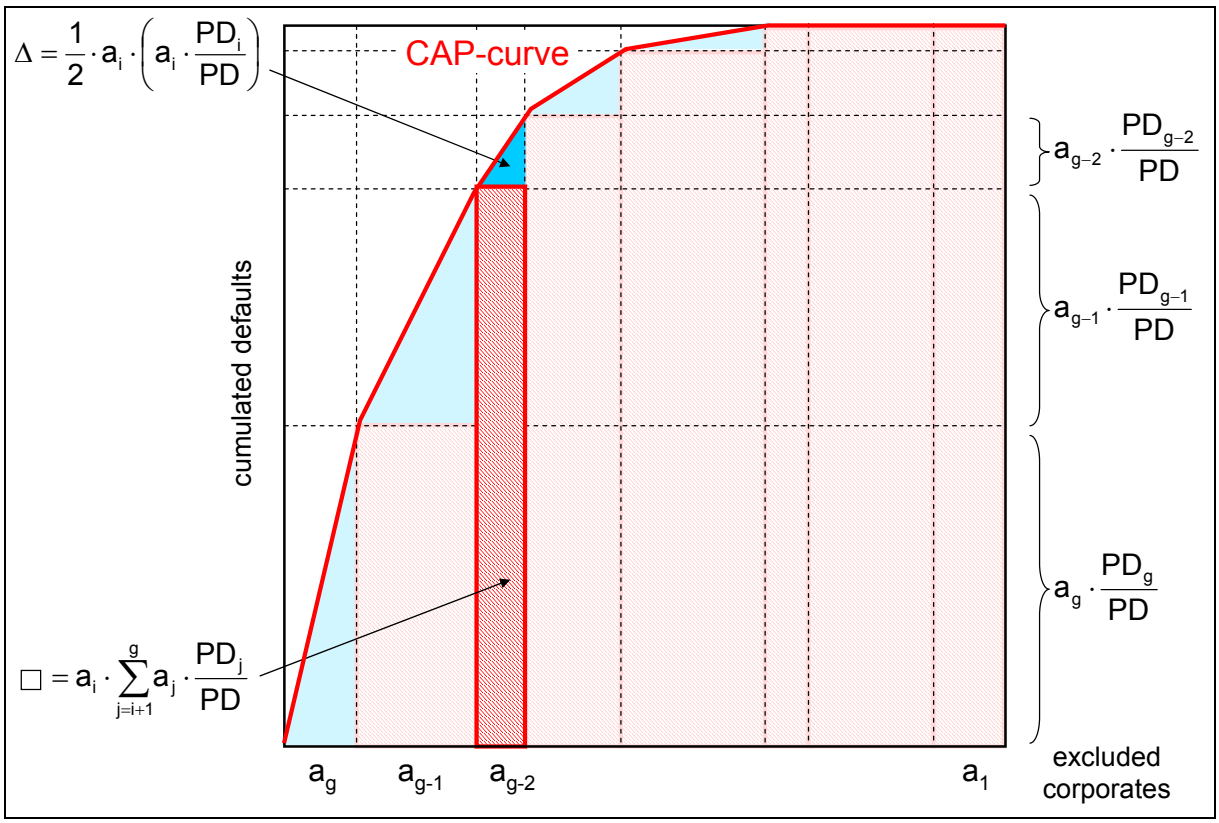

Figure 7: vertical decomposition of the area under the CAP curve AUC_{CAP}

F 14) $$\int_0^1 CAP(x)dx = \sum_{i=1}^{g} \left(\frac{1}{2} \cdot a_i \cdot \left(a_i \cdot \frac{PD_i}{PD} \right) + a_i \cdot \left(\sum_{j=i+1}^{g} a_j \cdot \frac{PD_j}{PD} \right) \right)$$ (cf. Figure 7)

with g…number of separate rating classes,
a_j … share of corporations of rating class j in all corporations,
PD_j … realized default rate in rating class j,

PD … average default rate

Given the following definition:

F 15) $\text{cumPD}_i \equiv \frac{\sum_{j=1}^{i} a_j \cdot \text{PD}_j}{\text{PD}}$ whereby cumPD_i stands for the share of the defaulters of rating classes 1..i in all defaulters and

F 16) $\text{cumPD}_g = 1$ it follows that

F 17) $\frac{a_i \cdot \text{PD}_i}{\text{PD}} = \text{cumPD}_i - \text{cumPD}_{i-1}$ By insertion to formula F 14 follows

F 18) $\int_0^1 \text{CAP}(x)dx = \sum_{i=1}^{g} a_i \cdot \left(\frac{\text{cumPD}_i - \text{cumPD}_{i-1}}{2} + (1 - \text{cumPD}_i) \right)$ and thereby

F 19) $\int_0^1 \text{CAP}(x)dx = \sum_{i=1}^{g} a_i \cdot \left(1 - \frac{\text{cumPD}_i + \text{cumPD}_{i-1}}{2} \right)$ with $\sum_{i=1}^{g} a_i = 1$

F 20) $\int_0^1 \text{CAP}(x)dx = 1 - \frac{1}{2} \sum_{i=1}^{g} a_i \cdot (\text{cumPD}_i + \text{cumPD}_{i-1})$ An insertion to formula F 13 yields:

F 21) $\text{AR}_{\text{CAP}} = \frac{2 \cdot \left(1 - \frac{1}{2} \sum_{i=1}^{g} a_i \cdot (\text{cumPD}_i + \text{cumPD}_{i-1}) \right) - 1}{1 - \text{PD}}$ and thereby

F 22) $$\boxed{\text{AR}_{\text{CAP}} = \frac{1 - \sum_{i=1}^{g} a_i \cdot (\text{cumPD}_i + \text{cumPD}_{i-1})}{1 - \text{PD}}}$$

The measures presented above are equally well suited for evaluating rating systems with continuous and discrete scores (or predicted default probabilities).[87] Transforming continuous rating scales to discrete rating scales comes along with information losses that, however, are relatively minor under reasonable conditions (see chapter 2.3.4), so that the quality of insolvency predictions can be measured solely on the basis of rating class specific relative frequencies of occurrence and realized default rates.

Special note: In publications issued by MOODY'S Investors service a slightly different accuracy-ratio definition is applied (which leads to lower values compared with the values obtained by applying the conventional accuracy ratio definition, in particular for multiyear-prediction periods).[88]

F 23) $\text{AR}_{\text{CAP,MOODY'S}} = \text{A'} / 0{,}5$ (cf. Figure 5, right hand side)

The accuracy ratio as defined by MOODY'S does change only, when the underlying CAP-curve changes. It was designed to be invariant to proportional changes simultaneously affecting all rating classes (which would not alter the CAP-curve's shape) in order to provide a measure that is unaffected by changes in average default rates. [89]

---

[87] In case of continuous scores, g would be equal to the number of corporations with $a_i = 1/g$ for all i, i.e. each corporate would form its own rating class.

[88] see for instance MOODY'S (2005, p.11)

[89] "The benefit of not including the perfect foresight comparison is that the accuracy ratio will be invariant to changes in the aggregate default rate and will only change due to changes in the rating distribution and the

However, demanding default-rate-invariance - as defined above - is not very suggestive, as may be shown by following example: an obviously imperfect rating system, that separates 90% of all corporations into one class with a PD_1 of 0% and the remaining 10% of all corporations in another class with a PD_2 of 20% (average PD: 2%) would receive an identical $AR_{MOODY'S}$-score as a perfect rating system which is yielded, if all rating class default rates are quintupled, which also separated 90% of all corporations in one class with a PD_1 of 0% and the remaining 10% of the corporations with a PD_2 of 100% (average PD: 10 %).[90] It can be shown, both empirically and theoretically, that – counter the original intention – MOODY'S-measure is actually *stronger* (negatively) related with average default rates than the "conventionally defined" accuracy ratio.[91]

distribution of default rates. In practical terms, though, there is little empirical difference between the accuracy ratio of SOBEHART ET AL [(2000)] and that used in this Special Comment." CANTOR, MANN (2003, p.11) "*The accuracy ratio measures only relative accuracy, not absolute accuracy, and is invariant to proportional changes in marginal default rates*. The marginal default rate is the percent of issuers in any given rating category that subsequently default. If the marginal default rates for all rating categories change proportionally, neither the CAP plot nor the accuracy ratio changes at all.", CANTOR, MANN (2003, p.12).

[90] Although based on different samples, both rating systems evoke the same CAP-curve (but incidentally different ROC curves): in both cases 10% of all corporations have to be excluded in order to achieve a hit rate of 100%, so that in both cases MOODY'S-accuracy-ratios of 90.0% result (=(10%*100%/2+90%*100%-0,5)/0,5). According to the formula yielding the conventional accuracy ratio (see formula F 13) both values had to be "normalized" with 1/(1-PD) which would yield an AR-value of 91,8% in the first and 100% in the second place. Such, the *perfect rating* would not only get a better score than the imperfect rating, but it would also get the highest attainable value, 100%.

[91] Source: own studies. Based on MOODY'S (2004a) cohort rating performance data in the 1983-2003 period, there were found correlation coefficients of -0.591 / -0.682 for the interrelationship of the MOODY'S modified AR (forecast periods 1 year / 5 years) and the respective 1- and 5-year default rates. Correlation coefficients for the conventional AR and 1- and 5-year default rates were -0.494 / -0.437.
In addition, a simulation experiment was performed with 10,000 simulation runs using portfolios of 5,000 companies (which is roughly equivalent to the total number of corporations that currently are rated by S&P, see S&P (2005, p. 26))), whose rating class distribution and default behavior (cumulated 5 year default rate) were taken from the S&P (2004a) default study. Contrary to empirical examinations, in the experimental settings could be assured, that rating class specific default *probabilities* were constant throughout time (or better: throughout the different imulation runs) In each simulation run random realizations of default events for all companies of the portfolio were "diced" - according to the rating class specific default probabilities - and the resulting accuracy ratio and average default rate of the whole portfolio were determined.
In effect, conventional ARs and average default rates of the 10,000 simulation runs were practically uncorrelated, while the MOODY'S modified accuracy ratios exhibited a correlation coefficient of -18,3%.

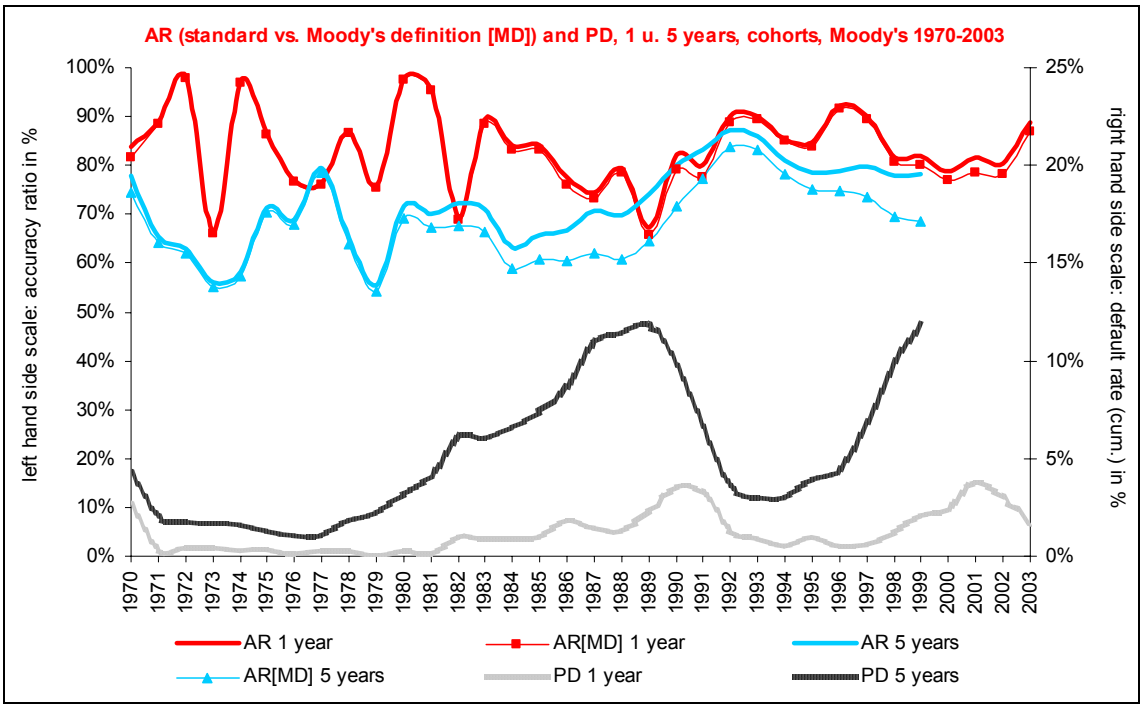

Figure 8: empirical time series of accuracy ratios for MOODY'S ratings (standard vs. MOODY'S definition) and default rates, MOODY'S cohorts, 1970-2003, forecast horizon: 1 year and 5 years

Other disadvantages that may be noted are, that the MOODY'S modified accuracy ratio is incompatible with the ROC-accuracy ratio, and that perfect rating systems cannot obtain MOODY'S values of 100%. For short forecast horizons, i.e. one year, differences between conventional and MOODY'S modified accuracy ratios are of rather negligible magnitudes (they differ by factor 1/(1-PD)), except in times of extraordinarily high default rates as in the early 1990ies or in 2001. Major differences and heavily varying magnitudes of differences in time, however, occur for longer forecast horizons (see Figure 8).[92]

[92] Owing to the rather small number of corporations that were rated by MOODY'S in the early 1970ies, time series of realized accuracy ratios is characterized by large fluctuations, see Figure 8. In 1970 only 1,031 companies were rated by MOODY'S, in 2004 nearly five times as many, see MOODY'S (2005, p.22f.).

2.3.3 Comparability with measures for categorial insolvency predictions

As was outlined in chapters 2.3.1 and 2.3.2, not only measures for categorial insolvency predictions are based on an analysis of errors of the types I and II but, after all, also ROC- and CAP-curves and the measures derived from them, such as the area-under-the-ROC-curve (AUC_{ROC}) or the accuracy ratio ($AR_{ROC/CAP}$).[93]

Albeit for every error-I-II-combination indefinitely many ROC-curves with differing AUC-values can be found that contain this particular combination,[94] it is possible to derive exact upper and lower limits for the measures that are related to these ROC-curves (for a derivation of the results see Appendix I). As the interval that is spanned by these limits is usually rather ample, additionally heuristic accuracy-ratio-estimators were developed and successfully empirically tested, that assume particular functional ROC-curve forms.

Being able to transform error-I-II-combinations into accuracy-ratio-values makes it possible to compare results of older insolvency prediction studies (see on this chapters 3.3 and 3.4), which are usually only reported in terms of single error-I-II-combinations on a *univalent* basis[95] and on a *uniform* basis with newer studies, that usually report their results in terms of AUC or accuracy ratios.

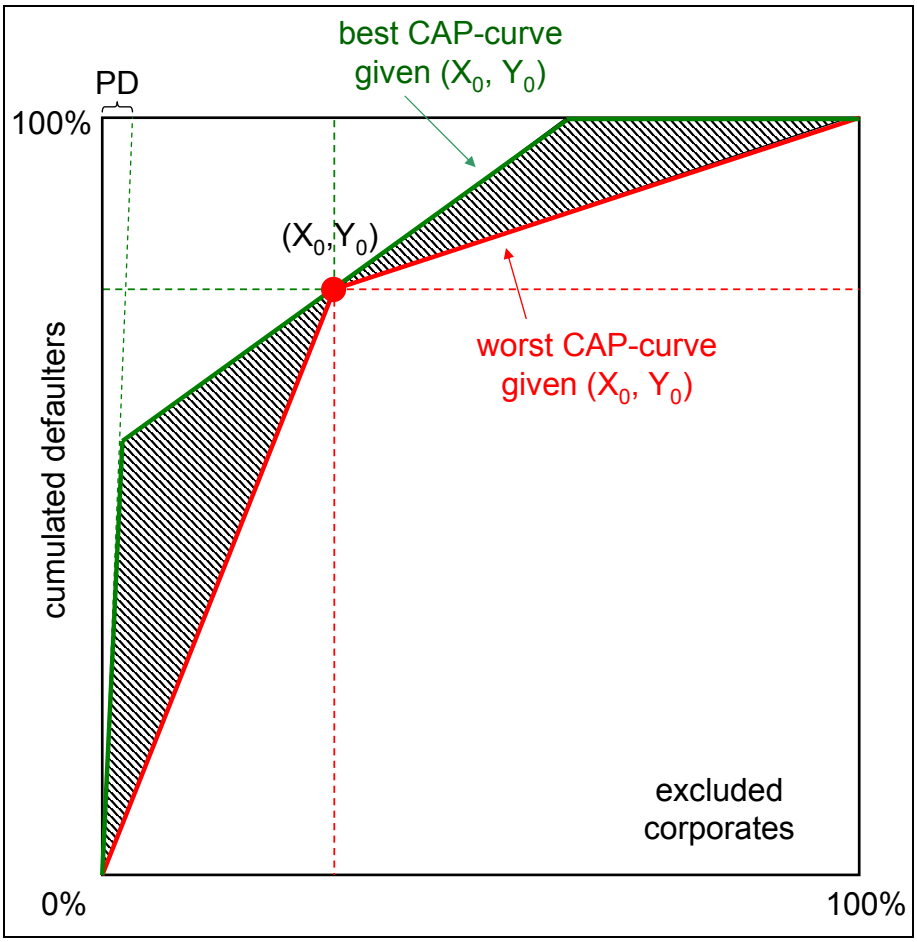

Figure 9: best and worst CAP-curve given (X_0, Y_0)

[93] Contrary to categorial insolvency predictions, ordinal predictions are not evaluated on the basis of one single error-I-II-combination but on the universe of all error-I-II-combinations attainable by the rating system by applying various thresholds.

[94] The only exception is the (0%; 0%) error-I-II-combination that is exclusively comprised by the "perfect" ROC-curve.

[95] Bivalent error-I-II-combinations, which prohibit direct comparisons, unless both errors of one rating model are bigger/smaller than those of another rating model, are transformed into a univalent accuracy ratio value.

It can be shown that, under plausible assumptions, the "worst possible" CAP-curve containing a given point (X_0, Y_0) in the CAP-diagram is the linear spline (0%; 0%) - (X_0; Y_0) – (100%; 100%). This CAP-curve corresponds to a CAP-curve of a rating model that can only differentiate *between* the two groups that are separated by X_0 – but without possessing any discriminative power *within* these groups – which essentially is commensurate with a categorial rating model with only one single threshold value.

The best possible CAP-curve containing (X_0, Y_0), on the other hand, is characterized as follows: starting from (0%; 0%) it proceeds along the dotted straight line (PD), which means that the corporations that are excluded first have a default rate of 100%, afterwards it linearly proceeds through (X_0; Y_0) until it intersects the 100%-hit-rate-line (see Figure 9). Slope and absolute term of the line have to be identified by an optimization approach. Subsequently the CAP-curve horizontally proceeds until it reaches (100%; 100%), i.e. the corporations finally excluded have a probability of default of 0%.

Formulas for obtaining best and worst possible AR values, given a combination of errors of types I and II (F_1 and F_2) or given a pair of CAP-coordinates (X_0, Y_0) are reported in Table A. Besides upper and lower limits for the accuracy ratio and AUC, four heuristic estimators for the accuracy ratio are given AR_{mv} (mean value of upper and lower limits of accuracy ratios attainable), AR_α, AR_β, and $AR_{\alpha\&\beta}$ (average of AR_α and AR_β), see also Appendix I and IV for derivation and empirical accomplishments of the various estimators.

For error-I-II-combinations with equally sized errors of type I and II, all four estimators give good predictions of the true accuracy ratio, for asymmetrical error values best results are achieved with $AR_{\alpha\&\beta}$.

	CAP-coordinates	errors of types I and II
AR_{min}	$\frac{Y_0 - X_0}{1-PD}$	$1-(F_1+F_2)$
AR_{max}	$\frac{1-PD-4X_0+4\cdot X_0\cdot Y_0+4\cdot PD\cdot Y_0-4PD\cdot Y_0^2}{1-PD}$	$1-4\cdot F_1\cdot F_2$
AR_{mv}	$\frac{AR_{max}+AR_{min}}{2}$	$\frac{AR_{max}+AR_{min}}{2}$
AR_α	$\frac{\log(X_0-Y_0\cdot PD)-\log Y_0-\log(1-PD)}{\log(X_0-Y_0\cdot PD)+\log Y_0-\log(1-PD)}$	$\frac{\log F_2-\log(1-F_1)}{\log F_2+\log(1-F_1)}$
AR_β	$\frac{\log(1-Y_0)-\log(1-PD-X_0+Y_0\cdot PD)+\log(1-PD)}{\log(1-Y_0)+\log(1-PD-X_0+Y_0\cdot PD)-\log(1-PD)}$	$\frac{\log F_1-\log(1-F_2)}{\log F_1+\log(1-F_2)}$
$AR_{\alpha\&\beta}$	$\frac{AR_\alpha+AR_\beta}{2}$	$\frac{AR_\alpha+AR_\beta}{2}$

Table A: Formulas for obtaining exact upper and lower limits and heuristic estimators AR_{mv}, AR_α and AR_β for accuracy ratios from CAP-coordinates/ combinations of errors of type I and II

Example: In a study conducted by ALTMAN (1968, p. 599; see also chapter 3.4) a rating model was presented, which – given a sample of each 33 corporations who stayed solvent / became insolvent within one year – could correctly classify 31 of the insolvent and 32 of the solvent corporations. The type I error rate therefore was 6.1% and the type II error rate 3.0%. Exact upper and lower limits for the accuracy ratio are thus 1-(6.1%+3.0%) = 90.9% and 1-4*6.1%*3.0%=99.3%. Estimated values for the accuracy ratio are AR_{mv}=95%, AR_α = 96.5%, AR_β=97.8%, and $AR_{\alpha\&\beta}$=97.2%.

As opposed to most other studies, ALTMAN (1968, p. 603) additionally also stated all individual score values for the (very few) corporations, partitioned into solvent and insolvent one. Therefore, *here* it was possible to reproduce the complete ROC-curve and calculate the respective accuracy ratio (98.7%) directly.

For comparison: the ratings of S&P and MOODY'S have historically achieved average one-year-accuracy-ratios of "only" about 85% (see also chapter 3.5).[96] However, the results of ALTMAN's (1968) study are based on a very small sample – and thus are statistically not very reliable (see chapter 3.1). Additionally, the performance of the model was measured on the same sample on that the model was calibrated before ("in-sample"). In later studies, in particular in those studies that were carried out by other authors, ALTMAN's model performed much worse (see chapter 3.4).

[96] The achieved accuracy ratio values of about 85% refer to the universe of all corporations that are rated by S&P or MOODY'S. The accuracy of agency ratings for US industrial corporations, see the sample definition of ALTMAN'S study in chapter 3.4, is considerably worse, see on this chapter 3.5.

2.3.4 Accuracy losses caused by discretizing continuous rating scales

Rating agencies and banks transform continuous (or "quasi-continuous"[97]) scores to a finite number of ordinal rating classes, for instance for investigating historical default and/or migration characteristics of their ratings.[98,99,100] Seen from the point of view of an external user, who only knows about the rating class of a company but not about its continuous rating score, the rating looses part of its information value. While all information concerning cross-class relative probabilities of default are maintained, all information concerning intra-class relative probabilities of defaults are sacrificed. [101,102].

No studies could be found that attempted to *quantify* the magnitude of such information losses. Two simulation studies, however, based on empirical default data showed, that under certain conditions, using ratings with continuous rather than discrete scales yields no *statistically significant* improvements (expressed in different measures).[103,104] These results are in so

---

97 The rating model developed by MOODY'S-KMV, for instance, is using up to 1,000 different rating classes (intervals of probabilities of default), see KEENAN, SOBEHART (1999, p.12)); or the CREDITREFORM-Bonitätsindex (a German business score) is stated on a scale that allows up to 500 different values, see SCHWAIGER (2002, p.16). The SMB-rating model family *RiskCalc*, for which exist a variety of localized models, delivers continuous outputs (probabilities of default), see for instance KOCAGIL ET AL. (2003, p. 30).

98 At the end of the 1990ies only four out of the 46 biggest US financial institutions had rating systems in operation with more than eight rating classes for performing loans; four institutions even had only one to three such rating classes. However, having a reasonable number of rating classes available does not imply already, that bank rating systems are reasonably refined: at one third of the banks examined, the largest rating class accounted for more than 50%, in some cases even for up to 80%, of all rated corporate customers. At only 15% of all banks the largest rating class accounted for less than 30% of all customers, see TREACY, CAREY (2000/1998, p. 902).
Basing on an examination of more than one hundred US commercial banks of different size classes, ENGLISH, NELSON (1998, p. 5) find, that about 2/3 of their new costumers were classified by banks with the respective most frequently used rating class. Bank ratings, however, were mapped on a 5ary rating scale, that was predetermined by the National bank authorities.
According to more current surveys, German commercial banks use bank internal rating systems with 8 rating grades (Dresdner Bank, 2002), 12 RG (Commerzbank, 2003), 25 RG (Volks- und Raiffeisenbanken, 2002), see FISCHER (2004, p. 165 and the literature there cited).

99 "Internal rating systems with larger numbers of grades are more costly to operate because of the extra work required to distinguish finer degrees of risk. Banks making heavy use of ratings in analytical activities are most likely to choose to bear these costs because fine distinctions are especially valuable in such activities (however, at least a moderate number of Pass grades is useful even for internal reporting purposes)." TREACY, CAREY (2000/1998, p. 902)

100 KRAHNEN, WEBER (2001, p.13): "[…] the central question for the definition of a rating system now remains, how fine a rating system should be, i.e., how many categories it should have. It could be as fine as the POD itself, being basically identical to POD, or it could map PODs into a finite number of categories. Of course, a rating system which models POD would be the most exact one. However, for quite a number of situations a less fine rating system would be sufficient and more appropriate in an organizational context. The fineness of a rating system cannot be considered independently from Backtesting […]. There is no use in defining a large number of rating categories, if a bank is not able to back-test consistently, due to lack of data."

101 MCQUOWN (1993, p. 8): "There are 19 different gradations at S&P, including the pluses and minuses. S&P's precision could be, therefore, no greater than 1 in 19. […] We suspect that the resolution of EDFs may be nearer 1 in 100. Banks, typically, use fewer than ten gradations, of which three may be non-performing."

102 MILLER (1998): "While the intensive quantification of market risk has led to measurements accurate to the basis point (and beyond), difficulties in quantifying credit risk have resulted in the practice of measuring this risk with far less precision. Indeed, financial institutions that develop their own internal measures of credit risk usually employ a '1' to '9' scale of creditworthiness for their exposures,[…] Even with the further refinement of 'notches' designated with a '+' or '-' the vast universe of credit risk is reduced to at most thirty buckets."

103 see KEENAN/ SOBEHART (1999, p. 12f.), FRERICHS, WAHRENBURG (2003, especially p. 3, 35)

far not very instructive, as if there are any difference in information value – and there *must* be some, as was shown by the above theoretical reasoning – they can be made significant at will at any desired confidence level, simply by varying sample size. Contrary to *statistical* examinations, in *simulation studies* sample scarcity is no issue at all.[105]

In Appendix IV two approaches are presented that estimate the order of magnitude of information losses that come with discretizing continuous rating scores to a discrete set of rating categories.

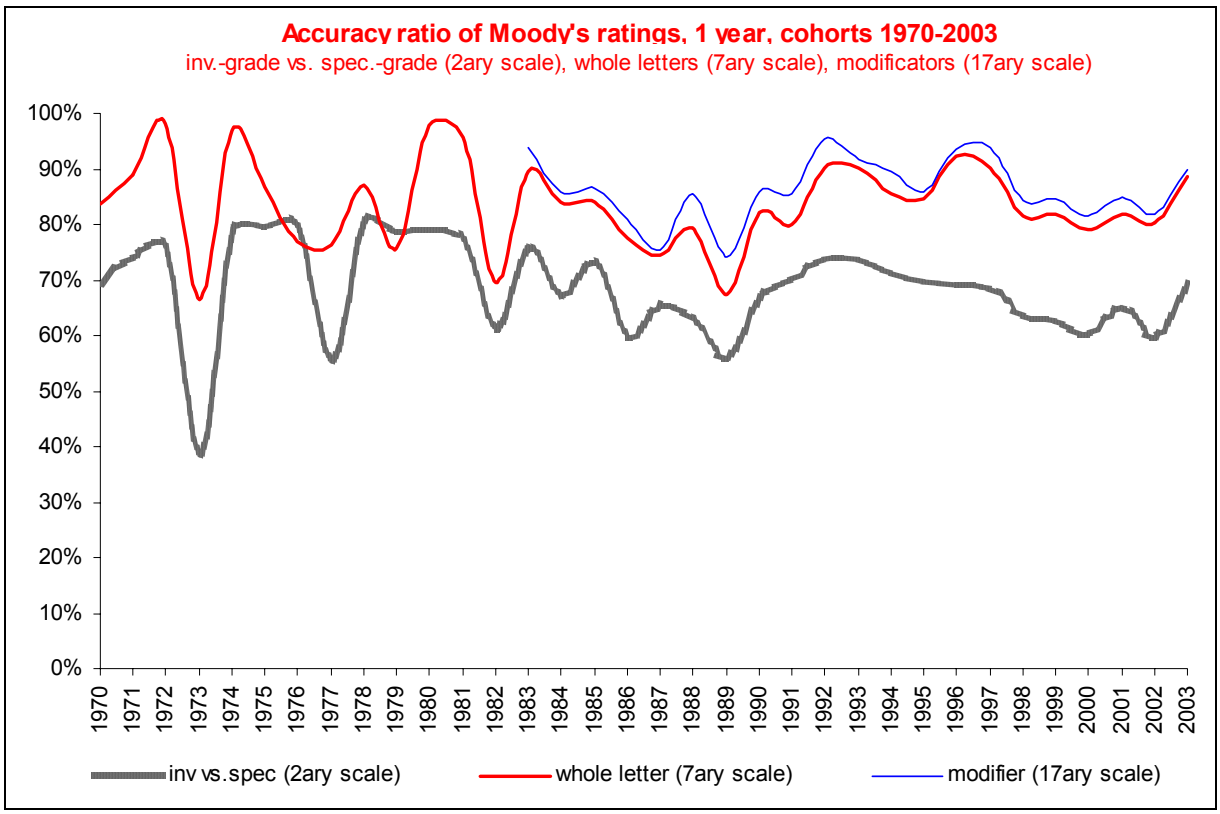

Figure 10: accuracy ratios for MOODY'S ratings, 1 year forecast horizon, 1970-2003, 2-, 7- and 17ary rating scales (source: own calculations)

To begin with, Figure 10 should convey a first notion of the empirical relevance of the issue. The figure displays time series data for the accuracy ratios of MOODY'S ratings for the 1970-2003 period for three distinct rating scales that differ in number of rating classes:[106]

- 2ary scale: *investment grade* (Aaa-Baa3) vs. *speculative grade* (Ba1-Caa/C)),
- 7ary scale: whole letter ratings (Aaa … Caa/C),
- 17ary scale: modified ratings (Aaa, Aa1, Aa2, A3, …, CCC/C).

In all years of the 1983-2003 period, the 7ary scale accuracy ratios were better than the respective values of the 2ary scale, and in all years the 17er scale values were even better than those of the 7ary scale. [107]

---

[104] A further, related study examines the relationship between rating scale fineness and regulatory equity capital requirements, see JANKOWITSCH, PICHLER, SCHWAIGER (2003).

[105] See on the inappropriateness of the concept of *statistical significance* in the context of simulation models SCHMEISER (2001, p. 43).

[106] source: own calculations based on MOODY'S cohort data (2004, p. 27-38)

[107] This does not reflect construction-conditioned features of the accuracy ratios calculated. See for instance the findings for 1976 and 1979. In both years, 2ary scale accuracy ratio exceeded 7ary scale accuracy ratios. Both years were characterized by extraordinary few defaults (1976: 2 defaults, 1979: 1 default), that coincidentally only affected the best *speculative* grade class Ba. Thus, comprehending the whole letter rating

In the 1983-2003 period the average accuracy ratio of the 2ary scale rating was 66.9% (with a standard deviation $\sigma = 5.4\%$)[108], while the 7ary scale rating achieved 83.1% ($\sigma = 6.0\%$) and the 17ary scale rating 86.2% ($\sigma = 5.7\%$).

Passing over from a 2ary to 7ary scale, thus adding only 5 rating classes, increases the accuracy ratio by 16.2 pp (percentage points) on average with a standard deviation of 3.8 pp. Passing over from a 7ary to a 17ary scale, thus adding twice as many more rating classes, only results in an average accuracy ratio increase by additional 3.2 pp ($\sigma = 1{,}7$ pp). By the procedures presented in Appendix IV it can be estimated, that by passing from a 17ary scale to a continuous scale the accuracy ratio should increase by another 0.5 pp (or stated conversely: under the prevailing circumstances using a 17ary scale instead of a continuous scale provokes information losses of about only 0.5 pp, or using a 30ary scale would provoke information losses of about only 0.1pp).

The total information losses mainly result from information losses that occur in the "worst" rating class (Caa/C). Discretizing rating scores in the domain of other rating classes, especially in the domain of investment grade rating classes, exerts comparatively little influence on the information value of the rating system.

---

classes AAA, Aa, A, and Baa into one single aggregate, *investment grade*, did not deteriorate the predictive value of the rating system because no single default occurred in any of the classes, while comprehending the classes Caa/C, B and Ba into one single class actually *improved* performance, because *stating no ordering,* is better than *stating a wrong ordering* [of default rates among speculative grades].

[108] See also CANTOR, MANN (2003, p. 19): "Much of the information content of the rating system comes from the ability to determine whether credits are investment grade or speculative grade." According to own examinations based on *whole letter rating classes*, the conventional cut-off Aaa-Baa vs. Ba-Caa/C (investment grade vs. speculative grade) does not always yield accuracy-ratio-maximizing values for Moody's cohorts from 1971-2003. For some cohorts, among them all cohorts as of 1992, best values are achieved for the dichotomous rating class division Aaa-Ba vs. B-Caa/C, i.e. by considering Ba-ratings as "revised investment grade" ratings. Based on pool data and whole letter ratings, the AR-maximizing cut-off is compatible with the conventional investment-grade vs. speculative grade dichotomy (Aaa-Baa vs. Ba-Caa/C) from 1971-1999, but from 2000 onward best results are achieved with a Aaa-Ba vs. B-Caa/C separation.
Based on *modified rating classes* AR-maximizing cut-offs for the various cohorts from 1983-2003 are positioned between Aaa-Baa2 vs. Baa3-Caa/C and Aaa-B1 vs. B2-Caa/C – the latter cut-off is valid as of 2001. According to the pool data, the AR-maximizing cut-off for modified rating classes is steadily located at Aaa-Ba2 vs. Ba3-Caa/C as of 1986.
If a two-class separation of a rating system is attempted, that is as discriminative as possible (here in terms of achieving the maximal attainable accuracy-ratio) the conventional investment grade vs. speculative grade dichotomy currently yields inferior results in all examined cases (pool vs. cohort data, whole letter ratings vs. modified ratings). Instead, at least the two "best" modified speculative grade rating classes, Ba1 and Ba2, should be assigned to the better one of the two segments ("revised investment grade"). For reasons of practicability, in particular for maintaining identical cut-offs for whole letter ratings and modified ratings, additionally also rating class Ba3 should assigned to the better one of the two segments. The "revised speculative grade" should comprise only rating classes B and Caa/C.

2.4 Accuracy measures for cardinal insolvency predictions

While ordinal insolvency predictions only order corporations according to their *relative default probabilities*, cardinal predictions *explicitly* state probabilities of default.

By ignoring their cardinal implications, probabilities of default may also be interpreted as simple ordering criteria. Therefore all measures and methods that were developed for assessing ordinal predictions may also be used for cardinal predictions:[109,110]

- *Resolution:* measures by how much *realized default rates* differ between different prognoses. Minimal resolution is given, when realized default rates are the same for all prognoses. Maximal resolution is achieved, when only default rates of 0% or 100% are occurring (for a graphical representation of *resolution* see Figure 3),
- *Discriminative power:* measures by how much prognoses of defaulters and non-defaulters differ.

Additionally in case of cardinal predictions also criteria may be tested that compulsory require the ex-ante specification of probabilities of default:

- *Calibration:* measures to what extend stated default *probabilities* are matched by realized default *rates*,
- *Unconditional bias:* states, by how much average forecasted *probability* of default differs from the actual average default rate,
- *Refinement:* measures by how much *forecasted default probabilities* differ from each other (cf. *resolution*). Minimal refinement is given then, when all predicted default probabilities are the same, maximal refinement is achieved, when only 0% and 100%-prognoses are stated (no matter how predictive they are).

As in the ordinal case, measures that simultaneously quantify some or all of the quality aspects of predictions mentioned above are labeled *accuracy measures*. Measures that quantify the accuracy of a rating system in relation to some reference measures are labeled *measures of relative accuracy* or *skill measures*.[111]

Measures that quantify only single of the aspects mentioned above, in particular *calibration,* will not be discussed in the following. Representatives of these measures are for instance:[112]

- *Grouped Brier score*[113,114]:

F 24) $$\text{Brier}_{\text{group}} = \frac{1}{g}\sum_{i=1}^{g}\left(\text{PD}_{i,\text{fore}} - \text{PD}_{i.\text{real}}\right)^2$$

with $\text{PD}_{i,\text{fore/real}}$ … forecasted/real default rate for rating class i,

---

109 See MURPHY, WINKLER (1992, p. 440) for examples of formal definitions of the aspects *resolution, discrimination, calibration, refinement, unconditional bias, accuracy* and *skill*.

110 Note that the ordinal measures Accuracy Ratio and Area under Curve that were presented in chapter 2.3.2 are being influenced by *resolution* and *discriminative power*, too.

111 see MURPHY, WINKLER (1992, p. 440)

112 For further test, such as binomial tests, χ^2-tests, or normal tests, that exclusively check for correct calibrations of the forecasted default probabilities see BASEL COMMITTEE (2005, p. 47ff.).

113 See on this FRERICHS, WAHRENBURG (2003, p. 16, own notation). In a simulation study the authors find, that the grouped Brier score is no suitable validation measure for rating systems, because it fails to reliably identify rating systems that are known to be "inferior" because of their designs.

114 Note, despite certain formal similarities, the *grouped Brier scores* fundamentally differs from the *Brier score* that is presented below. The *grouped Brier score* is only affected *calibration* issues and is completely unreceptive with respect to the discriminative power of forecasts.

g … number of rating classes [115,116],

- *"Rommelfanger-index"* [117]

F 25) $$F = \sum_{i=1}^{g} \Delta_i \cdot \frac{a_i + \hat{a}_i}{2} \cdot f_i$$

with $\Delta_i = \max(0; PD_{i,real} - PD_{i,fore})$ for i=1..g-1, respectively $\Delta_i = \max(0; PD_{i,fore} - PD_{i,real})$

for i=g, a_i, $\hat{a}_i$… relative amounts of all loans in the validation/ estimation sample,

f_i … "appropriate weighting factors"[118,119]

The accuracy measures for cardinal insolvency predictions presented in the following are based on a unitary rationale: they compare the individually forecasted probabilities of defaults $PD_{i,fore}$ with the realizations of the individual default events Θ_i (with Θ_i=1/ Θ_i =0 if corporate i defaults/ does not default) and "sconce" the differences that occur with different "punishment functions".

Contrary to categorial insolvency predictions, that only use the extreme predictions "default" vs. "non-default" (which may be either right or wrong), it may be less intuitive in case of default predictions stated in terms of *probabilities* of default, why deviations from individual default probabilities and default realizations should be "sconced" at all.

After all, predictions must adopt values between 0% and 100%, while default realizations can only be of the extremes 100% (default) or 0% (non-default). Even in case forecasted probabilities are "right", meaning that they are *correctly calibrated,* if fore instance 1%/ 5%/ 20% of the corporations are defaulting for which a default probability of 1%/ 5%/ 20% was forecasted, the predictions are being "punished", that is they don't get the best possible valuation. However, what is actually being "punished" in these situations is not a failure in *calibration*, but a failure in *discrimination*: a rating that had predicted a probability of default in 2003 for each and every German corporate of 1.35%, had been perfectly calibrated, but had been punished for its non-selective forecasts. The best possibly accuracy value would be achieved by a rating that stated a PD of 100% for those 1.35% of all corporations who really defaulted subsequently and a PD of 0% for the remaining corporations who did not default subsequently.[120]

---

[115] When determining the score value, it probably would be more suggestive to use weighting factors proportional to the number of corporations [or loan amounts] per rating class than to use uniform weighting factors.

$$Brier^{*}_{grouped} = \sum_{i=1}^{g} a_i (PD_{i,fore} - PD_{i,real})^2$$

with a_i … share of the corporations with rating i in all corporations

[116] As a result of a simulation study by FRERICHS, WAHRENBURG (2003, p.17) the authors find, that the grouped Brier score, as opposed to all other measures examined, was not suited to discriminate between "good" and "bad" rating systems when used as sole measure.

[117] DVFA (2004, p. 600, own notation)

[118] See DVFA (2004, p. 599, translation). There are no assertions made, what constitutes the "appropriateness" of weighting factors.

[119] Next to its orientation to measuring only *calibration* aspects*,* the index may also be criticized for its dependency on irrelevant variables (like the structure of the estimation sample). It may also be criticized for setting incentives for systematically misreporting default probabilities: In the rating classes 1..g-1 only too high and in rating class g only too low probabilities of defaults are being "punished", therefore systematically overstating (rating class 1..g-1) or understating (rating class g) default estimations is rewarded.

[120] see for instance KRÄMER (2003, p. 396f.)

Following two accuracy measures for cardinal insolvency predictions that differ regarding their "sconce" or "punishment functions" are considered:

- **logarithmic "punishment function":** CIE (conditional information entropy)[121]:
  - **F 26)** $CIE = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\left|PD_{i,fore} + \theta_i - 1\right|\right)$ with n … number of corporations,
    Note: CIE is only not defined in those cases, where a default occurs, although it was ruled out with certainty (Θ_i=1 *and* $PD_{i,fore}$=0) or in those cases where no default occurs, although it was forecasted with certainty (Θ_i=0 *and* $PD_{i,fore}$=1).
  - **F 27)** $E(CIE) = -\frac{1}{n}\sum_{i=1}^{n} PD_{i,real} \cdot \log PD_{i,fore} + (1 - PD_{i,real}) \cdot \log(1 - PD_{i,fore})$
  - for g discrete rating classes:
    F 28) $E(CIE) = -\sum_{i=1}^{g} a_i \left(PD_{i,real} \cdot \log PD_{i,fore} + (1 - PD_{i,real}) \cdot \log(1 - PD_{i,fore})\right)$
    with a_i … share of all corporations of rating class i in all corporations
  - **F 29)** $Skill_{CIE} = CIER = \frac{CIE_{PD} - CIE}{CIE_{PD}}$
    F 30) $CIE_{PD} = -(PD \cdot \log PD + (1 - PD) \cdot \log(1 - PD))$

    with CIER … conditional information entropy *ratio* [122]
    with CIE_{PD}…CIE-value of a "naïve" reference rating, that always forecasts the average probability of default PD[123,124],
    note: the term $CIE_{PD} - CIE$ is also referred to as KULLBACK-LEIBLER-distance[125] or Wealth-Growth-Rate-Pickup[126]

---

[121] *Entropy* is a concept that is borrowed from thermodynamics measuring the degree of a system's disorderliness. In the context of insolvency prediction, entropy is intended to quantify the degree of uncertainty that is associated with the distribution of probabilities of default that is asserted by a specific rating model for a given sample of corporations, cf. SOBEHART, KEENAN, STEIN (2000, p. 14). See SHANNON (2001/1948, p. 11f.) for an axiomatic foundation for using logarithmic punishment functions. The last of the three axioms, however, is not meaningfully applicable in case of corporate defaults which are usually modeled as variables that can adopt only two possible values ("default" vs. "non-default"). See also MATHESON, WINKLER (1976), KEENAN, SOBEHART (1999, p.9), and BASEL COMMITTEE (2005, p.44) for formula F 27 (own notation). See KRÄMER, GÜTTLER (2003, p. 12) for formula F 26.

[122] see KEENAN, SOBEHART (1999, p. 10)

[123] CIER … conditional information entropy ratio, SOBEHART, KEENAN, STEIN (2000, p. 14): "The CIER compares the amount of 'uncertainty' regarding default in the case where we have no model (a state of more uncertainty about the possible outcomes) to the amount of 'uncertainty' left over after we have introduced a model (presumably, a state of less ignorance),

[124] In the light of highly volatile default rates, at least in case of corporate bond markets, more than just "naivety" is required for correctly forecasting average default rates. See on this topic for instance KEENAN (1999), S&P (2004b, p. 3).

[125] see Basel COMMITTEE (2005, p. 30)

[126] see CANGEMI, SERVIGNY, FRIEDMAN (2003, p. 40)

- **squared "punishment function":** Brier score: [127]
  - **F 31)** $BS = \frac{1}{n}\sum_{i=1}^{n}\left(PD_{i,fore} - \theta_i\right)^2$ [128]
  - **F 32)** $E(BS) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real}\cdot\left(1-PD_{i,fore}\right)^2 + \left(1-PD_{i,real}\right)\cdot PD_{i,fore}^2$
  - for g discrete rating classes:

    F 33) $E(BS) = \sum_{i=1}^{g} a_i \cdot \left(PD_{i,real}\cdot\left(1-PD_{i,fore}\right)^2 + \left(1-PD_{i,real}\right)\cdot PD_{i,fore}^2\right)$
  - **F 34)** $Skill_{BS} = \frac{BS_{naive} - BS}{BS_{naive}}$ [129] with

    F 35) $BS_{naive} = PD\cdot\left(1-PD\right)^2 + \left(1-PD\right)\cdot PD^2 = PD\cdot\left(1-PD\right)$

Both "punishment functions" presented are *arbitrary* in the sense, that there is no direct link to the utility of the user(s) of the predictions (see also chapter 1). However, both measures behave "plausibly", so that one can at least assume a tight correlation with the users' utility: both scores "reward" *correctly calibrated*[130,131] and *discriminative* [132] forecasts. By appropriately transforming the resulting scores (for one example see Figure 11), relationships to other quality relevant aspects of cardinal insolvency predictions - like *resolution*, *refinement*, and *bias* - can be shown as well.[133]

---

[127] see BRIER (1950, p. 1), MURPHY, WINKLER (1992, p. 439, formula 7), KRÄMER, GÜTTLER (2003, p. 11), FRERICHS, WAHRENBURG (2003, p.14), OeNB (2004a, p. 123ff.), GRUNERT, NORDEN, WEBER (2005, p.517)

[128] In the notation that is used in the context of regression analyses, the Brierscore equals the "sum of the squared residuals" (RSS) divided by n, with RSS = $\Sigma\ Y_i^* - Y_i$, with Y_i^* … forecasted value of the variable to be explained and Y_i … actual parameter value of the variable to be explained, see for instance GUJARATI (1999/1992, p. 170ff.).

[129] In the notation that is used in the context of regression analyses, BS_{naiv} equals the sum of the total variation of the variable to be explained (TSS) divided by n. Therefore it holds, that $Skill_{BS}$ = (TSS – RSS)/ TSS, and thus $Skill_{BS} = r^2$, with r^2… sample coefficient of determination and r^2 = ESS/TSS with ESS=TSS-RSS, see for instance GUJARATI (1999/1992, p. 170ff.).

[130] This is a non-trivial statement. If for instance *absolute values* of differences are chosen as punishment function $|PD_{i,fore} - \theta_i|$, so stating $PD_{i,fore}$ = 0% for $E(PD_{i,real})$<50% and $PD_{i,fore}$ = 100% for $E(PD_{i,real})$<50% leads to lower expected punishments than to state the true expected probabilities $PD_{i,prog} = E(PD_{i,real})$, see Appendix II. See also Appendix II for a proof of the incentive compatibilities of the Brier score and the entropy score.

[131] Already BRIER (1950, p.2) mentioned *incentive compatibility* as one advantage of his score: "[the forecaster] is encouraged to state unbiased estimates of the probability of each event when he cannot forecast perfectly."

[132] Both measures achieve their best possible values only, when a rating systems gives probabilities of default of either 0% or 100% *and* is always right with these predictions.

[133] cf. MURPHY, WINKLER (1992)

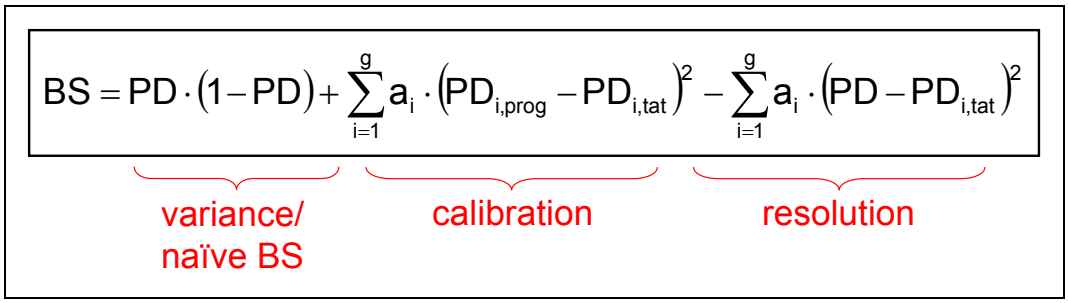

Figure 11: Brier score decomposition into the components variance – calibration - resolution[134]

One critical attribute of the Brier score and other cardinal measures that already becomes transparent in Figure 11, is their dependence from the average sample (or basic population) default rate. The bigger the environments variability (PD*(1-PD)), the *bigger (=worse)* the Brier score becomes (for a thorough analysis see Appendix III).

For correcting for this unwanted environment dependence of cardinal forecasting accuracy measures, it was suggested to use *skill measures*, which consider accuracy measures in relation to the accuracy of naïve forecasts in the same environment.[135,136] This dependence is *unwanted*, because it impairs the inter-sample performance comparability of accuracy measures if samples differ in terms of average default rates.[137]

However, empirically and theoretically (based on certain additional assumptions, see Appendix II), it can be shown, that skill scores, at least those for the Brier score and CIE, are environmental dependent, too. Paradoxically, while Brier score and CIE are signaling deteriorations of accuracy with increasing environmental variability, the respective skill scores are signaling improvements of (relative) accuracy.[138,139] The accuracy measures for ordinal insolvency predictions presented in chapter 2.3 do not exhibit these disadvantages.[140]

---

134 see MURPHY, WINKLER (1992, p. 439, formula 10, own notation)

135 WINKLER (1994, p. 1397): "The development of so called 'skill-scores' has been motivated by the desire to produce average scores that reflect the relative ability of forecaster rather than some combination of the forecaster's ability to and the situation's difficulty. These skill scores attempt to neutralize the contribution of the situation by comparing a forecaster's average score to the average score that an unsophisticated forecasting scheme would have obtained for the same set of forecasting situations."

136 see also KRÄMER (2003, p. 406)

137 Differences in default rates of different samples may for instance be due to differences in length of periods covered (cumulated default rates rise approximately linear in time), they may be due to covering different periods of the business cycle or due to covering groups of firms that inherently differ in terms of default risk, or they may just be the result of composing samples with unrepresentatively high share of defaulters, which is common use in insolvency prediction studies with comparatively small data sets.
18 from 31 insolvency studies listed in FALKENSTEIN, BORAL, CARTY (2003/2000, p. 14) use data sets of less than 100 corporations, 9 use 100-1,000 and only 4 more than 1,000 corporations. In 14 of the 18 (=78%) studies with less than 100 corporations, exactly 50% of the corporations covered were defaulters, the same holds for only 5 of the 9 (=56%) studies with 100-1,000 corporations and for none of the four studies with more than 1,000 corporations. In these four cases the shares of defaulters in all corporations were 5.0%, 6.6%, 9.7%, and 14.1%.

138 In a study of precipitation forecast accuracies of 20 weather stations, that were situated in regions with vastly different precipitation frequencies, WINKLER (1994, p. 1401f) found considerably high correlations between the Brier scores of the various weather stations' forecasts and their average precipitation frequencies (r=+0.87) – which was to be expected (speaking pictorially: "In a desert it is easy to correctly forecast a rainless day.") However, there was also found a positive correlation of Brier scores and precipitation frequencies of nearly the same magnitude (r ca. +0.80) (speaking pictorially: "In a desert it is difficult to forecast a high share of the few raining days.") – which is obviously counter the intention of skill scores, which aim at neutralizing (and not reverting) the impact of environmental differences to the accuracy measures.
Note: higher Brier scores correspond with *lower quality* forecasts, while higher skill scores correspond with *higher quality* forecasts.

Sometimes, the accuracy measures mentioned above are used for assessing the quality of ordinal or cardinal forecasts under the *fiction* of correct calibration, i.e. by ignoring the true values for $PD_{i,fore}$ and by setting ex-post $PD_{i,fore}$ to the realized default rates for every i. [141] In that case, formulas F 28 and F 33 can be simplified as following:

F 36) $$CIE_{cal} = -\sum_{i=1}^{g} a_i \left(PD_i \cdot \log PD_i + (1 - PD_i) \cdot \log(1 - PD_i)\right)$$

F 37) $$BS_{cal} = \sum_{i=1}^{g} a_i \cdot \left(PD_i \cdot (1 - PD_i)^2 + (1 - PD_i) \cdot PD_i^2\right)$$

F 38) $$BS_{cal} = \sum_{i=1}^{n} a_i \cdot PD_i \cdot (1 - PD_i)$$

The accuracy measures thus calculated are non-sensitive with respect to wrong calibrations (or *missing* calibrations – as in case of ordinal solvency predictions) – the middle term in Figure 11 disappears – and are measuring only a combination of environmental v*ariability* and *resolution*. These measures can also be interpreted as upper limits of accuracy the respective rating systems could at best achieve, i.e. in case of *perfect calibration*, given their resolutions and given the environmental variability.

Due to their dependence on the average sample default rates, these measures are still unsuited for portfolio spanning comparisons of methods. For an assessment of different methods based on the same portfolio they are not more informative than the measures for ordinal insolvency predictions (like AUC_{ROC} or AR) that were presented in the previous chapters.

However, in particular in case of intersecting ROC-/CAP-curves they could be considered as *additional* indicators for the relative quality of rating models, as in the end none of the diverse accuracy measures for insolvency predictions is directly linked to the users' utilities. The more indicators suggest the predominance of one particular rating model in a direct comparison with another rating model, the more the decision maker's certitude is strengthened that he chooses the right method (if he has to decide between two competing models without having the opportunity to combine their informational value). If, on the other hand, indicators are giving widely conflicting signals, it can be assumed that the decision maker is making *no material mistakes* if he opts for one of the models by chance or if he applies some subordinate criteria (for instance development costs or the degree transparency of the model) to enforce a decision.

---

[139] See Appendix III for a model based derivation of the results. See also CANTOR, MANN (2003, p. 12) and DVFA (2004, p. 599) concerning the (nearly perfect) environmental independence of the accuracy ratio.

[140] see Appendix III

[141] cf. KRÄMER, GÜTTLER (2003, p. 12)

3 Empirical findings concerning accuracy of insolvency predictions

3.1 Limitations on usefulness of empirical comparisons - Development of benchmarks for prediction accuracy measures

On purely theoretical grounds it can be shown, that applying one and the same rating model on different samples of corporations may yield substantially and systematically varying measures of predictive quality.[142] Therefore the quality of a rating model cannot be unequivocally characterized by a single value - and the relative quality of different rating models cannot be unequivocally inferred from comparisons when they are based on samples of corporations that differ in some relevant aspects which affect accuracy measures. Based on empirical and theoretical studies there have been identified several such variables – unrelated to the *prediction quality* of a rating model. As determinants for such sample specific "structural differences" have to be mentioned: [143]

- *Pool vs. cohort data:* When assessing the accuracy of insolvency predictions, depending on whether only data sets of corporations of the same period, e.g. 1994, were used (cohort data) or datasets of corporations from different periods (pools), e.g. 1970-2003, differences in measured accuracy have to be expected. Based on theoretical considerations, better values should be achieved with cohort data sets, as a rating system here "only" has to deliver a consistent (discriminative) ordering of corporations at one single point in time in order to achieve a good valuation. Pooled data based rating systems, on the other hand, additionally must provide orderings that have to be consistent throughout time, too.[144] These differences are accentuated, if the rating system is designed to give relatively *stable* ratings (point-in-time- (current-condition-) (PIT) vs. through-the-cycle-approach (TTC) [145,146]). Empirically, though, problems related to the usage of pool vs. cohort data seem to be of surprisingly minor relevance.[147] Besides that, for circumventing bottlenecks in the

---

142 see for instance HAMERLE, RAUHMEIER, RÖSCH (2003)

143 OENB (2004, p. 137f.) makes following demands on the quality of validation benchmark data samples: comparable data quality, uniform definitions of input variables (in particular in case of qualitative variables), consistency of target variables (definition of default), structural consistency (with respect to company sizes, regional distribution, industry classification, and legal forms).

144 Example: In order for a 2004 cohort based rating system to achieve a good valuation, corporations that were BBB-rated at 01/01/2004 should exhibit a much smaller default rate in 2004 than corporations that were BB-rates at 01/01/2004. The same holds for a 1970-2004 cohort based rating system, but additionally the one-year default rates of BBB-rated corporations with date of rating 01/01/2004 should be much smaller than the default rates of corporations that were BB-rated at 01/01/1970, or at 01/01/1980, etc. Essentially this implies, that pooled data based ratings must provide rating class specific default rates that are stable through time.

145 Based on a simulation model LÖFFLER (2004a, p. 709) finds dramatic performance differentials between TTC- and PIT-ratings.

146 For an extended review of both approaches see e.g. BASEL COMMITTEE (2005, p.10 ff.).

147 MOODY'S-ratings for instance achieved one- and five–year accuracy ratios based on pooled data for the 1983-2002 period of 82.6% and 71.0%, while the same period's average (issuer weighted) cohort accuracy ratios were only marginally better with values of 83.5% and 72.9%, see CANTOR, MANN (2003, p.19).
If MOODY'S data for 1970-1982 is included in the analysis, the sequence of performance between pool and cohort data does even reverse (!): the 1970-2003 average cohort accuracy ratio (83.4%) is marginally lower than the 1970-2003 pooled accuracy ratio (83.5%). Source: own analysis. Note: contrary to CANTOR, MANN (2003), the own analysis was based on 7-ary ratings. Further, the conventional definition of *accuracy ratio* was applied (and not Moody's definition), see chapter 2.3.2.
The sign of difference between the average cohort and pooled AR is in particular sensitive with regard to the inclusion or exclusion of the data of the 1973, 1976, 1977 and 1979 cohorts. All four cohorts were character-

provision of data for defaulted companies, practically all default studies are based on pooled data.[148]

- *Positioning of the observation data period in the default cycle*: It was noted, that in general the quality of insolvency predictions is (unfortunately) especially low, when it is especially important: namely in times of above average default rates (see also Figure 8). One possible explanation is, that from a portfolio of corporations that is made of "white, grey and black sheep"[149]only "black sheep" tended to become insolvent in good times, while in economically more adverse times both "black" and "grey sheep" tended to become insolvent, which reduces the measured discriminative power of insolvency predictions. [150,151] Because most default studies rest upon pooled data of relatively long periods (see above), biases resulting from this effect should be rather small.

- *Preselection of portfolios (in particular bank portfolios):* As in bank portfolios both corporations which are either especially *vulnerable* to defaults ("black sheep") and corporations that are exceptionally *stable* ("white sheep") are underrepresented, discriminative power of rating systems is negatively affected, as the rating systems essentially have to differentiate between "bright and dark grey sheep".[152]
  Especially vulnerable corporations are underrepresented, because banks do not accept *potential customers* from whom they expect unusually high insolvency hazards.[153] Additionally, *existing customers,* whose credit-worthiness has declined substantially ever since the initiation of the costumer relationship, are often bank internally transferred to "problem portfolios". *If* the assessment of bank data based rating models is executed by excluding these "problematical costumers" a further decrease in *measured* accuracy has to be expected.[154,155,156]

---

ized both by extremely low default rates and extremely low accuracy ratios, whereof average cohort performance was more negatively affected than pooled data performance.

148 In a meta-study conducted by AZIZ, DAR (2004, p. 35ff.) only one out of 82 insolvency prognosis studies was based on data of one single cohort. 69 studies (84%) were based on pooled data of at least five subsequent years, and 35 (42%) of the studies were actually based on pooled data of at least 10 subsequent years (!).

149 "Black sheep" is a German phrase which can best be translated as "rotten apple", although by loosing the intuitive color-related connotation, which is important in the context above. Accordingly, a "white sheep" would be a "good apple".

150 "Recessions both increase and broaden the base of defaulters, lowering the measured power of default models in these periods. […] In [good] times, the really bad firms (C) default, while moderate firms (B) and excellent firms (A) don't default. In bad times, however, section C still shows more defaults, but now section B is a gray area; a new class of firms that were previously almost never defaulting are now defaulting at low, but significant rates. This adds gray to a situation that was previously black and white, which the lower power of the model reflects." FALKENSTEIN, BORAL, KOCAGIL (2000, p. 22f)
Based on MOODY'S (2004) bond rating performance data, a correlation coefficient of -0,494 / -0,437 results for the interrelationship of accuracy ratios and average 1- and 5-year default rates for the 1983-2003 period's cohorts. Source: own analysis.

151 See on this also S&P (2004b, p. 14): „Trends in the one-year Gini ratio emerge during periods of both extremes in default pressure, […] In periods of high defaults, there tends to be greater variation with respect to how the defaults are distributed across the ratings spectrum, which reduces the Gini.", see also BALCAEN, OOGHE (2004, p. 31 and the literature there cited).

152 Besides *validation*, other phases of the development of rating models are affected from the pre-selection of portfolios, too. If the development sample (=training/ learning sample) of a statistical rating model only included data from non-rejected applicants, incorrectly calibrated probabilities of default have to be expected, when the model is applied to rejected applicants, see FEELDERS (2000, p. 1ff.).

153 For quantitative effects of censored customer acquisitions to predictive accuracy measures see KRAFT, KROISANDT, MÜLLER (2004, p. 7f.) and the literature there cited.

154 For the effects of excluding "problematic costumers" from estimation and validation samples see LEHMANN (2003, p.8f).

An underrepresentation of corporations with exceptionally *good* financial standings in bank loan portfolios, on the other hand, has to expected, as those corporations tend to be corporations, who have to take out no (or less than average many) bank loans, because of their above average endowments with equity capital, or because of their privileged access to alternative financial sources, like bond issuance (in particular in case of large corporations). When excluding 10% of the corporations with the *worst* ratings, accuracy ratio losses – measured at the remaining portfolio – of 10% to more than 50% (!) have to be expected. When excluding 10% of the corporations with the *best* ratings comparatively small accuracy losses of 2.5% or less have to be expected.[157]

- *Average (sample) default rates:* One major disadvantage of *cardinal* measures for insolvency predictions, see chapter 2.4, is their dependence on average (sample) default rates. However, the measures for ordinal insolvency predictions, see chapter 2.3, are *not* affected therefrom, see also Appendix III.
- *Industry classification:* Especially *discriminative* insolvency predictions are given by rating agencies for *financial corporations*. While MOODY's ratings achieve extraordinary good one-year accuracy ratios of 92.3% for financial corporations, the performance for non-financial corporations (which account for about 60% of the Moody's rated costumers, see Table H in chapter 3.5) is considerably worse with an accuracy value of only 80.5%.[158] It is interesting to note, that just financial corporations are excluded from most default studies (see also chapters 3.3 to 3.5). A more detailed industry classification within the group of *non-financial corporations* would *probably* not reveal dramatic differences with respect to prognosis accuracy.[159,160,161]
- *Size of corporations:* MOODY's ratings achieve on average one-year accuracy ratios of 84.4% for "big" corporations as compared to only 74.0% for "small" corporations.[162] Even more pronounced size-dependent differences in prognosis accuracy were noted in various empirical studies (see chapter 3.3).[163] As causations for the reduced predictive ac-

---

155 FALKENSTEIN, BORAL, CARTY (2003/2000, p. 23): "Many institutions transfer credits to special asset groups once a credit is placed in any of the regulator criticized asset categories. Once there, many institutions do not continue to spread the financial statements associated with these high risk borrowers, or the borrowers no longer submit them."

156 FALKENSTEIN, BORAL, KOCAGIL (2000, p. 7): "Much of the dearth in default data is due to the vagaries of data storage within financial institutions. Defaulting companies are often purged from the system after their troubles begin, which creates a sample bias in that the default probability implicit in current bank databases is invariably low, even for a non-recessionary period."

157 See also Appendix V for the respective formal and empirical analyses.

158 See MOODY'S (2004c, p.2). The AR-values stated refer to "historical averages" (probably 1983-2003).

159 In a sample of 30,000 corporations BLOCHWITZ, LIEBIG, NYBERG (2000, p. 28ff., see also chapter 3.3) found only marginal differences in accuracy ratios for their insolvency predictions based on linear discriminant analyses that were individually calibrated and validated for three industry aggregates (trade, manufacturing, others). The respective AR-values were 55.8%, 60.0% and 54.4%.

160 In a sample of 50,000 corporations DWYER, KOCAGIL, STEIN (2004, p. 17f.) could increase the discriminative power of their rating system only marginally (but statistically significant) from 54.4% to 55.1% by including industry specific ratios for nine different industry aggregates (However, the discriminative power of the ratings systems' performance within the industry aggregates was not stated).

161 With a rating model that was developed and validated based on accounting data of 19.500 Austrian firms, KOCAGIL ET AL. (2003, p. 20) find following, only slightly varying industry specific accuracy ratios: AR construction = 58.6%, AR industrials = 59.1%, AR services = 54.0%, AR trade = 57.1%.

162 See MOODY'S (2004c, p.2). The AR-values stated, refer to "historical averages" (probably 1983-2003). No explicit definitions for "big" or "small" were given.

163 See on this also KOCAGIL ET AL. (2003, p. 20). The same model achieved AR-values of 51.1% for corporations with revenues between 0.5-5m EUR, 59.3% for corporations with 5-25m EUR and 64.6% for corporations with revenues that exceeded 25m EUR.

curacy for smaller companies (who are less often quoted on stock exchanges and who have less often traded debt) were mentioned: worse *quality* of financial statements [in terms of *correctness*, not in terms of *credit-worthiness*] [164,165,166] (see also the next paragraph), non-availability of capital market data [167] but also lower quality of *default information*[168,169] which are essential ingredients for the calibration and validation of rating models.[170,171]

- *Data quality:* "Bad" quality data, i.e. due to missing or wrong information about the financial conditions of the corporations examined or their default states, negatively impact the accuracy of insolvency prediction models, that are developed with these data and adulterates validation results.[172,173]

---

164 STEIN ET AL (2003, p. 5): "An important result of these structural differences, are differences in the availability of good quality data on which to develop and test default models. Because most middle-market firms are not issuers of public securities, they are not required to report details of their financial statements on a regular basis as public firms do. These firms typically report such information to their lenders but it is not generally available to the marketplace in most countries. Furthermore, the quality of these reported financial statements both with respect to data accuracy and accounting rigor is typically inferior to that in the public markets."

165 FALKENSTEIN, BORAL, CARTY (2003/2000, p. 77): "Accounting statements are less noisy for rated companies, and this is reflected in the far greater number of audited statements for public companies as opposed to private companies. Adding noise to the input variables clearly weakens the power of any model to predict from them."

166 BOHN, AVORA, KORABLEV (2005, p. 14), however, qualify the conclusions that can be drawn thereof: „On the other hand, larger firms […] usually operate in multiple segments. This makes their financial ratios more difficult to interpret.".

167 STEIN ET AL (2003, p. 5): "[…] Since, by definition, private firms do not have publicly traded equity and debt, price series of these financial assets are not available for individual firms. This implies that even if the firm-specific details of these companies were publicly known, price discovery reflecting the incorporation of these risks does not take place. Thus, various asset pricing-based approaches to default risk that have enjoyed wide success and acceptance for public firms cannot be directly applied to private ones."

168 FALKENSTEIN, BORAL, CARTY (2003/2000, p. 77): "Yet, some of this loss of power can also be explained by the fact that defaults are measured better for rated firms than unrated public firms, and for public firms vs. private firms. More of the 'goods' in the unrated universes are mislabelled, but unfortunately we do not know which ones."

169 By merging MOODY's and KMV's databases it could for instance be estimated, that MOODY'S (pre-merger) data base contained only one third (!) of all defaults that pertained to corporations with revenues of less than 1 m US$ that were covered by MOODY's database, see DWYER, STEIN (2003, p. 7, 9). MOODY'S database is being used for developing, calibrating and validating insolvency predicting models for small and medium-sized enterprises.

170 FALKENSTEIN, BORAL, KOCAGIL (2000, p. 12): "Size is a notorious correlate with various inputs, most significantly the quality of financial statements and our measurement of default. Larger companies tend to have audited statements that are of better quality. More importantly, perhaps, is that our measure of default is more accurate for the larger firms. While we do our best to make sure that companies in our database are truly defaulted or non-defaulted companies, and in fact exclude more data than we use because of this effort, inevitably we do make some misidentifications. Therefore, size and data quality correlate positively. "

171 BOHN, ARORA, KORABLEV (2005, p. 14) explain the *declined* (!) performance of a particular financial ratio based model when applied to larger firms as follows: „larger firms have more sophisticated financial statements since they usually operate in multiple segments. This makes their financial ratios more difficult to interpret."

172 see STEIN ET AL (2003, p. 30f)

173 For possible quantitative effects of "managing data quality" by "data cleaning" (deleting incomplete or presumably wrong data records) see DWYER, KOCAGIL, STEIN (2004, p. 8, 19ff.): increase in 1-year (5-years-) accuracy ratios from 48.2% (40.1%) to 51.7% (45.5%) or ESCOTT, GLORMANN, KOCAGIL (2001b, p.19): increase in 1-year accuracy ratios from 59.7% to 70.9%. In this case, however, it is questionable in how far the "data cleaning" criteria applied were really addressing *data quality* problems - rather than just biasing the sample by removing some groups of firms, where the rating model had often made faulty predictions before.

- *Regional origin:* Predictive power of agency ratings does considerably vary with regional origin of the corporations under consideration. It was noted, that agency ratings for European corporations are quite more selective than those for US-American corporations.[174] Whether these differences can really be traced back to *region specific* peculiarities, like different accounting systems or insolvency laws[175], or whether the quality of rating processes of the rating agencies' local branches does vary[176] - or whether the differences that were found, can be *completely* explained by other "structural sample differences" (such as coverages of different phases of insolvency cycles, sizes of corporations, or industry classifications[177], …) is unknown.

Restrained from "structural" aspects of the *sample* used, also methodical aspects of the *forecast method* or *rating model* itself can impact measured accuracy, like:

- *Underlying definition of default*[178]: Accuracy measures of rating models can vary depending on the *definitions of default* that were chosen. However, empirical relevance of this issue seems to be rather small[179,180,181], at least when certain alternative *objective* definitions of default are being used (delay of payment [typically at least 90 days], restructuring, insolvency).[182,183,184] For theoretical reasons it would be more appropriate, not to consider

---

174 One-/Five-year-accuracy-ratios for US-American corporations with S&P-ratings are 82%/74%, while those for European corporations are 94%/84%, see S&P (2004b, p. 11). In cases of US-American MOODY'S rated corporations AR-values of 81.0%/66.6% are achieved [AR values were calculated according to MOODY'S own definition] and for European corporations 95.3%/92.4%, see MOODY'S (2004c, p. 2).

175 CANTOR (2004, p. 3): „Regional distinctions, however, are critically important when applying this global methodology. Local expertise is likely to be quite valuable in rendering judgments about the meaning of financial statements, the macroeconomic and financial environment, and potential sources of support. Moreover, regional bankruptcy regimes may influence the incentives of issuers to service their debts and the incentives of other market participants to provide financial support in times of distress."

176 CANTOR (2004, p. 3): "Why have European ratings provided more powerful rank orderings of credit risk compared to American ratings? Is it because Moody's analytical practices are better in Europe or because relative risk is simply easier to judge in Europe? The following exhibit presents data which suggests the correct answer may be 'some of each.'"

177 See data given in CANTOR (2004, p. 9). Aside from Aaa-rated corporations, European MOODY'S rated *industrial firms* are usually much bigger (two to four times) than American MOODY'S rated industrial firms.
According to the data given in BASEL COMMITTEE (2000c, p. 33f.) the share of financial corporations (banks plus insurances companies) in all MOODY'S rated US corporations is only 30%, while for European MOODY'S rated corporations it is 66%! The respective ratios for S&P are 52% and 69%.
All in all, European agency rated corporations are bigger than American corporations and are more often associated with financial industries. On a univariate basis, both factors have been found important in explaining differences in predictive power of subgroups of corporations.

178 See on this topic in particular BALCAEN, OOGHE (2004, p. 21ff.).

179 In a study based on 35,000 medium-sized Austrian enterprises HAYDEN (2003, p. 33) showed, that statistical models that were calibrated on *bankruptcies*, practically had the same discriminatory power when validated on defaults defined as *credit rescheduling events* or on defaults defined as *delays of payments events* as statistical models that were specifically calibrated on these default events.

180 see on this also GRICE, DUGAN (2001, p. 154ff.)

181 More important than the exact definition of default seem so be, whether *any* or only the individually *first* default shall be forecasted, see LEHMANN (2003, p.8): "The analysis was also carried out with (not first-time) LLP [loan-loss-provision] […] as definition of default. A number of 'easily classifiable' observations entered, the performance of the rating system rose by a considerable amount. Yet, this is rather trivial. Usually, defaulted loans enter a separate monitoring process. The bank is most of all interested in the 'surprises' in its non-default loan portfolio. The true capabilities of a credit rating system show in the prediction of first-time LLP, not the extrapolation of past LLP. The definition of the default criterion has great impact on the results. Therefore, studies with a different default criterion cannot be compared easily."

182 For the various empirically used definitions of default see S&P (2004, p. 7f), MOODY'S (2004, p. 3). See also the tables in chapters 3.3 to 3.5.

delay of payments as default events, because *delays of payments* do not inherently cause any adversities to the creditors.[185,186] At the best, *delays of payments* are good predictors for future debt losses, so that they should rather be used as *explanatory variables*, but not as *variables to be explained*. Considering practical issues, including *delays of payments* (but also other events, such as *loan restructuring*) as default events raises serious questions of data availability – even if bank internal data is accessible.[187] Potentially fatal, however, is the default definition according to Basel II, that defines the existence of default events, amongst others, already then, when a bank assumes, that an obligor "is *unlikely* to pay its credit obligations to the banking group in full […]"[188]. As indications of the unlikeliness of full repayments are considered, amongst others, when "the bank *makes a charge-off* or account-specific provision resulting from a *significant perceived decline in credit quality* subsequent to the bank taking on the exposure." When establishing the accuracy of default predictions - with the above mentioned definition of *default* – it is not (only) being tested how well bank ratings are suited for predicting bankruptcies, but (also) how well *current* bank ratings can predict *future* bank ratings (see "unlikely" and "perceived decline in credit quality"), which is rather a test of rating *stability* than a test for *predictive power* of rating, or how well *current bank ratings* can predict their *future bank actions* ("the bank makes a charge-offs"). If these actions are triggered by future values of the same rating system whose predictive power is being examined, predicting own future actions can also be rather considered to be a test of *rating stability* than a test of *predictive power*!

- *Aspired temporal stability of ratings*:[189,190] By pursuing other, potentially rivaling or conflicting goals, in particular by attempting to stabilize ratings (see the notes above concern-

---

183 The economic impact of *defaults*, however, seems to be sensitive with respect to the default definition chosen, see VARMA, CANTOR (2005, p. 32f., p. 43) for an examination on bond recovery rates for seven different (initial) default events. In case of *distressed exchanges* or *missed principal* defaults, bond holders can realize recovery rates that are on average more than 35 percentage points higher than in case of *chapter 7* defaults. In about 85% of all 1,084 examined defaults, however, *missed interest payments* or *chapter-11*-defaults are involved, whose default rates differ by comparatively insubstantial 10 percentage points on a univariate basis (and even less on a multivariate basis).

184 KOCAGIL, AKHAVEIN (2001, p. 5, formatting added): „The discussion about the definitions of default included within the proposals appears to have centered around when a firm would be considered to have defaulted, and hence the impact on aggregate default rate numbers and PDs. *There has been less discussion on how different default definitions might impact the variables used within internal rating tools. Our understanding is that this is because, as our own experience shows, the factors that can predict default are generally the same, whether the definition of default is 90 days past due or bankruptcy* (in fact many of the definitions contained within BIS II are steps on the road to bankruptcy/insolvency)."

185 In an empirical examination GUPTON, STEIN (2005, p. 22) find, that in case of bank loans, 20% to 50% of all "delay-of-payment-defaults" do not cause any economic damages to the concerned banks, without the banks having to utilize collateral or to restructure loans. For that reason, many banks do not consider and record such events as defaults. However, such nonuniform data recording behavior inevitably results in mismatches in measured default rates- and in conceptual incompatibilities of other Basel II risk parameters, such as LGD (loss given default). See on this also BASEL COMMITTEE (2005, p. 63) and NORDEN, WEBER (2005, p. 48f.).

186 see on this also KOCAGIL, AKHAVEIN (2001, p. 5), KOCAGIL ET AL. (2003, p. 6)

187 see FALKENSTEIN, BORAL, KOCAGIL (2000, p. 7): "If a company has its loans restructured in such a way that there is an adverse effect upon the lender, such as moving payments back in time without any compensation, in general we do not capture this as a default. This is not our intention, as adverse restructurings are part of Moody's corporate definition of default. It is a limitation of the data: rarely are such restructurings properly recorded in internal systems."

188 BASEL COMMITTEE (2004, §452f, formatting added)

189 "Many financial market participants - investors, regulators and issuers - desire stable ratings. However, while reflecting an aversion to volatility per se, their desire for rating stability also reflects the view that *more stable* ratings are *more accurate* ratings with respect to the relative fundamental credit risk of a borrower", see CANTOR, MANN (2003, p. 15), "Some investors […] highly value rating stability to avoid unex-

ing the OTC-approach), *accuracy* of ratings may be considerably negatively affected.[191] Annotation: The aspired *stability of individual rating notes* should not be confounded with the *stability of the rating system*. The latter quality factor measures, whether the predictive accuracy of the rating system diminishes throughout time (but with constant time horizon), see also the remarks concerning *pool- vs. cohort data*.[192]

- *Refinement of rating scale:* The more imprecise a rater communicates his insolvency predictions, for instance by using only few discrete rating grades instead of continuous scores or default probabilities, the smaller is the discriminative power of the ratings. When using a 7ary scale (see S&P notation *AAA, AA, A, BBB, BB, B, CCC/C*) information losses, measured in accuracy ratio, of a magnitude of 2% - 3% (1 – 2.5 percentage points) have to be expected compared to a rating model with a continuous scale. Using a 17ary instead of a continuous scale comes along with information losses of about 0.5% (< 0.5 percentage points) (see chapter 2.3.4 and Appendix IV).

SWETS (1988, 1289ff.) states following four qualifications concerning validity and reliability of accuracy measures, that are violated by some of the examples stated above:

(1) "*Adequacy of truth:* The tester should know with certainty for every item in the test sample whether it is positive or negative. Incorrectly classifying test items will probably depress measures of accuracy." (ibid) Related problems are expected to occur with respect to *data quality* and definition of default (see above).

(2) "*Independence of truth determination and system operation:* The truth about sample items should be determined without regard to the system's operation". (ibid) Otherwise the system's performance is likely to be overstated. Related problems have to be expected in relation with *subjective definitions of default*.

(3) "*Independences of test sample and truth determination*. Procedures used to establish the truth should not affect the selection of cases. Thus the quest for adequate truth may bias the sample of test cases, perhaps resulting in an easier sample than is realistic." (ibid) See for instances the comments on *preselected portfolios*, in this case however, biasing the test samples resulted in worse than realistic samples.

(4) "Representativeness of the sample*: The sample should fairly reflect the population of cases to which the [...] system is usually applied."* (ibid) Following issues are relevant here: *positioning of the observation data period in the default cycle* and composition of the test sample with respect to corporations *legal forms*, *industry classifications*, *sizes*, and *regional origins*.

When ever possible, comparisons of different rating models' performances should always take place based on the same sample and based on the same methodological premises (like

---

pected portfolio revisions.", see FONS (2002, p. 4), "Moody's believes that our ratings system-management practices, as set forth above, are desired by both issuers and investors. Issuers want stability in ratings and the opportunity to make changes in their financial condition, if possible, to avoid changes in ratings.", see ibid (p. 12).

[190] CANTOR, MANN (2003, p. 1): "Moody's corporate bond ratings are intended to be 'accurate' and 'stable' measures of relative credit risk, as determined by each issuer's relative fundamental creditworthiness and without reference to explicit time horizons. Moody's performance should therefore be measured by both rating accuracy (the correlation between ratings and defaults) and rating stability (the frequency and magnitude of ratings changes).

[191] As one positive consequence resulting from increased rating stability, a reduction in transaction cost for those investors can be expected, who are pursuing rating-dependent investment strategies (e.g. due to regulatory restrictions), see LÖFFLER (2004a).

[192] see NORDEN, WEBER (2005, p. 41)

definition of default, attempted stabilization, fineness of rating scale, etc.). However, for empirical comparisons, these options are in many cases just not available, in particular not, when ratings cannot be independently reproduced by an outside researcher, for instance because details of the rating model are kept secret[193], because the ratings are influenced subjectively by the rater as the cases arise[194], or because the rating model utilizes data that is not publicly available.[195] But also in these cases it would be desirable to make portfolio- and rating model spanning comparison, considering all known disturbance variables in as much as possible.

Fortunately, many studies do not only present results for one particular rating method, but also results for other methods as well – based on the same sample and methodological premises - so that at least here direct comparisons are possible. The rating model that was most often chosen for such reference purposes, is the so called ALTMAN'S Z-score rating model (see chapter 3.4). Often univariate predictive accuracies of single ratios are reported, too (see chapter 3.3).

It also has to be noted, that none of the rating models that are used in reality can perfectly forecast future default events. At best *stochastic* statements can be made. Therefore empirically measured accuracy values have to be interpreted as realizations of random events, for which confidence intervals may be given based on formal methods and assuming certain approximations or based on numerical methods (simulation).

The only accuracy measure that is considered subsequently for the reasons outlined in chapters 2.3 and 2.4 is the *accuracy ratio*.

A very conservative[196] estimation of the *standard deviation* (the standard deviation is subsequently used for inferring confidence intervals by assuming normally distributed AR-values) of the accuracy ratio, in the following referred to as *approximation 1*, with N_D defaulters and N_{ND} non-defaulters is given as follows: [197]

F 39) $\hat{\sigma}^2_{AR} \le 4 \cdot \frac{AUC_{ROC} \cdot (1 - AUC_{ROC})}{\min(N_D, N_{ND})}$ substituting AUC_{ROC} by AR (see F 8) yields:

F 40) $\hat{\sigma}^2_{AR} \le \frac{1 - AR^2}{\min(N_D, N_{ND})}$ (*approximation 1*)

---

193 see for instance ALTMAN, HALDEMAN, NARAYANAN (1977, p. 12): "The actual coefficients […] for the seven variables cannot be reported due to the proprietary nature of the ZETA model […]." or ALTMAN, MARCO, VARETTO (1994, p. 512): "The coefficients of all the functions are protected by secrecy for the purpose of safeguarding the investments […] made in research, testing and database creating."

194 S&P (2003b, p. 17): "There are no formulae for combining scores to arrive at a rating conclusion. Bear in mind that ratings represent an art as much as a science. A rating is, in the end, an opinion."

195 The rating model of PLATTNER (2002, p. 50f) for instance is using a dummy variable (as one of its 27 (!) explaining variables) that is set to 1, if the corporations' house bank believes, that the respective customer is affected by "temporary liquidity problems" and to 0 otherwise. For researchers outside the bank, however, it is practically impossible to correctly specify this variable – and to make things worse, just this variable is the most influential of all variables contained in the model, see ibid (p.46).

196 *Conservative* here means that the values thus calculated are bigger than the real standard deviation.

197 see STEIN (2002, p. 19), BASEL COMMITTEE (2005, p.41) with $\hat{\sigma}^2_{AR} = 4 \cdot \hat{\sigma}^2_{AUC}$

A more efficient, formally only marginally more elaborate approximation for the standard deviation of the Accuracy Ratio, in the following referred to as *approximation 2*, is given as follows:[198]

F 41) $\hat{\sigma}^2_{AR} \le \frac{4}{3 \cdot N_D \cdot N_{ND}} \cdot \left[(2N_{ND} + 1) \cdot AUC_{ROC} \cdot (1 - AUC_{ROC}) - (N_{ND} - N_D)(1 - AUC_{ROC})^2 \right]$ with

F 42) $\hat{\sigma}^2_{AR} \le \frac{1}{3 \cdot N_D \cdot N_{ND}} \cdot \left[(2N_{ND} + 1) \cdot (1 - AR^2) - (N_{ND} - N_D)(1 - AR)^2 \right]$ (*approximation 2*)

An unbiased determination of the standard deviation of the accuracy ratio can be obtained with following formula, subsequently referred to as *E-H-T (2003)*, which however is extremely cumbersomely to implement: [199]

F 43) $\hat{\sigma}^2_{AR} = \frac{1}{(N_D - 1)(N_{ND} - 1)} \left[1 + (N_D - 1) \cdot \hat{P}_{D,D,ND} + (N_{ND} - 1) \cdot \hat{P}_{ND,ND,D} - 4(N_D + N_{ND} - 1) \left(A - \frac{1}{2} \right)^2 \right]$ with

F 44) $\hat{\sigma}^2_{AR} = \frac{1}{(N_D - 1)(N_{ND} - 1)} \left[1 + (N_D - 1) \cdot \hat{P}_{D,D,ND} + (N_{ND} - 1) \cdot \hat{P}_{ND,ND,D} - (N_D + N_{ND} - 1) \cdot AR^2 \right]$ (*E-H-T (2003)*)

The terms $P_{D,D,ND}$ and $P_{ND,\,ND,D}$ and their determinants are calculated as follows:

F 45) $P_{D,D,ND} = P(S_{D,1}, S_{D,2} < S_{ND}) + P(S_{ND} < S_{D,1}, S_{D,2}) - P(S_{D,1} < S_{ND} < S_{D,2}) - P(S_{D,2} < S_{ND} < S_{D,1})$ und

F 46) $P_{ND,ND,D} = P(S_{ND,1}, S_{ND,2} < S_D) + P(S_D < S_{ND,1}, S_{ND,2}) - P(S_{ND,1} < S_D < S_{ND,2}) - P(S_{ND,2} < S_D < S_{ND,1})$

where $S_{D,1}$ and $S_{D,2}$ / $S_{ND,1}$ and $S_{ND,2}$ are score values for two corporations randomly chosen from the sample of defaulters/ non-defaulters.

- with $P(S_{D,1}, S_{D,2} < S_{ND}) = \sum_{i=1}^{g}$ (share of non-defaulters with a rating of i) * (share of defaulters with a rating worse than i)^2, thus

  F 47) $P(S_{D,1}, S_{D,2} < S_{ND}) = \sum_{i=1}^{g} \left(a_i \cdot \frac{1 - PD_i}{1 - PD} \cdot \left(\sum_{j=i+1}^{g} a_j \cdot \frac{PD_j}{PD} \right)^2 \right)$ (own computation)

- with $P(S_{ND} < S_{D,1}, S_{D,2}) = \sum_{i=1}^{g}$ (share of non-defaulters with a rating of i) * (share of defaulters with a rating better than i)^2, thus

  F 48) $P(S_{ND} < S_{D,1}, S_{D,2}) = \sum_{i=1}^{g} \left(a_i \cdot \frac{1 - PD_i}{1 - PD} \cdot \left(\sum_{j=1}^{i-1} a_j \cdot \frac{PD_j}{PD} \right)^2 \right)$ (own computation)

- with $P(S_{D,1} < S_{ND} < S_{D,2}) = P(S_{D,2} < S_{ND} < S_{D,1})$ and $P(S_{D,1} < S_{ND} < S_{D,2}) = \sum_{i=1}^{g}$ (share of non-defaulters with a rating of i) * (share of defaulters with a rating better than i) * (share of defaulters with a rating worse than i), thus

  F 49) $P(S_{D,1} < S_{ND} < S_{D,2}) = P(S_{D,2} < S_{ND} < S_{D,1}) = \sum_{i=1}^{g} \left(a_i \cdot \frac{1 - PD_i}{1 - PD} \cdot \left(\sum_{j=1}^{i-1} a_j \cdot \frac{PD_j}{PD} \right) \cdot \left(\sum_{j=i+1}^{g} a_j \cdot \frac{PD_j}{PD} \right) \right)$

  (own computation)

---

[198] see STEIN (2002, p. 19) with $\hat{\sigma}^2_{AR} = 4 \cdot \hat{\sigma}^2_{AUC}$

[199] see ENGELMANN, HAYDEN, TASCHE (2003, p. 10, formula 10) and BASEL COMMITTEE (2005, p.40) with $\hat{\sigma}^2_{AR} = 4 \cdot \hat{\sigma}^2_{AUC}$

Analogously holds:

- **F 50)** $P(S_{ND,1},S_{ND,2} < S_D) = \sum_{i=1}^{g}\left(a_i \cdot \frac{PD_i}{PD} \cdot \left(\sum_{j=i+1}^{g} a_j \cdot \frac{1-PD_j}{1-PD}\right)^2\right)$ (own computation)

- **F 51)** $P(S_D < S_{ND,1},S_{ND,2}) = \sum_{i=1}^{g}\left(a_i \cdot \frac{PD_i}{PD} \cdot \left(\sum_{j=1}^{i-1} a_j \cdot \frac{1-PD_j}{1-PD}\right)^2\right)$ (own computation)

- **F 52)** $P(S_{ND,1} < S_D < S_{ND,2}) = P(S_{ND,2} < S_D < S_{ND,1}) = \sum_{i=1}^{g}\left(a_i \cdot \frac{PD_i}{PD} \cdot \left(\sum_{j=1}^{i-1} a_j \cdot \frac{1-PD_j}{1-PD}\right) \cdot \left(\sum_{j=i+1}^{g} a_j \cdot \frac{1-PD_j}{1-PD}\right)\right)$ (own computation)

In a simulation experiment based on a portfolio of 10,000 corporations, a given distribution $a_1..a_g$ of corporations over g rating classes 1..g and 15 different PD-vectors[200] $PD_1..PD_g$, standard deviations for the accuracy ratio were obtained resting upon 200 simulation runs per setting (i.e. per PD-vector). The values thus obtained are compared with *approximations 1* and *2* and the exact (unbiased) method *E-H-T (2003)* (see Figure 12):

Standard deviations obtained with approximation 1 are oversized by factor 2.3 to 1.8, standard deviations obtained with approximation 2 by factor 2.3 to 1.4. As was to be expected, there were no statistically significant differences between standard deviations obtained by simulation and by formula E-H-T (2003).

Note: As can be seen from the respective formulas, the only input needed by both *approximation* methods are the expected accuracy of the rating model (stated in AUC_{ROC} or AR) and the number of defaulters and non-defaulters, while the exact (unbiased) E-H-T(2003) estimator requires rating class specific data concerning default rates and frequencies as well. These data, however, are not available for most insolvency studies.

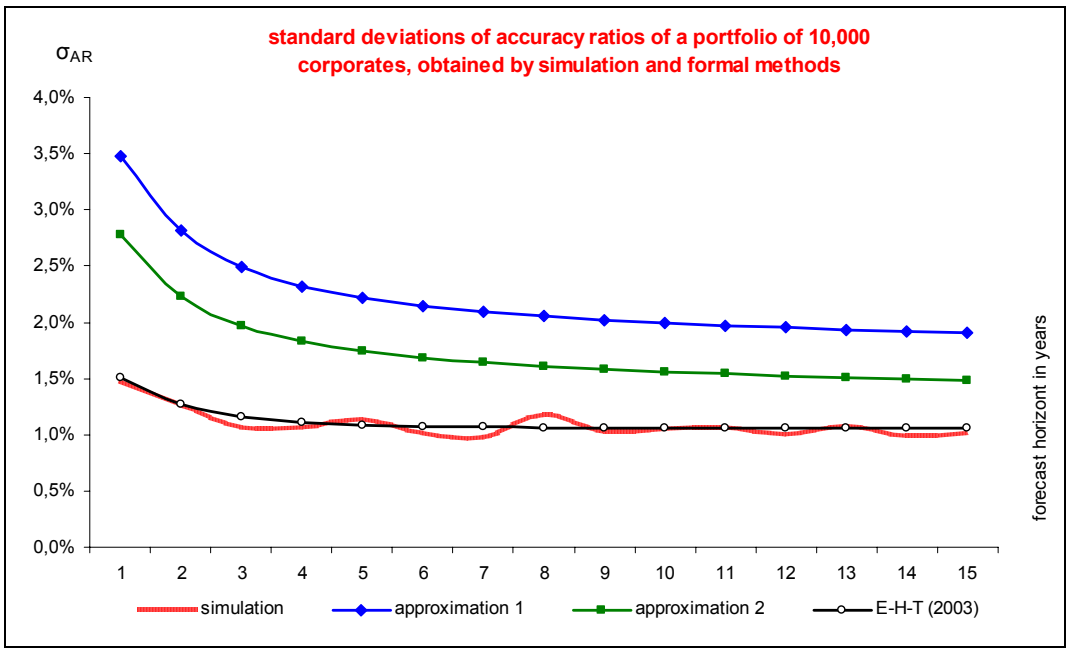

Figure 12: standard deviations of accuracy ratios of a portfolio of 10,000 corporations, obtained by simulation, two approximations, and one exact method; 15 different forecast periods; 200 simulation runs per setting

[200] For the simulations a 17ary rating scale was used. The distribution of companies over the various rating classes was chosen according to S&P (2004). Rating class specific default probabilities were chosen according to the respective historical, cumulated 1-15-years default rates, see S&P (2004, p. 13).

If accuracy ratio values are (approximately) normally distributed[201], confidence intervals with a confidence level of α can be constructed as follows:[202]

F 53) $$CI_\alpha = \left[AR \pm \sigma_{AR} \cdot \Phi^{-1}\left(\frac{1+\alpha}{2}\right) \right]$$

with Φ^{-1} … reverse function of the standard GAUSSIAN distribution

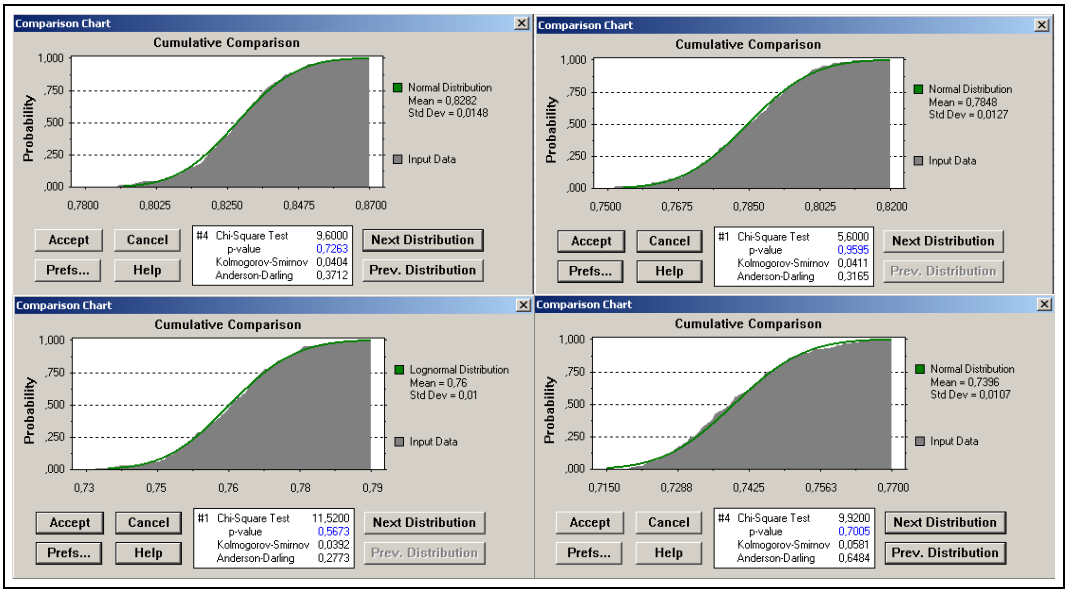

Figure 13: distributions of accuracy ratios obtained by simulation (gray areas) and GAUSSIAN distributions, fitted by maximum-likelihood-estimations (lines)

For portfolios of the size and default distributions that were used in the simulation experiment, the distribution of accuracy ratios can appropriately be modeled by GAUSSIAN distributions (see Figure 13)[203], so that reliable confidence intervals can also be obtained by the formulas given in F 44 to F 53. For smaller portfolios or portfolios with unusually small probabilities of default, however, the GAUSSIAN distribution (or the *distribution of POISSON*[204])

---

201 For theoretical reasons, it is impossible that the accuracy ratio of a given portfolio is exactly normally distributed. First, in a portfolio with a finite number of corporations only a finite number of different accuracy ratio values are attainable – while the GAUSSIAN distribution is a continuous function, which therefore can adopt indefinitely many different possible values. Second, by definition accuracy ratio values are limited to values between -1 and +1 while there are no limits for the GAUSSIAN distribution.

202 see for instance ENGELMANN, HAYDEN, TASCHE (2003, p. 10)

203 The four graphs in Figure 13 show empirical distributions of the accuracy ratio values that were obtained for the first 4 of the 15 PD-vectors which represent rating class specific probabilities of default for forecast horizons of 1 to 4 years. Also displayed in the respective graphs are measures that show the quality of the maximum likelihood fittings according to three test methods, χ^2-, KOLMOGOROV-SMIRNOV- and ANDERSON-DARLING-test. According to the χ^2-test all fittings with p-values bigger than 0.5 are considered to be good, according to the KOLMOGOROV-SMIRNOV-test p-values smaller than 0.03 and according to the ANDERSON-DARLING-test p-values smaller than 1.5 (see DECISIONEERING (2000, p. 141)). According to the χ^2-test and ANDERSON-DARLING-test all of the four fittings that are displayed in Figure 13 are evaluated as good, while according to the KOLMOGOROV-SMIRNOV-test all fittings marginally miss a good evaluation.

204 Utilizing the POISSONIAN limit theorem, it follows that the distribution of POISSON can be used for modeling *rare events*. In case of portfolios with small rating class specific probabilities of default and "sufficiently large" (HARTUNG (1991, p. 122), translation) rating classes, rating class specific defaults can be modeled with the distribution of POISSON with parameters $\lambda_i = n_i \cdot PD_i$, although the binomial distribution with parameters n_i and PD_i would be the most direct choice for modeling rating class specific defaults. However, on the contrary to binomial distributions with different PD-parameters, the sum of random variables that follow the distribution of POISSON also follows the distribution of POISSON with $\lambda_\Sigma = \Sigma\lambda$. For $\lambda \geq 9$ the distribution of POISSON can be approximated by a GAUSSIAN distribution, see HARTUNG (1991, p. 213).

might not be appropriate,[205] so that confidence intervals can only be determined numerically (i.e. per simulation).

When comparing accuracy results that were empirically achieved by different rating systems, the uncertainty of measured accuracy values that result by random realizations of individual defaults has to be considered. In the example chosen, even with a sample size of 10,000 corporations, there is a 5% chance that a random realization of the measured accuracy ratio is differing by more than 3 percentage points from the expected value (which however is only known in simulation studies - but not in real life applications).[206]
In a portfolio of 5,000 companies, which about equals the current number of S&P or MOODY'S rated issuers[207], confidence interval breadths would rise by factor $\sqrt{2}$ (to ±4,2 percentage points for α=95%); and in a portfolio of "only" 1,600 corporations, which is roughly equal to the number of S&P and MOODY'S rated corporations in the beginning of the 1980ies even by factor 2,5 (to ±7,4% percentage points for α=95%).

In the following chapters the results of various studies are presented and compared. Insofar these data were available, the specific parameter values of the above mentioned *influencing factors* (size of database, temporal and industry origin, sizes of enterprises, usage of bank portfolio data yes/no) were given – but they were *not* used for "correcting" the measured accuracy ratio values, as no accepted correcting mechanisms is known.

The performance of the various rating models shall not only by compared in relation to each other (see chapter 3.5), but also in relation to three benchmark rating models, that can be implemented with minimal means and whose performance has been extensively studied.

- In chapter 3.2 it is examined (for the German basic population of corporations), how accurate insolvency predictions are, whose only inputs are the legal forms and industry classifications of firms.
- Chapter 3.3 summarizes the results of empirical studies that examined the insolvency predictive value of single ratios derived from firms' financial statements.
- In chapter 3.4 the performance of a *multivariate* insolvency prediction model is examined, that served as benchmark model in many insolvency prediction studies.

The performance of these three benchmark models is meant to form more meaningful lower thresholds of the minimal acceptable predictive quality of rating models than naïve predictions with expected accuracies of 0% do.

---

[205] If a portfolio consisted of n corporations with identical probabilities of default PD, the number of defaults, which actually follows a binomial distribution, could – according to the theorem of DE MOIVRE/ LAPLACE – well be approximated by a GAUSSIAN distribution with $\mu = n \cdot PD$ and $\sigma = \sqrt{n \cdot PD \cdot (1-PD)}$, if $n \cdot PD \cdot (1-PD) \geq 9$, see HARTUNG (1991, p. 201).

For the respective default rate would apply: $\mu_{rate} = PD$ and $\sigma_{rate} = \sqrt{\frac{PD \cdot (1-PD)}{n}}$

For small values for PD, i.e. $\mu = n \cdot PD \approx n \cdot PD \cdot (1-PD)$, a good approximation by the GAUSSIAN distribution is possible, if the expected value of defaults is at least 9. For PD=1%, for instance, the minimal portfolio size would be $\frac{9}{1\% \cdot 99\%} \approx \frac{9}{1\%} = 900$ corporations (with identical probabilities of default).

[206] With a confidence level α=95%, $-\Phi^{-1}(2.5\%) = \Phi^{-1}(97.5\%) \approx 1.96$, σ_{AR} = 1.5%, and μ_{AR}=82.9% a confidence interval with $CI_{95\%} = [82.9\% \pm 2.9\%] = [80.0\%; 85.8\%]$ results.

[207] On 01/01/1985 (01/01/2004) 1,620 (4,810) corporations were rated by MOODY'S and 1,650 (5,189) by S&P, see MOODY'S (2005, p.35) and S&P (2005, p. 26).

3.2 Benchmark I: Attainable accuracy by taking into account legal status and industry classification of corporations

The purpose of this chapter is to establish the attainable prediction accuracy of rating model that is only based on publicly available insolvency statistics.

In the insolvency statistics issued by the STATISTISCHES BUNDESAMT, the German Federal Statistical Office, univariate break-downs of insolvency *frequencies* partitioned by industry classification, legal status, federal states, age, and size of enterprises (numbers of employees) are published. Additionally, insolvency *rates* are published by corporations partitioned by industry classification, legal status and federal states.[208] If insolvency *frequencies* and *rates* are known, it is possible to derive the total number and relative frequencies of corporations (including non-defaulted corporations) of the respective groups.[209] Thus, if the various groups are sorted according to realized insolvency rates, the univariate *ex-post*-prediction accuracy, measured for instance in accuracy ratios, of the criteria mentioned above can be obtained.

However, if the intention is to derive insolvency *predictions,* a method to infer an ordinal relationship based on information, which is available at the beginning of the period for which insolvencies shall be predicted must be found. For this purpose, in the following realized insolvency rates of the preceding year will be used. Although, more sophisticated methods are conceivable, it can be shown empirically that the quality of *ordinal* insolvency predictions could only marginally be improved, as historical orderings of groups are excellent estimators for future years' orderings (see Figure 18 in anticipation of this chapter's results).

For being able to obtain discriminative insolvency predictions based on industry classification, legal status, or home federal state of corporations, two preconditions must be fulfilled:
(I) there must be "notable" differences in realized insolvency rates of the groups that are defined by applying the above mentioned criteria and (II) these difference must be relatively stable (or more general. these differences must be *predictable*) year-on-year.

These aspects are investigated in the following. In Figure 14 time series of insolvency rates partitioned for 14 industry sector are given for 1994-2003, in Figure 15 time series for corporations of different legal status for 1980-2003, and in Figure 16 for different federal states for 1980-2003 are given.[210] Altogether up to 3.3 m corporations per year (in 2003) are covered by these statistics.[211]

---

[208] See for instance STATISTISCHES BUNDESAMT (2004a, 2004c). Additionally here can be found partitionings of insolvencies according to *applicants for insolvency* (creditors vs. obligors) or for *causes of insolvency* (over-indebtness, illiquidity, …)

[209] Number of corporations $_i$ = insolvencies $_i$ / insolvency rate $_i$, if insolvency rate $_i > 0\%$.

[210] Data source for univariate analyses: 1980-1998 GÜNTERBERG, WOLTER (2003) and for 1999-2003 STATISTISCHES BUNDESAMT (2004b).

[211] With to exceptions, on of it of major importance (see below), the total number of corporations covered is essentially equal to the turnover tax statistics that covers all corporations liable for turnover taxes with revenues of more than 16,617 EUR per year.
Additionally 350,000 GmbH (ltd) and 7,000 AG [plc] were considered, in particular holding companies, who are not liable for turnover taxes and thus are not covered by turnover tax statistics, see STATISTISCHES BUNDESAMT (2004a, p. 20f). Official insolvency statistics, however, are inconsistent in this respect: when calculating legal status dependent insolvency rates, these non-turnover-tax-liable corporations were included (see ibid.), but they were excluded when calculating industry specific or all-corporations default rates (while the total number of *insolvencies* remained constant), which can be shown by reverse projections based on insolvency *frequencies* and *rates* (S. 19, 20, 41) and which was confirmed via telephone inquiry by the STATISTISCHES BUNDESAMT. Therefore, the average all-corporations default rate for Germany is overstated in

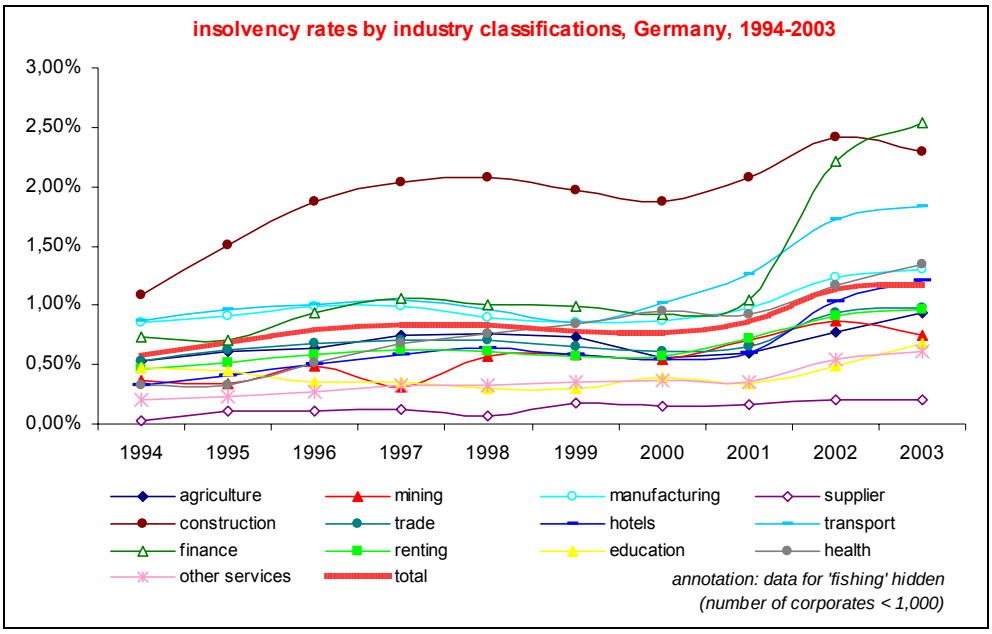

Figure 14: insolvency rates by industry classification, Germany, 1994-2003[212]

acc. ratio	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**mv**
ex-ante		25.5%	26.7%	25.8%	25.4%	25.4%	24.7%	25.0%	20.9%	19.8%	**24.3%**
ex-post	23.3%	25.6%	26.7%	25.9%	25.6%	25.4%	25.1%	26.0%	21.9%	19.8%	**24.5%**

Table B: univariate predictive accuracy (AR) of insolvency predictions based on industry classifications, by ex-post-sorting (realized default rates of respective years) and ex-ante-sorting (realized default rates of preceding years), mv ... mean value

Pronounced differences between insolvency rates of various industries[213] are to be noticed (see Figure 14). The ordinal relationship between the various industries' default rates is relatively stable, even for very long periods of time: from the five industries with the highest insolvency rates in 1994, four are still among the "worst five" industries in 2003. From the five

official statistics by about 12% and was in 2003 not 1.35% (ibid., p. 20) or 1.34% (STATISTISCHES BUNDESAMT (2004c), but 1,19% (own calculations).
For correcting these inconsistencies there were used factors for projecting from the number of revenue-tax-liable GmbH and AG to the entirety of all GmbH and AG. It was assumed that these factors (GmbH: 1.83, AG: 2.07), which were determined based on 2003 data, were the same throughout time and for all industries. Larger deviations compared to the (definitely overstated) official industry specific insolvency rates of the STATISTISCHES BUNDESAMT will therefore in particular appear in industries with above average shares of GmbH (or AG) in all companies: mining and quarrying: 1.07% (STATISTISCHES BUNDESAMT (2004c)) vs. 0.75% (own calculations) and manufacturing: 1.60% (STATISTISCHES BUNDESAMT (2004c)) vs. 1.30% (own calculations). Smaller deviations will appear in industries with below average shares of GmbH (and AG): hotels and restaurants 1.25% vs. 1.20% or health and social work: 1.43% vs. 1.35%.

[212] source: own calculations based on GÜNTERBERG, WOLTER (2003, p. 142f) 1994-1998 data (small trades deducted) and STATISTISCHES BUNDESAMT (2004b) 1999-2003 data (industry specific corrections of under recordation of GmbH and AG).

[213] Following sectors were covered (official denominations and, where required, own abbreviations): A: agriculture, hunting and forestry (*agriculture*), B: fishing, C: mining and quarrying (*mining*), D: manufacturing, E: electricity, gas and water supply (*supplier*), F: construction, G: wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods (*trade*), H: hotels and restaurants (*hotels*), I: transport, storage and communication (*transport*), J: financial intermediation (*finance*), K: real estate, renting and business activities, etc. (*renting*), M: education, N: health and social work (*health*), O: other community, social and personal service activities (*other services*).

"best" industries (industries with the lowest insolvency rates) in 1994, three are still among the five best industries in 2003[214]

Permanent and materially above-average default rates are only occurring in *construction.* Only in 2003 one other, relatively minor industrial sector (at least in terms of numbers of corporations)[215] was characterized by even higher default rates (*financial intermediation*[216]). Permanent below-average default rates were characteristic for *electricity, gas and water supply; education*; and *other community, social and personal service activities*.
In Table B attainable accuracy-ratios for one-year insolvency predictions are given whose only explanatory variable is the firms' industry classification.[217] Presented are both ex-post classifications and ex-ante predictions.

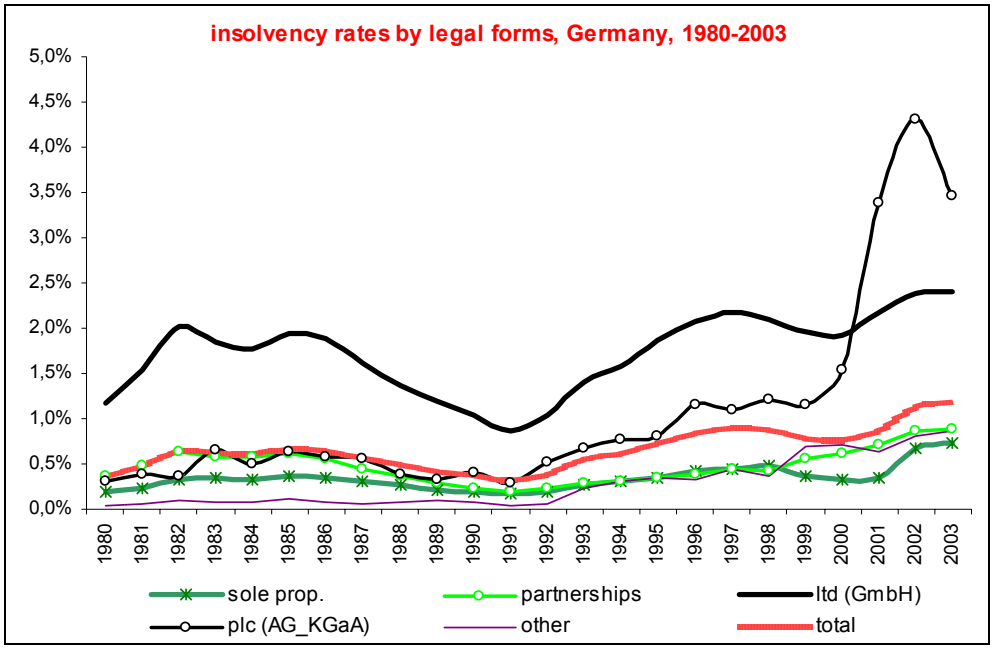

Figure 15: insolvency rates by legal forms, Germany, 1980-2003[218]

AR	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**	**1991**	**…**
ex-post		41.4%	40.6%	37.9%	37.9%	38.3%	38.4%	38.0%	38.4%	38.9%	38.6%	38.6%	…
ex-ante	39.8%	41.4%	40.6%	37.9%	37.9%	38.3%	38.4%	38.0%	38.4%	38.9%	38.6%	38.6%	…

[214] Because only insolvency predictions with forecast horizons of one year shall be examined, the year-on-year stability of default rates is more important than multi-year examinations: average changes in ranks were 0.94 ranks per sector and year. In 45% of all cases ranks remained the same, in 32% ranks changed by one position, in 15% by two positions and in only 7% of all cases by more than two positions. By randomly assigning ranks, average change in ranks would be 4.64 ranks per sector and year and in 67% of all cases ranks changed by more than two positions.

[215] Only 0.5% of the 3.3 m corporations belong to the group of *financial intermediation*. The other negative outlier group, *construction*, accounts for remarkable 11.4% of all corporations and thus has a much bigger impact on the predictive value of industry classifications.

[216] Included are for instance *commercial banks* (WZ2003-industry classification code 65.12), *investment companies* (WZ 65.23.1), *insurance industry* (WZ 66) or *insurance salesmen* (WZ 67.20.1).

[217] For calculating accuracy ratios are needed: industry specific realized default rates, see Figure 14, shares of the various industries in all corporations and an asserted ordering of industries with respect to expected default rates.

[218] Source: own calculations based on GÜNTERBERG, WOLTER (2003, p. 144f) for 1994-1998 data (small trade deducted) and STATISTISCHES BUNDESAMT (2004b) for 1999-2003 data.

1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	mv
40.0%	38.8%	39.1%	39.3%	37.9%	37.5%	36.1%	35.2%	41.5%	42.2%	30.6%	28.1%	**38.0%**
40.0%	38.8%	39.1%	39.3%	38.4%	37.5%	36.1%	39.4%	41.5%	42.5%	30.6%	28.1%	**38.3%**

Table C: univariate predictive accuracy (accuracy ratio) of insolvency predictions based on legal status, by ex-post-sorting and ex-ante-sorting, mv..mean value

In Figure 15, see above, time series of insolvency rates for five different legal forms (*ltd* - limited liability companies), *plc* - public limited companies and partnerships partly limited by shares (KGaA), *part* - partnerships (OHG-unlimited company), KG - limited partnership), *sole* - sole proprietorships and *other* - other legal forms)[219] are displayed for the 1980-2003 period.

Remarkable are the permanently above average default rates for limited liability companies. Although their share in all companies was only 25% in 2003, they accounted for 51% of all insolvencies in 2003. Thus, their insolvency rate is about twice as high as the average default rate of all companies and about three times as high as the insolvency rate of all *other* companies (in some of the years it is actually up to 5.5 times as high). Even higher insolvency rates do only occur among public limited companies as of 2001, whose share in all corporations is comparatively minor (0.4%).

The final variable that was considered for univariate explanations of corporate insolvencies was *federal state affiliation*, see Figure 16. Starting with the German reunification in 1990, there is a striking divergence in insolvency events between old and new federal states (former West-Germany vs. former East-Germany), in particular from 1990-1997. The rank order of insolvency rates within new and old federal states is relatively stable. Within the new federal states Thüringen and Brandenburg achieve comparatively low and Sachsen-Anhalt and Berlin-East comparatively high insolvency rates. Within the old federal states noteworthy differences between Northern and Southern states become apparent. While the Southern states Baden-Württemberg, Bayern, Hessen, Rheinland-Pfalz - and as of 1999 also Saarland - score well, Northern states (and the special case Berlin-West) Nordrhein-Westfalen, Schleswig-Holstein, Bremen, Niedersachsen - and as of 2001 also Hamburg - come off badly.

[219] Hereunto belong for instance *Vereine (associations)* or *Genossenschaften (co-operations)*. Only 1.6% of all corporations are organized in such legal forms.

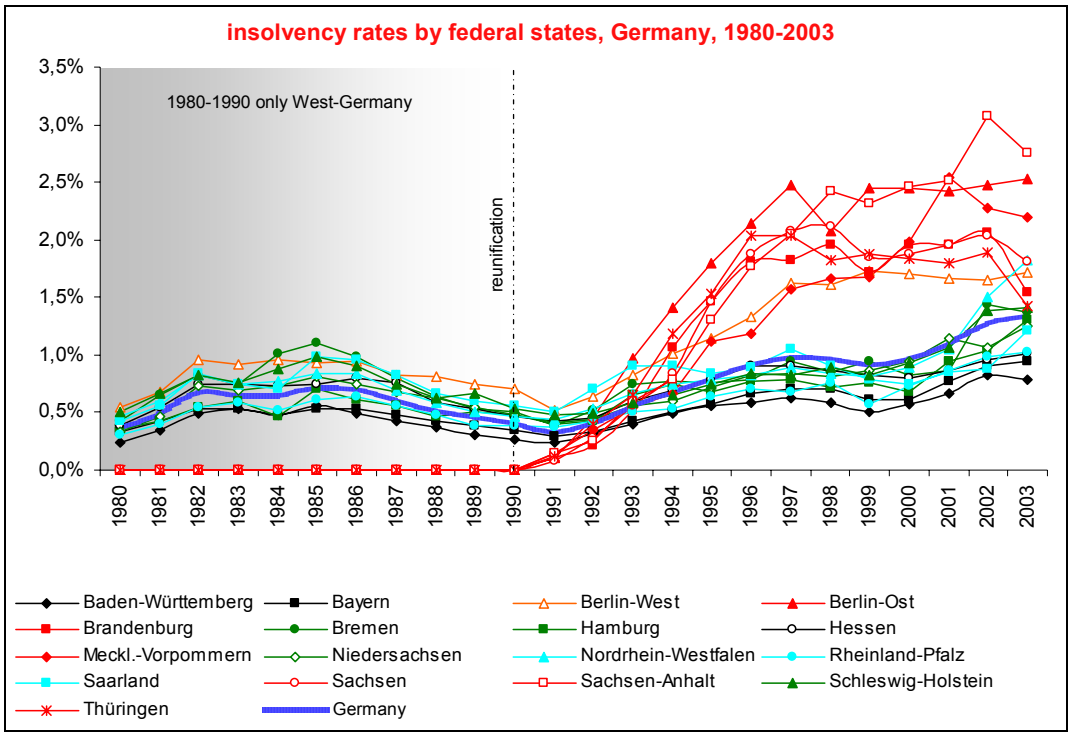

Figure 16: insolvency rates by federal states, Germany, 1980-2003[220]

AR	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	…
ex-post		11.7%	12.0%	8.7%	12.3%	10.6%	12.0%	11.1%	10.8%	12.7%	12.7%	19.1%	…
ex-ante	13.2%	11.8%	12.1%	8.7%	12.8%	11.7%	12.6%	11.7%	11.4%	13.0%	12.9%	19.8%	…

1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	mv
13.1%	9.7%	11.5%	16.8%	20.4%	21.2%	22.0%	24.4%	23.8%	21.5%	18.2%	18.5%	**15.4%**
14.0%	12.0%	14.4%	17.8%	20.5%	21.6%	22.4%	25.0%	24.5%	21.6%	19.9%	19.4%	**16.0%**

Table D: univariate predictive accuracy (accuracy ratio) of insolvency predictions based on federal state affiliation, by ex-post-sorting and ex-ante-sorting, mv ... mean value

Besides *univariate* predictive accuracy, *multivariate* predictive accuracy is a matter of particular interest. For its calculation absolute numbers and insolvency frequencies for groups of corporations that are identical with respect to their parameter values concerning industry classification *and* legal status *and* federal states (or alternatively individual datasets of corporations that contain these three criteria) were needed. Though, such data is not centrally provided for Germany by the STATISTISCHES BUNDESAMT. The STATISTISCHES BUNDESAMT, however, does provide cross-classified tables with *industry classification* and *legal status* insolvency data for the 1999-2003 period. [221]

[220] source: own calculations based on GÜNTERBERG, WOLTER (2003, p. 151ff) for 1994-2001 data (small trade deducted) and STATISTISCHES BUNDESAMT (2004b) 2001-2003 data; extrapolation: split-up of insolvencies Berlin, total into Berlin-West and Berlin-Ost (East) as of 2001; extrapolation of total number of companies domiciled in Berlin-West with West-German growth rate of total number of corporations as of 1991.

[221] STATISTISCHES BUNDESAMT (2004b)

The further advancements are largely analogous to the univariate analyses. The only difference is, that this time not only 14 groups (industry sectors) or 6 groups (legal status)[222] are formed, but 84 (=14*6) groups that each represent unique combinations of industry classifications *and* legal forms. Given about 3.3 m corporations and 40,000 insolvencies a year (2003) *on average* 39,000 corporations and 470 insolvencies remain per group. For preventing outlier values and division by zero, all groups that contained less than 1,000 corporations in 1999 were merged within their respective legal forms (as *legal form* was a more predictive univariate indicator of insolvencies than industry classification)[223], so that the number of distinctive groups was reduced to 61.[224] With 517.130 corporations (15.6%) the group most frequently represented was *sole proprietorships- trade*, the smallest group with only 479 corporations (0.015%) was the "remnant group" *other legal forms – other industries*.

Subsequently Figure 17 depicts insolvency rates for all corporations classified by legal forms *and* industries for 1999-2003: [225]

[222] As opposed to the data sets that were utilized for univariate analyses (1980-2003), the data set that was used for multivariate analyses allowed a separate treatment of GmbH&Co KG which form a subgroup of partnership companies with unusually high insolvency rates (cf. Figure 17).

[223] Merged were e.g. the groups *sole proprietorships –fishery* (716 corporations) and *sole proprietorships – mining* (591 corporations) forming the group *sole proprietorships – other industries*.

[224] Only 0.3% of all corporations were affected by merging groups. Most of the mergers referred to the generally weakly occupied industries *fishery* and *mining* and to the weakly occupied legal forms GmbH&CoKG and AG/KGaA (plc).

[225] Industry groups within legal forms were sorted with descending average 1999-2003 insolvency rates.

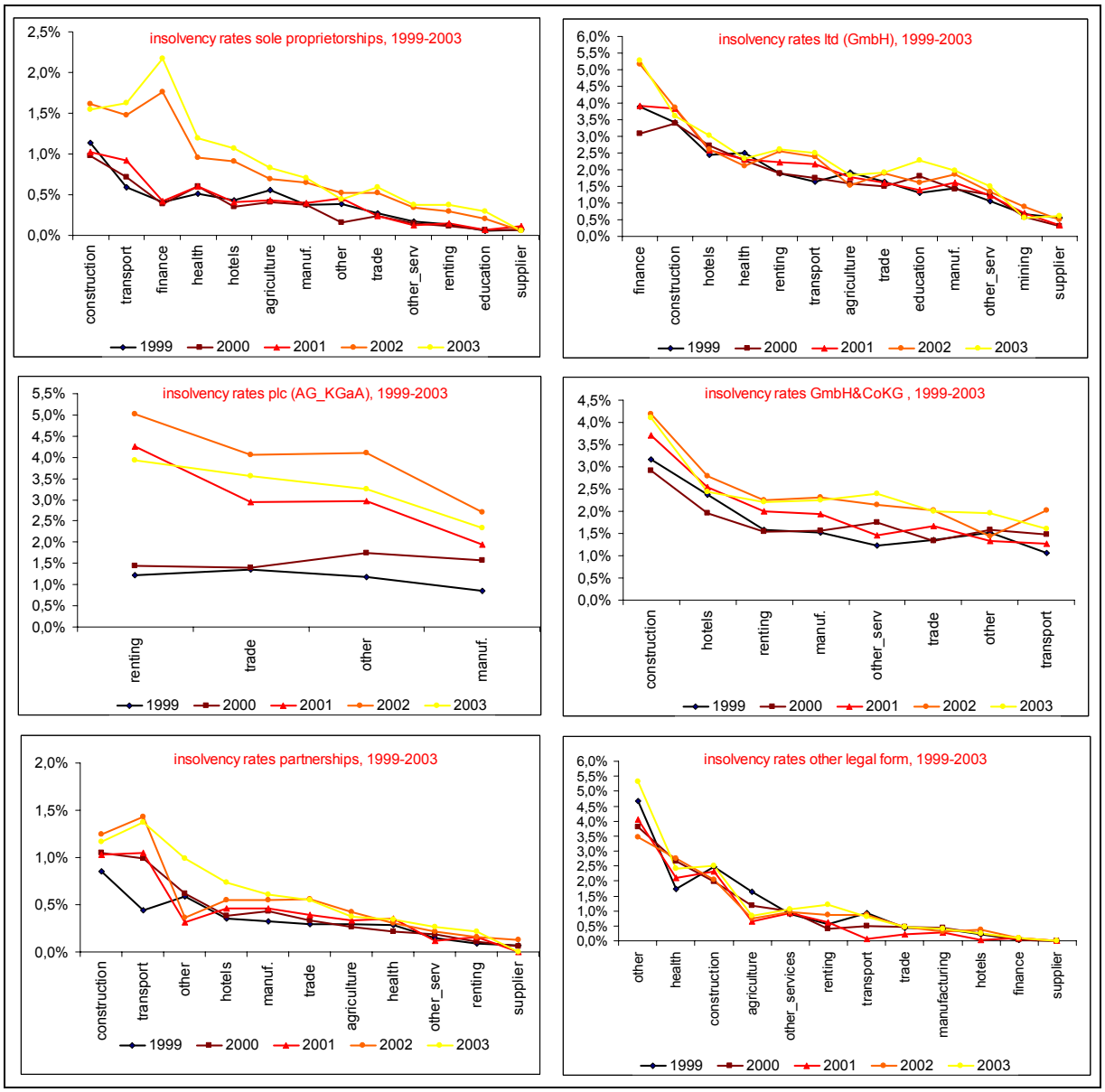

Figure 17: insolvency rates by legal forms and industries, Germany, 1999-2003

The graph shows, that:

- material differences occur in default rates of corporations of different industries within the same legal forms (see the steep slopes of the legal form specific graphs for all legal forms except AG/KGaA (plc)[226]),
- material differences occur in default rates of corporations within the same industries but different legal forms (see also the scale differences in Y-axes of the six legal form specific diagrams),
- relative industry specific default rates within legal forms are very stable throughout time (see the largely *non-intersecting* year-graphs within the six legal form specific diagrams) and

[226] In case of public limited companies just those industries were weakly represented, that were often related with far above average default rates (construction, hotels, finance) or with far below average default rates (supplier, other services).

- at least for *limited liability companies*, *partnerships* and *other legal forms* also *absolute* default rates are relatively stable, whereas there seems to be a structural break for *sole proprietorships* between 2001 and 2002, which might be caused by changes of insolvency law in 1999[227,228], that leads to materially and proportionally increased insolvency rates for all industries. In cases of GmbH&Co.KG and AG/ KGaA insolvency rates are not stationary either, but they seem to rise rather steadily and proportionally throughout time.

Despite a few outliers, the 1999-2003 rank orders of the 61 industry-legal status groups are very stable. Rank correlation coefficients for adjacent years vary between 0.924 and 0.975, which is not only highly statistically significant,[229] but already close to perfect stability.

From the numerically more important groups (share in all companies > 1%) permanently bad scores are obtained by *ltd – construction* (rank $_{\text{mv 1999-2003}}$=58.0 of 61 [mv…mean value], $PD_{\text{mv 1999-2003}}$=3.6%, share in all corporations=3.9%), *ltd – renting [and services]*[230] (rank=51.4; PD=2.2%, share=6.6%) and *ltd – transport* (rank=49.8; PD=2.1%, share=1.1%). Permanently good scores were achieved from *partner – renting [and services]* (rank = 6.8; PD=0.15%, share=3.9%), *sole – renting [and services]* (rank=8.4; PD=0.21%, share=15.5%) and *sole – other services* (rank=14.6; PD=0.23%, share =6.5%).

The accuracy ratio values of ordinal insolvency predictions that are attainable by taking into account firms' industry classifications and legal forms are displayed in Table E. Presented are both, AR-values for ex-post [classifications] and ex-ante predictions.

AR	**1999**	**2000**	**2001**	**2002**	**2003**	**mv**
ex-post		53.8%	54.7%	44.5%	44.5%	**49.4%**
ex-ante	53.0%	54.1%	55.1%	45.2%	45.2%	**50.5%**

Table E: multivariate predictive accuracy (accuracy ratio) of industry classification and legal status, by ex-post- and ex-ante-sortings, mv ... mean value

Graphical representations of the accuracy ratio time series for the various univariate prognoses and the multivariate industry-legal status-rating are given in Figure 18, while in Figure 19 (ex-post) CAP-curves[231] for the industry-legal status-rating are given.

---

227 see for instance STATISTISCHES BUNDESAMT (2004, p.5ff)

228 see also PLATTNER (2002, p. 37, translation): "The interpretation of the development since the implementation of the reformed insolvency law is obscured, as a part of the newly introduced consumer insolvencies is caused by entrepreneurial activity. In an economical sense, consumer insolvencies may be corporate insolvencies. (With the reformation of the reformation, which came into force as from 1st of December 2001, consumer insolvencies will be more narrowly restricted to private individuals.) Further, more insolvencies are being captured. Henceforth, also GbR [civil law associations] are subject to insolvency laws."

229 Based on 100,000 simulation runs the 99%-quantil of rank correlations at *random* placing is 29.8%. The 95%-quantile is 21.3%.

230 To the industry group "renting [and services]" belong following sectors: *real estate activities* (WZ 70), *renting of machinery and equipment without operator and of personal and household goods* (WZ 71), *computer and related activities* (WZ 72), *research and development* (WZ 73), *other business activities* (WZ 74).

231 The CAP curves are based on "ex-post-predictions", as there are only minor differences in quality (measured in accuracy ratios) between ex-post- and ex-ante-predictions, while for 1999 only ex-post-predictions could be given (as realized group specific default rates for the preceding year, 1998, were not available.)

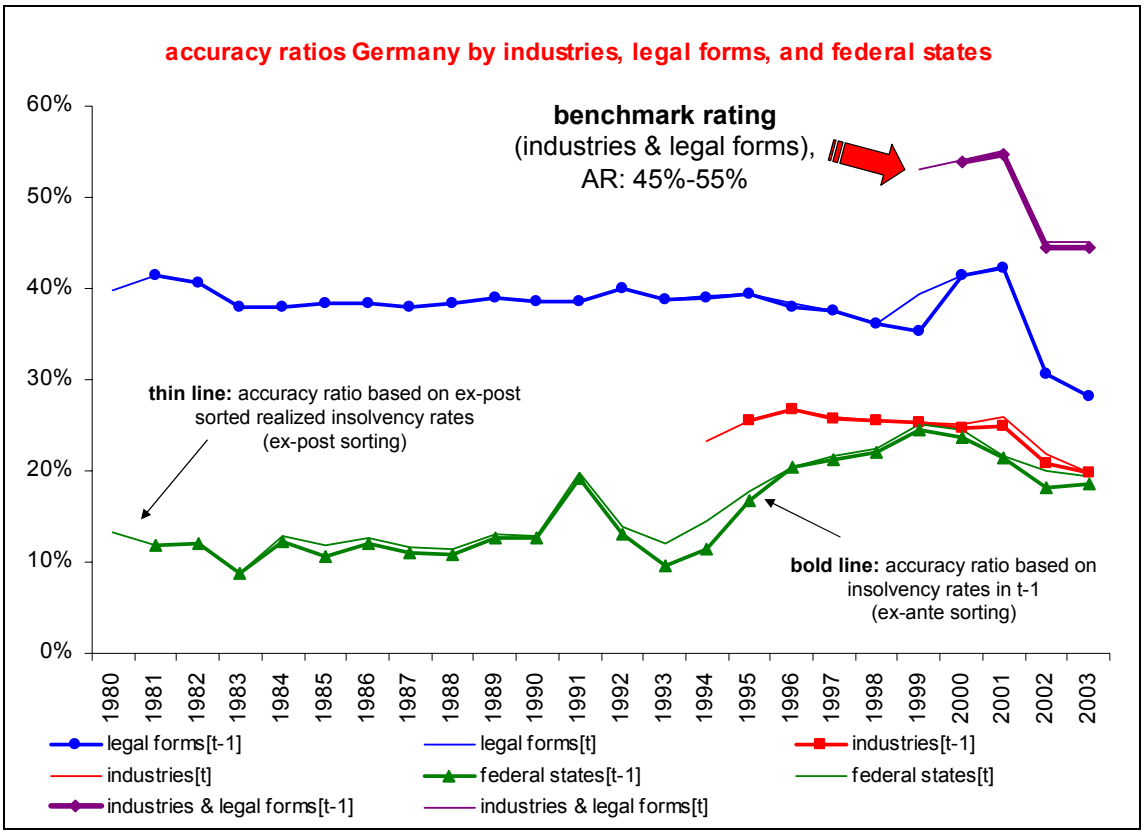

Figure 18: time series of predictive accuracy (measured in accuracy ratio) of ordinal insolvency predictions based on industry classifications, legal forms and federal state affiliations (univariate) and combination of industry classification and legal status, by ex-ante- and ex-post-sortings (t-1 vs. t)

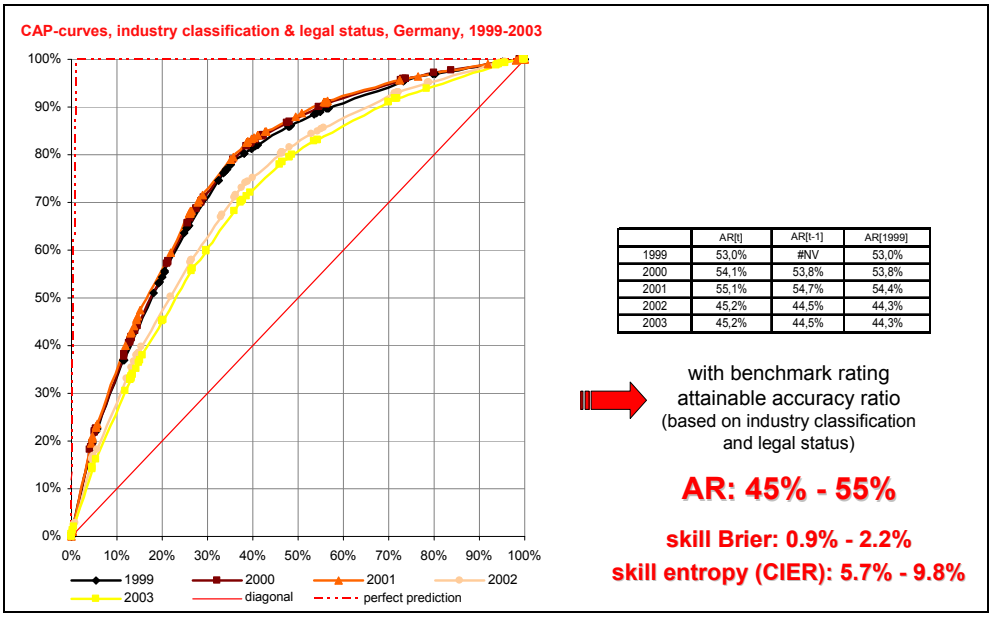

Figure 19: one-year-forecast CAP-curves (ex-post), 1999-2003 and accuracy-ratio-values (ex-post and ex-ante) for the industry – legal status benchmark rating

Altogether, it was shown that by considering only industry classifications and legal forms of corporations, accuracy ratios for insolvency predictions of 45%-55% (mean value 1999-2003: 50%) can be achieved. For rating models that are based on the same sample (representative sample of German corporations of all industries and legal forms), in particular for bureau or business scores like the CREDITREFORM-Bonitätsindex (see chapter 3.5) – this is a more rea-

sonable minimum level of just acceptable forecast accuracy than a fictive "naïve rating" with an expected accuracy ratio of only 0%.

Annotation: The predictive accuracy of the industry-legal-status-rating could *possibly* be improved by implementing more detailed industry classifications. Most important leverages for this are likely to be found by subdividing the currently biggest industry aggregates – which might be *heterogeneous* concerning default rates of their various subgroups[232]. The currently biggest industry aggregates are *manufacturing* (10.6% of all corporations), *construction* (11.4%), *trade* (24.3%), *renting and services* (27.0%) and *other services* (9.1%). It is also feasible, that the information content of the ratings could be improved by including domiciles (federal states) of corporations.[233]

Considering even further variables that were listed in the beginning of the chapter - like size of enterprises (see Figure 20[234,235]) or age of enterprises - probably would not be recommendable within the same statistical framework, as group sizes would become too small.[236]

Either some of the groups would have to be pooled[237], or additional criteria would have to be integrated via other statistical methods, such as linear or logit regression[238] (with considerably fewer interaction terms, if any).

---

[232] If they were not heterogeneous at all, nothing would be lost by subdividing these aggregates. Only the number of industry – legal status groups would rise.

[233] So for instance, insolvency rates of Saxonian corporations of given industry sectors *and* legal forms consistently differ from the respective German rates (they are higher), source: own calculations based on data from STATISTISCHES LANDESAMT DES FREISTAATES SACHSEN (2004). That is, differences in federal state specific default rates cannot be (completely) explained by differences in the distribution of certain industries or legal forms, which means, that adding *federal state* as explanatory variable should increase the predictive accuracy of the industry-legal-status-rating.

[234] The graphs were created utilizing data given in BLOCHWITZ, LIEBIG, NYBERG (2000, p. 12 and Appendix 2, p.3). The above mentioned authors' study is based on 140,000 financial statements of German corporations, whose financial statements were consigned to the DEUTSCHE BUNDESBANK between 1994 and 1999.
Although insolvency rates for the six separate groups differ by a factor of about 10, the predictive value of *enterprise size* is relatively low [measured on a an issuer weighted basis] (AR = 15,2%) and roughly conforms with the informational value of federal state domiciliation (cf. Table D).
One reason for the low informational value, despite the huge inter-group differences of insolvency rate, of *enterprise size* is, that most firms of the sample are characterized either by slightly above average (groups A and B) or slightly below average (group C) insolvency rates. The three remaining groups that are characterized by more *extreme* insolvency rates are relatively sparsely represented. As can be seen in Figure 20, right hand side, over 80% of all corporations are either in group A, B, or C.
Presumably, information value of *enterprise size* is much higher on a dollar-volume-weighted basis instead of an issuer-weighted basis.

[235] Comparable values for the univariate discriminatory power of *enterprise size* were found in HAYDEN (2003) for a sample of 35,000 Austrian corporations with revenues between 0.4 m to 75 m EUR (see ibid., variable 61 (net sales/CPI), AR = 11% (average by applying three different definitions of default)).

[236] When using 14 different industry sectors, 6 different legal forms, 16 federal states, (e.g.) 5 enterprise size classes and (e.g.) 3 age classes, 20,000 groups resulted. Given 3.3 m corporations and 40,000 defaults, *on average* only 165 corporations and 2.0 insolvencies per year remained.

[237] Owing to comparable industry specific insolvency rates (see Figure 17) legal form groups *GmbH&Co.KG* and *ltd* could be pooled without much loss in predictive accuracy. Dividing Federal states into only 3 instead of 16 groups might turn out to be sufficient (former East-German states vs. former West-German states (North) vs. former West-German states (South)). It also might be sufficient to consider only 2 distinctive age and size groups.

[238] In particular in cases of the cardinally scaled variables *enterprise size* and *age*, using regression equations seem to be a more promising approach.

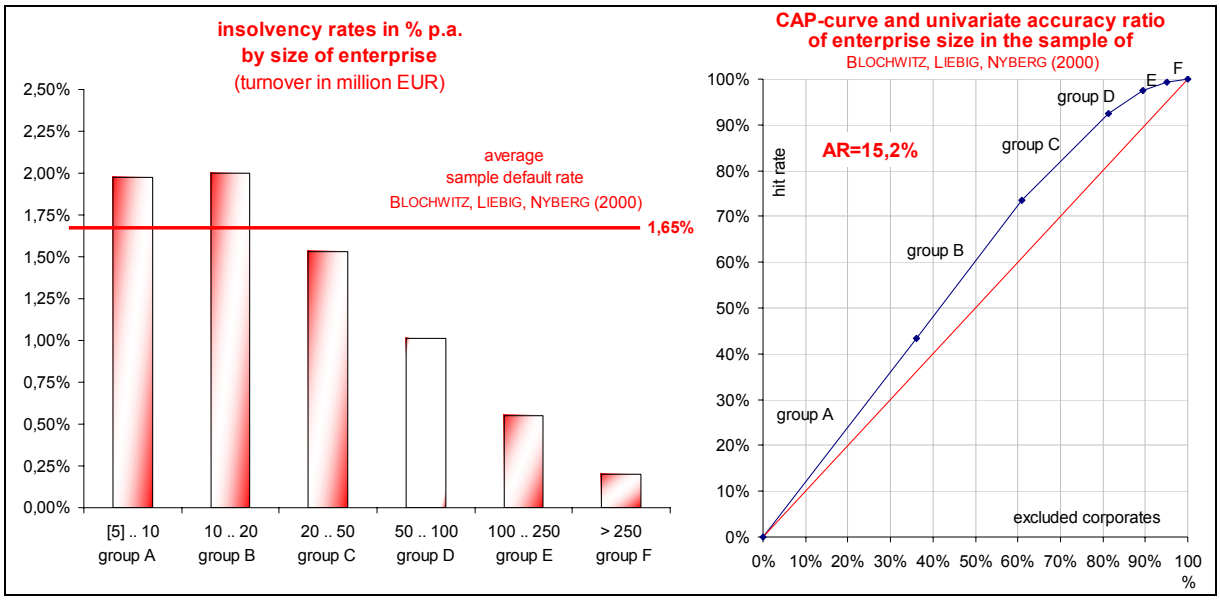

Figure 20: insolvency rates by size of enterprise (turnover in million EUR) in an empirical study (left hand size); CAP-curve and univariate discriminative power of enterprise size (CAP-Kurve) (right hand side), source: own calculations

3.3 Benchmark II: univariate discriminative power of financial ratios

Besides case-by-case decisions by loan and credit officers that are more or less guided by bank-internal rating-rules and fully formalized evaluations of "soft factors" such as *management quality* or *market position*[239], bank rating systems typically substantially rest upon formalized systems for evaluating "hard facts", in particular on statistical analyses of ratios that are derived from firms' financial statements and profits and losses accounts.[240,241]

Commercial rating models which aim at providing cheap, automated ratings for small and medium sized companies[242] and most statistical insolvency prediction models that are presented in scientific journals, are exclusively based on financial ratios. Additionally, in some of the models also variables, which were presented in the last chapter, such as *size of enterprise*, *industry classification, legal status*, or *regional provenance* are incorporated.

Even though rating agencies officially *don't* concede, that the ratings they render are essentially based on formal ratio analyses, too,[243] it was shown,

(I) that ratings (or *changes* in ratings) of the major rating agencies can at least be fairly well *reproduced*[244] or *forecasted* by simple statistical ratio-based models and

(II) that simple ratio models do achieve predictive accuracies that are comparable or superior to the accuracy that agencies' ratings are achieving on the *same samples*.[245,246]

---

[239] see for instance GRUNERT, NORDEN, WEBER (2005)

[240] see for instance BASEL COMMITTEE (2000b, p. 17ff.)

[241] ROMEIKE, WEHRSPOHN (2004b, p. 9, own translation): "In general, the bigger the bank's financial volume is, the more influential 'soft factors' are in determining the ratings. But also for multi-million dollar loans, soft factors keep their role as modifying variables. The basis is still established by (mathematical) methodical analyses of quantitative data, in particular firms' financial data."

[242] see ESCOTT, GLORMANN, KOCAGIL (2001b, p. 20)

[243] S&P (2003, p. 53). "The ratio medians are purely statistical, and are not intended as a guide to achieving a given rating level. [..] Ratios are helpful in broadly defining a company's position relative to rating categories. They are not intended to be hurdles or prerequisites that should be achieved to attain a specific debt rating. Caution should be exercised when using the ratio medians for comparisons with specific company or industry data because of major differences in method of ratio computation, importance of industry or business risk, and impact of mergers and acquisitions." etc. or S&P (2003, p. 17) "There are no formulae for combining scores to arrive at a rating conclusion. Bear in mind that ratings represent an art as much as a science."

[244] Although, the method used by AMATO, FURFINE (2004, p. 2666, see panel A) seems to be poorly calibrated (there are to few 'extreme' prognoses like AAA, AA or B and CCC/C) it achieves, by utilizing seven ratios, a reasonably precise reproduction of STANDARD AND POOR'S 7ary ratings (n=10,002): in 53% of all cases, the S&P rating is specified correctly, in 43.5% of all cases it is missed by exactly one grade, in 3.3% by two and in only 0.1% of all cases by more than two grades.
For comparison: on a 7ary scale the ratings of S&P and Moody's exactly match in 71% of all cases, in 28% they differ by one grade, in 1.1% by two and in 0.1% by more than two grades (own calculations based on GÜTTLER (2004, Appendix B)).
Based on seven ratios (plus dummy variables for legal years) BLUME, LIM, MACKINLAY (1998, p. 1404, n = 7,324) create ratings that in 57.2% - 40.2% - 2.3% - 0.3% of all cases deviate by 0 – 1 – 2 – 3 grades from S&P ratings.
The respective congruence values for S&P's and MOODYS'S ratings on a 4ary scale (as only investment grade classes AAA, AA, A, and BBB were considered) are 78.9% - 20.5% - 0.5% - 0.0% (own calculations based on GÜTTLER (2004, Appendix B)).

[245] see CAREY, HRYCAY (2001), ALTMAN, RIJKEN (2004), FONS, VISWANATHAN (2004)

[246] It's a moot point whether rating agencies really utilize non-public information, which is a frequently expressed presumption (see for instance WHITE (2001, p. 13ff.) or FONS (2002, p. 5)), and thus are *in principle*

As second benchmark – next to the industry-legal-status-rating that was presented in the last chapter, therefore the discriminatory power of ratios will be examined, that turned out to be especially discriminative in empirical studies.

With exception of the study by BEAVER (1966), who is regarded as founder of (univariate) ratio based insolvency prediction[247], in order to facilitate a statistically meaningful comparison, in the following only such studies were considered that based on an evaluation of several thousand corporations (see the remarks concerning width of confidence intervals of accuracy ratios in chapter 3.1).

Following details can be found in the respective columns of Table F:

- *Study:* author(s) and year of release of the study,
- *Number of corporations:* The number of corporations that were included in the studies is given separately for defaulted and non-defaulted companies. Sample sizes of validation samples[248], if available, are also given separately for default and non-default observations, as confidence intervals of the precision measures under consideration, accuracy ratios, do dependent thereof (and not just on the total number of corporations).
- *Database:* Data sources that were used by the studies are mentioned (bank portfolio, external databases, publications by rating agencies, etc.) as well as geographic origin, industry classification and legal status of the included companies.
- *Other:* Amongst others, in this column the applied definitions of *default* are given.
- *Prognosis accuracy:* In this column accuracy ratios of the ratios and rating models that were analyzed in the studies are given. If there were no accuracy ratio measures published in the respective studies, they were estimated, usually based on the $AR_{\alpha\&\beta}$-estimator (see chapter 2.3.3 and Appendix I). Estimated AR-values were

---

able to make more accurate insolvency predictions than external analysts. See on this MOODY's (2000, p. 9): "A common misconception is that rating agencies, through privileged access to senior management of an issuer, always gain insights that are not available through any other means. While Moody's meets with management of most issuers of securities it rates, it does not view this practice as indispensable to the formation of an accurate rating opinion. Moody's finds that meetings with senior management tend to accelerate the analytic process, because much of the information needed can be gathered quickly from one source and the answers to certain questions easily obtained, but not to change the substance of such process. If such a meeting is not available, the analyst usually has access to public filings, industry publications, and information from the issuer's competitors, suppliers, and customers through its normal course of business."
See on this also MAHONEY (2002b, p. 4): "Rating agencies routinely request nonpublic data in the course of their surveillance activities. However, unlike accounting firms, rating agencies have no authority to demand such data, and indeed many firms do not provide requested data. […] They can only work with the information which has been disclosed or which management has elected to provide."

247 RÖSLER (1988, p. 102), FALKENSTEIN, BORAL, CARTY (2003/2001, p. 9), BALCAEN, OOGHE (2004, p. 8)

248 Validation of rating models is typically carried out based on data, which was *not* already used for *parameterizing* (also *learning, training,* or *developing*) the rating model that shall be evaluated. Strictly speaking, a validation that is based on such data does not test the *predictive power* of a model, but test its ability to *reproduce the input data*. Depending on the flexibility of the model's structure (which depends on the number of included variables and parameters or additional restrictions (for instance concerning the algebraic signs certain parameters may or may not adopt)) improvements in reproduction of learning samples may be accompanied by decreased "out-of-sample" performance. ("over-fitting"), see e.g. STEIN (2002, p. 14f.), OENB (2004a, p. 48), DWYER, KOCAGIL, STEIN (2004, p. 26).

indicated by asterisks. Also given are 95%-confidence-intervals based on *approximation 2* (see chapter 3.1).[249]

[249] *Approximation 2* was used, because the data at hand did not permit the implementation of the exact formula E-H-T-(2003) that was given in chapter 3.1. Based on the examinations that were presented in the same chapter, values obtained by *Approximation 2* are divided by 1.6 for yielding more realistic values. Empirically measured (or calculated) accuracy ratios were implemented in formula F 42 but were truncated at 80%, which limits the impact of positively biased accuracy ratios on the size of the confidence interval.

Table F: studies that examined univariate insolvency predictive power of financial ratios; *to be continued on the following pages*

study	number of corporations	database	other	prognosis accuracy
BEAVER (1966)	79 defaulters, 79 non-defaulters (paired sample by industry sector, and size), 706 observations [financial statements], no separate validation sample[250]	defaulters: "MOODY'S Industrial Manual": usually big stock companies, non-defaulters: sampled from "12,000 Leading U.S. corporations" 1954-1964, USA, industrial firms	error statistics for 30 financial ratios, 7 groups of financial ratios, (p. 106), forecast: 1..5 years, def. of default: insolvency (59/79), bond default (3/79), bank account overdraft (1/79), non-payment of dividends for preferred stocks (16/79) (!)	**cash flow / total debt:** eI: 21.5%, eII: 5.1% (1 year), → AR: 73.4%-95.6%, $AR_{\alpha\&\beta}$=89%, **$AR_{actually}$ = 93.1% (*)** [251], **net income / total assets:** eI: 16.5%; eII: 11.4% (1 year) → AR: 72.1%-92.5%, **$AR_{\alpha\&\beta}$=86% (*)** **current ratio:** eI: 30.4%, eII: 8.9% (1 year) → AR: 60.7%-89.2%, **$AR_{\alpha\&\beta}$=80% (*)** $CI_{95\%,AR}$ = ± 7 PP (*)
KEENAN, SOBEHART (1999), SOBEHART, KEENAN, STEIN (2000)	530 defaulters, 9,000 corporations 54,000 observations; walk-forward-testing [252]	1989-1999, big companies ("MOODY'S proprietary databases")	AR data for one financial ratio and three/ five rating models, forecast: 1 year def. of default?	**AR distance to default** ("Merton model variant"): **67%** **AR return on assets: 53%** $CI_{95\%,AR}$ = ± 3 PP (*)
BLOCHWITZ, LIEBIG, NYBERG (2000)	2,300 defaulters, at least 30,000 non-defaulters, 139,685 observations training sample: data 1994-1996, no separate validation sample (?)	DEUTSCHE BUNDESBANK financial statements database, "representative for German middle market borrowing" (p.12), revenue at least 5 m EUR, 1993 – 1999; industrials: 44%, trade: 42.6%, construction: 5.5%, farming: 0.5%, other: 7.3%, several legal forms	6 ratios, industrial sector specific analysis (p.28-30), 3 rating models, forecast: 1 year def. of default: initiation of insolvency proceedings	**AR capital recovery ratio**: **53.7%** (manufacturing), **44.0%** (other enterprises), **47.9%** (trade), **AR equity ratio**: **44.8%** (other, trade), **AR return on equity: 40.4%** (manuf.), **37.5%** (trade), **24.9%** (other) **AR ROCE: 45.2%** (manuf.) **AR net interest rate: 44.4%** (manuf.) $CI_{95\%,AR}$ = ± 1.5 PP (*)

[250] Note: In case of univariate analyses, the absence of a validation sample does not cause severe issues of "over-fitting", because no model parameters are actually "fitted" to the empirical data of the learning sample as in the case of multivariate analyses. The only bias that might occur is, that – the more variables are tested - the more likely it is, that the best performing variable performs best on the given portfolio just by chance, and therefore is likely to perform worse on a holdout portfolio.

[251] See chapter 2.3.3 and Appendix I for determining upper and lower limits for the accuracy ratio values and $AR_{\alpha\&\beta}$. Only for one financial ratio, *cash flow to total debt,* an additional bar chart with 22 intervals for solvent and insolvent corporations was given, see BEAVER (1966, p. 92ff.), from which an accuracy ratio value of 93.1% could be obtained.

[252] "Walk-forward-testing" is a data parsimonious validation method, which abolishes the conventional strict separation of rating data sets into training and validation samples. Initially, a rating model is parameterized based on historical data by a specific date t_0 and is validated on the default information for t_1. Subsequently the model is re-estimated – by including t_1-data – and validated on t_2 data and so forth, see SOBEHART, KEENAN, STEIN (2000, p. 2f.) and STEIN (2002, p. 15f.). Therefore, what is actually being tested by this approach is not the predictive power of a concrete, fully parameterized rating model, but the predictive power of a rating methodology (like linear discriminant analysis, logit analysis, etc.) in the given environment.

study	*number of corporations*	*database*	*Other*	*prognosis accuracy*
FALKENSTEIN, BORAL, CARTY (2003/2000)	public firms (big plc): 15,805 corporations, 1,529 defaults, 130,019 observations [fin. statements], private firms (SMB): 24,718 corporations, 1,621 defaulters, 115,351 observations; training sample: up to and including 1995 data, includes ca. 25% of the default and 50% of the non-default data (p. 23)	public firms: MOODY'S default database: US, Canada, Compustat, 1980-1999; (CRD) private company data: "predominantly private, middle market corporations provided by financial institutions" (p.22), USA, Canada (?), industry sectors: industry 22%, services 27%, trade 31%, construction 6%, other 14%	AR data for 2 financial ratios, PD-profiles for 10 financial ratios, ca. 10 different rating models/ model variations forecast: 1 and 5 years, def. of default: 90 days past due, credit written down, classified as non-accrual, insolvency	**AR liabilities / assets: 61.9%** (public firms) **AR liabilities / assets: 44.8%** (private firms) **AR net income / assets: 61.5%** (public firms) **AR net income / assets: 39.9%** (private firms) $CI_{95\%,AR} = \pm 2$ PP (*) [both public and private firms; basis: complete sample ?]
SOBEHART, STEIN, MIKITYANSKA, LI (2000)	1,406 defaulters, 14,447 corporations "walk-forward-testing"	MOODY'S proprietary default database, MOODY'S proprietary ratings database, COMPUSTAT, IDC, non-financial US firms [detailed into 30 sectors], (big) plc, 1980-1999	default rates for certain quantiles for 9 different variables (p.11), AR data for one financial ratio and 5 rating models, forecast: 1 year; def. of default: "bankruptcy, chapter 11, distressed exchange, indenture modified, dividend omission, missed principal and/or interest payments"	**AR distance to default** ("Merton model variant"): **67%** **AR return on assets: 53%** $CI_{95\%,AR} = \pm 2$ PP (*)
ESCOTT, GLORMANN, KOCAGIL (2001a,b)	development sample: 485 defaulters, 4,866 corporations 11,427 observations, (1987-1992) "exceptionally good data quality" (p.5, translation); validation sample: "representative and realistic" (p.5, translation); 1,000 defaulters, 20,000 corporations, 100,000 observations (1992-1999)	bank portfolio data [development + validation] (p. 4) non stock exchange listed corporations of various industry sectors (no financials), Germany; no affiliated groups, no state ownership, no holdings or property developers; industrials: 52%, trade: 27%, services: 13%, construction: 5%, other 2% [validation sample];	graphical representations of shares of defaulters/ non-defaulters for 8 groups; for 9 ratios (p.11-15) forecast: 1 year, 5 years def. of default: insolvency, distressed exchange, moratorium, note protest	**AR debt coverage: 56%** (*), **AR net indebtness: 54%** (*), **AR trade creditors ratio: 53%** (*), **AR EBITD: 48%** (*), **AR liabilities structure: 44%** (*) **AR profit on sales: 43%** (*) **AR equity ratio: 39%** (*) **AR personnel expenses on sales: 3%** (*) **AR sales growth: 1%** (*)[253] (based on development sample), $CI_{95\%,AR} = \pm 3.5$ PP (*)

[253] Insolvent corporations are overrepresented among corporations with sizably above average as well as below average sales growth rates. Therefore the discriminative power of the variable *sales growth* could be substantially improved by applying appropriate transformations (like squaring down or using absolute values of growth rates).

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
SHUMWAY (2001)	300 defaulters, 3,182 corporations, 39,745 observations (training sample: 1962-1983: 1,822 corporations, 118 defaulters)	intersection of Compustat Industrial File and CRSP Daily Stock Return File for NYSE and AMEX stocks, defaulters: Wall Street Journal Index, Capital Changes Reporter, Compustat Research File, Directory of Obsolete Securities, Nexis, 1962 -1992, no financials;	deciles data for hit rates, 1 financial ratio, 9 rating models (p. 118, 120, 122) forecast: 1 year def. of default: initiation of insolvency proceedings	**AR net income / total assets: 66.4%** (*) $CI_{95\%,AR}$ = ± 5.5 PP (*) [basis: validation sample ?]
KOCAGIL, IMMING, GLORMANN, ESCOTT (2003)	2,156 defaulters, 19,524 corporations, 83,613 observations, partition in estimation and validation sample?	pooled data sets of the two biggest Austrian banks, predominantly small private firms (0.5-1m EUR: 29%, 1-5m EUR: 41%, 5-25m EUR: 20%, …), same exclusions as in ESCOTT, GLORMANN, KOCAGIL (2001a,b), industrials: 27%, trade: 33%, services: 17%, construction: 12%, other 11%;	graphical representations of shares of defaulters/ non-defaulters for 8 groups; for 8 ratios (p.13-17) forecast: 1 year, 5 years def. of default: inexplicit (see p.6ff): delay-in-payment [90 days], insolvency, loan loss provision ?	**AR equity-liabilities ratio 43%** (*) **AR liabilities structure 25%** (*) **AR sales revenue 37%** (*) **AR debt service 33%** (*) **AR cashflow/ liabilities 39%** (*) **AR liabilities/revenue ratio 37%** (*) **AR revenue growth 3%** (*) [non-transformed] **AR current assets structure 22%** (*) (based on development sample ?)[254], $CI_{95\%,AR}$ = ± 3 PP (*)
ENGELMANN, HAYDEN, TASCHE (2003)	3,000 defaulters, 300,000 observations validation sample: 825 defaulters, 200,000 observations	DEUTSCHE BUNDESBANK financial statements database (see above), Germany, 1994-1999	AUC-data for 3 financial ratios (p. 15, 16) and 3 rating models, confidence intervals for AUC forecast: 1 year def. of default: insolvency	**AR ordinary business income / total assets: 57.7%** (*) **AR change in (net sales / total assets): 22.5%** (*) **AR current assets/ total assets: 7.1%** $CI_{95\%,AR}$ = ± 1.5 PP (*) [basis: complete sample?]

[254] It is assumed, that 1/3 of the default data is used for the development and 2/3 for the validation sample - as in ESCOTT, GLORMANN, KOCAGIL (2001a,b).

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
HAYDEN (2003)	defaulters – observations (total vs. validation sample. [1998-1999]): insolvency-sample: 1,024 -124,479 (393 – 27,388) restructuring sample: 1,459 – 48,115 (528 – 15,683) delay in payment sample: 1,604 – 16,797 (433 – 4,649)	3 Austrian banks, ÖNB, SMB research Austria; exclusion of various inconsistent or incomplete datasets, no big corporations, minimum sales: 0.36 m EUR, 1987-1999, 81% limited liability companies, 14% limited partnerships, 4% single-ownership-companies, 1% general partnerships; 29% industrials, 25% services, 33% trade, 12% construction, 1% farming	AR data for 65 ratios [sorted by 10 ratio groups] and 3 default definitions (!) forecast: 1 year def. of default: (1) insolvency vs. (2) restructuring vs. (3) delay-in-payment [90 days]	**AR equity / assets: 44.1%** (*), **AR ordinary business income / assets: 41.1%** (*) **AR ordinary business income / operating income: 40.3% (*)** **AR cash flow / (liabilities-advances): 39.4% (*)** **AR retained earnings/assets: 39.0% (*)** [displayed are the averages of the AR-values for the three different default definitions] $CI_{95\%,AR}$ = ± 2.5 PP, ± 2 PP, ± 2 PP (insolvency, restructuring, delay-in-payment) (*) [basis: complete sample?]

(*) source of accuracy ratio data: own calculations, usually based on data concerning error rates of types I and II

Disregarding BEAVER's (1966) study, which was based on a rather small sample and an unusual definition of default, following statements concerning insolvency predictive power of single financial ratios can be made:

- The most discriminative financial ratios for small and medium-sized businesses, which predominantly emanate from ratio groups *capital structure*, *profitability,* and *debt coverage*, achieve accuracy ratios of at least 45%, partially up to 55%.
- Univariate insolvency predictions based on financial ratios are yielding for large stock companies accuracy values that exceed the respective values for SMB by about 10 percentage points. The highest discriminative power was achieved by a financial ratio ("distance-to-default"[255]) which was based on market data (and not on financial statements data) and which was the only ratio, that incorporated some measure of *risk* (here: volatility of market value).

Comparison with the results of the industry-legal-status-rating:

It is interesting to note, that the forecast accuracy of the industry-legal-status-rating roughly equals the univariate predictive forecast value of the most discriminative financial ratios– at least in cases of small and medium-sized businesses (but not for large stock companies).

If, however, the industry-legal-status-rating would be applied to the same samples, which where used for determining predictive values of the single financial ratios, its forecast quality would be worse than the performance it achieved on the basic population of all German corporations (cf. chapter 3.2), because the samples that were used were usually more homogeneously, in many cases even *perfectly* homogeneously (stock companies samples), with respect to *legal status*, which was the most important explaining variable of the industry-legal-status-rating. In particular, sole proprietorships and partnerships were usually heavily underrepresented compared to the basic population of all German companies.

Also for as far as industry classification is concerned, a further decrease in power of the industry-legal-status-rating is to be expected. Only one of the nine studies cited above was *exclusively* based on one single sector ("industrials"), while all other studies contained corporations of practically all other sectors – only *financial corporations* were explicitly excluded in most cases. However, in most of the studies, industrial sectors with "extreme" default rates – both high (construction) and low (services) – were underrepresented, while sectors with average default rates, in particular *industrials,* were notedly overrepresented.[256]

Altogether, it may be concluded that financial ratios are more precise insolvency predictors than industry classification and legal status. However, for establishing those ratios, financial statements of the corporations have to be available– which might form a much harder restriction in some fields of application than the knowledge of a corporate's industry sector and legal status, which are often already evident from the company's name.

---

[255] see SOBEHART ET AL (2000, p. 9): distance to default = (MVA – DP) / Vol_{MV}, with MVA ... market valued assets, DP ... default point, Vol_{MV} ... volatility of market value with DP = short-term debt + 0.5 * long-term debt.

[256] According to the data given in chapter 3.2, only about 11% of all 3.3 million German corporations are industrials, another 11% are building companies, 24% are trading companies and 46% of all corporations are operating in some of the service sectors (industry sectors H, K-O) and 7% in other sectors.

3.4 Benchmark III: ALTMAN'S Z-score

After BEAVERS (1966) pathbreaking *univariate* insolvency prognosis study (see chapter 3.3) was published, a further development of financial ratio based insolvency prediction methods towards *multivariate* statistical models was merely a matter of time.[257] Already shortly after, this development was carried out by ALTMAN (1968) [258], who devised the first *multivariate* insolvency prediction model for corporations[259], the so-called *Z-score model.*

Contrary to *univariate* models, *multivariate* models are not limited by the predictive power of single ratios. They can take advantage of the combined predictive power of ratio *combinations* – possibly also by implementing ratios, which have only low or no predictive value at all on a *univariate* basis. [260] And, though they are not based on an explicit and complete theory, they at least provide standardized methods for selecting and aggregating ratios that shall be included in a forecast model. [261]

As ALTMAN's Z-score model still enjoys great popularity in finance literature[262] and is still implemented in some of today's commercial rating models[263], and because its capacity was tested in so many empirical studies, its applicability as third benchmark for evaluating the predictive quality of rating models shall be evaluated subsequently. In fact, the *Z-score model* is rather about a *family* of models, as already three different versions of it exist (see below).

ALTMAN's Z-score model is based on a *multivariate linear discriminant analysis*, a categorial statistical procedure,[264] which segregates corporations into two groups ("presumably solvent" vs. "presumably insolvent").[265]

---

[257] Already BEAVER (1966) recognized this avenue, but could not find satisfactory solutions himself: "Suggestions for future research: The analysis conducted here has been a univariate analysis – that is, it has examined the predictive ability of ratios, one at a time. It is possible, that a multiratio analysis, using several different ratios […] would predict even better than the single ratios. Some preliminary efforts have undertaken to develop multiratio models, but the results have not been very encouraging in the sense, that the best single ratio appears to predict about as well as the multiratio models." (ibid, p. 100)

[258] See for instance SOBEHART ET AL (2000, p.6), FALKENSTEIN, BORAL, CARTY (2003/2000, p.9), FRERICH, WAHRENBURG (2003, p.4), BALCAEN, OOGHE (2004, p.11).

[259] However, at that time, multivariate analyses were already used for consumer credit evaluations, see ALTMAN (1968, p. 591).

[260] See for instance ALTMAN (1968, p. 594): "The variable finally established did not contain the most significant variables, amongst the twenty-two original ones, measured independently." and ibid (p. 595f.) concerning variable X_5 – which, however, was no longer considered in later model revisions: "This final ratio is quite important because […] it is the least significant ratio on a univariate basis. In fact, based on the statistical significance measure, it would not have appeared at all. However, because of its unique relationship to other variables in the model, the Sales/ Total assets ratio ranks second in its contribution to the overall discriminatory ability of the model."

[261] cf. ALTMAN (2002, p.8)

[262] FALKENSTEIN, BORAL, CARTY (2003/2000, p. 74): "The most well-known quantitative model for private firms in the U.S. is Altman's Z-score. Virtually every accounting or financial analysis book uses Z-score to demonstrate how financial statement data can be translated into an equation that helps predict default. […]"

[263] The Z-score method is being used by the rather miniscule German rating agencies CONFIRM GmbH and IKU, who offer their services for 1,200,- EUR resp. 312,- EUR per rating, see ROMEIKE, WEHRSPOHN (2004, p. 18, 27, 29).

[264] See for instance KÜTING, WEBER (2004/1993, p. 363), DIMITRAS, ZANAKIS, ZOPOUNIDIS (1996), DEUTSCHE BUNDESBANK (1999, p. 58), OeNB(2004, p. 41f.).

[265] ALTMAN (1968), however, derives a *trisection* from the model output. He assumes the existence of a middle interval that he depicts as "grey zone" (or "zone of ignorance" or "gray area", see ibid, p. 606). In the discriminant analysis system of the DEUTSCHE BUNDESBANK (1999, p.55) this interval is referred to as "B-area" or "indifference area".

The predictive power of the model, however, is not evaluated on this 0/1 classification, but on the underlying discriminant score, the "Z-score" itself, see also chapter 2.3.[266]

Based on a very small sample of 33 insolvent and 33 solvent corporations (for more details see Table G) ALTMAN's (1968) original model correctly classified 31 of the insolvent and 32 of the non-insolvent corporations (with a forecast horizon of one year). Following discriminant function was estimated:

F 54: $Z\text{-score} = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$ [267] with

X_1 … working capital/total assets,
X_2 … retained earnings/total assets,
X_3 … earnings before interest and taxes/total assets,
X_4 … market value equity/book value of total liabilities,
X_5 … sales/total assets.[268]

Corporations with Z-score values of less than 1.81 are classified as highly vulnerable to defaults, while corporations with Z-score values greater than 2.67 resp. 2.675 [269] are rather invulnerable, at least for a period of one year.

In order to extend the applicability of the Z-score model to non-listed companies, in a subsequent model revision, called Z'-score, variable X_4 was replaced by *book value* of equity/ book value of total liabilities. Following, all coefficients were reestimated:

F 55: $Z'\text{-score} = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X'_4 + 0.998X_5$ [270] with

X_1 … X_3 and X_5 … see above,
X'_4 … *book value* equity/book value of total liabilities

Because variable X_5 was considered as too "industry specific", it was abandoned in a later model revision that aimed at extending the applicability of the model to non-industrial firms. All remaining coefficients were reestimated again:

F 56: $Z''\text{-score} = 6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X'_4$ [271] with

X_1 .. X_4 … see Z'-score.

According to this score, all corporations with a Z"-score value below 1.10 are classified as vulnerable.[272] In the following, Table G gives a review of empirical studies that measured the predictive capability of the Z-score model in one of its three versions. Where available, univariate predictive power of the respectively best ratios and rating models are given, too, in order to control for sample specific opaqueness.

---

[266] See ALTMAN, SAUNDERS (1998, p.1737) for a procedure for mapping Z-scores to S&P-grades based on 750 rated US-corporations.

[267] In ALTMAN (1968, p. 594) following formula is given: $Z\text{-score} = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5$. In this formula, however, variables X_1 to X_4 are interpreted as absolute percentage values (a value of *33%*, for instance, has to be entered as *33* instead of *0.33*), see on this and on formula F 54 ALTMAN (2000/1968, p. 12f).

[268] Variables X_2, X_3, and X_4 can be interpreted as proxies „for historic, current, and future profitability", see Altman, Rijken (2004, p. 2686)

[269] cf. Altman (1968, p. 602) vs. Altman (2002, p.18)

[270] ALTMAN (2000/1968, p.25)

[271] ALTMAN, SAUNDERS (1998, p. 1737f.): "The Z"-score model is a four variable version of the Z-score approach. It was designed to reduce distortions in credit scores for firms from different industries."

[272] see ALTMAN (2002, p. 22).

Table G: survey of studies that examined predictive power of the various Z-score versions, *to be continued on next page*

study	number of corporations	database	other	Z/Z'/Z"	prognosis accuracy
ALTMAN (1968), ALTMAN (2000/ 1968)	33 defaulters, 33 non-defaulters (paired sample by industry sector and size]; training sample[273])	manufacturing companies, USA, 1946-1965, total assets: 0.7 m - 25 m US$ (on average 6.4 m US$), public limited companies	def. of default: initiation of insolvency proceedings (1968), payment default to public held debt (2000/1968)	Z Z' Z"	eI.: 6.1%, eII. 3.0%, → **AR Z-score**: 90.9% -99.3%, $AR_{\alpha\&\beta}$=97%, **$AR_{actually}$ =98.7%** (*) eI.: 9.1% eII.: 3.0%, → AR **Z'-score**: 87.9% - 98.9%, **$AR_{\alpha\&\beta}$=96%** (*), **AR Z"-score: 70% - 80% (?)** [274] $CI_{95\%,AR}$ = ± 10.5 PP (*)
KEENAN, SOBEHART (1999), SOBEHART, KEENAN, STEIN (2000)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	Z' Z"	**AR Z'-score: 43%** **AR Z"-score: 53%** **AR best single financial ratio: 53%** (return on assets [=NI/assets]) **AR best model: 73%** ("nonlinear model", "MOODY'S model") $CI_{95\%,AR}$ = ± 3 PP (*)
SOBEHART, STEIN, MIKITYANSKA, LI (2000)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	Z(?) Z"(?) (p. 14)	**AR Z-score (?): 56%** **AR Z"-score (?): 53%** **AR best single financial ratio: 53%** (return on assets) **AR best model: 73%** (MOODY'S Public Firm) $CI_{95\%,AR}$ = ± 2 PP (*)

[273] The given accuracy measures refer to the *training sample* and are therefore positively biased, in particular because of the small sample size in relation to the large number (22) of variables that were examined (ibid., p. 594). Though ALTMAN (1968, p. 601f.) was considering two validation samples, he did not simultaneously examine them regarding to classification errors of types I and II. The validation sample for examining errors of type II was clearly not representative (non-insolvent corporations with *temporary profitability problems*).

[274] In ALTMAN (2002, p. 18) three further own studies on the Z-score's performance were cited that found, for different samples of insolvent corporations, following errors of type I when applying cut-off scores of 2.675/ 1.81: 1969-1975: 86 corporations, error type I: 18%/ 25%; 1976-1995: 110 corporations, error type I: 15%/ 22%; 1997-1999: 120 corporations, error type I: 6%/ 16%. No information was given concerning data sources for defaulted companies, their legal forms, industry sectors and sizes of enterprises. Also missing was data concerning non-insolvent corporations - probably no non-insolvent corporations were included in the studies at all. It was mentioned, however, that "In 1999, the proportion of U.S. industrial firms that had Z-scores below 1.81 was over 20%." (ibid, p.18). Assuming, that also in the other evaluation periods only 20% of the solvent manufacturing U.S. corporations had Z-scores below 1.81 and assuming that the defaulted companies that were examined were representative for the basic population of all defaulted manufacturing companies following accuracy ratios result for the three studies: 55%-80% ($AR_{\alpha\&\beta}$=71%), 58%-82% ($AR_{\alpha\&\beta}$=74%) and 64%-87% ($AR_{\alpha\&\beta}$=79,5%) [own calculations based on error rates of type I and II].

study	*number of corporations*	*Database*	*other*	*Z/Z'/Z"*	*prognosis accuracy*
FALKENSTEIN, BORAL, CARTY (2003/2000)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	Z' Z" (p.75)	public firms, 1 year (p.86f.): **AR Z'-score: 51.3%** **AR best single financial ratio: 61.5%** (net income / assets) **AR Z"-score: 62.5%** **AR best model: 78.5%** "percentiles and their squares" private firms, 1 year (p.86f.): **AR Z'-score: 33.1%** **AR best single financial ratio: 44.8%** (net income / assets) **AR Z"-score: 45.5%** **AR best model: 54.1%** ("RiskCalc") $CI_{95\%,AR}$ = ± 2 PP / ± 2 PP [public firm, private firm] (*)
FALKENSTEIN, BORAL, KOCAGIL (2000)	defaulters/ corporations/ observations: US private: 1,393/ 33,964/ 139,060; Australian private: 1,447 / 27,712 / 79,877; Canada private: 271 / 4,472 / 18,538; resampling[275]	SMB, industrials: 25%, trade: 35%, services: 17%, other: 11%; many corporations with missing industry classification, 1989-1999	def. of default: 90 days past due, bankruptcy, placement on internal non-accrual list, write-down	Z" (p.20)	**AR Z"-score: 42.0%** (US, 1-2 years) **AR best model: 53.7%** ("RiskCalc US") (US, 1-2 years) **AR Z"-score: 27.7%** (Australia, 1-2 years) **AR best model: 39.7%** ("RiskCalc Australia") (Australia, 1-2 y.) **AR Z"-score: 49.7%** (Canada, 1-2 years) **AR best model: 58.3%** ("RiskCalc Canada") (Canada, 1-2 y.) $CI_{95\%,AR}$ = ± 2 PP / ± 2 PP / ± 4.5 PP [US, Australia, Canada] (*)[276]
KOCAGIL, AKHAVEIN (2001)	1,143 defaults, 41,557 firms, 170,503 obs., no separate hold-out-sample ?	Japanese private firms, exclusion of listed companies, financials, real estate, public sector, 1994-2000	def. of default: 90 days past due, bankruptcy, non-accrual list, write-down	Z' Z"	**AR Z'-Score: 41.8%** **AR Z"-Score: 30.9%** **AR best model: 69.4%** (RiskCalc Japan) $CI_{95\%,AR}$ = ± 2 PP / ± 2 PP

[275] "Resampling", here referred to as "cross-validation", is another data parsimonious model development and validation method that does not strictly separate between training and validation samples. By randomly choosing from the original dataset, two subsets are created: one development sample (here: 80%) and one validation sample (her 20%). By repeatedly creating new development and validation samples, confidence intervals for the variables under consideration, in particular for the accuracy ratio, can be derived. Resampling does also enable to examine the overall stability of the rating methodology (which financial ratios were chosen in each simulation run, which coefficients were estimated): "It is not a test of [the model] directly, but instead the general methodology, and gives a good indication of its robustness.", see ibid, p. 20.
On resampling/ bootstrapping see also STEIN (2002, p. 17ff.)

[276] FALKENSTEIN, BORAL, KOCAGIL (2000, p. 18) state following values as *"standard errors"* for the accuracy ratio (but possibly refer to the s*tandard deviation* of the *AUC*, see formula on page 26 (ibid.)): 1.8 PP, 1.5 PP, and 3.6 PP [US, Australia, Canada].

study	*number of corporations*	*database*	*other*	*Z/Z'/Z"*	*prognosis accuracy*
SHUMWAY (2001)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	Z (p. 27)	**AR Z-score: 52.4%** (*) **AR best single financial ratio: 66.4%** (*) (net income / total assets) **AR best DA-model with ALTMAN variables: 70.4%** (*) **AR best model with ALTMAN variables: 76.6%** (*) **AR best model: 83.7%** (*) ("Accounting and Market") $CI_{95\%,AR}$ = ± 4.5 PP (*)
STEIN, KOCAGIL, BOHN, AKHAVEIN (2003)	3,123 defaulters, 43,950 corporations, 326,216 observations (USA and Canada) up to and including 1999 no separation between development and validation sample	MOODY'S KMV's CRD (source: 20 banks from USA (76%) and Canada (24%)), SMB, sales 0.5 m – 8.8 m - 287,5 m in US$ (5%-50%-95% quantile), 1986-2001, sectors: industrials 26%, trade: 34%, services: 15%, construction: 9%, farming: 6%	def. of default: delay in payment >90d, non-accrual status, special loss provisioning/ reservation, charge-off, troubled debt restructuring, or internal designation of 'doubtful' for repayment.	(Z) Z" (p.8)	**AR Z"-score: 39%** (*) (from graph. representation, p.12) **AR best model: 52.5%** (*) (RiskCalc US/Canada) $CI_{95\%,AR}$ = ± 1.5 PP (*) p. 8: "Note that there are actually two Z-score models, a four variable model and a five variable model. In general, we have found the four variable model to outperform the five variable model and thus we present only the four variable model's results."
KOCAGIL, IMMING, GLORMANN, ESCOTT (2003)	see chapter 3.3	see chapter 3.3	see chapter 3.3	Z" (?)	**AR Z"-score (?): 34.5%** **AR best single financial ratio: 43%** (*) (equity-liabilities-ratio) **AR best model: 54.7%** (RiskCalc Austria)
DWYER, KOCAGIL, STEIN (2004)	3,764 defaulters, 51,000 corporations, 225,000 observations (USA and Canada) walk-forward-test	MOODY'S KMV'S CRD (see above), 1989-2002, no financials, insurances, real estates, non-profit or state-run corporations	def. of default: probably same as above	Z"(?)	**AR Z"-score (?): 42.3%**, **AR best model: 57.0%** (EDF RiskCalc V3.1) $CI_{95\%,AR}$ = ± 1.25 PP (*)

(*) source of accuracy ratio data: own calculations, usually based on data concerning error rates of types I and II

With exception of ALTMAN's (1986) original study and later studies, cited in ALTMAN (2002), that were carried out by himself, empirical results consistently show very poor results for the Z-score model(s)– not only compared to other rating models, but also compared with simple financial ratios.

Even though, ALTMAN'S Z-score model originally was calibrated on U.S. stock companies data with financial statements that date back 40 to 60 years (as seen from today), univariate predictive power of the variables that were included in the model, with exception of variable X_5, which was abandoned in a later model revision, is by no means bad, when for instance applied to financial statements of present international[277] and German/Austrian small and medium-sized businesses of various legal forms. [278,279] Bad, however, is the calibration of the model's coefficients.[280] It is actually so bad, that this *multivariate* model does not add any predictive value to the *univariate* predictive value of simple financial ratios, such as net income / total assets.

All in all, the Z-score model is *no suitable benchmark* for evaluating the performance of insolvency prediction models. Therefore, it won't be used as reference method in the following tables in chapter 3.5.

More reasonable benchmarks are formed by univariate predictive performance of sample specific most discriminative ratios or by the performance of ratios that have empirically turned out to be very discriminative (see chapter 3.3). In order to assess sample specific opaqueness, it would be desirable, if model developers also stated the predictive accuracy that can be achieved by simple, uncalibrated multivariate models.[281] Another useful benchmark would be a comparison with the performance of the rating models that are given in the next chapter. From the 24 studies that are presented there, 8 studies include complete documentations of the rating models that were used.[282] Six models thereof are neither exclusively based on variables that require (bank specific) insider information nor are too region-specific. These six models can be easily reverse engineered and applied to any validation dataset.

It might well be, that a part of the continuing popularity of ALTMAN's model among rating model developers, who still use it as a benchmark rating model, stems from the fact that ALTMAN model is so easy to beat.

---

277 See also DIMITRAS, ZANAKIS, ZOPOUNIDIS (1996, p. 496ff.) for a survey of the 35 most often used financial ratios based on a review of 59 rating models from 47 studies. Four of the five ALTMAN-ratios are among the top 10 ranks (!), with X_1: #1, X_2: #9, X_3: #4, X_4: #24, X_5: #10.

278 Based on data from HAYDEN (2003, p. 14 and 18) following rank order of univariate predictive power of 65 ratios was determined on a dataset of Austrian SMB (mean value of accuracy ratios for three different default definitions): X_1 (HAYDEN no. 15): rank 22/65, X_2 (HAYDEN no. 58) rank 11/65, X_3 (HAYDEN no. 45) rank 18/65, X_4 (HAYDEN no. 2) rank 1/65 (!) (As ALTMAN'S ratio "equity/ liabilities" was not included in HAYDEN's list, the ratio "equities/assets" (HAYDEN no. 2) was used instead. This ratio should be ordinally equivalent to "equity/ liabilities" and therefore should have the same predictive accuracy), X_5 (HAYDEN no. 42) rank 38/65. The ALTMAN's variables' relative performance is worse, when measurements are based on a sample of *German* SMB, see HAYDEN (2002, p. 73f.). ALTMAN'S variables $X_{1..5}$ there only achieve ranks 31, 63 (probably due to data errors), 11, 16, and 25 (of 65), source: own examinations.

279 Conversely, FRERICHS, WAHRENBURG (2003, p. 3, 10) find that "none of the financial ratios used for the Altman Z"-score is chosen in a stepwise selection procedure based on the Deutsche Bundesbank database". They, however, don't give a survey of the univariate predictive accuracy of the 49 examined ratios.

280 see in particular the results of SHUMWAY's (2001) study.

281 An exemplary study in this respect is given in FALKENSTEIN, BORAL, CARTY (2003/2000).

282 Some of the studies actually incorporate more than just one completely documented model. In the above context, only one model per study is counted.

3.5 Accuracy of real insolvency prediction models

In the following tables, surveys of the performance of several insolvency prediction models are given. For improving the comparability of the results, the studies were separated into three blocks: in Table H only such studies are listed, that examined large stock companies, Table I reproduces studies at international small and medium sized businesses, while in Table J studies on German/ Austrian SMB are presented. Additionally to the information that was already given in the tables of the previous chapters, it is also specified, whether the underlying rating models were completely documented in the respective studies, so that a recreation of the rating models – and thus a direct application to other samples is possible.
With exception of the studies from BEAVER (1967) and ALTMAN (1968), which are solely interesting for historical reasons, only such studies were listed that based on samples of at least 1,000 corporations.

Table H: survey of insolvency prediction studies for large stock companies; *to be continued on the following pages*

study	number of corporations	database	other	prognosis accuracy
BEAVER (1966)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR cash flow / total debt**: **93.1% (*)** $CI_{95\%,AR}$ = ± 7 PP (*) → only the univariate predictive power of financial ratios is being examined (no rating model),
ALTMAN (1968, 2000/1968)	see chapter 3.4	see chapter 3.4	def. of default: see chapter 3.4	for AR Z-, Z'- and Z"-score and $CI_{95\%,AR}$ see chapter 3.4; 3 discriminant analysis models with 5, 5, 4 variables → all three models are completely documented, model 1 requires capital market data
OHLSON (1980)	105 defaulters, 2,058 non-defaulters no separate validation sample	1970-1976, listed stock companies, industrials, USA, non-defaulters: COMPUSTAT; defaulters: Wall Street Journal Index	17 error-I-II-combinations per model (p. 130), def. of default: insolvency	2 logit models, 9 variables **AR model 1: 81.6%** (*) **AR model 2: 86.7%** (*) $CI_{95\%,AR}$ = ± 5.5 PP (*) → both models are completely documented
ZMIJEWSKI (1984)	1,600 non-defaulters, 81 defaulters thereof validation sample: 800 non-defaulters, 41 defaulters	1972-1978, all American or NY Stock Exchange listed corporations with industry sector SIC-code < 6,000 [i.e. without transportation, finance, real estates, services], default information: Capital Changes Reporter, Wall Street Journal Index, COMPUSTAT Research File	def. of default: initiation of insolvency proceedings	probit model, 3 variables panel A: mv eI: 17.1%, mv eII: 2.2% (p.71) **AR unweighted probit model:** 81%-98,5%, **$AR_{\alpha\&\beta}$= 94% (*)** $CI_{95\%,AR}$ = ± 9 PP (*) → the models presented ("model families") are completely documented

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
KEENAN, SOBEHART (1999)	see chapter 3.3, analog: SOBEHART, KEENAN, STEIN (2000), SOBEHART, STEIN, MIKITYANSKA, LI (2000)	see chapter 3.3	def. of default: see chapter 3.3	**AR return on assets: 53%** **AR distance to default:** ("Merton model variant") **67%** **AR MOODY'S Public Firm** ("nonlinear model", "MOODY'S model"): **73%,** $CI_{95\%,AR}$ = ± 3 PP (*) → the 8+1 variables used are given in SOBEHART ET AL (2000, p. 10), capital market data is required, details for aggregation are missing
FALKENSTEIN, BORAL, CARTY (2003/2000)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR liabilities / assets: 61.9%** (public fi.) **AR RiskCalc [Can/US]: 76.5%** (public fi.) $CI_{95\%,AR}$ = ± 2 PP [public firm] (*) → the 10 variables used are given, details for aggregation are missing [non-parametrical transformation+probit]
SHUMWAY (2001)	see chapter 3.3	see chapter 3.3	see chapter 3.3	discriminant analysis and logit model, 3 – 6 variables **AR net income / total assets: 66.4%** (*) **AR model with ALTMAN-variables: 76.6%** (*) **AR ZMIJEWSKI: 68.9%** (*) **AR "Accounting and Market**": **83.7%** (*) $CI_{95\%,AR}$ = ± 4,5 PP (*) → all 12 models are completely documented, some models require market data (amongst others, also the one model that achieved the best performance)
CAREY, HRYCAY (2001)	defaulters – observations: 1970-1987: 69 – 7,641 1988-1993: 82 – 3,348 1994-1998: 63 – 3,253 training sample: 1970-1987	defaulters: 01/1999 release of MOODY'S Corporate Bond Default Database; non-defaulters: corporations rated by MOODY'S, 06/1999 release of Compustat (Data von 1970-1998), US-corporations, non-financials	rating class specific defaults (absolute and relative), def. of default: see MOODY's	logit model, 4 variables **AR** (*) [283] 1970-87 1988-93 1994-98 **1970-98** **MOODY'S ratings:** 80.0% 67.4% 70.7% **75.4%** **logit model:** 86.4% 78.4% 75.0% **82.8%** $CI_{95\%,AR}$ (*) ± 7 PP ± 6,5 PP ± 8 PP **± 4 PP** → the logit-model is completely documented

[283] Accuracy ratio values for Moody's ratings were determined based on a 7ary scale (see ibid, p. 244, 267), accuracy ratio values for the logit model of CAREY, HRYCAY (2001) were determined based on a 10ary scale (see ibid, p. 220, 266).

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
KEALHOFER (2003)	(I): "all identified defaults of non-financial companies with public debt ratings from 1979 to 1990" (p.34); 657 corporations p.a., 1,066 corporations in total[284] (II): 121 defaults; on average 1,347 different corporations p.a., 1,579 different corporations in total out-of-sample-Test (?)	(I): all non-financial corporations with public S&P ratings, 1979-1990 (II): all non-financial corporations with public S&P a*nd/or* (?) MOODY's rating, 1990-1999	(I): 10 combinations of hit rates and corporations excluded, per model (graphs) (II): continuous CAP-curves for all methods (graphs), def. of default: delay in payment, restructuring (similar to S&P definition)	(I) **AR S&P's ratings: 60%** (*) **AR KMV EDF: 68%** (*) $CI_{95\%,AR}$ = ± 6,5 PP (*) (II) **AR S&P's ratings (implied): 68%** (*) **AR MOODY'S ratings (implied): 68%** (*), **AR KMV EDF: 85%** (*) $CI_{95\%,AR}$ = ± 5.5 PP (*) → incomplete naming of included variables, details for aggregation are missing
FONS, VISWANATHAN, (2004)	training sample: 14,176 observation, 371 defaults, (1989-2000): validation sample: 1,378 corporations, 60 defaults (2001) 1,167 corp. p.a. (median)	US non-financials with MOODY'S-rating (large listed companies), 1989-2002, accounting data: COMPUSTAT	PD-profiles for all used six ratios def. of default: „missed interest payment, filing for bankruptcy or completing a distressed exchange …"	**AR interest coverage 68%** [285] **AR** 1999 2000 2001 2002 2003 **total** **Bond implied ratings** 80% 74% 86% 89% 91% **83.3%** **MDP implied ratings** 74% 65% 80% 77% 83% **76.0%** **MOODY'S ratings** 72% 59% 76% 81% 88% **74.3%** $CI_{95\%,AR}$ = ± 7.5 PP (*) [only 2001 data], $CI_{95\%,AR}$ = ± 3.5 PP (*) [1989-2001 data], all six used variables are given; MDP: aggregation analog to FALKENSTEIN, BORAL, CARTY (2003/2000) [non-linear transformation + probit]

[284] Based on MOODY'S (2005, p. 12f.) data, the average issuer weighted default rate in 1979-1990 was 1.48% p.a. There are no S&P statistic prior to 1981 available, but according to S&P (2004, p.7), for 1981-1990 an average default rate of 1.31% can be determined. If one assumes a default rate of 1.4% p.a. for the KEALHOFER (2003) sample with 657 corporations per year, for 1979-1990 about 12*1.4%*657=110 defaults in 12*657=7,884 observations have to be expected.

[285] Calculated according to ibid (p. 11): „Indeed, our testing found that a model containing only interest coverage is nearly 90% as accurate as the complete model. A model containing interest coverage and leverage together displays more than 95% of the accuracy of the complete model.".

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
STANDARD & POOR'S (2004)	10,438 corporations, 66,500 observations, thereof 1,170 defaults out-of-time test[286]	1981-2003, geographic origin (2000): [287] USA 74%, Europe, Middle East, Africa 19%, Asia 3%, Latin Am. 4% industry sectors (2000): banks 12%, insurance 42% [sic!, see footnote above], industrials 46%	def. of default: see S&P (2004, p.7f.)): payment defaults on financial liabilities, distressed exchanges	**AR S&P's ratings: 83%** [288] $CI_{95\%,AR}$ = ± 1.5 PP (*) **by regions:** AR USA: 82%, AR European Union: 94% → details for aggregation are missing [289]
MOODY'S (2004)	63,500 observations, thereof 1,150 defaults out-of-time test	1983-2003; geographic origin (2000): [290] USA 67%, Europe, Middle East, Africa 19%, Asia 10%, Latin America 5%, industry sectors (2000): banks 29%, insurances 11%, industrials 60% size of enterprises (sales, *median* (!) >> 1 billion US$ [291]	def. of default: see MOODY'S (2004, p.3)): payment defaults on financial liabilities, insolvencies, distressed exchanges	**AR MOODY's ratings: 85.0%** (*) $CI_{95\%,AR}$ = ± 1.5 PP (*) **by regions:** AR North America: 81.0%, AR Europe: 95.3%, AR Asia/ Pacific: 84.7% **by (asset) size:** AR "large" corporations: 84.4%, AR "small" corporations: 74.0% **by industry sectors:** AR financial corporations: 92.3%, AR non-financial corporations: 80.5% **for information: AR sovereigns: 87.0%** [292] → details for aggregation are missing

[286] At each particular date S&P had to make a default forecast (assign a rating), only those defaults were known (and could be part of S&P's development samples), that had occurred *before* that particular date. In case S&P is regularly revising its rating methodology, this approach corresponds to the "walk-forward-method" (see above). For a description of *out-of-time testing* and other testing methods see SOBEHART, KEENAN, STEIN (2000, p. 1ff.).

[287] Data is based on BASEL COMMITTEE (2000c, p.33f). Concerning rating composition by industry sectors differing information for 1980-2001 can be found at S&P (2002, p.14): financial institutions 19.5%, consumer/service sector 13.5%, aerospace/automotive/capital goods/metal 11%, insurance/real estate 10%, utilities 9%, leisure time/media 8%, health care/chemicals 6%, energy/natural resources 5.5%, telecommunications 4.5%, transportation 4.5%, high tech/computers/office equipment 4%, forest/building products/home builders 4%.

[288] for the AR data see S&P (2004b, p. 11)

[289] "There are no formulae for combining scores to arrive at a rating conclusion. Bear in mind that ratings represent an art as much as a science. A rating is, in the end, an opinion.", S&P (2003b, p. 17)

[290] The data is based on BASEL COMMITTEE (2000c, p.33f). Concerning rating composition by regions there are slightly differing data given in MOODY'S (2004c) for 2004: North America 63%, Europe 24%, Asia/ Pacific 12%.

[291] see CANTOR (2004, p.9) for median-enterprise sizes by regions and rating classes.

Table I: survey of insolvency prediction studies for international SMB; *to be continued on the following page*

study	number of corporations	database	other	prognosis accuracy
VARETTO (1998)	training sample: á 1,920 "sound" and "unsound" corporations validation sample: á 449 "sound" and "unsound" corporations	Italian industrial corporations, source: Italian Central Bank and 50 commercial banks, 1982-1995	def. of default: inconsistent use and incomplete explanation of the term "unsound"	error I =4.7%, error II =6.5% **AR genetic algorithm:** 88.9% - 98.8% **$AR_{\alpha\&\beta}$ = 96%** (*) $CI_{95\%,AR}$ = ± 2.5 PP (*) → the variables used, were not given
FALKENSTEIN, BORAL, KOCAGIL (2000)	see chapter 3.4	see chapter 3.4	def. of default: see chapter 3.4	AR **RiskCalc US: 53.7%** (US-corporations) AR **RiskCalc US: 57.7%** (Canadian corporations) AR **RiskCalc US: 36.8%** (Australian corporations) AR **RiskCalc Australia: 39.7%** (Australian corporations) AR **RiskCalc Canada: 58.3%** (Canadian corporations) $CI_{95\%,AR}$ = ± 2 PP / ± 2 PP / ± 4.5 PP [US, Austr., Can.] (*) → the 8 variables used are named, details for aggregation are missing
FALKENSTEIN, BORAL, CARTY (2003/2000)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR liabilities / assets: 44.8%** (private firms) **RiskCalc [Can/US]: 54.1%** (private firms) → 8 variables, see above $CI_{95\%,AR}$ = ± 2 PP [private firms] (*)
WESTGAARD, WIJST (2001)	70,000 corporations (1996 cohort), 2,000 defaults (1998!)[293], thereof each 50% for development and validation sample	complete count of all Norwegian limited companies, 1995-1999 (ca. 100,000 corporations p.a.), minimum total assets: 12,500 EUR, Dun & Bradstreet register of bankruptcies, Norsk Lysningsblad, diverse data clearings,	def. of default: insolvency	error I =5%, error II < 50% [294] **AR logit model:** at least 45.0%- 90.0%, **$AR_{\alpha\&\beta}$>74%** (*) $CI_{95\%,AR}$ = ± 2 PP (*), → the model is completely documented and contains 10 variables; some of the variables are Norway specific (regional origin)

[292] Region, size class, and industry sector specific accuracy ratio values were quoted from MOODY'S (2004c, p.2). Owing to MOODY's idiosyncratic accuracy ratio definition, *conventional a*ccuracy-ratio-values are somewhat bigger than stated in the table above (see chapter 2.3.2).

[293] Being in possession of such an embracing data set (see above), restraining analysis to *just one* cohort is hardy comprehensible. The given explanation suggests, that validation results may be severely positively biased: "In the empirical analysis the 1996 accounting data and the 1998 bankruptcy data are used. Several combinations have been tried, but using the 1996-1998 data set appeared to give the best results.", see ibid, p. 344

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
KOCAGIL, AKHAVEIN (2001)	see chapter 3.4	see chapter 3.4	def. of default: see chapter 3.4	**AR RiskCalc Japan: 69.4%** $CI_{95\%,AR}$ = ± 2 PP / ± 2 PP → the 7 variables used are named, details for aggregation are missing
STEIN, KOCAGIL, BOHN, AKHAVEIN (2003)	see chapter 3.4	see chapter 3.4	def. of default: see chapter 3.4	AR **KMV Private Firm Model (PFM): 45%** (*) AR **RiskCalc US/Canada: 52.5%** (*) $CI_{95\%,AR}$ = ± 1.5 PP (*) → the variables used, were not given
DWYER, KOCAGIL, STEIN (2004)	see chapter 3.4	see chapter 3.4	def. of default: see chapter 3.4	AR **EDF RiskCalc V3.1: 57.0%** AR **RiskCalc [V1.0]: 49.5%** AR **PFM: 46.1%** $CI_{95\%,AR}$ = ± 1.25 PP (*) → the 9-11 variables used are named, details for aggregation are missing

[294] In the original text (ibid, p. 347) only very few quantile values concerning estimated default rates for defaulters and non-defaulters were given. These data imply that 95% of the defaulters had an (estimated) PD of > 0.6%, while 50% of the non-defaulters had an estimated PD of < 0.48%. Unfortunately, it was not given how many non-defaulters had an estimated PD of < *0.6%.*

Table J: survey of insolvency prediction studies German/ Austrian SMB, *to be continued on the following pages*

study	number of corporations	database	other	prognosis accuracy
FRITZ, HOSEMANN (2000)	2,580 "good accounts", 1,019 "bad accounts", monthly account data thereof 75% training sample (2,699 corporations), 25% validation sample (900 corporations)	corporate current account data of Deutsche Bank AG (?) costumers with annual turnover of between 2.5 and 25 m EUR, diverse data clearings,	def. of default: first time losses on loans	**AR linear discriminant analysis #2: 62%** (*) **AR k-nearest neighbors: 60%** (*) **AR genetic algorithm: 62%** (*) **AR neural network #2: 61.5%** (*) **AR decision tree: 54.5%** (*) $CI_{95\%,AR}$ = ± 4.5 PP (*) → the 11-12 variables that were used in the models were partially named; variables are based on monthly values concerning account-keeping (utilization of credit line, overdrafts, …); details for aggregation are missing
BLOCHWITZ, LIEBIG, NYBERG (2000),	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR capital recovery ratio**: **53.7%** (manufacturing), **44,0%** (other enterprises), **47.9%** (trade), **AR BUNDESBANK discriminant analysis: 57.4%** **AR KMV Private Firm Model: 59.7%** **AR BUNDESBANK discriminant analysis + expert system: 68.0%** [295] $CI_{95\%,AR}$ = ± 1.5 PP (*) → the 4-6 variables are named (discriminant analysis), coefficients are not given
ESCOTT, GLORMANN, KOCAGIL (2001b)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR debt coverage: 56%** (*) [development sample] **RiskCalc Germany: 59.7%** [validation sample] $CI_{95\%,AR}$ = ± 2.5 PP (*) → the 9 used variables are named, details for aggregation are missing

[295] In another study, see DEUTSCHE BUNDESBANK (1999, p. 60f), which was not included in the survey because of missing details concerning sample size, the DEUTSCHE BUNDESBANK discriminant analysis model achieved a considerably better classification result than in BLOCHWITZ, LIEBIG, NYBERG (2000): error I: 10.4% + 15.8%/2 = 18.3%; error II: 12.8%+ 19.2%/2 = 22.4% → AR: 59%-84%, **$AR_{\alpha\&\beta}$ =75%**. However, in that case the downstream expert system could improve the ratings' accuracy only marginally by three percentage points: error I: 16.5%+6.7%/2 = 19.9%, error II: 15.0% + 5.0%/2 = 17.5% → AR: 63%- 86%, **$AR_{\alpha\&\beta}$ =78%**.

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
SCHWAIGER (2002)	11,610 corporations, 1999 cohort (?) (p.437), thereof ca. 200 defaulters (1.88%), out-of-time-Test	sample: "representative for Austrian SMB" (p.434, translation), sales between 1 m and 50 m EUR, "representative" coverage of industry sectors (p.440)	default rates and distribution of corporations among 12 rating classes def. of default: CREDITREFORMINDEX >500 [296]	**AR Bonitätsindex CREDITREFORM Österreich (partial sample): 67.4%** (*) $CI_{95\%,AR}$ = ± 5 PP (*) → the 15 used variables are given, some variables can not be derived from financial statements ("credit assessment" [sic!], "payment behavior of the company's customers", …), details for aggregation are missing
LAWRENZ, SCHWAIGER (2002)	2.5 m corporations, thereof ca. 49,000-53,000 defaulters (2.0%-2.1%) p.a. out-of-time-Test	"practically the basic population [of German corporations]" (p.6, translation), 1998-2000	default rates for cohorts and pool 1998-2000, distribution of corporations among 12 rating classes as of 01/01/2001, def. of default: see above	**AR Bonitätsindex CREDITREFORM Deutschland (basic population): 50.1%** (ibid, p. 27), **51.8%** (*) (own calculation) [297] $CI_{95\%,AR}$ = ± 0.35 PP (*) [p.a.] → see above
PLATTNER (2002)	3,162 observations of a *stratified sample* (p.42); thereof 165 defaulters (?)[298], training sample = validation sample	portfolio of 30,000 KfW-debtors, different industry sectors and legal forms, 1994-1998	def. of default: insolvency	**logit model** (27 variables): eI: 27.0% eII 5.5% AR: 67.5% -94%, **$AR_{\alpha\&\beta}$=86%** $CI_{95\%,AR}$ = ± 4.5 PP (?)(*) → the model is completely documented, but comprises the variable *"appraisement of liquidity by house bank"*,
KOCAGIL, IMMING, GLORMANN, ESCOTT (2003)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR cashflow/ liabilities 39%** (*) **AR equity-liabilities-ratio 43%** (*) **AR RiskCalc Austria: 54.7%** () $CI_{95\%,AR}$ = ± 3 PP (estimation sample) / ± 2 PP (validation sample) (*) → the used 8 variables are given, details for aggregation are missing (logit model, based on transformed data)

[296] A CREDITREFORM-Bonitätsindex of 500-600 points implies the existence of "massive delays in payment" up to "hard negative characteristics", see SCHWAIGER (2002, p.438).

[297] If one assumes that all three cohorts, for which only rating class specific default rates were given, were characterized by the same *distribution* of corporations among the 12 rating classes as that from 01/01/2001, following cohort accuracy ratio values resulted: AR_{1998} = 52.9%, AR_{1999}=50.3%, AR_{2000}=51.9%.

[298] This value was estimated resting upon a later, unpublished study by PLATTNER.

study	*number of corporations*	*database*	*other*	*prognosis accuracy*
LEHMANN (2003)	400 defaulters, 19,600 non-defaulters, resampling, test sample: 19,000 non-defaulters, 67 defaulters	German SMB of a German commercial bank, different industry sectors, "about two thirds of the companies reported annual turnovers up to 5 million EUR." (S.7)	def. of default: first-time loan loss provision	**AR financial ratios rating (I): 43.6%** (*) **AR checking account (II): 54.2%** **AR analyst (soft factor) rating (III): 43.8%** **AR combination of (I) and (II): 58.4%** **AR combination of (I), (II) and (III): 62.4%** → models are not documented $CI_{95\%,AR}$ = ± 9PP (*) [299]
ENGELMANN, HAYDEN, TASCHE (2003)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR ordinary business income / total assets: 57.7%** (*) **AR best logit model: 62.4%** [300] $CI_{95\%,AR}$ = ± 2.5 (*) [301] → all three logit models (with 4 variables) are completely documented
HAYDEN (2003)	see chapter 3.3	see chapter 3.3	def. of default: see chapter 3.3	**AR equity / assets: 44.1%** (*), **AR best logit model: 58.9%** $CI_{95\%,AR}$ = ± 3.5 PP, ± 3 PP, ± 3.5 PP (insolvency, restructuring, delay in payment) (*) [302] → all three logit models (with 6-9 variables) are completely documented

[299] Based on a resampling study, LEHMANN (2003, p. 22) determines a standard deviation of the ROC_{AUC} of 2.11 PP. Assuming Gaussian distributed *Accuracy Ratios,* this corresponds with a 95%-confidence interval for the accuracy ratio of ± 2.11 PP * 2 * 1.96 = ± 8.3 PP.

[300] Based on the same dataset, HAYDEN (2002, p. 75ff.) achieves an AR-value of 70% with a logit model with 12 variables.

[301] Based on a resampling (or *bootstrapping*) analysis, ENGELMANN, HAYDEN, TASCHE (2003, p. 18) determine an AR-confidence interval of ±2,49 PP for model 1.

[302] Based on data given by HAYDEN (2003, p. 33), following 95%-confidence intervals result: $CI_{95\%,AR}$ = ± 2.4 PP, ± 4.0 PP, ± 4.9 PP (insolvency, restructuring, delay in payment). [The confidence interval for the insolvency samples probably was determined based on the complete sample, not just on the validation sample.]

Annotation to Table H, insolvency prediction studies on large stock companies
It is interesting to note, that the enormous effort commercial rating agencies are putting into their rating processes [303] – and the enormous fees they charge in return[304,305] – are not adequately reflected by the *accuracy* of their prognoses.[306] On principle, the one-year-accuracy ratios of about 85% they achieve are quite good or even above-average. But the cited studies exclusively base on U.S. non-financial corporations – and just here, i.e. on a comparable basis, rating agencies come off particularly bad.[307]

Annotation to Table J, insolvency prediction studies on German/ Austrian SMB
Similar unfavorable findings with respect to the relative accuracy of their insolvency predictions have to be noted for commercial rating models for (German and Austrian) SMB.

In a current market study[308], 3 out of 15 rating models for German SMB were evaluated as being "very good": CREDITREFORM-Bilanzrating, CREDITREFORM-Bonitätsindex and MOODY'S RiskCalc Germany. However, the predictive accuracy of the CREDITREFORM-*Bonitätsindex* [Germany] is even worse than the predictive accuracy of the simple industry-legal status rating (!), that was presented in chapter 3.2 – and which is based on the same sample of corporations.[309,310] With an accuracy ratio of 60%, the multivariate, financial ratio based model *MOODY'S RiskCalc Germany* –achieves a somewhat better predictive quality - based on a different sample, though. But the model's predictive accuracy is only marginally higher than the predictive accuracy of the best single financial ratio, that was used within the model, and it is not higher than the predictive accuracy of some of the competing rating models that are available for free, and which were based on comparable samples in terms of origin of data and sample size (see Table J).[311] Likewise, the rating model CREDITREFORM-*Bilanzrating*[312] achieves only an accuracy ratio of 60% as well.[313]

---

303 See for instance S&P (2003b) who describe elements of their rating processes on more than 100 pages, while describing simple statistical models, such as discriminant or logit analysis requires just a few lines.

304 WHITE (2001, p. 14): "Both Moody's and S&P have the following 'list prices' for the requested ratings: 3.25 basis points on [bond] issues up to $500 million, with a minimum fee of $25,000 and a maximum of $125,000 (S&P) or $130,000 (Moody's); both charge an additional 2 basis points on amounts above $500 million (S&P caps the amount at $200,000; it also has a one-time fee of $25,000 for first-time issuers). Both offer negotiated rates for frequent issuers and offer quarterly charges on amounts outstanding for issuers of commercial paper."

305 TREACY, CAREY (2000/1998, p. 911): "S&P's fee for rating a public corporate debt issue ranges from $25,000 to more than $125,000, with the usual fee being 0.0325 percent of the face amount of the issue. Fees are a reflection of the substantial resources the agencies typically devote to producing each rating, especially the initial rating.", see also CANTOR, PACKER (1994, p. 4)

306 Representatives of rating agencies would try to qualify the relevance of these findings by pointing out to the better relative performance of agency ratings for longer forecast horizons, to the greater stability of agency ratings when compared to other forecasts (market or accounting based rating models) or would demand to include additional features of the agencies' rating systems into the accuracy analysis, such as outlook or watchlist status of their ratings, see CANTOR, MANN (2003, p. 25,27), FONS, VISWANATHAN (2004, p. 10), and HAMILTON (2004, p. 11)

307 See on this in particular the studies of CAREY, HRYCAY (2001) and KEALHOFER (2003).

308 see ROMEIKE, WEHRSPOHN (2004a and 2004b)

309 In 1998-2000 the CREDITREFORM-Bonitätsindex achieved AR-values of 51%, see Table J, while the industry-legal status rating achieved AR-values of 53%-55% in 1998-2001, see Table E in chapter 3.2. Only in 2002 and 2003 its predictive accuracy decreases to values of around only 45%.

310 costs: yearly membership fee: 400 – 500 EUR plus 16 – 20 EUR per rating, see ROMEIKE, WEHRSPOHN (2004b, p.23)

311 costs: 'individual price system' with a basic charge of 20,000US$, see ROMEIKE, WEHRSPOHN (2004b, p.24)

312 Contrary to the CREDITREFORM-Bonitätsindex the CREDITREFORM-Bilanzrating model requires the existence of financial statements.

313 costs: 549,- EUR per rating see ibid. (p.22), source of accuracy-ratio-value: own calculations based on the graphical representation of the CAP-curve given in ibid (p. 20).

4 Conclusion

It has been the purpose of this discussion paper to contribute in improving the comparability of forecast quality measures of insolvency prediction studies. For being able to learn from *good,* i.e. *highly predictive*, insolvency prediction models, it is first of all necessary, to correctly identify them. Owing to the great number of empirically used measures for gauging *predictive quality*, this is by no way a trivial task. In the discussion paper at hand, it was shown that ordinal predictive quality measures, in particular the *accuracy ratio*, are suited best for sample spanning comparisons. Also several methods for extending the range of application of these measures were presented.

A sample spanning comparison of accuracy measures is substantially aggravated by the impact of both systematic and unsystematic "disturbance factors". Of outstanding importance in this respect is the influence of *enterprise size,* respectively of those variables that are closely associated with it in empirical data samples, as *data quality* and possible *preselection biases*. Not at least owing to disclosure requirements for large and listed companies, high-quality financial statement and default databases are available for developing and validating insolvency prediction models for *this* group. Additionally, for listed companies capital market data is available - like actual stock and volatility of market value of equity or bond prices – which can serve as a valuable auxiliary input to rating models. The situation in cases of small and medium businesses is less advantageous. Besides aspects of (formal) quality of financial statements, that is already worse than in case of large companies, serious deficits can result from the fact that most insolvency prediction models for SMB are based on *bank* portfolios, which raises serious issues of preselection, as was outlined in chapter 3 and shown in Appendix V.

With accuracy ratio values ranging from 75% to more than 90% the surveyed studies show that insolvencies for *large corporations* can be predicted fairly well with financial ratio based rating models alone. The precision achieved is formidable, compared with insolvency prediction for SMB (50%-60%), private costumers (over 50%) [314], a simple industry-legal status rating (50%), univariate predictive power of single ratios (55%-65% [when applied to large corporation data sets]) – but also when compared with the quality of predictions of completely different fields of application (50%-95%).[315]

Some of the completely documented and free of charge rebuildable rating models, that were presented in the survey, are achieving better prediction results than prevailing commercial rating models.

---

[314] For comparison: based on a validation sample of 410,000 data sets, the SCHUFA-Score (2001), a rating model for private costumers, achieves an accuracy-ratio of 53%, see FAHRMEIR, HENKING, HÜLS (2002, p. 27). Contrary to FALKENSTEIN, BORAL, CARTY (2003/2000, p.9) a superiority in terms of data availability in case of private customers compared to corporate customers does not result in a superiority of insolvency prediction models that are derived from the respective data sets. See also the comparably poor performance of the CREDITREFORM Bonitätsindex for which millions of solvent corporations' datasets and ten thousands of insolvent corporations' datasets are available.

[315] Rain forecasts achieve accuracy ratios of 54%-80% (median: 64%); aptitude tests, for instance for predicting college graduation, achieve AR-values of 60%-86%. In case of medical imagery, classical chest films achieve AR-values of around 96% and computed-tomography-examinations (brain) AR-values of 94%, while AR-values obtained by polygraph lie detectors (both in "field studies" and laboratory experiments) are around 73%, see SWETS (1988, p. 1288ff. and the literature there cited). Predictions of subsequent offenses of violent delinquents achieve AR-values of about 50% and AIDS-tests AR-values of 84% to 94%, see SWETS, DAWES, MONAHAN (2000, p. 10 and 16) [all values were given in AUC]. Note: some of the examples cited above do not measure the quality of *predictions*, but the quality of *classifications.*

Bibliography

ALTMAN, E.I. (1968): "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", in Journal of Finance, vol. 23 (4), pp. 589-610, September 1968

ALTMAN, E.I., HALDEMAN, R.G., NARAYANAN, P. (1977): "ZETA analysis: A new model to identify bankruptcy risk of corporations", in Les Cahiers de Recherché, 1977, pp. 1- 45; likewise published in Journal of Banking and Finance, vol. 1 (1), pp. 29- 54, 1977

ALTMAN, E.I., MARCO, G., VARETTO, F. (1994): "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)", in Journal of Banking and Finance, vol. 18, pp. 505-529, 1994

ALTMAN, I., SAUNDERS, A. (1998): "Credit Risk Measurement: Developments over the last 20 years", in Journal of Banking and Finance, vol. 21, pp. 1721-1742, 1998

ALTMAN, E. I.(2000/1968): "Predicting financial distress of companies: revisiting the Z-score and Zeta ® models", Working Paper, Stern School of Business, New York University, http://www.stern.nyu.edu/~ealtman/Zscores.pdf, (25.04.2005), 07/2000

ALTMAN, E. I. (2002): "Revisiting Credit Scoring Models in a Basel 2 Environment", Working Paper, Stern School of Business, New York University, http://www.stern.nyu.edu/fin/workpapers/papers2002/pdf/wpa02041.pdf (22.2.2005), 05/2002

ALTMAN, E.I., RIJKEN, H.A. (2004): "How rating agencies achieve rating stability", in Journal of Banking and Finance, vol. 28, pp. 2679-2714, 2004

AMATO, J.D. FURFINE, C.H. (2004): "Are credit ratings procyclical?", in Journal of Banking and Finance, vol. 28, pp. 2641-2677, 2004

AZIZ, M.A., DAR, H.A. (2004): "Predicting Corporate Financial Distress: Whither do We Stand?", http://www.lut.ac.uk/departments/ec/Reasearchpapers/2004/JOBF.pdf (20.1.2005), University of Loughborough Working Paper, 2004

BALCAEN, S. OOGHE, H. (2004): "35 Years of studys on Business Failure: An Overview of the Classic Statistical Methodologies and their Related Problems", Vlerick Leuven Gent Working Paper Series 2004/15, http://www.vlerick.be/research/workingpapers/vlgms-wp-2004-15.pdf (5.1.2005), 2004

BANGIA, A., DIEBOLD, F. X., SCHUERMANN, T. (2002): "Ratings Migration and the Business Cycle, With Application to Credit Portfolio Stress Testing", in Journal of Banking and Finance, vol. 26, pp. 445-474, 2002

BASEL COMMITTEE ON BANKING SUPERVISION (ED.) (2000a): "Summary of Responses Received on the Report 'Credit Risk Modelling: Current Practices and Applications'", Bank for International Settlements, http://www.bis.org/publ/bcbs71.pdf (24.8.2004), May 2000

BASEL COMMITTEE ON BANKING SUPERVISION (ED.) (2000b): "Range of Practice in Banks' Internal Ratings Systems", Discussion Paper, Bank for International Settlements, http://www.bis.org/publ/bcbs66.pdf (28.8.2004), January 2000

BASEL COMMITTEE ON BANKING SUPERVISION (ED.) (2000c): "Credit Ratings and Complementary Sources of Credit Quality Information", Working Paper #3, http://www.bis.org/publ/bcbs_wp3.pdf (30.8.2004), 2000

BASEL COMMITTEE ON BANKING SUPERVISION (ED.) (2001): "The internal ratings-based approach: supporting document to the New Basel Capital Accord", Consultative document, Bank for International Settlements, http://www.bis.org/publ/bcbsca05.pdf (24.8.2004), January 2001

BASEL COMMITTEE ON BANKING SUPERVISION (ED.) (2004): „Internationale Konvergenz der Kapitalmessung und Eigenkapitalanforderungen, Überarbeitete Rahmenvereinbarung", Übersetzung der Deutschen Bundesbank, http://www.bundesbank.de/download/-bankenaufsicht/pdf/eigenkapitalempfehlung_de.pdf (7.9.2004, in German), English version available at http://www.biz.org/publ/bcbs107.pdf (5.4.2005), June 2004,

BASEL COMMITTEE ON BANKING SUPERVISION (ED.) (2005): "Studies on the Validation of Internal Rating Systems", Working Paper No. 14, http://www.bis.org/publ/-bcbs_wp14.pdf (25.5.2005), revised version, 05/2005

BASEL COMMITTEE: see BASEL COMMITTEE ON BANKING SUPERVISION

BEAVER, W.H. (1966): "Financial Ratios as Predictors of Failure", Empirical Research in Accounting: Selected studys, in a supplement to Journal of Accounting Research (4), pp. 71- 111, 1966

BLÖCHLINGER, A., LEIPPOLD. M. (2005/2004): "Economic Benefit of Powerful Credit Scoring", National Centre of Competence in Research Financial Valuation and Risk Management, University of Zürich, Working Paper No. 216, http://www.nccr-finrisk.unizh.ch/media/pdf/wp/WP216_2.pdf (27.5.2005), 04/2005

BLOCHWITZ, S., LIEBIG, T., NYBERG, M. (2000): "Benchmarking Deutsche Bundesbank's Default Risk Model, the KMV® Private Firm Model® and Common Financial Ratios for German Corporations", Workshop on Applied Banking Research, BASEL COMMITTEE ON BANKING SUPERVISION, http://www.bis.org/bcbs/oslo/liebigblo.pdf (16.8.2004), 2000

BLOCHWITZ, S., HOHL, S., TASCHE, D., WEHN, C.S. (2004): "Validating Default Probabilities on Short Time Series", http://www.chicagofed.org/publications/capital_and_market_-risk_insights/2004/validating_default_probabilities.pdf (1.6.2005), Working Paper, Federal Reserve Bank of Chicago, 05/2004

BLUME, M.E., LIM, F., MACKINLAY, A.C. (1998): "The Declining Credit Quality of U.S. Corporate Debt: Myth or Reality", in Journal of Finance, vol. LIII, No. 4, pp. 1389-1413, August 1998

BOHN, J., ARORA, N., KORABLEV, I. (2005): "Power and Level Validation of the EDF™ Credit Measure in North America", Moody's KMV, Working Paper, http://www.moodyskmv.-com/research/whitepaper/EDF_Validation_NorthAmerica.pdf (24.5.2005), 03/2005

BRIER, G.W. (1950): "Verification of forecasts expressed in terms of probability", in Monthly Weather Review 78, pp. 1 – 3, 1950

CANGEMI, B., SERVIGNY, A. DE, FRIEDMAN, C. (2003): "Standard & Poor's Credit Risk Tracker for Private Firms, Technical Document", S&P Working Paper, http://www.-standardandpoors.co.jp/spf/pdf/rev_CRTTechDocument20031126.pdf (3.3.2005), 2003

CANTOR, R., PACKER, F. (1994): „The Credit Rating Industry", Quarterly Review, Federal Reserve Bank of New York, vol. 19 (2), pp. 1-26, 1994

CANTOR, R., C. MANN (2003): "Measuring the Performance of Corporate Bond Ratings", Special Comment, Report #77916, Moody's Investor's Services, 04/2003

CANTOR, R. (2004): "Measuring the Quality and Consistency of Corporate Ratings across Regions", Moody's Investors Service, Special Comment, Report # 89168, 11/2004

CAREY, M. S. HRYCAY, M. (2001): "Parameterizing Credit Risk Models with Rating Data", in Journal of Banking and Finance, vol. 25 (1), pp. 197-270, 2001

CROSBIE, P., BOHN, J. (2003): "Modeling Default Risk", Modeling Methodology, Moody's KMV, http://www.moodyskmv.com/research/whitepaper/ModelingDefaultRisk.pdf (16.7.2004), 2003

CROUHY, M., GALAI, D., MARK, R. (2001): "Prototype risk rating system", in Journal of Banking and Finance, vol. 25, pp. 47-95, 2001

DECISIONEERING (ED.) (2000): "Crystal Ball 2000, User Manual®", Decisioneering, Inc, Denver, Colorado, 2000

DEUTSCHE BUNDESBANK (ED.) (1999): "Zur Bonitätsbeurteilung von Wirtschaftsunternehmen durch die Deutsche Bundesbank", in DEUTSCHE BUNDESBANK, Monatsbericht Januar 1999, pp. 51-64 (in German), English Version available at http://www.bundesbank.de/-download/volkswirtschaft/mba/1999/199901mba_art03_credworth.pdf (5.4.2005), January 1999

DEUTSCHE BUNDESBANK (ED.) (2001): "Die neue Baseler Eigenkapitalvereinbarung (Basel II)", in DEUTSCHE BUNDESBANK, Monatsbericht April 2001, pp. 15-44 (in German), English version available at at http://www.bundesbank.de/download/volkswirtschaft/-mba/2001/200104mba_art01_baselaccord.pdf (22.6.2005), April 2001

DEUTSCHE BUNDESBANK (ED.) (2003): "Validierungsansätze für interne Ratingsysteme", in Deutsche Bundesbank, Monatsbericht September 2003, pp. 61-74 (in German), English version available at http://www.bundesbank.de/download/volkswirtschaft/mba/2003/-200309_en_rating.pdf (5.4.2005), September 2003

DEUTSCHE VEREINIGUNG FÜR FINANZANALYSE, KOMMISSION RATING STANDARDS, ARBEITSKREIS 2 „VALIDIERUNG" (DVFA, ED.) (2004): „DVFA – Validierungsstandards", in Finanz Betrieb 09/2004, pp. 596-601 (in German), 2004

DIMITRAS, A.I., ZANAKIS, S.H., ZOPOUNIDIS, C. (1996): "A survey of business failures with an emphasis on prediction methods and industrial applications, Theory and methodology", in European Journal of Operational Research, vol. 90, pp. 487-513, 1996

DVFA: see DEUTSCHE VEREINIGUNG FOR FINANZANALYSE

DWYER, D.W., STEIN, R. M. (2003): "Inferring the Default Rate in a Population by Comparing Two Incomplete Default Databases", Technical Report #021216, Moody's KMV, 03/2003

DWYER, D. W., KOCAGIL, A. E., STEIN, R. M. (2004): "The Moody's KMV EDF™ RiskCalc™ v3.1 model, Next-Generation Technology for Predicting Private Firm Credit Risk", http://www.moodyskmv.com/research/whitepaper/EDF_RiskCalc_v3_1.pdf, (16.7.2004), Moody's KMV Company, 05/2004

ELTON, E. J., GRUBER, M. J., AGRAWAL, D., MANN, C. (2004): "Factors affecting the valuation of corporate bonds", in Journal of Banking and Finance, vol. 28, pp. 2747-2767, 2004

ENGELMANN, B., HAYDEN, E., TASCHE, D. (2003): "Measuring the Discriminative Power of Rating Systems", Deutsche Bundesbank, Discussion Paper, Series 2: Banking and Financial supervision, No 01/2003, http://www.bundesbank.de/bankenaufsicht/banken-aufsicht_diskussionspapiere.en.php (19.7.2004), 2003

ENGLISH, W. B., NELSON, W. R. (1998): "Bank Risk Rating of Business Loan", Board of Governors of the Federal Reserve System FEDS, Paper No. 98-51, http://ssrn.com/-abstract=148753 (22.5.2005), 12/1998

ESCOTT, P., GLORMANN, F., KOCAGIL, A.E. (2001a): „Moody's RiskCalc™ for nicht börsennotierte Unternehmen: Das Deutsche Modell" Moody's Investors Service, Rating Methodology (in German), 06/2001

ESCOTT, P., GLORMANN, F., KOCAGIL, A.E. (2001b): "RiskCalc™ for Private Companies: The German Model", Moody's KMV, Modeling Methodology, http://www.moodyskmv.-com/research/whitepaper/720441.pdf (18.02.2005, in German), English version available at http://riskcalc.moodysrms.com/us/research/crm/720431.pdf (5.4.2005), 11/2001

FAHRMEIR, L., HENKING, A., HÜLS, R. (2002): „Methoden zum Vergleich verschiedener Scoreverfahren am Beispiel der SCHUFA-Scoreverfahren", in Risknews 11/2002, pp. 20-29 (in German), 2002

FALKENSTEIN, E., BORAL, A., KOCAGIL, A. E. (2000): "RiskCalc™ For Private Companies II: More Results and the Australian Model", Moody's Investors Service, Rating Methodology, Report # 62265, 12/2000

FALKENSTEIN, E., BORAL, A., CARTY, L. (2003/2000): "RiskCalc™ For Private Companies", Modeling Methodology, Moody's KMV, 2003

FEELDERS, A.J. (2000): "Credit Scoring and Reject Inference with Mixture Models", in International Journal of Intelligent Systems in Accounting, Finance and Management, vol. 9, pp. 1-8, 2000

FISCHER, A. (2004): „Qualitative Merkmale in bankinternen Ratingsystemen: eine empirische Analyse zur Bonitätsbeurteilung von Firmenkunden", Uhlenbruch Verlag, Bad Soden am Taunus, 2004, PhD thesis, University of Münster (in German), 2004

FONS, J. S. (2002): "Understanding Moody's Corporate Bond Ratings and Rating Process", Moody's Investors Service, Report # 74982, 05/2002

FONS, J. S., VISWANATHAN, J. (2004): "A User's Guide to Moody's Default Predictor Model: an Accounting Ratio Approach", Moody's Investors Service, Report # 90127, 12/2004

FRANKS, J., SERVIGNY, A. DE, DAVYDENKO, S. (2004): "A Comparative Analysis of the Recovery Process and Recovery Rates for Private Companies in the U.K., France, and Germany", STANDARD AND POOR'S Risk Solution, 06/2004

FRERICHS, H., WAHRENBURG, M. (2003): "Evaluating internal credit rating systems depending on bank size", Working Paper Series: Finance and Accounting, Johann Wolfgang Goethe-Universität Frankfurt Am Main, No. 115, http://www.wiiw.de/publikationen/-Evaluationinternalcreditrating898.pdf (25.8.2004), 09/2003

FRITZ, S., HOSEMANN, D. (2000): "Restructuring the Credit Process: Behaviour Scoring for German Corporations", in International Journal of Intelligent Systems in Accounting, Finance and Management, vol. 9, pp. 9–21, 2000

GRICE, J.S., DUGAN, M.T. (2001): "The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher", Review of Quantitative Finance and Accounting, vol. 17, pp. 151-166, 2001

GRUNERT, J., NORDEN, L., WEBER, M. (2005): "The role of non-financial factors in internal credit ratings", in Journal of Banking and Finance, vol. 29, pp. 509-531, 2005

GUJARATI, D. (1999/1992): "Essentials of Econometrics", Irwin/McGraw-Hill, 2nd edition, 1999

GÜNTHER, T., GRÜNING, M. (2000): „Einsatz von Insolvenzprognoseverfahren bei der Kreditwürdigkeitsprüfung im Firmenkundenbereich", in Die Betriebswirtschaft, Heft 1/2000, pp. 39-59 (in German), 2000

GÜNTERBERG, B., WOLTER, H.-J. (2003): „Unternehmensgrößenstatistik 2001/2002 - Daten und Fakten", IfM-Mat. Nr. 157, Bonn, 2003, http://www.ifm-bonn.org/dienste/dafa.htm (30.11.2004, in German), 2003

GUPTON, G.M., STEIN, R. M. (2002): "LossCalc™: Model for Predicting Loss Given Default (LGD)", Moody's KMV Company, Modeling Methodology, 2002

GUPTON, G. M., STEIN, R. M. (2005): „LossCalc V2: Dynamic Prediction of LGD Modeling Methodology", Moody's KMV, Working Paper, http://www.moodyskmv.com/-research/whitepaper/LCv2_DynamicPredictionOfLGD.pdf (25.5.2005), 01/2005

GÜTTLER, A. (2004): "Using a Bootstrap Approach to Rate the Raters", Working Paper Series: Finance and Accounting, Johann Wolfgang Goethe-Universität Frankfurt, No. 132/ 2004, http://opus.zbw-kiel.de/volltexte/2004/2343/pdf/835.pdf (11.1.2005), forthcoming in Financial Markets and Portfolio Management, 10/2004

HAMERLE, A., RAUHMEIER, R., RÖSCH, D. (2003): "Uses and Misuses of Measures for Credit Rating Accuracy", Version 04/2003, Working Paper, University of Regensburg, http://-defaultrisk.com/pdf__files/Uses_n_Misuses_o_Measures_4_Cr_Rtng_Accrc.pdf (3.1.2005), 2003

HAMILTON, D.T. (2004): "Rating Transitions and Defaults Conditional on Watchlist, Outlook and Rating History", Moody's Investors Service, Special Comment, Report # 81068, 02/2004

HARTUNG, J. (1991): "Statistik: Lehr- und Handbuch der angewandten Statistik", Oldenbourg Verlag, München, Wien, 8th edition (in German), 1991

HAYDEN, E. (2002): "Modeling an Accounting-Based Rating System for Austrian Firms", PhD thesis, faculty of business and computer sciences, University of Vienna, www.bwl.univie.ac.at/bwl/fiwi3/members/hayden/diss.pdf (9.5.2005), 2002

HAYDEN, E. (2003): "Are Credit Scoring Models Sensitive With Respect to Default Definitions? Evidence from the Austrian Market", Working Paper, University of Vienna http://www.bwl.univie.ac.at/bwl/fiwi3/members/hayden/DefaultDefinitions.pdf (16.8.2004), 2003

HULL, J.C. (2003): "Options, futures and other derivatives", 5th edition, Prentice-Hall, Upper Saddle River, 2003

HUSCHENS, S., HÖSE, S. (2003): „Sind interne Ratingsysteme im Rahmen von Basel II evaluierbar? – Zur Schätzung von Ausfallwahrscheinlichkeiten durch Ausfallquoten“, in Zeitschrift für Betriebswirtschaft zfb, vol. 73 (2), pp. 139-168 (in German), 2003

JANKOWITSCH, R., PICHLER, S., SCHWAIGER, W. S. A. (2003): “Rating Granularity and Basel II Capital Requirements”, TU Wien, Working Paper, 11/2003, http://www.wu-wien.ac.at-/inst/ikw/hp/download/publ/basel.pdf (10.2.2005), under review in Journal of Banking and Finance, 2003

JORDÁO, F., STEIN, R.M. (2003): “What is a more powerful model worth?”, Technical Report, Report #030124, Moody’s KMV Company, 2003

KEALHOFER, S. (2003): “Quantifying Credit Risk I: Default Prediction”, in Financial Analysts Journal January/February 2003, pp. 30–44, 2003

KEENAN, S. C. (1999): “Predicting Default Rates: A Forecasting Model for Moody's Issuer-Based Default Rates”, Moody’s Investors Service, Special Comment, Report # 47729, 08/1999

KEENAN, S.C., SOBEHART, J.R. (1999): “Performance Measures for Credit Risk Models”, Moody’s Investors Service, Research Report # 1-10-10-99, 1999

KOCAGIL, A. E., AKHAVEIN, J. D. (2001): “Moody's RiskCalc™ for Private Companies: Japan”, Moody’s Investors Service, Rating Methodology, Report # 73072, 12/2001

KOCAGIL, A. E., IMMING, R., GLORMANN, F., ESCOTT, P. (2003): “RiskCalc™ For Private Companies: The Austrian Model”, Moody’s KMV, Modeling Methodology, http://www.moodyskmv.com/research/whitepaper/atmethod_german.pdf (10.05.2005) (in German), 11/2003

KRAFT, H., KROISANDT, G., MÜLLER, M. (2004): “Redesigning Ratings: Assessing the Discriminatory Power of Credit Scores under Censoring”, Fraunhofer Institut for Techno- and Wirtschaftsmathematik (ITWM), Working Paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=302137 (3.3.2005), 09/2004

KRAHNEN, J.P., WEBER, M. (2001): “Generally accepted rating principles: A primer”, in Journal of Banking and Finance, vol. 25 (1), pp. 3-23, 2001

KRÄMER, W. (2003): „Die Bewertung und der Vergleich von Kreditausfall-Prognosen“, in Kredit und Kapital, vol. 36 (3), pp. 395-410 (in German), 2003

KRÄMER, W., GÜTTLER, A. (2003): “Comparing the accuracy of default predictions in the rating industry: The case of Moody’s vs. S&P”, Technical Report-Reihe des SFB 475 Nr. 23 (Universität Dortmund), http://www.wiwi.uni-frankfurt.de/schwerpunkte/-finance/wp/332.pdf, (13.4.2005), 2003

KÜTING, K., WEBER, C.-P.: (2004/1993): „Die¬ Bilanzanalyse, Lehrbuch zur Beurteilung von Einzel- and Konzernabschlüssen“, 6th edition (in German), Schäffer-Poeschel, Stuttgart, 2004

LAWRENZ, J., SCHWAIGER, W.S.A. (2002): „Bank Deutschland: Aktualisierung der Quantitative Impact Study (QIS2) von Basel II“, in Risknews 01/2002, pp. 5-30 (in German), 2002

LEE, W.-C. (1999): "Probabilistic Analysis of Global Performances of Diagnostic Tests: Interpreting the Lorenz Curve-Based Summary Measures", in Statistics in Medicine, vol. 18, pp. 455-471, 1999

LEHMANN, B. (2003): "Is It Worth the While? The Relevance of Qualitative Information in Credit Rating", EFMA 2003 Helsinki Meetings, http://ssrn.com/abstract=410186 (2.1.2005), 04/2003

LÖFFLER, G. (2003): "Avoiding the Rating Bounce: Why Rating Agencies are Slow to React to new information", Working Paper Universität Ulm, http://www.mathematik.uni-ulm.de/dof/papers/ratingbounce.pdf (1.12.2004), to be published in Journal of Economic Behavior and Organization, 2003

LÖFFLER, G. (2004a): "An anatomy of rating through the cycle", in Journal of Banking and Finance, vol. 28, pp. 695-720, 2004

LÖFFLER, G. (2004b): "Ratings versus market-based measures of default risk in portfolio governance", in Journal of Banking and Finance, vol. 28, pp. 2715-2746, 2004

MAHONEY, C. (2002a): "The Bond Rating Process in a Changing Environment", Moody's Investors Service, Special Comment, Report #73741, 01/2002

MAHONEY, C. (2002b): "The Bond Rating Process: A Progress Report", Moody's Investors Service, Rating Policy, Report #74079, 02/2002

MATHESON, J. E., WINKLER, R. L. (1976): "Scoring rules for continuous probability distributions", in Management Sciences, vol. 22 (10), 1976

MCQUOWN, J.A. (1993): "A Comment on Market vs. Accounting-Based Measures of Default Risk", KMV Working Paper, http://www.moodyskmv.com/research/whitepaper/-A_Comment_on_Market_vs_Accounting_Based_Measures_of_Default_Risk.pdf, (16.7.2004), KMV Corporation, 1993

MILLER, R.M. (1998): "A Nonparametric Test for Credit Rating Refinements", in Risk Magazine 08/1998

MOODY'S (ED.) (2000): "Moody's Investors Service Response to the Consultative Paper Issued by the Basel Committee on Bank Supervision 'A New Capital Adequacy Framework'", Moody's Investors Service, Special Comment, 03/2000

MOODY'S (ED.) (2004a): "Default & Recovery Rates of Corporate Bond Issuers, A Statistical Review of Moody's Ratings Performance, 1920-2003", Moody's Investors Service, Special Comment, 01/2004

MOODY'S (ED.) (2004b): "Moody's Rating Symbols & Definitions", Moody's Investors Service, Report #79004, 08/2004

MOODY'S (ED.) (2004c): "The Performance Of Moody's Corporate Bond Ratings: December 2004 Quarterly Update", Moody's Investors Service, 12/2004

MOODY'S (ED.) (2005): "Default and Recovery Rates of Corporate Bond Issuers, 1920-2004", Moody's Investors Service, 01/2005

MURPHY, A.H., WINKLER, R.L. (1992): "Diagnostic verification of probability forecasts", in International Journal of Forecasting, vol. 7, pp. 435-455, 1992

NANDA, S., PENDHARKAR, P. (2001): "Linear Models for Minimizing Misclassification Costs in Bankruptcy Prediction", in International Journal of Intelligent Systems in Accounting, Finance and Management, vol. 10, pp. 155–168, 2001

NORDEN, L., WEBER, M. (2005): „Möglichkeiten und Grenzen der Bewertung von Ratingsystemen durch Markt und Staat", in Zeitschrift für betriebswirtschaftliche Forschung (ZfbF), special edition 52, p. 31-54 (in German), 2005

OeNB: see ÖSTERREICHISCHE NATIONALBANK

OHLSON, J.A. (1980): "Financial Ratios and the Probabilistic Prediction of Bankruptcy", in Journal of Accounting Research, vol. 18 (1), pp. 109-131, 1980

ÖSTERREICHISCHE NATIONALBANK (OeNB, ED.) (2004a): „Ratingmodelle und -validierung", Leitfadenreihe zum Kreditrisiko (in German), English version available at http://www.oenb.at/en/img/rating_models_tcm16-22933.pdf (5.4.2005), Wien, 2004

ÖSTERREICHISCHE NATIONALBANK (OeNB, ED.) (2004b): „Kreditvergabeprozess und Kreditrisikomanagement", Leitfadenreihe zum Kreditrisiko (in German), English version available at http://www.oenb.at/en/img/credit_approval_process_tcm16-23748.pdf (5.4.2005), Wien, 2004

ÖSTERREICHISCHE NATIONALBANK (OeNB, ED.) (2004c): „Neue quantitative Modelle der Bankenaufsicht", Leitfadenreihe zum Kreditrisiko (in German), English version available at http://www.oenb.at/en/img/new_quantitative_models_of_banking_supervision_tcm16-24132.pdf (5.4.2005), Wien, 2004

PLATTNER, D. (2002): „Warum Firmen Pleite machen, Der Einfluss finanzieller Kennziffern und anderer Faktoren auf die Insolvenzwahrscheinlichkeit kleiner and mittlerer Unternehmen", in KfW-Beiträge Nr. 28, August 2002, pp. 37-54 (in German), 2002

RISKMETRICS (1996): "RiskMetrics – Technical Document", 4th edition, Morgan Guaranty Trust Company, New York, 1996

ROMEIKE, R., WEHRSPOHN, U. (2004a): "Rating-Software im Test", in RATINGaktuell 06/2004, pp. 10-19 (in German), 2004

ROMEIKE, F., WEHRSPOHN, U. (2004b): „Marktstudie Rating-Software im Test", http://www.cre-germany.com/Artikel/Rating-Software-2004.pdf (15.12.2004, in German), published in extracts in RATINGaktuell 06/2004, 2004

RÖSLER, J. (1988): „Die Entwicklung der statistischen Insolvenzdiagnose", in HAUSCHILD, J. (ED.) (1988): „Krisendiagnose durch Bilanzanalyse", Verlag Dr. Otto Schmidt KG, Köln, 1988, pp. 102-115 (in German), 1988

S&P: see STANDARD AND POOR'S

SCHMEISER, B. W. (2001): „Some Myths and Common Errors in Simulation Experiments", in PETERS, B.A., SMITH, J. S., MEDEIROS, D. J., ROHRER, M. W. (ED., 2001): „Proceedings of the 2001 Winter Simulation Conference", Piscataway NJ, IEEE press, pp. 39-46, http://www.informs-sim.org/wsc01papers/006.PDF (30.5.2005), 2001

SCHWAIGER, W.S.A. (2002): „Auswirkungen von Basel II auf den österreichischen Mittelstand nach Branchen und Bundesländern", in Österreichisches Bankarchiv 06/2002, pp. 433-446 (in German), 2002

SCOTT, D. W. (1992): "Multivariate Density Estimation: Theory, Practice and Visualization", John Wiley, New York, 1992

SHANNON, C.E. (2001/1948): "A Mathematical Theory of Communication", in Bell System Technical Journal, vol. 27, 1948, pp. 379–423, 623–656, reprinted in Mobile Computing and Communications Review, vol. 5 (1), pp. 3 -55, 2001

SHUMWAY, T. (2001): "Forecasting Bankruptcy More Accurately: A Simple Hazard Model", in Journal of Business, vol. 74 (1), pp. 101-124, 2001

SOBEHART, J.R., KEENAN, S.C., STEIN, R.M. (2000): "Benchmarking Quantitative Default Risk Models: A Validation Methodology", Moody's Investors Service, Rating Methodology, Report # 53621, 03/2000

SOBEHART, J. R., STEIN, R. M., MIKITYANSKA, V., LI, L. (2000): "Moody's Public Firm Risk Model: A Hybrid Approach to Modeling Short Term Default Risk", Moody's Investors Service, Rating Methodology, Report #53853, 03/2000

SOMERS, R. H. (1962): "A new asymmetric measure of association for ordinal variables", American Sociological Review; Dec 1962, vol. 27 (6), pp. 799-811, 1962

STANDARD AND POOR'S (ED.) (2002): "Ratings Performance 2001", Special Report 02/2002, The McGraw Hills Companies, 2002

STANDARD AND POOR'S (ED.) (2003a): "Ratings Performance 2002, Default, Transition, Recovery, and Spreads", Special Report 02/2003, The McGraw Hills Companies, 2003

STANDARD AND POOR'S (ED.) (2003b): "Corporate Ratings Criteria", The McGraw Hills Companies, 2003

STANDARD AND POOR'S (ED.) (2004a): "Ratings Performance 2003", Special Report 03/2004, The McGraw Hills Companies, 2004

STANDARD AND POOR'S (ED.) (2004b): "S&P Quarterly Default Update & Rating Transitions", The McGraw Hills Companies, 10/2004

STANDARD AND POOR'S (ED.) (2005): "Annual Global Corporate Default Study: Corporate Defaults Poised to Rise in 2005", Global Fixed Income Research, The McGraw Hills Companies, 2005

STATISTISCHES BUNDESAMT (ED.) (2004a): „INSOLVENZEN IN DEUTSCHLAND 2003, STRUKTUREN UND ENTWICKLUNGEN", http://www.destatis.de/presse/deutsch/pk/2004/insolvenzen_2003_i.pdf (04.06.2004, in German), Statistisches Bundesamt — Pressestelle, Wiesbaden, 2004

STATISTISCHES BUNDESAMT (ED.) (2004b): special enquiry concerning corporate insolvencies of selected industries by legal status, 1999-2003, Martin Bemmann, 20.12.2004

STATISTISCHES BUNDESAMT (ED.) (2004c): „Insolvenzen insgesamt and Insolvenzhäufigkeiten von Unternehmen nach ausgewählten Wirtschaftszweigen, Rechtsformen und Ländern, Deutschland", http://www.destatis.de/basis/d/insol/insoltab1.php (15.2.2005, in German), 2004

STATISTISCHES LANDESAMT DES FREISTAATES SACHSEN (2004): special enquiry concerning corporate insolvencies of selected industries by legal status in Sachsen, Kamenz, Martin Bemmann, 2.11.2004

STEIN, R.M. (2002): "Benchmarking Default Prediction Models, Pitfalls and Remedies in Model Validation", Moody's KMV, Report #030124, 2002

STEIN, R.M., KOCAGIL, A.E, BOHN, J., AKHAVEIN, J. (2003): "Systematic And Idiosyncratic Risk In Middle-Market Default Prediction: A Study Of The Performance Of The RiskCalc™ and PFM™ Models", Moody's Investors Service, Special Comment, Report #77261, 02/2003

STEIN, R. M. (2005): „The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing", in Journal of Banking and Finance, vol. 29, pp. 1213-1236, 2005

STEINER, M., HEINKE, V. G. (1996): "Ratingagenturen an nationalen und internationalen Finanzmärkten", (in German), in BÜSCHGEN, H. E., EVERLING, O. (ED.) (1996): „Handbuch Rating", Gabler Verlag, Wiesbaden, p. 580 – 627, 1996

SWETS, J. A. (1973): "The Relative Operating Characteristic in Psychology", in Science, vol. 182, pp. 990-1.000, 1973

SWETS, J. A., PICKET, R. M. (1982): „Evaluation of Diagnostic Systems, Methods from Signal Detection Theory", Academic Press, series in cognition and perception, New York et al., 1982

SWETS, J. A. (1988): "Measuring the Accuracy of Diagnostic Systems", in Science, vol. 240, pp. 1285- 1293, 1988

SWETS, J.A., DAWES, R.M., MONAHAN, J. (2000): "Psychological Science Can Improve Diagnostic Decisions", in Psychological Science in the Public Interest, vol. 1 (1), pp. 1-26, 2000

TREACY, W. F., CAREY, M. S. (2000/1998): "Credit Risk Rating at Large U.S. Banks", Federal Reserve Bulletin, 11/1998, pp. 987-921, also published in "Credit risk rating systems at large US banks", in Journal of Banking and Finance, vol. 24, pp. 167-201, 2000

VARETTO, F. (1998): "Genetic algorithms applications in the analysis of insolvency risk", in Journal of Banking and Finance, vol. 22, pp. 1421-1439, 1998

VARMA, P., CANTOR, R. (2005): "Determinants of Recovery Rates on Defaulted Bonds and Loans for North American Corporate Issuers: 1983-2003", in Journal of Fixed Income, vol. 14 (4), pp. 29-44, 2005

WESTGAARD, S., WIJST, N. VAN DER (2001): "Default probabilities in a corporate bank portfolio: a logistic model approach", in European Journal of Operational Research, Vol 135, pp. 338-349, 2001

WHITE, L. J. (2001): "The Credit Rating Industry: An Industrial Organization Analysis", New York University, Center for Law and Business, Research Paper No. 01-001, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=267083 (22.4.2005), 04/2001

WINKLER, R. L. (1994): "Evaluating Probabilities: Asymmetric Scoring Rules", in Management Science, vol. 40 (11), pp. 1395-1405, 1994

Appendix I: Upper and lower limits of accuracy ratio values

Intention: It shall be determined, what discriminative power, measured in accuracy ratio, an ordinal prediction method may posses at least/ at most if only one CAP-curve point (X_0, Y_0) is known or when only one combination of errors of types I and II is know.

X_0 … percentage of exluded companies,

Y_0 … hit rate (=100%-error of type I)

Technical constraints

In a CAP-diagram the CAP curve must proceed trough points (0%; 0%) and (100%; 100%). It must be continuous and weakly monotonic, i.e. $CAP(x)' \geq 0$. Further, $CAP(x)' \leq 1/PD$, as for excluding 100% of all defaulters, at least PD% of all corporations have to be excluded.

Textual constraints

Subsequently it is assumed, that the rating model underlying a given CAP-curve fulfills a fundamental prerequisite of rating models, namely that better predictions are accompanied by lower default probabilities. Thus, only *concave* CAP curves are considered, i.e. $CAP(x)' \geq 0$ and $CAP(x)'' \leq 0$.

In point (0%; 0%) the CAP-curve's slope has to be 1.0 or more, otherwise, with non-increasing slopes, point (100%; 100%) could not be reached.

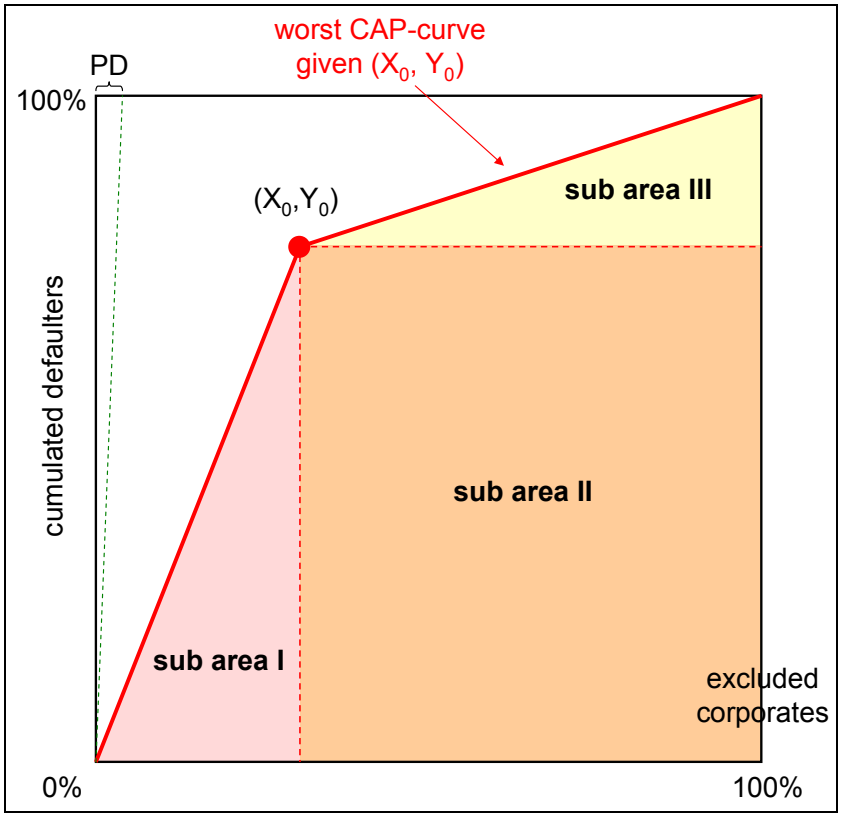

Figure 21: surface area decomposition for determining minimal AUC_{CAP}

Determining lower limits of accuracy ratios

The "worst possible" [316] CAP-curve that passes through CAP-diagram point (X_0, Y_0) and which complies with all technical and textual constraints, is the linear spline (0%; %) (X_0, Y_0) – (100%; 100%), see Figure 21.

This CAP-curve corresponds with a prediction model, that can only differentiate *between* but *not among* the two groups that are separated by X_0, i.e. the ordinal model that creates this CAP-curve does not add any value to a categorial model with one only cut-off score yielding (X_0, Y_0).

F 57) $$AUC_{CAP,min} = \frac{X_0 \cdot Y_0}{2} + (1 - X_0) \cdot Y_0 + \frac{(1 - X_0) \cdot (1 - Y_0)}{2}$$

F 58) $$AUC_{CAP,min} = \frac{X_0 \cdot Y_0}{2} + Y_0 - X_0 \cdot Y_0 + \frac{1 - X_0 - Y_0 + X_0 \cdot Y_0}{2}$$

F 59) $$\boxed{AUC_{CAP,min} = \frac{1 - X_0 + Y_0}{2}}$$ or alternatively

F 60) $AUC_{CAP,min} = \frac{1}{2} + \frac{Y_0 - X_0}{2}$ with area under the main diagonal = ½. Thus, the area above the diagonal line and below the CAP-curve is $\frac{Y_0 - X_0}{2}$.

The difference Y_0-X_0 is also referred to as "vertical distance", because it accords to the vertical distance of the point (X_0; Y_0) to the main diagonal. The maximal vertical distance of a *ROC*-Curve is an empirically used quality measure on its own.[317]

By using the relation between AUC_{CAP} and AR_{CAP} that were outlined in chapter 2.3.2 it follows that[318]

F 61) $$AR_{min} = \frac{2AUC_{CAP,min} - 1}{1 - PD}$$

F 62) $$AR_{min} = \frac{1 - X_0 + Y_0 - 1}{1 - PD}$$

F 63) $$\boxed{AR_{min} = \frac{Y_0 - X_0}{1 - PD}}$$ with Y_0–X_0 measuring the vertical distance of (X_0; Y_0) to the main diagonal (see above).

---

[316] The *worst possible* CAP-curve is the CAP-curve with the smallest surface area (AUC_{CAP}) of all admissible CAP-curves.

[317] see Lee (1999, p.462).

[318] As ROC-curve and CAP-curve based accuracy ratios are identical, in the following no CAP- or ROC-indices are used in connection with accuracy ratios. They are used only in connection with AUC-measures.

In the following, $AUC_{CAP,min}$ and AR_{min} are expressed as functions of errors of types I and II (instead of as functions of CAP-coordinates X_0 and Y_0).

F 64) $Y_0 = 1 - F_1$ with $F_{1/2}$ .. error [failure] of type I/II

F 65) $F_2 = \frac{\text{non} - \text{defaulters excluded at } (X_0, Y_0)}{\text{all non} - \text{defaulters}} = \frac{X_0 - (1 - F_1) \cdot PD}{1 - PD}$ and thus

F 66) $X_0 = F_2 \cdot (1 - PD) + (1 - F_1) \cdot PD$ Insertion to formula F 59 gives:

F 67) $AUC_{CAP,min} = \frac{1 - (F_2 \cdot (1 - PD) + (1 - F_1) \cdot PD) + (1 - F_1)}{2}$

F 68) $\boxed{AUC_{CAP,min} = \frac{(1 - F_1 - F_2) \cdot (1 - PD) + 1}{2}}$ Insertion to formula F 63 yields:

F 69) $AR_{min} = \frac{1 - F_1 - F_2 \cdot (1 - PD) - (1 - F_1) \cdot PD}{1 - PD}$

F 70) $\boxed{AR_{min} = 1 - (F_1 + F_2)}$ with

F 71) $AUC_{ROC} = \frac{AR + 1}{2}$ and formula F 70 the area under the ROC-curve is given by:

F 72) $\boxed{AUC_{ROC,min} = 1 - \frac{F_1 + F_2}{2}}$

Determining upper limits of accuracy ratios

The *best possible* (i.e. *surface area maximizing*) admissible CAP-curve that is passing through (X_0, Y_0) must fulfill following conditions: starting from (0%; 0%) it proceeds along the dotted straight line (PD), which means that the corporations that are excluded first have a default rate of 100%, afterwards it linearly proceeds through (X_0; Y_0) until it intersects the 100%-hit-rate-line (see Figure 9). Subsequently the CAP-curve horizontally proceeds until it reaches (100%; 100%), i.e. the corporations finally excluded have a probability of default of 0%. The slope and absolute term of the center line have to be identified by an optimization approach that is presented below.

Proof: It has to be shown, that the centre piece of the surface area maximizing CAP-curve has to be a straight line. This will be shown, by proving that both CAP-curve sections right and left hand side of (X_0, Y_0) must be straight lines with the same slope.

- Following condition has to be met by the slope, *a*, of the CAP-curve in (X_0, Y_0): $\frac{Y_0}{X_0} \geq a \geq \frac{1 - Y_0}{1 - X_0}$ - otherwise, given the technical and textual restrictions, points (0%; 0%) and (100%; 100%) could not be "reached" starting from (X_0; Y_0).
- Because the tangent of a (weakly) concave function is always proceeding above (or on) the function itself, the right hand part of the AUC (=right hand side integral of the CAP-curve) – is, for a given *a*, surface area maximizing then, when it is linear. Likewise it can be shown, that the left hand side gradient of the surface maximizing CAP-curve has to be linear, too.

- Both left hand and right hand side sections of the surface area maximizing CAP-curve at (X_0; Y_0) must be characterized by the *same slope*: given a left hand side CAP-curve, the area under the CAP-curve right hand side from X_0 is the greater, the steeper the right hand side straight line is. But as the right hand side straight line must not be steeper then the left hand side straight line (concavity requirement), AUC is maximized than, when the straight lines that meet in (X_0; Y_0) have the same slope, i.e. are lying on *one* line.

Slope and absolute value of this straight line are determined subsequently via an optimization approach. For that, the area under the CAP curve is divided into four sub areas I..IV as shown in Figure 22. Additionally, auxiliary variables X_u, Y_u and X_r are introduced.

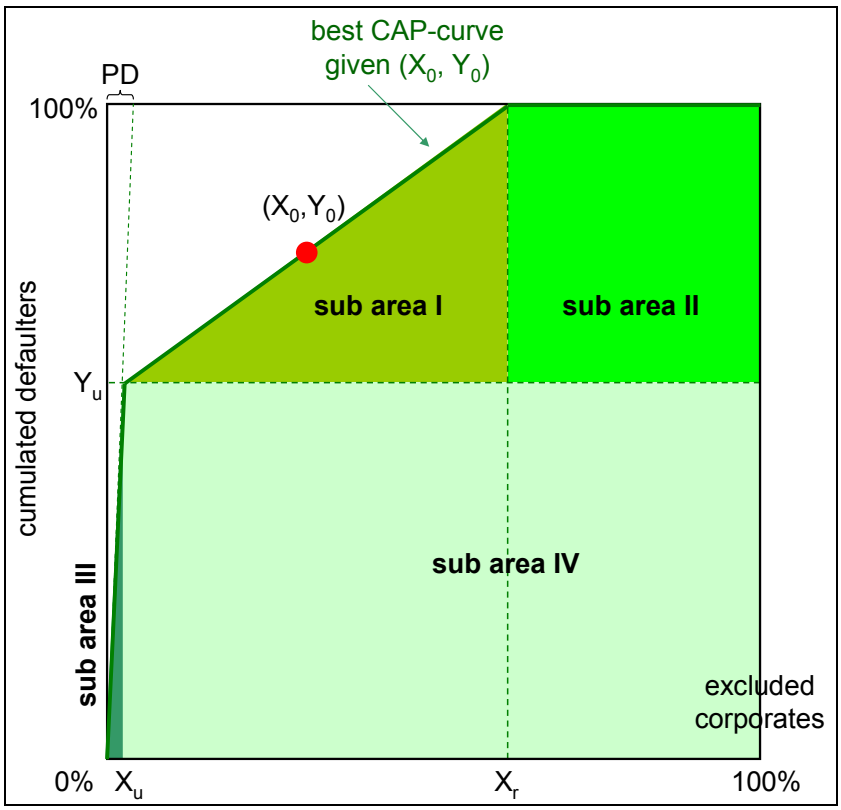

Figure 22: surface area decomposition for determining maximal AUC_{CAP}

X_u … x-coordinate of the point of intersection of the middle CAP-section and the PD-line,
Y_u … y-coordinate of the point of intersection of the middle CAP-section and the PD-line,
X_r … x-coordinate of the point of intersection of the middle CAP-section and the 100%-hitrate-line,
a … slope of the middle part of the CAP-curve
b … absolute value of the middle part of the CAP-curve

By summing sub areas I...IV to $AUC_{CAP,max}$ it follows:

F 73) $$AUC_{CAP,max} = (X_r - X_u)\cdot(1-Y_u)/2 + (1-X_r)\cdot(1-Y_u) + X_u \cdot Y_u/2 + (1-X_u)\cdot Y_u$$

F 74) $$AUC_{CAP,max} = \frac{X_r}{2} - \frac{X_u}{2} - \frac{X_r \cdot Y_u}{2} + \frac{X_u \cdot Y_u}{2} + 1 - X_r - Y_u + X_r Y_u + \frac{X_u \cdot Y_u}{2} + Y_u - X_u \cdot Y_u,$$

F 75) $$AUC_{CAP,max} = -\frac{X_r}{2} - \frac{X_u}{2} + \frac{X_r \cdot Y_u}{2} + 1$$

It holds that (for illustration see also Figure 22):

F 76) $\frac{X_u}{Y_u} = \frac{PD}{100\%}$

F 77) $X_u = Y_u \cdot PD$

From the linear equation for the middle part of the CAP-curve follows that:

F 78) $aX_u + b = Y_u$

F 79) $aX_0 + b = Y_0$

F 80) $aX_r + b = 1$ Inserting formula F 77 into formula F 78 yields:

F 81) $a \cdot Y_u \cdot PD + b = Y_u$

F 82) $b = Y_u \cdot (1 - a \cdot PD)$

F 83) $Y_u = \frac{b}{1 - a \cdot PD}$ From formula F 80 follows that:

F 84) $X_r = \frac{1-b}{a}$ By insertion of formula F 83 into formula F 78 it follows that

F 85) $X_u = \frac{b \cdot PD}{1 - a \cdot PD}$ From formula F 79 follows:

F 86) $a = \frac{Y_0 - b}{X_0}$ From formulas F 83 and F 86 follows:

F 87): $Y_u = \frac{b}{1 - \frac{Y_0 - b}{X_0} \cdot PD} = \frac{X_0 \cdot b}{X_0 - Y_0 \cdot PD + b \cdot PD}$ Formulas F 85 and F 86 yield:

F 88): $X_u = \frac{b \cdot PD}{1 - \frac{Y_0 - b}{X_0} \cdot PD} = \frac{X_0 \cdot b \cdot PD}{X_0 - Y_0 \cdot PD + b \cdot PD}$ Formulas F 84 and F 86 yield:

F 89) $X_r = \frac{1-b}{\frac{Y_0 - b}{X_0}} = X_0 \cdot \frac{1-b}{Y_0 - b}$ By eliminating auxiliary variables X_r, X_u and Y_u by inserting formulas F 87, F 88 and F 89 into F 75 it follows:

F 90) $AUC = -\frac{X_0}{2} \cdot \frac{1-b}{Y_0 - b} - \frac{X_0 \cdot b \cdot PD}{2 \cdot (X_0 - Y_0 \cdot PD + b \cdot PD)} + \frac{X_0}{2} \cdot \frac{1-b}{Y_0 - b} \cdot \frac{X_0 \cdot b}{X_0 - Y_0 \cdot PD + b \cdot PD} + 1$

F 91) $AUC = \frac{(-X_0 + bX_0) \cdot (X_0 - Y_0 \cdot PD + b \cdot PD) - (X_0 \cdot b \cdot PD) \cdot (Y_0 - b) + bX_0^2 - b^2 X_0^2}{2 \cdot (Y_0 - b)(X_0 - Y_0 \cdot PD + b \cdot PD)} + 1$

F 92) $AUC = \frac{b^2 (2 PD X_0 - X_0^2) + b (-PD X_0 + 2 X_0^2 - 2 PD X_0 Y_0) - X_0^2 + PD X_0 Y_0}{2 (Y_0 - b) (b PD + X_0 - PD Y_0)} + 1$

Subsequently, this term is maximized by b:[319]

$$\textbf{F 93)}\quad \frac{\partial AUC}{\partial b} = \frac{-PD\,X_0 + 2X_0^2 + 2b\,(2PD\,X_0 - X_0^2) - 2PD\,X_0\,Y_0}{2\,(-b+Y_0)\,(b\,PD + X_0 - PD\,Y_0)} - \frac{PD\,(-X_0^2 + b^2\,(2PD\,X_0 - X_0^2) + PD\,X_0\,Y_0 + b\,(-PD\,X_0 + 2X_0^2 - 2PD\,X_0\,Y_0))}{2\,(-b+Y_0)\,(b\,PD + X_0 - PD\,Y_0)^2} + \frac{-X_0^2 + b^2\,(2PD\,X_0 - X_0^2) + PD\,X_0\,Y_0 + b\,(-PD\,X_0 + 2X_0^2 - 2PD\,X_0\,Y_0)}{2\,(-b+Y_0)^2\,(b\,PD + X_0 - PD\,Y_0)}$$

$$\textbf{F 94)}\quad \frac{\partial AUC}{\partial b} = \frac{X_0\,((-1+b)\,X_0 + PD\,(-b+Y_0))\,(X_0\,(1+b-2Y_0) - PD\,(b-Y_0)\,(-1+2Y_0))}{2\,(X_0 + PD\,(b-Y_0))^2\,(b-Y_0)^2}$$

$$\textbf{F 95)}\quad \frac{\partial AUC}{\partial b} == 0$$

$$\textbf{F 96)}\quad b_{opt} = \frac{X_0 - PD\cdot Y_0 - 2\cdot Y_0\cdot X_0 + 2\cdot PD\cdot Y_0^2}{-PD - X_0 + 2\cdot PD\cdot Y_0} \quad [320]$$

$$\textbf{F 97)}\quad AUC_{CAP,max} = \frac{\left(\frac{X_0 - PD\,Y_0 - 2X_0\,Y_0 + 2PD\,Y_0^2}{-PD - X_0 + 2PD\,Y_0}\right)^2 (2PD\,x - x^2) + \frac{X_0 - PD\,Y_0 - 2X_0\,Y_0 + 2PD\,Y_0^2}{-PD - X_0 + 2PD\,Y_0}\,(-PD\,x + 2x^2 - 2PD\,x\,y) - x^2 + PD\,x\,y}{-2\left(\frac{X_0 - PD\,Y_0 - 2X_0\,Y_0 + 2PD\,Y_0^2}{-PD - X_0 + 2PD\,Y_0}\right)^2 PD - 2\,\frac{X_0 - PD\,Y_0 - 2X_0\,Y_0 + 2PD\,Y_0^2}{-PD - X_0 + 2PD\,Y_0}\,x + 4\,b\,PD\,y + 2\,x\,y - 2PD\,y^2} + 1$$

By simplifying the above term it follows:

$$\textbf{F 98)}\quad \boxed{AUC_{CAP,max} = 1 - \frac{PD}{2} - 2X_0 + 2\cdot X_0\cdot Y_0 + 2PD\cdot Y_0 - 2PD\cdot Y_0^2}$$

From formulas F 61 and F 98 it follows:

$$\textbf{F 99)}\quad AR = \frac{2 - PD - 4X_0 + 4\cdot X_0\cdot Y_0 + 4\cdot PD\cdot Y_0 - 4PD\cdot Y_0^2 - 1}{1 - PD}$$

$$\textbf{F 100)}\quad \boxed{AR_{max} = \frac{1 - PD - 4X_0 + 4\cdot X_0\cdot Y_0 + 4\cdot PD\cdot Y_0 - 4PD\cdot Y_0^2}{1 - PD}}$$

For information:

$$\textbf{F 101)}\quad X_r = 2X_0 - 2\cdot PD\cdot Y_0 + PD$$

$$\textbf{F 102)}\quad X_u = 2\cdot PD\cdot Y_0 - PD$$

$$\textbf{F 103)}\quad Y_u = 2\cdot Y_0 - 1$$

[319] Derivations, nulls of derivation and simplifications of the respective terms were determined with *Mathematica 5*.

[320] The algebraic sign of the second derivation of AUC with respect to b is negative. Thus, the solution found is a maximum.

F 104) $$a = \frac{Y_0 - 1}{-PD - X_0 + 2 \cdot PD \cdot Y_0}$$

F 105) $$b = Y_0 - \frac{Y_0 - 1}{-PD - X_0 + 2 \cdot PD \cdot Y_0} \cdot X_0$$

If $AUC_{CAP,min}$ and AR_{min} shall be determined as functions of errors of types I and II, following equations result from applying formulas F 64, F 66 and F 98:

F 106) $$AUC_{CAP,max} = 1 - \frac{PD}{2} - 2((1-F_1)\cdot PD + F_2 \cdot (1-PD)) + 2 \cdot ((1-F_1)\cdot PD + F_2 \cdot (1-PD)) \cdot (1-F_1) + 2PD \cdot (1-F_1) - 2PD \cdot (1-F_1)^2$$

F 107) $$AUC_{CAP,max} = 1 - \frac{PD}{2} - 2 \cdot PD + 2 \cdot PD \cdot F_1 - 2 \cdot F_2 - 2 \cdot PD \cdot F_2 + (-2 \cdot PD + 2 \cdot PD \cdot F_1 - 2 \cdot F_2 - 2 \cdot PD \cdot F_2) \cdot (-1 + F_1) + 2 \cdot PD - 2PD \cdot F_1 - 2PD \cdot (1 - 2F_1 + F_1^2)$$

F 108) $$AUC_{CAP,max} = 1 - \frac{PD}{2} + -2 \cdot PD \cdot F_1 + 2 \cdot PD \cdot F_1^2 - 2 \cdot F_2 \cdot F_1 + 2 \cdot PD \cdot F_2 \cdot F_1 + 2 \cdot PD - 2 \cdot PD \cdot F_1 - 2 \cdot PD + 4 \cdot PD \cdot F_1 - 2 \cdot PD \cdot F_1^2$$

F 109) $$AUC_{CAP,max} = 1 - \frac{PD}{2} - 2 \cdot F_2 \cdot F_1 + 2 \cdot PD \cdot F_2 \cdot F_1$$

F 110) $$\boxed{AUC_{CAP,max} = 1 - 2 \cdot F_1 \cdot F_2 \cdot (1-PD) - \frac{PD}{2}}$$

From formulas F 61 and F 110 follows:

F 111) $$AR_{max} = \frac{2 \cdot \left(1 - 2 \cdot F_1 \cdot F_2 \cdot (1-PD) - \frac{PD}{2}\right) - 1}{1-PD}$$

F 112) $$AR_{max} = \frac{2 - 4 \cdot F_1 \cdot F_2 \cdot (1-PD) - PD - 1}{1-PD}$$

F 113) $$AR_{max} = \frac{1 - PD - 4 \cdot F_1 \cdot F_2 \cdot (1-PD)}{1-PD}$$

F 114) $$\boxed{AR_{max} = 1 - 4 \cdot F_1 \cdot F_2}$$

From formulas F 71 and F 114 follows:

F 115) $$AUC_{ROC,max} = \frac{1 - 4 \cdot F_1 \cdot F_2 + 1}{2}$$

F 116) $$\boxed{AUC_{ROC,max} = 1 - 2 \cdot F_1 \cdot F_2}$$

Determining heuristic estimators for accuracy ratios

If, given a combination of errors of types I and II, not only intervals formed by the upper and lower limits of possible accuracy ratios shall be stated, but also *univalent* estimators, the average value of the upper and lower accuracy ratio limits lends itself to that:

F 117) $$AR_{mv} \equiv \frac{AR_{min} + AR_{max}}{2}$$

Obviously, this is only a *heuristic* estimator, for which, at the time being, not more and not less can be claimed than that it always delivers accuracy ratio values *within* the range of admissible values (for empirical examinations see the following pages). Further heuristic measures for estimating accuracy ratios based on single combinations of errors of types I and II are AR_α, AR_β and $AR_{\alpha\&\beta}$ (average value of AR_α and AR_β), that base on on the parametrical ROC-curve functions $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta(x)=1-(1-x)^{1/\beta}$. A subsumption is given in Table K:

	CAP-coordinates	errors of type I and II
lower limit	$AUC_{CAP,min} = \frac{1 - X_0 + Y_0}{2}$	$AUC_{ROC,min} = 1 - \frac{F_1 + F_2}{2}$
	$AR_{min} = \frac{Y_0 - X_0}{1 - PD}$	$AR_{min} = 1 - (F_1 + F_2)$
upper limit	$AUC_{CAP,max} = 1 - \frac{PD}{2} - 2X_0 + 2 \cdot X_0 \cdot Y_0 + 2PD \cdot Y_0 - 2PD \cdot Y_0^2$	$AUC_{ROC,max} = 1 - 2 \cdot F_1 \cdot F_2$
	$AR_{max} = \frac{1 - PD - 4X_0 + 4 \cdot X_0 \cdot Y_0 + 4 \cdot PD \cdot Y_0 - 4PD \cdot Y_0^2}{1 - PD}$	$AR_{max} = 1 - 4 \cdot F_1 \cdot F_2$
AR_{mv}	$AR_{MW} \equiv \frac{AR_{max} + AR_{min}}{2}$	$AR_{MW} \equiv \frac{AR_{max} + AR_{min}}{2}$
AR_α	$AR_\alpha = \frac{\log(X_0 - Y_0 \cdot PD) - \log Y_0 - \log(1 - PD)}{\log(X_0 - Y_0 \cdot PD) + \log Y_0 - \log(1 - PD)}$	$AR_\alpha = \frac{\log F_2 - \log(1 - F_1)}{\log F_2 + \log(1 - F_1)}$
AR_β	$AR_\beta = \frac{\log(1 - Y_0) - \log(1 - PD - X_0 + Y_0 \cdot PD) + \log(1 - PD)}{\log(1 - Y_0) + \log(1 - PD - X_0 + Y_0 \cdot PD) - \log(1 - PD)}$	$AR_\beta = \frac{\log F_1 - \log(1 - F_2)}{\log F_1 + \log(1 - F_2)}$
$AR_{\alpha\&\beta}$	$AR_{\alpha\&\beta} \equiv \frac{AR_\alpha + AR_\beta}{2}$	$AR_{\alpha\&\beta} \equiv \frac{AR_\alpha + AR_\beta}{2}$

Table K: formulas for obtaining exact upper and lower limits and heuristic estimators AR_{mv}, AR_α and AR_β for accuracy ratios and AUC from CAP-coordinates/ combinations of errors of type I and II

The suitability of the various heuristic measures presented, is examined subsequently based on empirical data of all nine models that were presented in chapter 3.5 and for which the data required for determining ROC-curves was available[321,322], see Figure 23 to Figure 31:

- The left hand side graphs in the following figures display the empirical ROC-curves and accuracy-ratio-equivalent ROC-curves according to the ROC_α- and ROC_β-methods. Thus, the left hand side graphs show, how well the shapes of empirical ROC-curves can be approximated either by calibrated ROC_α- or ROC_β-curves.
- The right hand side graphs show the accuracy-ratio-values that were estimated with the various methods for all combinations of errors of types I and II of the empirical ROC-curves. *True accuracy ratios* are displayed as well, see the bold horizontal lines. These graphs give a first impression of the precision and stability of the various estimators.

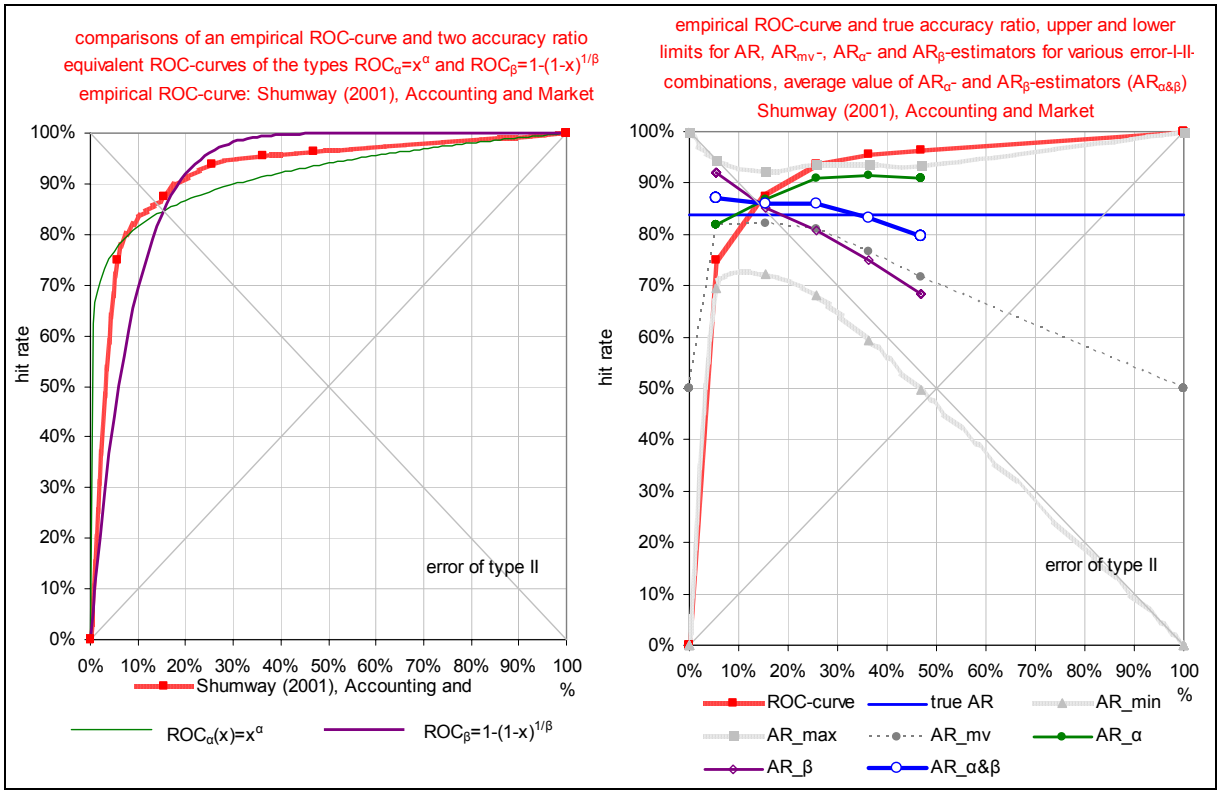

Figure 23: left hand side: empirical ROC-curve SHUMWAY (2001), *Accounting and Market* and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

[321] Following information are *alternatively* needed for obtaining ROC-curves:
- (graphical) representations of ROC- or CAP-curves,
- rating class specific default rates and shares of corporations,
- "ROC-coordinates" (various combinations of errors of types I and II.

[322] Although, the respective data were available, the BEAVER (1967) and ALTMAN (1968) studies were not examined owing to their unusually small samples.

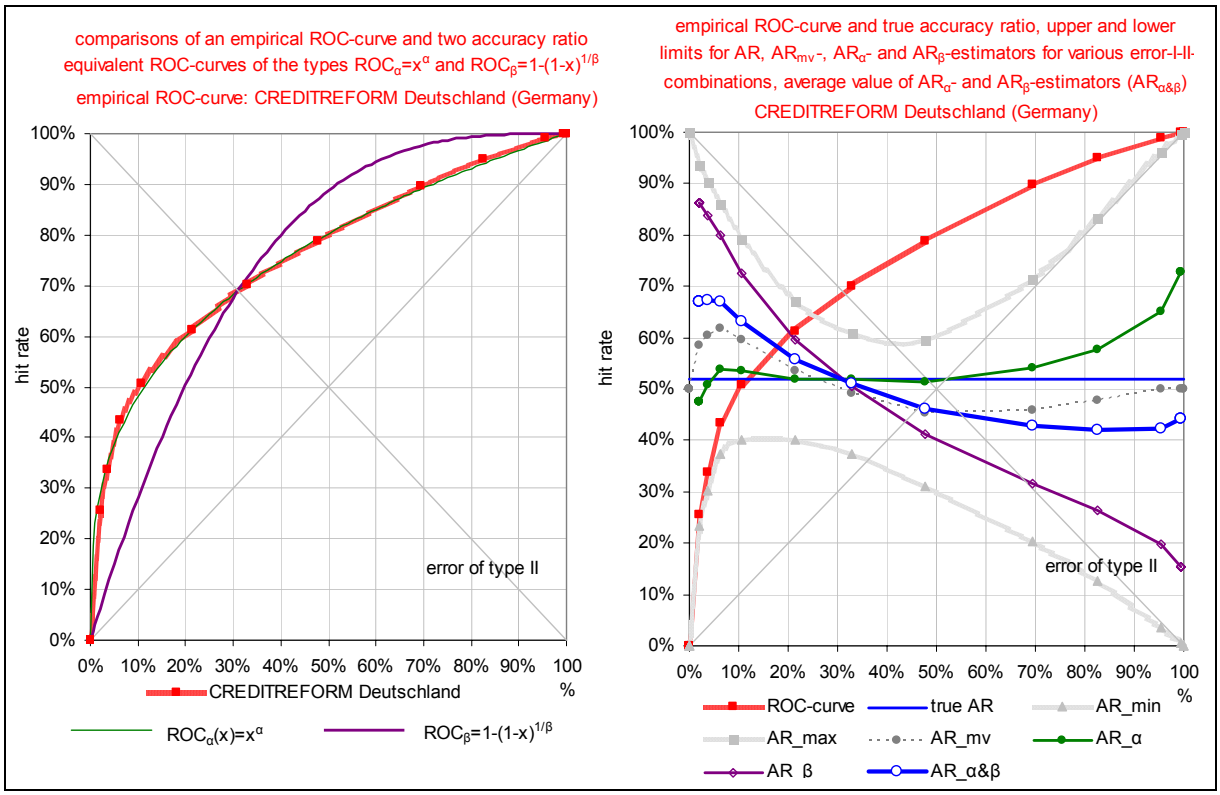

Figure 24: left hand side: empirical ROC-curve CREDITREFORM-Bonitätsindex Deutschland (based on data from LAWRENZ, SCHWAIGER (2002)) and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

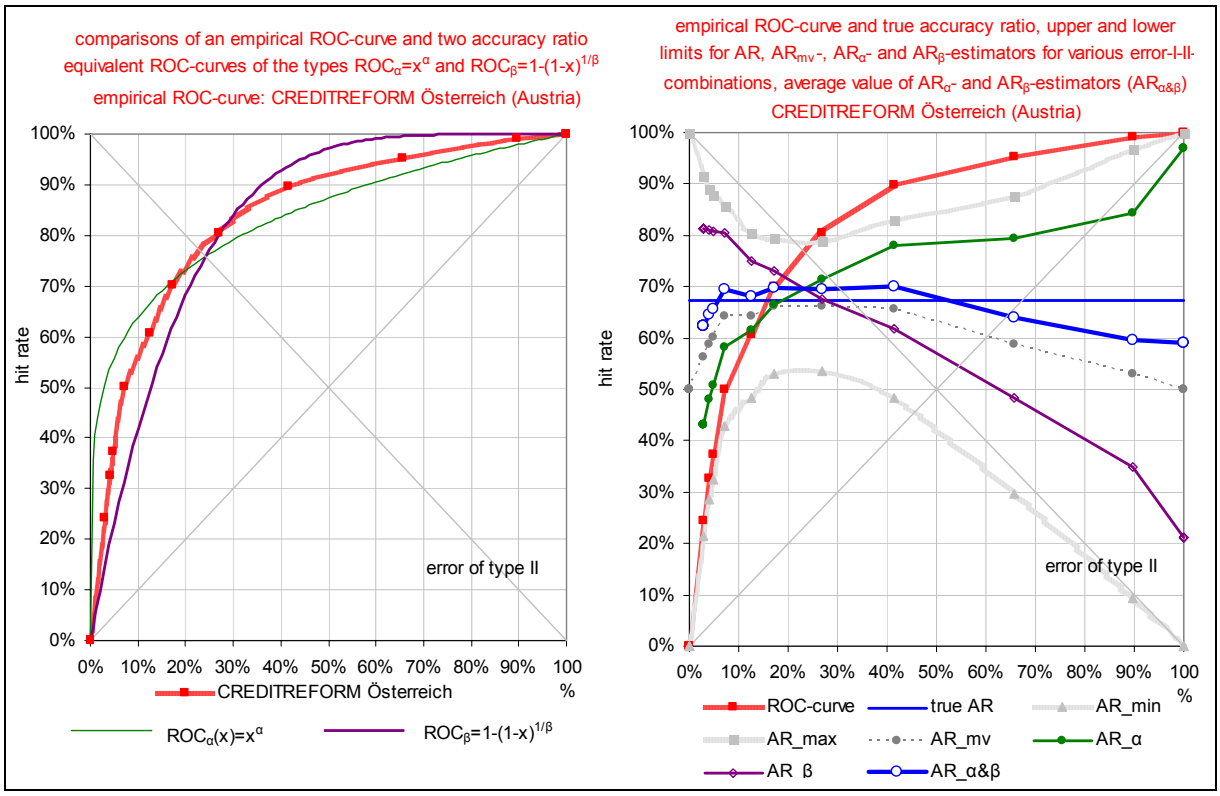

Figure 25: left hand side: empirical ROC-curve CREDITREFORM-Bonitätsindex Österreich (Austria) (based on data from SCHWAIGER (2002)) and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

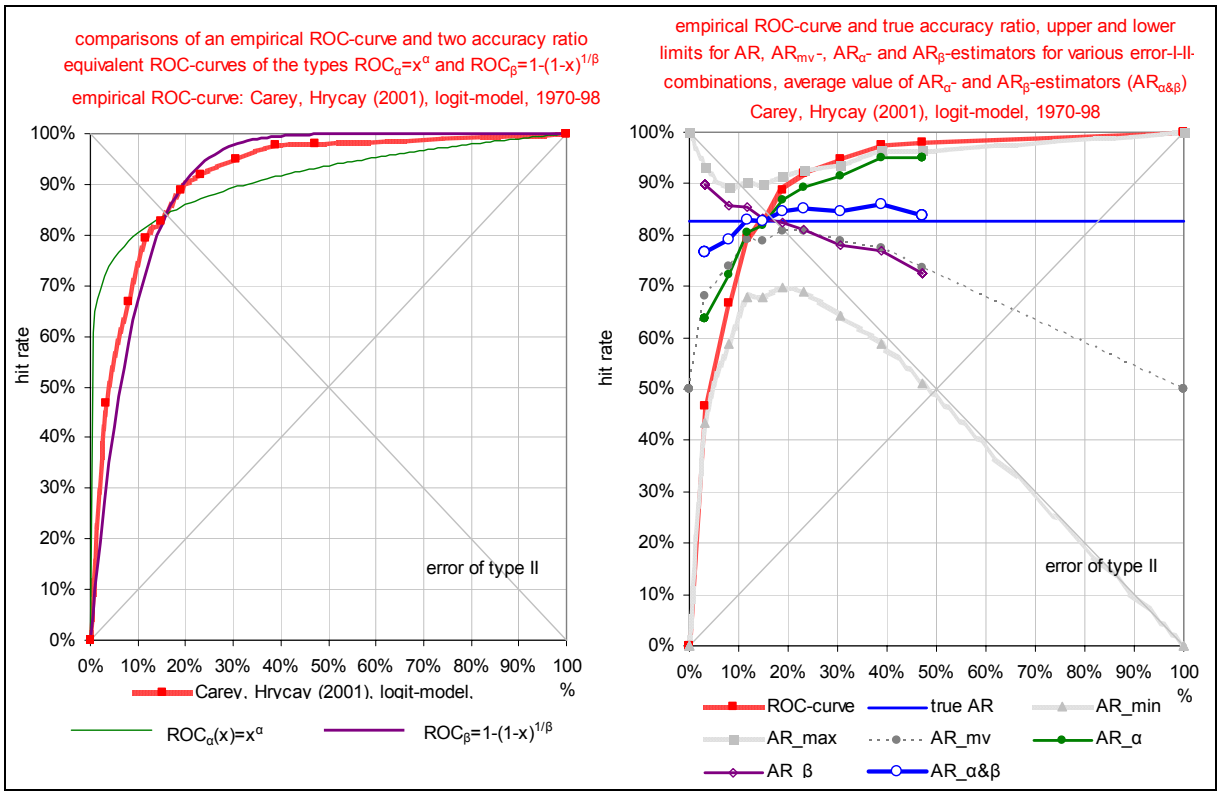

Figure 26: left hand side: empirical ROC-curve CAREY, HRYCAY (1980), logit model 1970-1998 and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

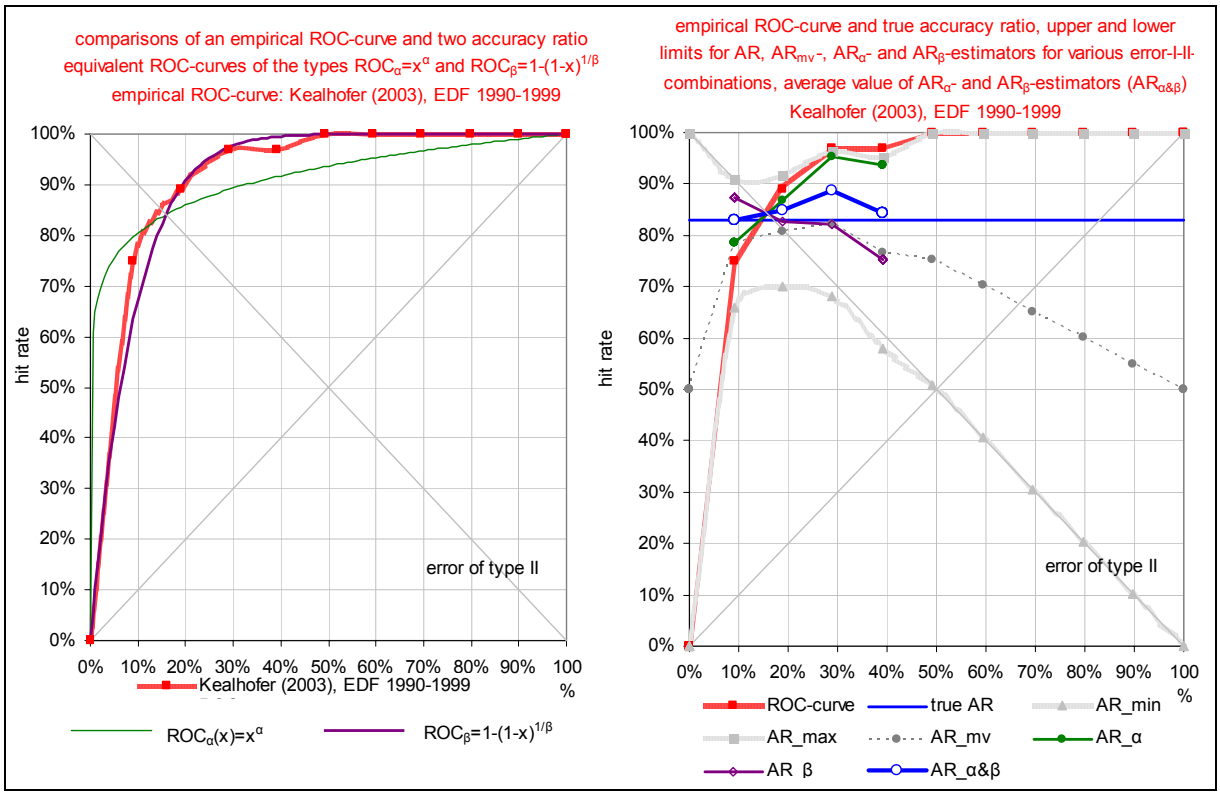

Figure 27: left hand side: empirical ROC-curve KEALHOFER (2003), EDF 1990-1999 and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

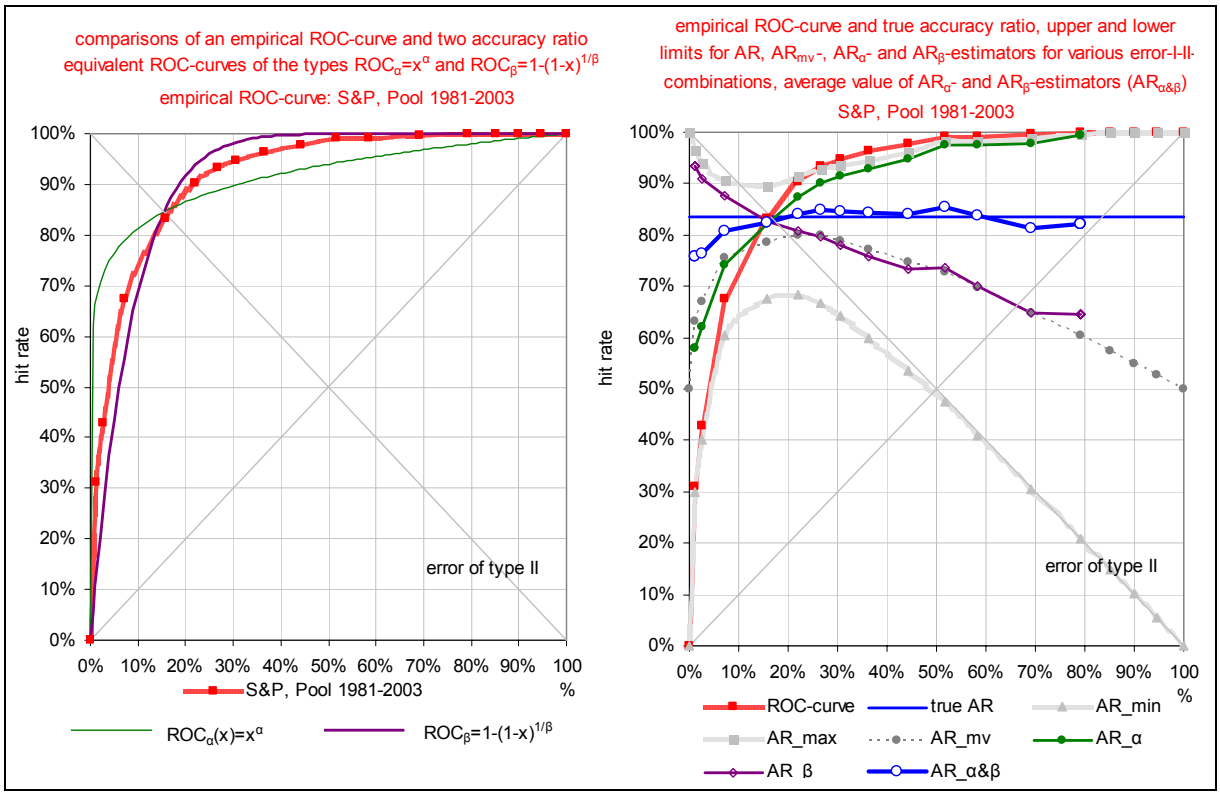

Figure 28: left hand side: empirical ROC-curve S&P ratings, pool 1981-2003 and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

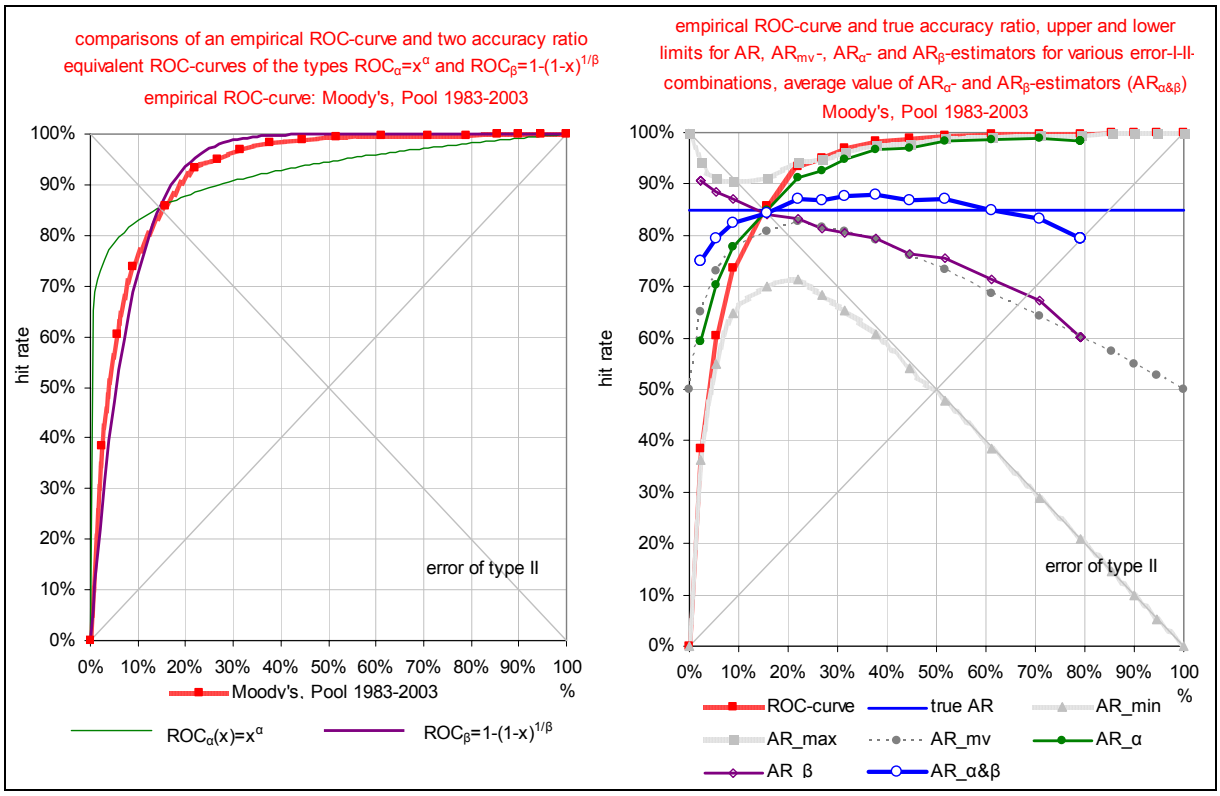

Figure 29: left hand side: empirical ROC-curve MOODY'S ratings, pool 1983-2003 and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

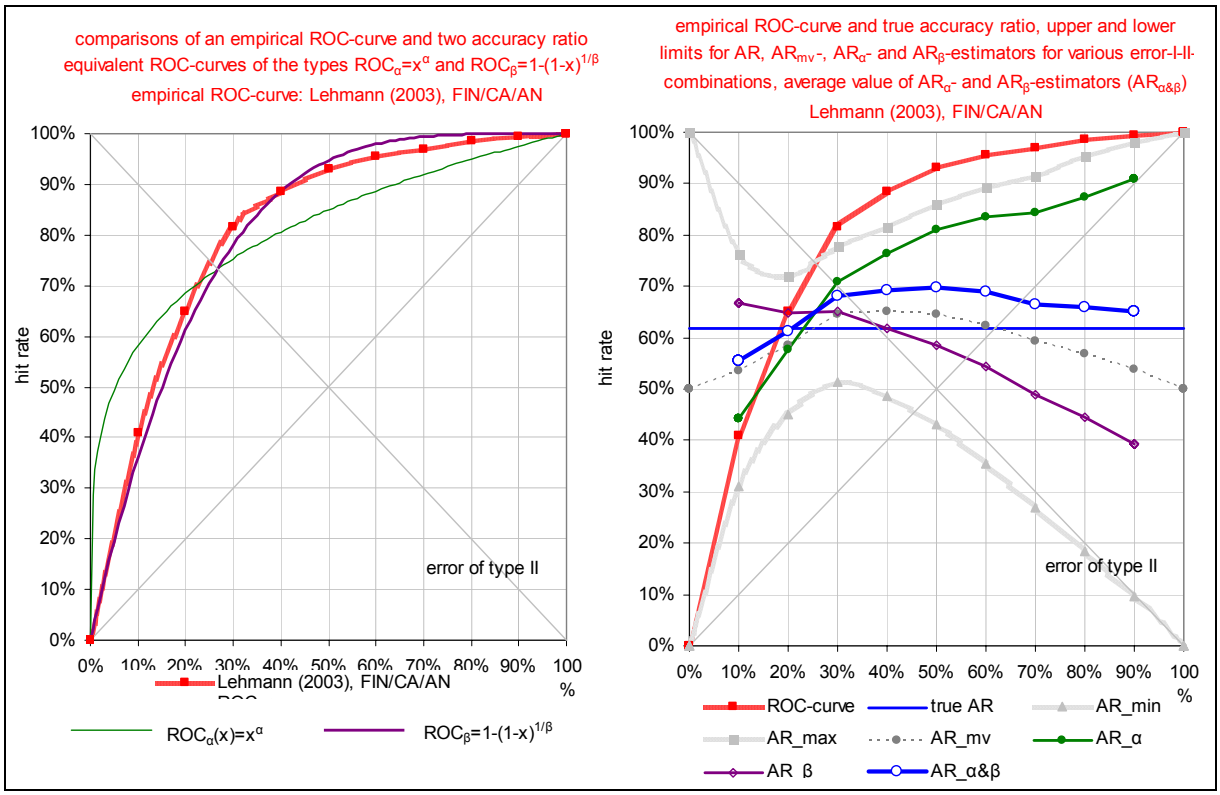

Figure 30: left hand side: empirical ROC-curve LEHMANN (2003), FIN/CA/AN and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

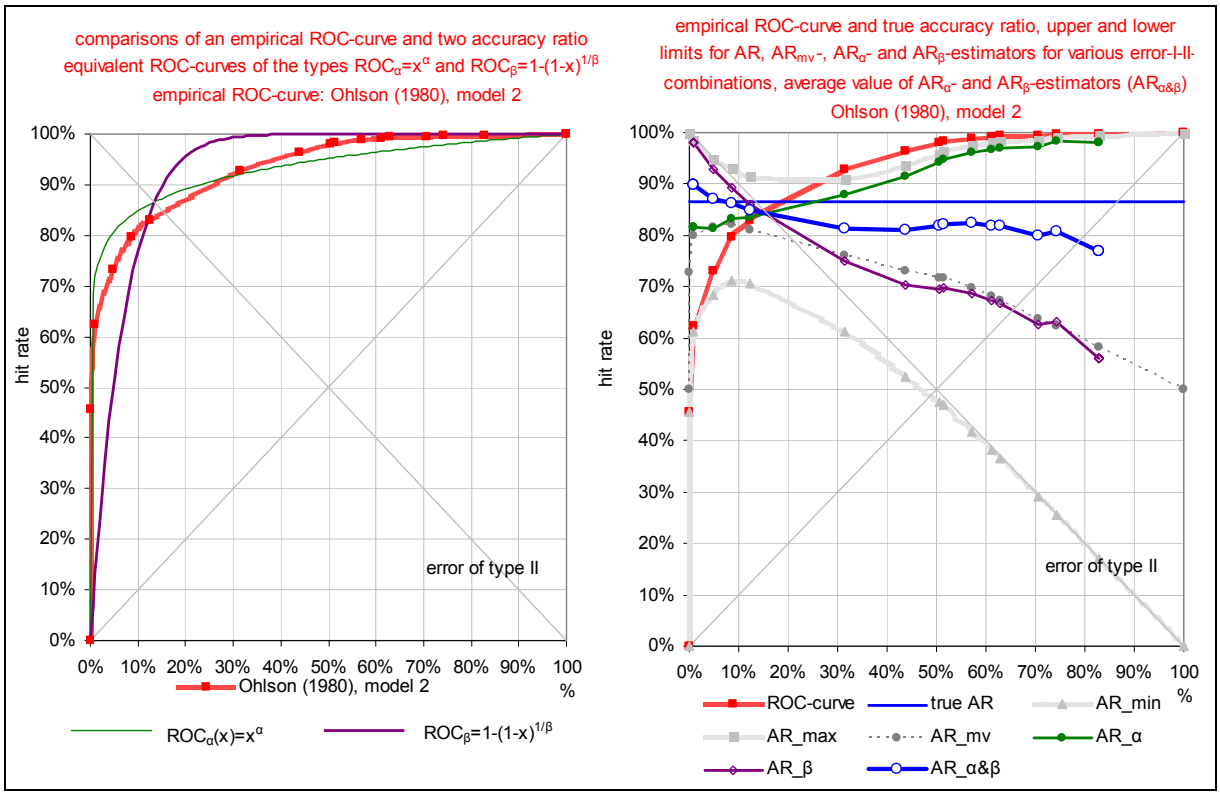

Figure 31: left hand side: empirical ROC-curve OHLSON (1980), model 2 and accuracy ratio equivalent ROC-curves $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta=1-(1-x)^{1/\beta}$; right hand side: empirical ROC-curve and true accuracy ratio, upper and lower limits for AR, AR_{mv}-, AR_α- and AR_β-estimators for various error-I-II-combinations, average value of AR_α- and AR_β-estimators ($AR_{\alpha\&\beta}$)

The examinations show that:

- only few empirical ROC-curves do correspond well with α-type ROC-curves, $ROC_\alpha(x)=x^\alpha$, one striking exception is the ROC-curve of the CREDITREFORM Deutschland Bonitätsindex that is given in Figure 24, all other accuracy ratio equivalent ROC_α-curves, however, are steeper than the empirical ROC-curves both for very small and very large errors of type II, while they are too flat for medium errors of type II,
- some of the empirical ROC-curves can be described quite well by ROC_β-curves, with $ROC_\beta=1-(1-x)^{1/\beta}$, see Figure 26 to Figure 30 – in all these and all remaining cases, however, accuracy ratio equivalent ROC_β-curves are too flat for both small and large errors of type II and too steep for medium errors of type II – which is directly opposed to the deviations that could be observed in case of ROC_α-curves,
- in the remaining cases (see Figure 23, Figure 25, Figure 31) empirical ROC-curves are positioned in about the middle of the space that is spanned by accuracy-ratio-equivalent ROC_α- and ROC_β-curves.

Altogether, the examinations showed that the shapes of empirical ROC-curves are characterized by a considerable diversity. Some of them are matched well by α-type curves, others by β-type curves and still others by a mixture of both types. As α-type curves can be transformed to β-type curves via reflection at the secondary diagonal, there is probably no functional form with *only one parameter* that is able to include both α- and β-type curves as special cases and still allows a good fit to empirical ROC-curves.

For the reasons stated above, the heuristic accuracy ratio estimators AR_α and AR_β should exhibit systematic biases both for very large and very small errors of type II – though, with opposite signs, so that the average of both measures, $AR_{\alpha\&\beta}$, should be a more suitable estimator.

ROC_α-curves are highly discriminative for "bad" corporations, i.e. they are able to identify groups of corporations with considerably high default rates. In the range of "good" corporations, however, they are characterized by a rather low discriminative power (see the comparatively steep, nearly linear progression of the respective ROC-curves for large errors of type II). ROC_β-curves, on the other hand, perform relatively poor among "bad corporations" (see the comparatively flat, nearly linear progression of the ROC-curves for small errors of type II), but they are highly discriminative for good corporations, because they are able to identify large groups of corporations with very low default rates.

Annotation: If more than just *one* error-I-II-combination were available for inferring the accuracy ratio of a rating model, one *additional* parameter (besides α or β) could be estimated, which stated by how much the respective ROC-curve (that has to be interpolated) corresponds to either one of the two extremes (α- or β-type).

Concerning the quality of the various heuristic accuracy ratio estimators, examinations based on the nine empirical ROC-curves revealed the following:

- AR_{mv}: The average value of the upper and lower limit for the accuracy ratio converges to 50% both for very small and very large errors of type II (as in the respective cases AR_{max} converges to 100% and AR_{min} to 0%) and thus is, at least for all rating models with an accuracy ratio that exceeds 50%, negatively biased. Also for error-I-II-combinations with identical errors of types I and II (see intersection points of secondary diagonals and ROC-curves) AR_{mv} tends to be too small (here: in 8 of 9 cases), but misses the true accuracy ratios by only -1 to -5 percentage points,
- AR_α: In all nine cases AR_α underestimates the true accuracy ratio for small errors of type II and overestimates it for large errors. In the range of identical errors of types I and II, the true accuracy ratio is hit very precisely [only in one case the deviation exceeds +/- 3 PP],

- AR_{β}: Directly opposed to AR_{α}, AR_{β} overestimates the true accuracy ratios for small and underestimates them for large errors of type II. In the range of identical errors of types I and II, the true accuracy ratio is hit very precisely [the deviation never exceeds +/- 3 PP],
- $AR_{\alpha\&\beta}$: In accordance with AR_{α} and AR_{β}, $AR_{\alpha\&\beta}$ very precisely hits true accuracy ratios in the range of identical errors of types I and II. But contrary to AR_{mv}, AR_{α}, and AR_{β} there is no identifiable, uniform bias pattern for $AR_{\alpha\&\beta}$ in respect with sizes of errors of type II. For small errors of type II $AR_{\alpha\&\beta}$ overestimates the true accuracy ratio in three cases and underestimates it in six cases. For very large errors of type II $AR_{\alpha\&\beta}$ overestimates true accuracy ratios in one case and underestimates it in five cases[323]. The recurring estimation errors in case of extreme errors of type II are an order of magnitude smaller than those for AR_{α} and AR_{β}. In general, $AR_{\alpha\&\beta}$ gives very precise accuracy ratio estimations for a wide range of error-I-II-combinations.

Conclusion: For error-I-II-combinations with identical errors of types I and II, all heuristic estimators, AR_{mv}, AR_{α}, AR_{β}, and $AR_{\alpha\&\beta}$, predict true accuracy ratios quite well (only AR_{mv} is – marginally – negatively biased). In case of asymmetric error-I-II-combinations, however, AR_{mv}, AR_{α}, and AR_{β} exhibit systematic and substantial biases, whereas $AR_{\alpha\&\beta}$ still gives very precise estimations of true accuracy ratios. Thus, compared with all other estimators, $AR_{\alpha\&\beta}$ only possesses advantages and no disadvantages and therefore should be the only choice.

In the surveys in chapters 3.3 to 3.5 $AR_{\alpha\&\beta}$ is used as the sole estimator for transforming error-I-II-combinations to accuracy ratios.

[323] In the remaining cases no error-I-II-combinations with large errors of type II were available.

Appendix II: Incentive compatibility of various measures

Intention: It shall be shown that the expected values of the conditional information entropy (CIE, logarithmic sconce function) and the Brier score (quadratic sconce function) are minimized (=exhibit their best values) then, when a rater states his true subjective probabilities as forecasted probabilities (incentive compatibility). For illustrative purposes, one non-incentive compatible example - based on a simple absolute value sconce function - is presented.

Conditional information entropy (CIE)

F 118) $$CIE = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\left|PD_{i,fore} + \theta_i - 1\right|\right)$$

Note: CIE is only not defined in those cases, where a default occurs, although it was ruled out with certainty (Θ_i=1 *and* $PD_{i,fore}$=0) or in those cases where no default occurs, although it was forecasted with certainty (Θ_i=0 *and* $PD_{i,fore}$=1).

F 119) $$E(CIE) = -\frac{1}{n}\sum_{i=1}^{n} PD_{i,real} \cdot \log PD_{i,fore} + (1 - PD_{i,real}) \cdot \log(1 - PD_{i,fore}) \quad \text{for all i=1,...,n,}$$

F 120) $$\frac{\partial E(CIE)}{\partial PD_{i,fore}} = \frac{1}{n} \cdot \left(\frac{-PD_{i,real}}{PD_{i,fore}} - \frac{(1 - PD_{i,real})}{1 - PD_{i,fore}} \cdot (-1) \right) \quad \text{[for all i]}$$

F 121) $$\frac{\partial E(CIE)}{\partial PD_{i,fore}} = \frac{1}{n} \cdot \frac{-PD_{i,real} \cdot (1 - PD_{i,fore}) + (1 - PD_{i,real}) \cdot PD_{i,fore}}{PD_{i,fore} \cdot (1 - PD_{i,fore})},$$

F 122) $$\frac{\partial E(CIE)}{\partial PD_{i,fore}} = \frac{1}{n} \cdot \frac{PD_{i,real} - PD_{i,real}}{PD_{i,fore} \cdot (1 - PD_{i,fore})},$$

F 123) $$\frac{\partial E(CIE)}{\partial PD_{i,fore}} = 0 \text{ for } PD_{i,fore} = PD_{i,real}$$

F 124) $$\frac{\partial\left(\frac{\partial E(CIE)}{\partial PD_{i,fore}}\right)}{\partial PD_{i,fore}} = \frac{1}{n} \cdot \frac{(PD_{i,fore} - PD_{i,real})^2 + PD_{i,real} \cdot (1 - PD_{i,real})}{(PD_{i,fore} - 1)^2 PD_{i,fore}^2} > 0 \text{ for } PD_{i,fore}=PD_{i,real},$$

i.e. E(CIE) is minimized, when $PD_{i,fore}$=$PD_{i,real}$ for every i [324]

Smaller CIE-values correspond with *better* predictions. The smallest – and thus best – attainable CIE-value is zero and is being achieved only, when all forecasted default probabilities are either 0% or 100% and when they are always correct. But also in case of not perfectly discriminative subjective probabilities, CIE is minimized, when true subjective probabilities are stated.

[324] The marginal solutions $PD_{i,fore}$=0% and $PD_{i,fore}$ =100% are not examined, because E(CIE) is not defined for them (except for $PD_{i,real}$=0 or $PD_{i,real}$=1), see notes above. Unless $PD_{i,real}$=1/0, if $PD_{i,fore}$ approaches 1/0, CIE approaches infinity. Thus, the local optimum at $PD_{i,fore}$=$PD_{i,real}$ is also the global optimum.

Brier score

F 125) $$BS = \frac{1}{n}\sum_{i=1}^{n}\left(PD_{i,fore} - \theta_i\right)^2$$

F 126) $$E(BS) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real}\cdot\left(1-PD_{i,fore}\right)^2 + \left(1-PD_{i,real}\right)\cdot PD_{i,fore}^2$$

F 127) $$E(BS) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real} - 2PD_{i,real}\cdot PD_{i,fore} + PD_{i,real}\cdot PD_{i,fore}^2 + PD_{i,fore}^2 - PD_{i,fore}^2\cdot PD_{i,real}$$

F 128) $$E(BS) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real} - 2PD_{i,real}\cdot PD_{i,fore} + PD_{i,fore}^2$$

F 129) $$E(BS) = \frac{1}{n}\sum_{i=1}^{n}\left(PD_{i,fore} - PD_{i,real}\right)^2 + PD_{i,real}\cdot\left(1-PD_{i,fore}\right)$$

F 130) $$\frac{\partial E(BS)}{\partial PD_{i,fore}} = \frac{2}{n}\cdot\left(PD_{i,fore} - PD_{i,real}\right) \quad \text{for all } i=1,\ldots,n$$

F 131) $$\frac{\partial E(BS)}{\partial PD_{i,fore}} = 0 \text{ for } PD_{i,fore} = PD_{i,real}$$

F 132) $$E(BS_{min}) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real}\cdot\left(1-PD_{i,real}\right)^2 + \left(1-PD_{i,real}\right)\cdot PD_{i,real}^2$$

F 133) $$E(BS_{min}) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real}\cdot\left(1-PD_{i,real}\right)$$

F 134) $$\frac{\partial\left(\frac{\partial E(BS)}{\partial PD_{i,fore}}\right)}{\partial PD_{i,fore}} = \frac{2}{n} > 0$$, i.e. E(BS) is minimized, if $PD_{i,fore}=PD_{i,real}$ for all i

Examination of marginal solutions ($PD_{i,fore}=0$ or $PD_{i,fore}=1$ in F 126) yield:

F 135) $$E(BS, PD_{i,fore} = 0) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real} \geq E(BS_{min})$$

F 136) $$E(BS, PD_{i,fore} = 1) = \frac{1}{n}\sum_{i=1}^{n} 1 - PD_{i,real} \geq E(BS_{min})$$

See comments on CIE: The Brier score is minimized, if true subjective probabilities are stated, i.e. it is incentive compatible, too.

A score based on an absolute value "sconce function" as example for a non incentive compatible score

F 137) $$A = \frac{1}{n}\sum_{i=1}^{n}\left|PD_{i,fore} - \theta_i\right|$$

F 138) $$E(A) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - PD_{i,real}\right)\cdot PD_{i,fore} + PD_{i,real}\cdot\left(1 - PD_{i,fore}\right)$$

F 139) $$E(A) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,fore} - PD_{i,fore}\cdot PD_{i,real} + PD_{i,real} - PD_{i,real}\cdot PD_{i,fore}$$

F 140) $$E(A) = \frac{1}{n}\sum_{i=1}^{n} PD_{i,fore} - 2\cdot PD_{i,fore}\cdot PD_{i,real} + PD_{i,real}$$

F 141) $$\frac{\partial E(A)}{\partial PD_{i,fore}} = \frac{1}{n}\left(1 - 2\cdot PD_{i,real}\right)$$

F 142) $$\frac{\partial E(A)}{\partial PD_{i,fore}} > 0 \quad \text{for} \quad PD_{i,real} < 50\% \quad \text{and} \quad \frac{\partial E(A)}{\partial PD_{i,fore}} < 0 \quad \text{für} \quad PD_{i,real} > 50\%$$

F 143) $$E(A)_{PD_{i,fore}=0 \,\wedge\, PD_{i,real}<50\%} = \frac{1}{n}\sum_{i=1}^{n} PD_{i,real}$$

F 144) $$E(A)_{PD_{i,fore}=PD_{i,real} \,\wedge\, PD_{i,real}<50\%} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - PD_{i,real}\right)\cdot PD_{i,real} + PD_{i,real}\cdot\left(1 - PD_{i,real}\right)$$

F 145) $$= \frac{1}{n}\sum_{i=1}^{n} PD_{i,real}\cdot 2\cdot\left(1 - PD_{i,real}\right)$$

F 146) $$2\cdot\left(1 - PD_{i,real}\right) > 1 \quad \text{for } PD_{i,real} < 50\%$$

If subjective probabilities of default, $PD_{i,real}$, are smaller than 50%, then the resulting expected score value that is based on an absolute-value-sconce-function is the better (=lower), the smaller the respective *forecasted* probability, $PD_{i,fore}$, is. Therefore, this score does not reward unbiased probabilities but is rewarding the statement of probabilities that are small as possible, i.e. 0%.

Stating true subjective probabilities is punished, if probabilities of default $PD_{i,tat}$ are very small, giving the unbiased prediction $PD_{i,fore} = PD_{i,real}$ leads to expected sconce values that are twice as large as those when stating $PD_{i,fore} = 0\%$.

Annotation: For $PD_{i,real} > 50\%$ stating always $PD_{i,fore}=100\%$ is being awarded.

Appendix III: Dependency of various measures of prediction accuracy on the average default rate

Intention: Based on a simple model, dependencies of the following measures for predictive quality from the average sample default rate shall be examined: Brier score, skill-Brier score, entropy, skill-entropy, accuracy ratio.

It is assumed, that a simple two-class reference rating system is - under all *environmental conditions* (i.e. *sample default rates*) - able to separate all corporations of a sample into two groups, whereby the first group shall be exclusively composed of non-defaulters and shall have a share in all corporations of exactly *1-a*. The share of the second group shall be *a* (*independent* from the "environment") and thus the probability of default of these corporations will be *PD/a* (*dependent* from the "environment"). It is assumed, that the rating model is correctly calibrated.

In the following, *environmental dependence* (in the above sense) is examined for the various predictive quality measures. [325]

Brier score

If correctly calibrated, the Brier score (see formula F 37) is given by:

F 147) $BS = \sum_{i=1}^{g} a_i \cdot PD_i \cdot (1 - PD_i)$ here: with g=2, PD_1=0%, a_1=1-a, PD_2=PD/a, a_2=a

F 148) $BS = (1-a) \cdot 0 + a \cdot \frac{PD}{a} \cdot \left(1 - \frac{PD}{a}\right)$

F 149) $BS = PD \cdot \left(1 - \frac{PD}{a}\right)$

F 150) $\frac{\partial BS}{\partial PD} = 1 - \frac{2PD}{a}$

F 151) $\frac{\partial BS}{\partial PD} > 0 \quad \text{für} \quad PD < \frac{a}{2}, \text{ d.h. } \frac{PD}{a} < 50\%$

Interpretation: For as long as the probability of default in the second class is smaller than 50%, every increase in the average probability of default, PD, results in an increase (=deterioration) of the Brier score. The strength of this *undesired dependence* can be quantified with a subsequent analysis of elasticity.

The elasticity of the Brier score with respect to PD states, by how many percent the Brier score rises, if the average default rate rises by one percent [*not percentage point*] and is defined as follows:

[325] *Environmental independence* as interpreted in the sense of CANTOR, MANN (2003, p. 12) is given, when an accuracy measure is invariant with respect to proportional changes in rating class specific default rates, see on this also DVFA (2004, p. 599). See the annotations in chapter 2.3.2 for criticism on this concept. For an alternative definition of *environmental independence* see SWETS (1988, p. 1286), see also OENB (2004a, p. 117f.). According to that definition, an accuracy measure would be considered *environmentally independent*, if it was independent from the share of defaulters in the test sample (whereby defaulters and non-defaulters are randomly drawn from their respective basic populations).

F 152) $$\varepsilon_{BS,PD} = \frac{\left(\frac{\partial BS}{BS}\right)}{\left(\frac{\partial PD}{PD}\right)} = \frac{\left(\frac{\partial BS}{\partial PD}\right)}{\left(\frac{BS}{PD}\right)}$$

F 153) $$\varepsilon_{BS,PD} = \frac{1-\frac{2PD}{a}}{\frac{PD\cdot\left(1-\frac{PD}{a}\right)}{PD}}$$

F 154) $$\varepsilon_{BS,PD} = \frac{\frac{a-2PD}{a}}{\left(\frac{a-PD}{a}\right)}$$

F 155) $$\varepsilon_{BS,PD} = \frac{a-2PD}{a-PD}$$

F 156) $$\lim_{PD\to 0} \varepsilon_{BS,PD} = \frac{a}{a} = 1$$

Interpretation: For small PD the elasticity of the Brier score with respect to PD, $\varepsilon_{BS,PD}$, is approximately equal to 1, i.e. for small PD the Brier score nearly linearly rises with increasing probabilities of default. This implies that there is a *strong environmental dependence* of the Brier score. This *strong environmental dependence* could also be confirmed empirically (here even in a literal sense, see chapter 2.4) and implicates that Brier score values of samples with differing default rates cannot be directly compared.
For such purposes the usage of skill-measures was suggested in order to correct for environmental dependence (see chapter 2.4), which are formed by dividing the accuracy measures that a certain model achieves in a given environment by the same measures that a *naïve rating model* achieves in the same environment. A *naïve rating model* is a model, that always states the same (but unbiased) probability of default PD for every corporate.

Skill-Brier score

If correctly calibrated, the Skill-Brier score (see formula F 34, F 35, and F 37) is given by:

F 157) $$Skill_{BS} = 1 - \frac{BS}{PD\cdot(1-PD)}$$ from formula F 149 follows:

F 158) $$Skill_{BS} = 1 - \frac{PD\cdot\left(1-\frac{PD}{a}\right)}{PD\cdot(1-PD)}$$ by canceling down PD and expanding the first term with (1-PD) follows:

F 159) $$Skill_{BS} = \frac{1-PD-1+\frac{PD}{a}}{1-PD}$$

F 160) $$\mathrm{Skill}_{BS} = \frac{\mathrm{PD} \cdot \left(\frac{1}{a} - 1 \right)}{1 - \mathrm{PD}}$$

F 161) $$\mathrm{Skill}_{BS} = \frac{\frac{1}{a} - 1}{\frac{1}{\mathrm{PD}} - 1}$$

F 162) $$\frac{\partial \mathrm{Skill}_{BS}}{\partial \mathrm{PD}} = \frac{\frac{1}{a} - 1}{-\left(\frac{1}{\mathrm{PD}} - 1 \right)^2 \cdot -\mathrm{PD}^2}$$ for 0%<PD<100% follows:

F 163) $$\frac{\partial \mathrm{Skill}_{BS}}{\partial \mathrm{PD}} = \frac{\frac{1}{a} - 1}{(1 - \mathrm{PD})^2} > 0$$

Interpretation: Every increase in the average probability of default, PD, results in an increase (=*improvement*) of the skill-Brier score. This is a surprising result because an increase in PD leads to a simultaneous *deterioration* of the underlying Brier score. This result is rather unsatisfying, because – counter to its original intention – the skill-Brier-score is obviously environmentally dependent, too.

The strength of this *undesired dependence* will be quantified with the subsequent analysis of elasticity.

F 164) $$\varepsilon_{\mathrm{SkillBS,PD}} = \frac{\left(\frac{\partial \mathrm{Skill}_{BS}}{\mathrm{SkS}_{\mathrm{Brier}}} \right)}{\left(\frac{\partial \mathrm{PD}}{\mathrm{PD}} \right)} = \frac{\left(\frac{\partial \mathrm{Skill}_{BS}}{\partial \mathrm{PD}} \right)}{\left(\frac{\mathrm{Skill}_{BS}}{\mathrm{PD}} \right)}$$

F 165) $$\varepsilon_{\mathrm{SkillBS,PD}} = \frac{\frac{\frac{1}{a} - 1}{(1 - \mathrm{PD})^2}}{\frac{\left(\frac{\frac{1}{a} - 1}{\frac{1}{\mathrm{PD}} - 1} \right)}{\mathrm{PD}}} = \mathrm{PD} \cdot \frac{\frac{1}{\mathrm{PD}} - 1}{(1 - \mathrm{PD})^2}$$

F 166) $$\varepsilon_{\mathrm{SkillBS,PD}} = \frac{1}{1 - \mathrm{PD}} > 0$$

F 167) $$\lim_{\mathrm{PD} \to 0} \varepsilon_{\mathrm{SkillBS,PD}} = 1$$

Interpretation: The skill-Brier score is the bigger (=better), the bigger PD is. For small PD, the elasticity of the skill-Brier score $\varepsilon_{\mathrm{Skill,,PD}}$ is nearly equal to 1, i.e. for small PD the skill-Brier score nearly linearly rises with increasing probabilities of default, which implies a *strong environmental dependence*.

Conditional Information Entropie (CIE)

If correctly calibrated, CIE (see formula F 36) is given by:

F 168) $\mathrm{CIE}_{\mathrm{cal}} = -\sum_{i=1}^{g} a_i \left(\mathrm{PD}_i \cdot \log \mathrm{PD}_i + \left(1-\mathrm{PD}_i\right) \cdot \log\left(1-\mathrm{PD}_i\right)\right)$ here with g=2, PD_1=0%, a_1=1-a, PD_2=PD/a, a_2=a (see assumptions at the beginning of this chapter):

F 169) $$\mathrm{CIE} = -\left((1-a)\cdot(0+1\cdot\log(1)) + a\cdot\left(\frac{\mathrm{PD}}{a}\cdot\log\left(\frac{\mathrm{PD}}{a}\right) + \left(1-\frac{\mathrm{PD}}{a}\right)\cdot\log\left(1-\frac{\mathrm{PD}}{a}\right)\right)\right)$$

F 170) $$\mathrm{CIE} = -\left(\mathrm{PD}\cdot\log\left(\frac{\mathrm{PD}}{a}\right) + (a-\mathrm{PD})\cdot\log\left(\frac{a-\mathrm{PD}}{a}\right)\right)$$

F 171) $$\frac{\partial \mathrm{CIE}}{\partial \mathrm{PD}} = -\left(\log\left(\frac{\mathrm{PD}}{a}\right) + \mathrm{PD}\cdot\frac{a}{\mathrm{PD}}\cdot\frac{1}{a} - \log\left(\frac{a-\mathrm{PD}}{a}\right) + (a-\mathrm{PD})\cdot\frac{a}{a-\mathrm{PD}}\cdot\frac{1}{a}\cdot(-1)\right)$$

F 172) $$\frac{\partial \mathrm{CIE}}{\partial \mathrm{PD}} = -\left(\log\left(\frac{\mathrm{PD}}{a}\right) + 1 - \log\left(\frac{a-\mathrm{PD}}{a}\right) - 1\right)$$

F 173) $$\frac{\partial \mathrm{CIE}}{\partial \mathrm{PD}} = \log\left(\frac{a-\mathrm{PD}}{\mathrm{PD}}\right)$$

For $\mathrm{PD} < \frac{a}{2}$ (i.e. PD/a < 50%) the enumerator inside the logarithm function is bigger than the denominator and thus $\frac{\partial \mathrm{CIE}}{\partial \mathrm{PD}} > 0$, while $\frac{\partial \mathrm{CIE}}{\partial \mathrm{PD}} \le 0$ for $\mathrm{PD} \ge \frac{a}{2}$.

Interpretation: For as long as the probability of default in the second rating class is smaller than 50%, every increase in PD is accompanied by an increase (i.e. *deterioration*) of the conditional information entropy, CIE.

The elasticity of CIE in respect to PD is given as follows:

F 174) $$\varepsilon_{\mathrm{CIE,PD}} = \frac{\left(\frac{\partial \mathrm{CIE}}{\mathrm{CIE}}\right)}{\left(\frac{\partial \mathrm{PD}}{\mathrm{PD}}\right)} = \frac{\left(\frac{\partial \mathrm{CIE}}{\partial \mathrm{PD}}\right)}{\left(\frac{\mathrm{CIE}}{\mathrm{PD}}\right)}$$

F 175) $$\varepsilon_{\mathrm{CIE,PD}} = \frac{\log\left(\frac{a-\mathrm{PD}}{\mathrm{PD}}\right)}{\frac{-\left(\mathrm{PD}\cdot\log\left(\frac{\mathrm{PD}}{a}\right) + (a-\mathrm{PD})\cdot\log\left(\frac{a-\mathrm{PD}}{a}\right)\right)}{\mathrm{PD}}}$$

F 176) $$\varepsilon_{\mathrm{CIE,PD}} = -\frac{\log\left(\frac{a-\mathrm{PD}}{\mathrm{PD}}\right)}{\log\left(\frac{\mathrm{PD}}{a}\right) + \frac{a-\mathrm{PD}}{\mathrm{PD}}\cdot\log\left(\frac{a-\mathrm{PD}}{a}\right)}$$

F 177) $$\varepsilon_{CIE,PD} = \frac{\log\left(\frac{PD}{a-PD}\right)}{\log\left(\frac{PD}{a}\right)+\frac{a-PD}{PD}\cdot\log\left(\frac{a-PD}{a}\right)}$$

with $\lim_{PD\to 0}\varepsilon_{CIE,PD} = 1$ [326]

Interpretation: For small PD, the elasticity of the CIE with respect to PD is nearly equal to 1, i.e. CIE nearly linearly rises with increasing probabilities of default for small PD, which implies a *strong environmental dependence*.

Skill-entropie (conditional information entropy ratio, CIER)

From formulas F 29f. and F 177 follows:

F 178) $$CIER = \frac{CIE_{PD} - CIE}{CIE_{PD}}$$

F 179) $$CIER = \frac{PD\cdot\log PD + (1-PD)\cdot\log(1-PD) - \left(PD\cdot\log\left(\frac{PD}{a}\right)+(a-PD)\cdot\log\left(\frac{a-PD}{a}\right)\right)}{PD\cdot\log PD + (1-PD)\cdot\log(1-PD)}$$

F 180) $$CIER = \frac{PD\cdot\log PD + (1-PD)\cdot\log(1-PD) - (PD\cdot\log PD - PD\cdot\log a + (a-PD)\cdot\log(a-PD) - (a-PD)\cdot\log a)}{PD\cdot\log PD + (1-PD)\cdot\log(1-PD)}$$

F 181) $$CIER = \frac{(1-PD)\cdot\log(1-PD) - (a-PD)\cdot\log(a-PD) + a\cdot\log a}{PD\cdot\log PD + (1-PD)\cdot\log(1-PD)}$$

with $enumerator_{CIER} = (1-PD)\cdot\log(1-PD) - (a-PD)\cdot\log(a-PD) + a\cdot\log a$ and
$denominator_{CIER} = CIE_{PD} = PD\cdot\log PD + (1-PD)\cdot\log(1-PD)$ follows:

F 182) $$\frac{\partial CIER}{\partial PD} = \frac{\frac{\partial enumerator_{CIER}}{\partial PD}\cdot denominator_{CIER} - enumerator_{CIER}\cdot\frac{\partial denominator_{CIER}}{\partial PD}}{denominator_{CIER}^{\,2}}$$

F 183) $$\frac{\partial CIER}{\partial PD} = \frac{1}{CIE_{PD}^{\,2}}\left(\cdot\begin{pmatrix} -1\cdot\log(1-PD) + (1-PD)\cdot\frac{1}{1-PD}\cdot(-1) \\ +\log(a-PD) + (-a+PD)\cdot\frac{1}{a-PD}\cdot(-1)\end{pmatrix}\cdot denominator_{CIER} - enumerator_{CIER}\cdot\ln\left(\frac{PD}{1-PD}\right)\right)$$

F 184) $$\frac{\partial CIER}{\partial PD} = \frac{\log(a-PD)\cdot denominator_{CIER} - \log(PD)\cdot enumerator_{CIER} + \log(1-PD)\cdot(enumerator_{CIER} - denominator_{CIER})}{CIE_{PD}^{\,2}}$$

F 185) $$\frac{\partial CIER}{\partial PD} = \frac{1}{CIE_{PD}^{\,2}}\cdot\begin{pmatrix} PD\cdot\log(PD)\cdot\log(a-PD) + (1-PD)\cdot\log(1-PD)\cdot\log(a-PD) \\ -(1-PD)\cdot\log(1-PD)\log(PD) + (a-PD)\cdot\log(a-PD)\log(PD) - a\cdot\log(a)\log(PD) \\ -(a-PD)\cdot\log(a-PD)\cdot\log(1-PD) + a\cdot\log(a)\cdot\log(1-PD) - PD\cdot\log(PD)\cdot\log(1-PD)\end{pmatrix}$$

F 186) $$\frac{\partial CIER}{\partial PD} = \frac{1}{CIE_{PD}^{\,2}}\cdot\begin{pmatrix} +\log(PD)\cdot\log(1-PD)\cdot(-1+PD-PD) \\ +\log(PD)\cdot\log(a)\cdot(-a) \\ +\log(PD)\cdot\log(a-PD)\cdot(PD+a-PD) \\ +\log(1-PD)\cdot\log(a-PD)\cdot(1-PD-a+PD) \\ +\log(a)\cdot\log(1-PD)\cdot a\end{pmatrix}$$

[326] calculated with *Mathematica 5*

F 187) $$\frac{\partial CIER}{\partial PD} = \frac{1}{CIE_{PD}{}^2} \cdot \begin{pmatrix} a \cdot \log(PD) \cdot \log(a - PD) + (1-a) \cdot \log(1-PD) \cdot \log(a-PD) + a \cdot \log(a) \cdot \log(1-PD) \\ -\log(PD) \cdot \log(1-PD) - a \cdot \log(PD) \cdot \log(a) \end{pmatrix}$$

The first term within the brackets is at least as big as the absolute value of the sum of the (negative) terms 4 and 5:

Proof: it has to be shown that $a \cdot \log(PD) \cdot \log(a-PD) > \log(PD) \cdot \log(1-PD) + a \cdot \log(PD) \cdot \log(a)$:

According to BERNOULLI's inequality[327] (generalized for real exponents) follows that:

F 188) $(1+x)^r < 1 + x \cdot r \quad$ for $0 < r < 1$ and for $x > -1$

For r=a with 0<a<1 and x = - PD/a with -PD/a > -1 follows:

F 189) $\left(1 - \frac{PD}{a}\right)^a < 1 - a \cdot \frac{PD}{a}$ and thus

F 190) $\left(\frac{a-PD}{a}\right)^a < 1 - PD$ and it follows that

F 191) $a \cdot (\log(a-PD) - \log(a)) < \log(1-PD)$

By extending the terms with log(PD), log(PD)<0 follows:

F 192) $a \cdot \log(PD) \cdot \log(a-PD) - a \cdot \log(PD) \cdot \log(a) > \log(1-PD) \cdot \log(PD)$ and thus

F 193) $a \cdot \log(PD) \cdot \log(a-PD) > \log(PD) \cdot \log(1-PD) + a \cdot \log(PD) \cdot \log(a)$ q.e.d.

F 194) $$\frac{\partial CIER}{\partial PD} > 0$$

And further: [328]

F 195) $\lim_{PD \to 0} CIER = 0$

F 196) $\lim_{PD \to a} CIER = 1$

F 197) $\lim_{PD \to 0} \frac{\partial CIER}{\partial PD} = \infty$

F 198) $\lim_{PD \to a} \frac{\partial CIER}{\partial PD} = \infty$

F 199) $\lim_{PD \to 0} \varepsilon_{CIER,PD} = 0$

F 200) $\lim_{PD \to a} \varepsilon_{CIER,PD} = \infty$

Interpretation: Analog to the skill-Brier score does the skill-entropy measures (CIER) improve with increasing average sample default rates. But for small PD, the elasticity of the CIER with respect to PD is nearly equal to 0, i.e. at least for small PD the skill-entropy is less environmental dependent than all other measures examined so far.

[327] see for instance http://en.wikipedia.org/wiki/Bernoulli%27s_inequality (26/04/2005). Jens Eisenschmidt pointed out the applicability of Bernoulli's inequality in this case.

[328] Formulas F 197 to F 200 were obtained with *Mathematica 5*.

Accuracy Ratio

For a two class rating system, whereby all defaulters are captured by the second class, which embraces a% of all corporations, the accuracy ratio is given as follows:

F 201) $$AR = \frac{a \cdot \frac{(0+1)}{2} + (1-a) \cdot 1 - \frac{1}{2}}{\frac{1}{2} - \frac{PD}{2}}$$

F 202) $$AR = \frac{a \cdot \frac{1}{2} + 1 - a - \frac{1}{2}}{\frac{1}{2} - \frac{PD}{2}}$$

F 203) $$AR = \frac{\frac{1}{2} - \frac{1}{2} \cdot a}{\frac{1}{2} - \frac{PD}{2}}$$

F 204) $$AR = \frac{1-a}{1-PD}$$

Annotation: An accuracy ratio of 100% (perfect rating) results, if every corporate of the second rating class is a defaulter.

F 205) $$\frac{\partial AR}{\partial PD} = \frac{1-a}{(1-PD)^2} > 0$$

F 206) $$\varepsilon_{AR,PD} = \frac{\left(\frac{\partial AR}{AR}\right)}{\left(\frac{\partial PD}{PD}\right)} = \frac{\left(\frac{\partial AR}{\partial PD}\right)}{\left(\frac{AR}{PD}\right)} = \frac{\left(\frac{1-a}{(1-PD)^2}\right)}{\frac{\left(\frac{1-a}{1-PD}\right)}{PD}}$$

F 207) $$\varepsilon_{AR,PD} = \frac{PD}{1-PD}$$

F 208) $$\lim_{PD \to 0} \varepsilon_{AR,PD} = \frac{0}{1} = 0$$

Interpretation: Every increase in PD is accompanied by an increase (=improvement) of the associated accuracy ratios. Thus, the accuracy ratio is an environmental dependent measure, too - at least when *environmental dependency* is defined in the above sense, but not in a sense that was outlined in chapter 2.3.2.
However, even though the accuracy ratios measure is *not completely environmental independent* in the above sense, it is at least *relatively insensitive* with respect to environmental influences ($\varepsilon_{AR,PD} \approx 0$ for small PD).

Annotation: For the MOODY'S-defined AR (see chapter 2.3.2) instead of formulas F 204 and F 207 following equations would result:

F 209) $$AR_{Moody's} = 1 - a$$

F 210) $$\varepsilon_{AR-Moody's,PD} = 0$$

In Table L and Table M formal results of this chapter are summarized. In the graphs following thereafter, Figure 32 to Figure 36, numerical examples for the five measures considered are given.

	Brier score	**skill-Brier score**	**accuracy ratio**
original measure	$BS = PD \cdot \left(1 - \frac{PD}{a}\right)$	$Skill_{BS} = \frac{\frac{1}{a} - 1}{\frac{1}{PD} - 1}$	$AR = \frac{1-a}{1-PD}$
first derivative with respect to PD	$\frac{\partial BS}{\partial PD} = 1 - \frac{2PD}{a}$ >0 for PD/a<50%	$\frac{\partial Skill_{BS}}{\partial PD} = \frac{\frac{1}{a} - 1}{(1-PD)^2} > 0$	$\frac{\partial AR}{\partial PD} = \frac{1-a}{(1-PD)^2} > 0$
elasticity with respect to PD	$\varepsilon_{BS,PD} = \frac{a - 2PD}{a - PD}$	$\varepsilon_{SkillBS,PD} = \frac{1}{1-PD}$	$\varepsilon_{AR,PD} = \frac{PD}{1-PD}$
elasticity for small PD	$\varepsilon_{BS,PD} \approx 1$	$\varepsilon_{SkillBS,PD} \approx 1$	$\varepsilon_{BS,PD} \approx 0$

Table L: various predictive accuracy measures (table 1 of 2), first derivatives and sensitivities with respect to PD for a simple two-class rating system,

	conditional information entropy (CIE)	**skill-entropy (CIER)**
original measure	$CIE = -\left(PD \cdot \log\left(\frac{PD}{a}\right) + (a - PD) \cdot \log\left(\frac{a - PD}{a}\right)\right)$	$CIER = \frac{(1-PD)\cdot\log(1-PD) - (a-PD)\cdot\log(a-PD) + a\cdot\log$ $PD\cdot\log(PD) + (1-PD)\cdot\log(1-PD)$
first derivative with respect to PD	$\frac{\partial CIE}{\partial PD} = \log\left(\frac{a - PD}{PD}\right)$ >0 for PD/a<50%	$\frac{\partial CIER}{\partial PD} = \frac{1}{CIE_{PD}^{\,2}} \cdot$ $\left(\begin{array}{l} a\cdot\log(PD)\cdot\log(a-PD) \\ +(1-a)\cdot\log(1-PD)\cdot\log(a-PD) \\ +a\cdot\log(a)\cdot\log(1-PD) \\ -\log(PD)\cdot\log(1-PD) \\ -a\cdot\log(PD)\cdot\log(a) \end{array}\right)$
elasticity with respect to PD	$\varepsilon_{CIE,PD} = \frac{\log\left(\frac{PD}{a-PD}\right)}{\log\left(\frac{PD}{a}\right) + \frac{a-PD}{PD}\cdot\log\left(\frac{a-PD}{a}\right)}$	(..) [329]
elasticity for small PD	$\varepsilon_{CIE,PD} \approx 1$	$\varepsilon_{CIER,PD} \approx 0$

Table M: various predictive accuracy measures (table 2 of 2), first derivatives and sensitivities with respect to PD for a simple two-class rating system,

[329] The elasticity of CIER with respect to PD was calculated with *Mathematica 5*. The respective term is very "bulky" and complex. As this term is not needed in the further analysis, it was not reproduced in the table above. The boundary value of this term for $\lim_{PD \to 0}$ was calculated with *Mathematica 5* as well and is given in the table above.

Brier score		a [share of companies of the second rating class]				
		50%	25%	10%	5%	1%
PD - probability of default	0,50%	0,5%	0,5%	0,5%	0,5%	0,3%
	1,00%	1,0%	1,0%	0,9%	0,8%	0,0%
	1,50%	1,5%	1,4%	1,3%	1,1%	---
	2,00%	1,9%	1,8%	1,6%	1,2%	---
	2,50%	2,4%	2,3%	1,9%	1,3%	---
	3,00%	2,8%	2,6%	2,1%	1,2%	---
	3,50%	3,3%	3,0%	2,3%	1,1%	---
	4,00%	3,7%	3,4%	2,4%	0,8%	---
	4,50%	4,1%	3,7%	2,5%	0,5%	---
	5,00%	4,5%	4,0%	2,5%	0,0%	---
	10,00%	8,0%	6,0%	0,0%	---	---

Figure 32: some numerical values for the Brier score for various *a* and *PD*

skill-Brier score		a [share of companies of the second rating class]				
		50%	25%	10%	5%	1%
PD - probability of default	0,50%	0,5%	1,5%	4,5%	9,5%	49,7%
	1,00%	1,0%	3,0%	9,1%	19,2%	100,0%
	1,50%	1,5%	4,6%	13,7%	28,9%	---
	2,00%	2,0%	6,1%	18,4%	38,8%	---
	2,50%	2,6%	7,7%	23,1%	48,7%	---
	3,00%	3,1%	9,3%	27,8%	58,8%	---
	3,50%	3,6%	10,9%	32,6%	68,9%	---
	4,00%	4,2%	12,5%	37,5%	79,2%	---
	4,50%	4,7%	14,1%	42,4%	89,5%	---
	5,00%	5,3%	15,8%	47,4%	100,0%	---
	10,00%	11,1%	33,3%	100,0%	---	---

Figure 33: some numerical values for the skill Brier score for various *a* and *PD*

accuracy ratio		a [share of companies of the second rating class]				
		50%	25%	10%	5%	1%
PD - probability of default	0,50%	50,3%	75,4%	90,5%	95,5%	99,5%
	1,00%	50,5%	75,8%	90,9%	96,0%	100,0%
	1,50%	50,8%	76,1%	91,4%	96,4%	---
	2,00%	51,0%	76,5%	91,8%	96,9%	---
	2,50%	51,3%	76,9%	92,3%	97,4%	---
	3,00%	51,5%	77,3%	92,8%	97,9%	---
	3,50%	51,8%	77,7%	93,3%	98,4%	---
	4,00%	52,1%	78,1%	93,8%	99,0%	---
	4,50%	52,4%	78,5%	94,2%	99,5%	---
	5,00%	52,6%	78,9%	94,7%	100,0%	---
	10,00%	55,6%	83,3%	100,0%	---	---

Figure 34: some numerical values for the accuracy ratio for various *a* and *PD*

entropy (CIE)		a [share of companies of the second rating class]				
		50%	25%	10%	5%	1%
PD - probability of default	0,50%	2,8%	2,5%	2,0%	1,6%	0,7%
	1,00%	4,9%	4,2%	3,3%	2,5%	0,0%
	1,50%	6,7%	5,7%	4,2%	3,1%	---
	2,00%	8,4%	7,0%	5,0%	3,4%	---
	2,50%	9,9%	8,1%	5,6%	3,5%	---
	3,00%	11,3%	9,2%	6,1%	3,4%	---
	3,50%	12,7%	10,1%	6,5%	3,1%	---
	4,00%	13,9%	11,0%	6,7%	2,5%	---
	4,50%	15,1%	11,8%	6,9%	1,6%	---
	5,00%	16,3%	12,5%	6,9%	0,0%	---
	10,00%	25,0%	16,8%	0,0%	---	---

Figure 35: some numerical values for CIE for various *a* and *PD*

skill entropy (CIER)		a [share of companies of the second rating class]				
		50%	25%	10%	5%	1%
PD - probability of default	0,50%	11,0%	22,1%	36,9%	48,4%	78,0%
	1,00%	12,5%	25,0%	42,0%	55,3%	100,0%
	1,50%	13,5%	27,1%	45,7%	60,8%	---
	2,00%	14,3%	28,9%	49,0%	65,7%	---
	2,50%	15,1%	30,5%	51,9%	70,4%	---
	3,00%	15,8%	31,9%	54,7%	75,0%	---
	3,50%	16,4%	33,3%	57,3%	79,9%	---
	4,00%	17,0%	34,6%	59,9%	85,1%	---
	4,50%	17,6%	35,8%	62,5%	91,1%	---
	5,00%	18,1%	37,0%	65,1%	100,0%	---
	10,00%	23,0%	48,2%	100,0%	---	---

Figure 36: some numerical values for the skill entropy (CIER) for various *a* and *PD*

Within the scope of the assumptions chosen, it could be shown formally, that an increase of the average default rate:

- causes strong increases, i.e. *deteriorations*, of the Brier score and the conditional information entropy (CIE), and strong/ moderate increases, i.e. *improvements*, of the skill-Brier score/ skill-entropy (CIER),
- has practically no impact on the accuracy ratio (for small PD).

Owing to its strong environmental dependency, i.e. sensitivity with respect to the average default rate, Brier score, skill-Brier score and CIE are no suitable measures for comparing predictive quality of models, which are evaluated on the basis of samples with different default rates. A more suitable measure, with some reservations, is the skill-entropy (CIER), but least environmental dependent is the accuracy ratio.

Appendix IV: Estimating information losses attributable to discretization of continuous rating scales

Intention: Information losses shall be quantified, that result from discretizing continuous rating scores into discrete rating classes.

Further proceedings: For estimating the above mentioned information losses, two different approaches are chosen:

The first approach is based on *individual* probabilities of default that are interpolated based on rating class specific default rates. Afterwards, the information content, measured in accuracy ratio, of a portfolio with *individual* PD is compared to the information content of a portfolio with only *rating class specific* PD.

The second methods is assuming certain formal, parametric functional representations of ROC-curves and tests for a variety of parameter values, what percentage of the area under the ROC-curve, AUC_{ROC} (resp. AR), a rating system with only g discrete classes can achieve.

It turns out, that both approaches are yielding widely identical qualitative and quantitative results.

Procedure I: Interpolation of individual probabilities of defaults

The further proceedings according to procedure I are as follows:

1. In a simulation model a portfolio of n corporations is created (here n=5,000). The corporations are assigned to g rating classes (here g=17 with AAA=1, AA+=2, …, CCC/C=17) according to a given frequency distribution and are sorted by ascending rating classes.
2. Every corporate "inherits" the *rating class specific* probability of default of the rating class it belongs to,
3. The (expected) accuracy ratio of the given portfolio is determined (see formula F 22).
4. Individual probabilities of defaults (see point 2) are *smoothed* for allowing intra-class varying default rates: the smoothing is carried out by assigning to each corporate a weighted average of the (unsmoothed) individual default probabilities of its *b* "better" and b "worse" neighbors. For technical reasons, b additional corporations at the upper/ lower boundary have to be created with probabilities of default of PD_min/ PD_max.[330]

   Weighting factors $w_{i,j}$ do linearly decrease with increasing distance of corporate i's neigbors, which corresponds with a triangle core[331] estimator with core width 2*b-1:

---

[330] For reasons of plausibility, PD_min should not be bigger than the average default probability of the best rating class. In the simulation study, PD_min was set to 0%. The exact specification of PD_min is of no practical relevance for the model's results. Measurable impacts, however, are attributable to PD_max. Within the scope of the sensitivity analysis of the model, various values for PD_max were tried and their implications for the model variables under consideration (in particular their impact on the accuracy ratio) were analyzed. PD_max directly affects heterogeneity of individual default rates of the corporations within the worst rating class. For reasons of plausibility, it should not be smaller than the average default rate of the worst rating class.

[331] See SCOTT (1992, p. 133): "The quality of a density estimate is now widely recognized to be primarily determined by the choice of smoothing parameter, and only in a minor way by the choice of the kernel." Therefore, in the following only the influence of the smoothing parameter on the model's results is tested, but not the choice of the kernel function.
Although a good kernel function should be "smooth, clearly unimodal [, and] symmetric about the origin" (see ibid, p. 138) –recommendations that are fulfilled in case of triangle core estimators, even large devia-

F 211) $PD_{smoothed,i} = \sum_{j=i-b}^{i+b} w_{i,j} \cdot PD_j$ with

F 212) $w_{i,j} = \frac{1 - \frac{|i-j|}{b}}{\sum_{k=i-b}^{i+b}\left(1 - \frac{|i-k|}{b}\right)}$

with $w_{i,j}$ … weighting factor with $\sum_{j=i-b}^{i+b} w_{i,j} = 1$

Examples: if b=1 then $PD_{smoothed,i} = PD_i$ or if b=3 then

$$PD_{smoothed,i} = \frac{1}{12} \cdot PD_{i-2} + \frac{2}{12} \cdot PD_{i-1} + \frac{3}{12} \cdot PD_i + \frac{2}{12} \cdot PD_{i+1} + \frac{1}{12} \cdot PD_{i+2}$$

5. Subsequently, rating class specific calibration factors are applied in order to make sure, that average smoothed rating class PD are equal to average unsmoothed rating class specific PD (see point 2):

F 213) $PD_{smoothed\_calibrated,i} = PD_{smoothed,i} \cdot \frac{PD_{unsmoothed,\ rating\ class[i]}}{\text{mean value}\left(PD_{smoothed,rating\ class[i]}\right)}$

6. The (expected) accuracy ratio of the portfolio with corporations with smoothed and calibrated PD is determined and compared with the respective value that was obtained in point 3.

For the parameterizations that were performed in steps 1 and 2, historical default rates and current frequency distributions of corporations according to the 17ary rating scales of S&P (2004)/ MOODY'S (2004) were chosen alternatively. Thus, the model results are in the first instance directly applicable only to S&P and MOODY's ratings. With different parameterizations, however, the simulation model could cover any other rating systems as well, for instance a bank rating system with a 7ary rating scale and completely different rating class specific PD and frequency distributions.

Steps 4 to 6 are repeated for different specifications of the smoothing procedure. Thereby the influence of the parameters *PD_max* and *b*, which ultimately have to be chosen with some arbitrariness (although they can be *interpreted* intuitively) within certain plausible bandwidths, on the results finally obtained are quantified.

In Table N and Table O results for S&P's and MOODY'S data are given. The respective values show the ratios of the accuracy-ratio-values of the 17ary scale portfolio (see step 3) and the accuracy ratios that were obtained by portfolios with continuous probabilities of defaults (see step 6).

If b-values smaller than 100 are chosen (b=100 equals a smoothing core that for each corporate entails only +/- 2% of the neighboring corporations of the portfolio – whereby more distant corporations are receiving lower weighting factors than closer corporations), there is practically no smoothing of probabilities of default taking place, except at the boundaries of neighboring rating classes, because in most cases only corporations of the same rating class

---

tions of these demands cause only minor "efficiency losses". See SCOTT (1992, p. 140) for a survey of eight commonly used kernel functions and their respective efficiency measures. SCOTT (1992, p. 139) concludes „Therefore the kernel can be chosen for other reasons (ease of computation, differentiability, …) without undue concern for loss of efficiency.".

are included for determining weighted averages of PD.[332] Relatively large b-values (b>300, which equals a smoothing core bandwidth of +/- 6% of all corporations), on the other hand, result in implausible leaps in individual PD of corporations at the boundaries of neighboring rating classes. These leaps are caused by the calibration in step 5, which follows the smoothing. For $\lim b \to \infty$ the initial situation would reappear, i.e. there would be no PD-differentiation within rating classes at all, and PD of corporations of neighboring rating classes would increase by leaps and bounds according to the rating class specific default rates.

Owing to the subsequent calibrations, the influence of the parameter PD_max is relatively low for given values of b. In case of S&P data the relative precision of the 17ary scale portfolio in relation to the continuous scale portfolio only marginally deteriorates by 0.2 PP (percentage points) when PD_max dramatically changes from 30% to 100%.[333] Within more reasonable boundaries for PD_max, 40%-75% (as is indicated by the shaded area in Table N), its influence on the model results diminishes further to only 0.1 PP. In case of MOODY'S data the influence of PD_max is somewhat bigger with 0.6 PP (0.4 PP). The greater impact of PD_max in case of MOODY'S data is due to the fact, that MOODY'S "worst rating class" (Caa/C) entails nearly twice as many corporations than S&P's "worst rating class" (CCC/C) so that a greater heterogeneity of individual probabilities of default – which is affected by PD_max - within this class will have a greater impact in case of a portfolio with MOODY'S rating class frequency distribution.

S&P data		b (number of better and worse neighbors that have to be considered when smoothing probabilities of defaults (core width = 2b-1))					
		50	100	200	300	400	500
max. PD (uncalibr.)	30.0%	99.8%	99.7%	99.5%	99.4%	99.4%	99.4%
	40.0%	99.8%	99.6%	99.5%	99.4%	99.4%	99.4%
	50.0%	99.8%	99.6%	99.4%	99.4%	99.4%	99.4%
	75.0%	99.7%	99.5%	99.4%	99.3%	99.3%	99.3%
	100.0%	99.7%	99.5%	99.3%	99.3%	99.3%	**99.3%**

Table N: precision of the 17ary scale portfolio in relation to the continuous probability distribution portfolio for various specifications of the smoothing algorithm (maximum attainable PD and number of better and worse that have to be considered for smoothing probabilities of default), S&P (2004) data (historical default rates and current distribution of corporations among 17 rating scales), bold value: smallest value; shaded area: plausible area with plausible parameters

[332] In case of a 17ary scale, every rating class entails on average about 6% of all corporations.

[333] Depending on the chosen core width, essentially only corporations of the worst rating class (CCC/C or Caa/C) are affected by this parameter. Further limits to the influence of this variable are due to the calibration that follows in step 5.
If too low values for PD_max were chosen, with PD_max < $PD_{CCC/C}$, this would result in an implausible, U-shaped curve of probabilities of defaults within the group of the worst corporations.

MOODY'S data		b (number of better and worse neighbors that have to be considered when smoothing probabilities of defaults (core width = 2b-1))					
		50	100	200	300	400	500
max. PD (uncalibr.)	**30.0%**	99.8%	99.7%	99.5%	99.5%	99.5%	99.5%
	40.0%	99.8%	99.6%	99.4%	99.4%	99.4%	99.4%
	50.0%	99.7%	99.5%	99.3%	99.2%	99.3%	99.3%
	75.0%	99.6%	99.4%	99.1%	99.0%	99.1%	99.2%
	100.0%	99.5%	99.2%	98.9%	**98.9%**	99.0%	99.1%

Table O: precision of the 17ary scale portfolio in relation to the continuous probability distribution portfolio for various specifications of the smoothing algorithm (maximum attainable PD and number of better and worse that have to be considered for smoothing probabilities of default), MOODY'S (2004) data (historical default rates and current distribution of corporations among 17 rating scales), bold value: smallest value; shaded area: plausible area with plausible parameters

The following examinations are based on a "conservative" parameter combination, PD_max=75% and b=150, which rather overstates than understates the information losses that are attributable to discretizing continuous rating scales.

The resulting distributions of the smoothed and calibrated probabilities of default according to the S&P- and MOODY'S data are shown in Figure 37 and Figure 38.

The ensuing graphs show - in the right hand side parts of the respective graphs - the CAP-curves for both the 17ary scaled ratings and the continuous probability ratings. In the left hand side parts of the graphs is displayed, what share of the total information losses (of the 17ary scale compared to the continuous scale ratings) can be attributed to each single rating class. Deviating from the conventional proceedings, for obtaining these values, after step 5 all PD, except those of the corporations of the rating class under consideration, are reset to the rating class specific values (see step 2). Afterwards, the accuracy ratio of this portfolio is determined and compared both with the accuracy ratio of the 17ary scale and the continuous probability scale rating.

Annotation to Figure 39 and Figure 40: The marked differences in the share of information losses attributable to the worst rating class (CCC/C resp. Caa/C) between S&P and MOODY'S data can be largely referred to the heavily varying class sizes (S&P: ca. 3%, MOODY'S: ca. 6% of all corporations. For the respective class sizes see also Figure 37 and Figure 38). Obviously, the wider a rating class is, the bigger are the information losses that accrue, if all differences in individual probabilities of defaults are ignored by assuming rating class specific default rates.

Conclusion Procedure I

Using discrete 17ary rating scales instead of continuous rating scales causes information losses of about 0.3% - 0.7% (S&P) or 0.4% - 1.0% (MOODY'S) (see shaded values in Table N and Table O). – i.e. on average 0.6% [if a univalent result is desired].

The examinations show, that a meaningful reduction of the (already rather sparse) information losses can only be achieved by increasing the differentiation within the worst rating class (CCC/C, Caa/C). On the other hand, improving differentiation among investment grade corporations (AAA to BBB-, resp. Aaa to Baa3) practically has got no impact at all on the measured predictive quality.

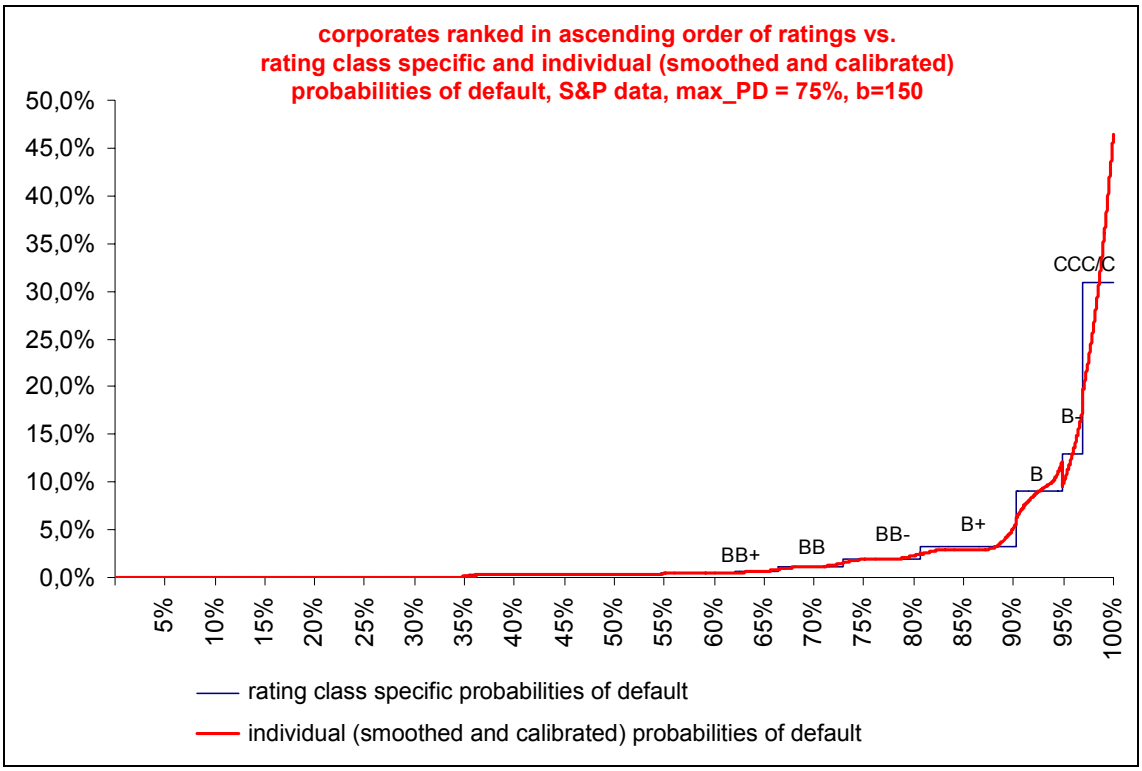

Figure 37: corporations ranked in ascending order of ratings vs. rating class specific and individual (smoothed and calibrated) probabilities of default, S&P data, max_PD = 75%, b=150

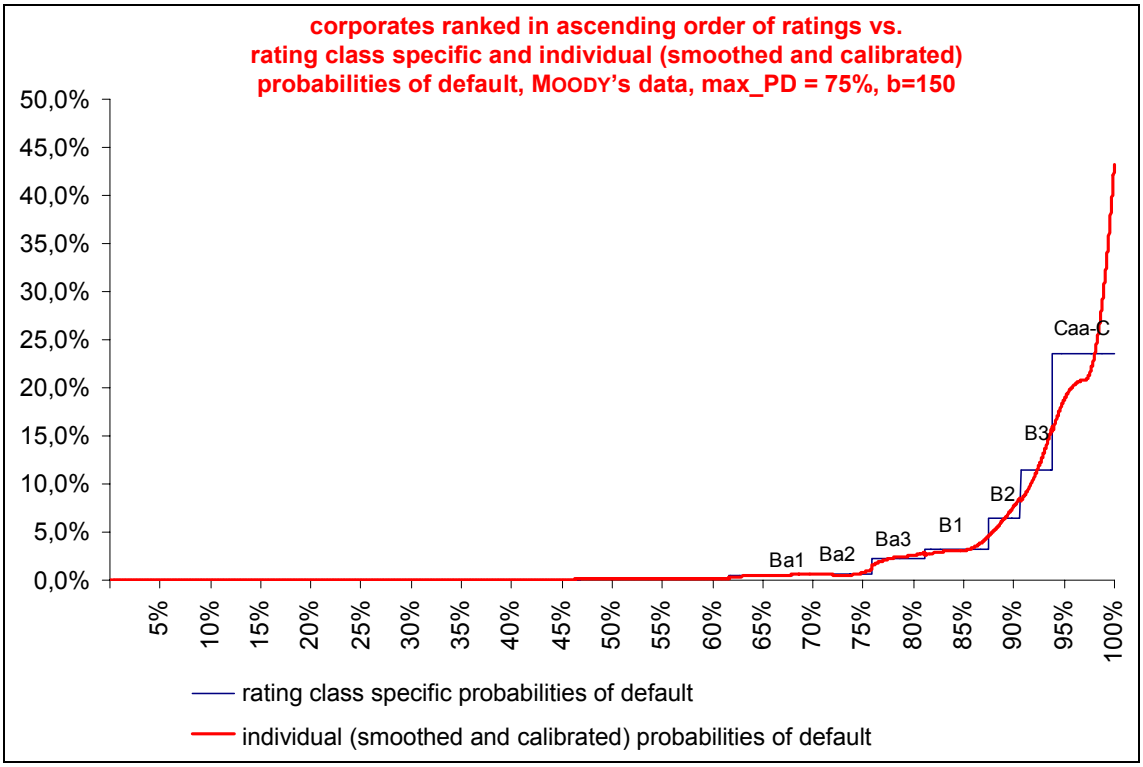

Figure 38: corporations ranked in ascending order of rating, smoothed and calibrated vs. rating class specific probabilities of default, MOODY'S data, max_PD = 75%, b=150

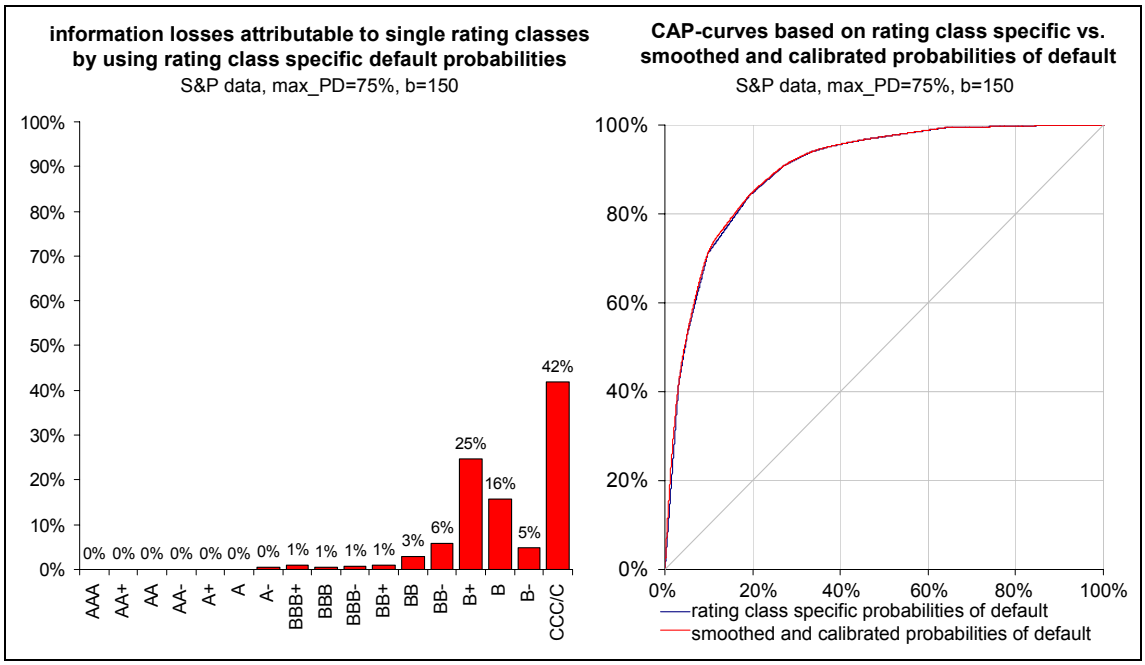

Figure 39: information losses attributable to single rating classes by using rating class specific (instead of individual) default probabilities (left hand side), CAP-curves based on rating class specific vs. smoothed and calibrated probabilities of default, S&P data, max_PD = 75%, b=150 (right hand side)

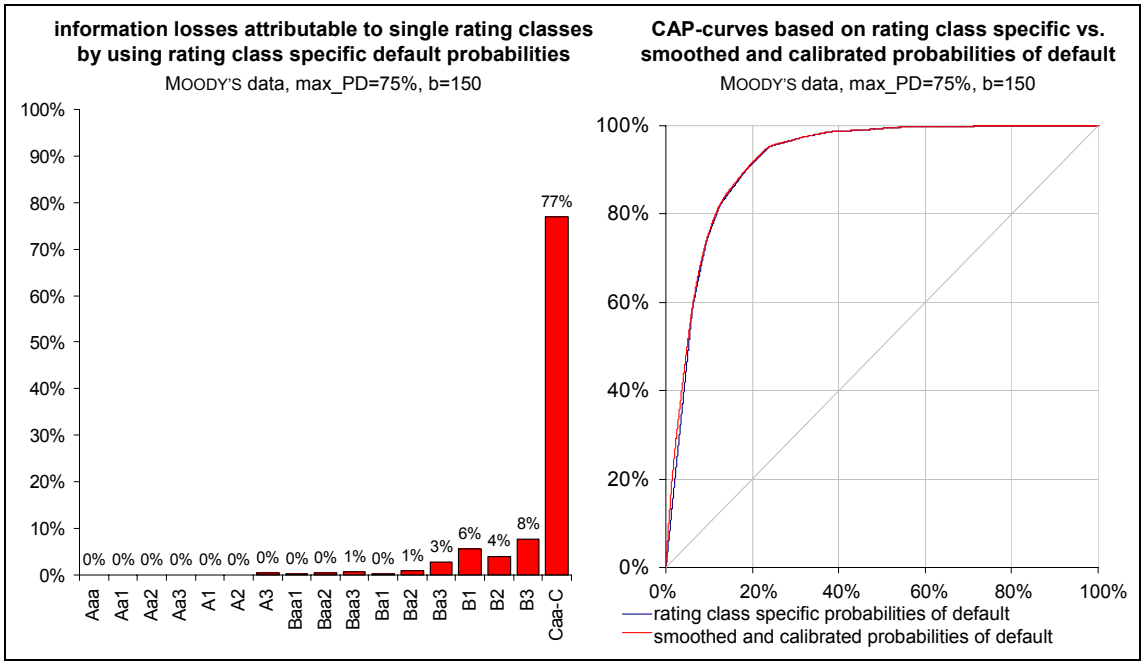

Figure 40: information losses attributable to single rating classes by using rating class specific (instead of individual) default probabilities (left hand side), CAP-curves based on rating class specific vs. smoothed and calibrated probabilities of default, MOODY'S data, max_PD = 75%, b=150 (right hand side)

Procedure II: Linear ROC-curve interpolation

The second procedure for estimating information losses that are attributable to discretizing rating scales does not (explicitly) model individual probabilities of default for subsequently deriving CAP- or ROC-curves and obtaining measures related with them. Instead, it models ROC-curves *directly*, by assuming certain formally manageable ROC-curve functions. Subsequently, the second procedure does determine how many percent of the "true" ("continuous") ROC-curve (AUC_{ROC}) – resp. how many percent of the true accuracy ratio – can be obtained with a rating system with only g rating classes. All parameter values for g ranging from 2 to 30 are examined numerically.

For the formal representation of ROC-curves two *parametrical* functions ("families of functions") are used, ROC_α and ROC_β, with parameters α and β that enable the specification of a multitude of various ROC-curves.

F 214) $ROC_\alpha(x) = x^\alpha$ with $ROC_\alpha(x)$=1-error of type I and x=error of type II

By definition $ROC_\beta(x)$-curves results from reflecting ROC_α–curves at the secondary diagonal. Thus, if in formula F 214 ROC(x) is replaced with 1-x and x with 1-ROC(x), and α with β, it follows:

F 215) $1 - x = (1 - ROC_\beta(x))^\beta$

F 216) $ROC_\beta(x) = 1 - (1 - x)^{1/\beta}$

Empirical ROC-curves can be described well by ROC_α- or ROC_β-curves or by mixtures of both types (see Appendix I).

For α=1, resp. β=1, the resulting ROC-curves follow the (primary) diagonal line in the false alert rate-hit rate-diagram (ROC-diagram), which correspond to the ROC-curves that "naïve rating models" would create. For α=0 and β=0, respective ROC-curves proceed along the exterior sides of the ROC-diagram ("perfect rating model"). For values for α resp. β between 0 and 1 curve progressions result that are similar to real ROC-curves (see Appendix I).

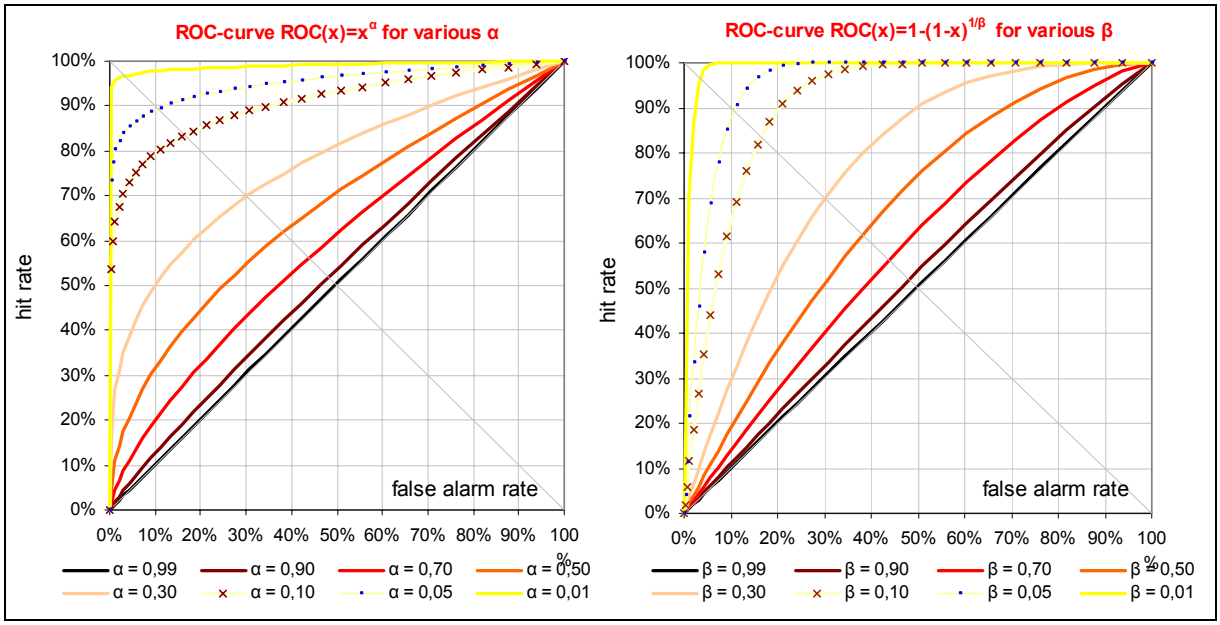

Figure 41: ROC-curves with $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta(x)=1-(1-x)^{1/\beta}$ for various parameter values for α and β

Given an error-I-II-combination, parameters α and β can be calculated as follows (cf. formulas F 214 and F 216)

F 217) $\alpha = \frac{\log(1-F_1)}{\log F_2}$ with F_1 … error (failure) type I and F_2 … error (failure) type II

F 218) $\beta = \frac{\log(1-x)}{\log(1-ROC_\beta(x))}$

F 219) $\beta = \frac{\log(1-F_2)}{\log F_1}$

The area under the ROC_α-curve, $AUC_{ROC,\alpha}$, and the associated accuracy ratio are given by:

F 220) $AUC_{ROC\alpha} = \int_0^1 ROC_\alpha(x)dx = \int_0^1 x^\alpha dx$

F 221) $AUC_{ROC\alpha} = \left|\frac{1}{\alpha+1}\cdot x^{\alpha+1}\right|_{x=0}^{x=1}$

F 222) $AUC_{ROC_\alpha} = \frac{1}{\alpha+1}$

F 223) $AR_\alpha = \frac{AUC_{ROC_\alpha} - \frac{1}{2}}{\frac{1}{2}} = \frac{\frac{1}{\alpha+1} - \frac{1}{2}}{\frac{1}{2}}$

F 224) $AR_\alpha = \frac{1-\alpha}{1+\alpha}$ and in combination with formula F 217:

F 225) $AR_\alpha = \frac{1 - \frac{\log(1-F_1)}{\log F_2}}{1 + \frac{\log(1-F_1)}{\log F_2}}$

F 226) $AR_\alpha = \frac{\log F_2 - \log(1-F_1)}{\log F_2 + \log(1-F_1)}$

By substituting ROC-coordinates F_1 and F_2 by CAP-coordinates X_0 and Y_0 according to formulas F 64 and F 66 follows:

F 227) $AR_\alpha = \frac{\log\left(\frac{X_0 - Y_0\cdot PD}{1-PD}\right) - \log Y_0}{\log\left(\frac{X_0 - Y_0\cdot PD}{1-PD}\right) + \log Y_0}$

F 228) $AR_\alpha = \frac{\log(X_0 - Y_0\cdot PD) - \log(1-PD) - \log Y_0}{\log(X_0 - Y_0\cdot PD) - \log(1-PD) + \log Y_0}$

The area under the ROC_β-curve, $AUC_{ROC,\beta}$, and the associated accuracy ratio are given by:

F 229) $AUC_{ROC\beta} = \int_0^1 ROC_\beta(x)dx = \int_0^1 1-(1-x)^{1/\beta}dx$

F 230) $$AUC_{ROC\beta} = \left| x + \frac{\beta}{1+\beta} \cdot (1-x)^{(1+\beta)/\beta} \right|_0^1$$

F 231) $$AUC_{ROC\beta} = 1 + \frac{\beta}{1+\beta} \cdot (1-1)^{(1+\beta)/\beta} - \left(0 + \frac{\beta}{1+\beta} \cdot (1-0)^{(1+\beta)/\beta} \right)$$

F 232) $$AUC_{ROC\beta} = 1 - \frac{\beta}{1+\beta}$$

F 233) $$AUC_{ROC\beta} = \frac{1}{1+\beta}$$ (cf. formula F 222)

F 234) $$AR_\beta = \frac{\frac{1}{1+\beta} - \frac{1}{2}}{2}$$

F 235) $$AR_\beta = \frac{1-\beta}{1+\beta}$$ (cf. formula F 224), in combination with formula F 219 follows:

F 236) $$AR_\beta = \frac{1 - \frac{\log(1-F_2)}{\log F_1}}{1 + \frac{\log(1-F_2)}{\log F_1}}$$

F 237) $$AR_\beta = \frac{\frac{\log F_1 - \log(1-F_2)}{\log F_1}}{\frac{\log F_1 + \log(1-F_2)}{\log F_1}}$$

F 238) $$AR_\beta = \frac{\log F_1 - \log(1-F_2)}{\log F_1 + \log(1-F_2)}$$ (cf. formula F 226)

By substituting ROC-coordinates F_1 and F_2 by CAP-coordinates X_0 and Y_0 according to formulas F 64 and F 66 follows:

F 239) $$AR_\beta = \frac{\log(1-Y_0) - \log\left(1 - \frac{X_0 - Y_0 \cdot PD}{1-PD}\right)}{\log(1-Y_0) + \log\left(1 - \frac{X_0 - Y_0 \cdot PD}{1-PD}\right)}$$

F 240) $$AR_\beta = \frac{\log(1-Y_0) - \log(1-PD-X_0+Y_0 \cdot PD) + \log(1-PD)}{\log(1-Y_0) + \log(1-PD-X_0+Y_0 \cdot PD) - \log(1-PD)}$$

Parameter values for α and β, that give rise to realistic accuracy-ratio-values, are for highly discriminative ratings in the range of about α = β = 0.1 (AR=82%), for rather below average ratings about α = β = 0.3 (AR=54%) and for very inefficient ratings about α = β = 0.5 (AR=33%) (see also chapter 3.5).

With the formulas above, exact AUC_{ROC} and accuracy ratio measures for continuous ratings can be calculated – and can subsequently be compared with ratings that principally base on the same error-I-II-trade-off, but which are based on a discrete number of rating classes (see formula F 22). Mathematically speaking, the attainable relative accuracy of a rating model with g discrete rating classes depends on how well a continuous ROC-curve can be approximated by a linear interpolation, i.e. by a "curve" that consists of g connected, linear sections

whose g-1 inner supporting points are positioned on the continuous score's ROC-curve (see Figure 42) while the two outer end points are positioned at (0%; 0%) and (100%; 100%).[334]

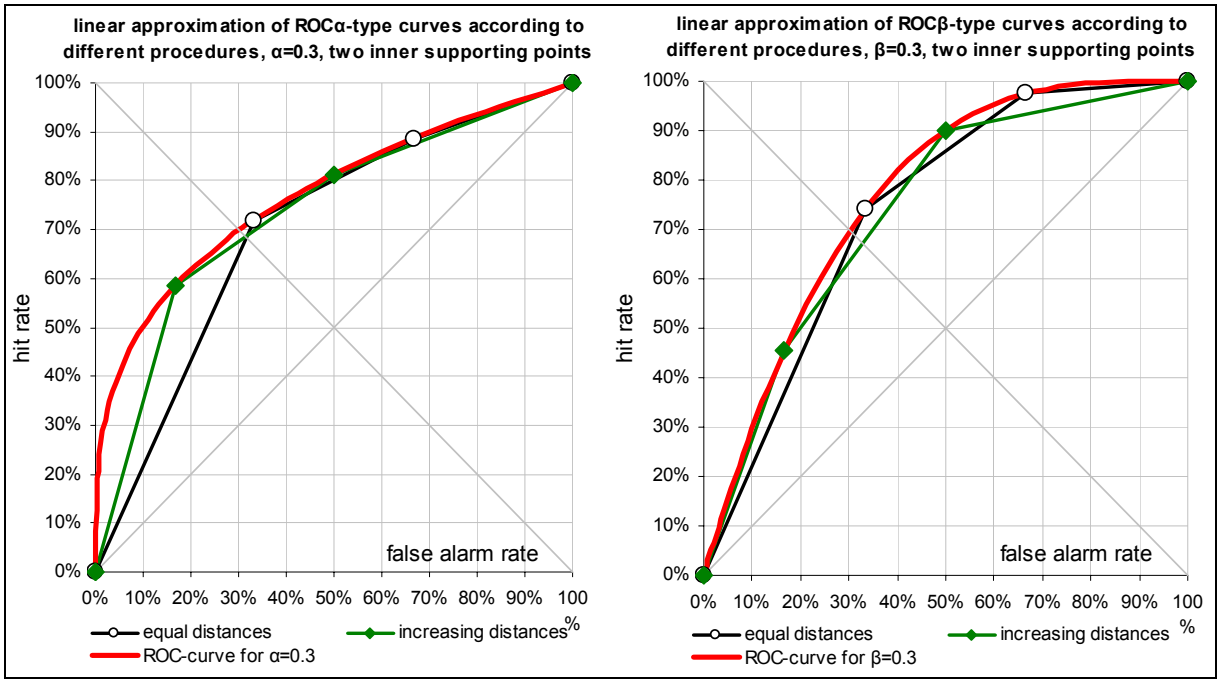

Figure 42: linear approximations of ROC_α- and ROC_β-type curves according to different procedures, $\alpha = \beta = 0.3$, two inner supporting points

However, one further input for the linear interpolation is needed – the localization of the abscissa values [false alert rates] that are used for determining the inner supporting points. In Figure 42 two examples of procedures for determining abscissa values (here for a rating model with only three different classes and therefore two inner supporting points) are presented – that are applied both for ROC_α- and for ROC_β-curves.

The first procedure creates three rating classes, whereby each rating class is assigned the same share in all false-alerts (here: 33.3% per rating class). The second procedure subdivides rating classes such, that the share in all false-alert-rates each rating class comprises rises linearly, i.e. the second class comprises twice as many false alert rates as the first and the third three times as many as the first (in general: i-th class comprise i times as many false alerts as the first class), so that shares in false alerts of 16.6%, 33.3% and 50% result.

In Figure 43 relative accuracy ratio values are shown for various α- and β-type curves and for 0 to 30 inner supporting points (1..31 discrete rating classes). Highlighted are the respective values for 2, 7 and 17 rating classes (which correspond to 1, 6 and 16 inner supporting points). All rating classes comprise equal shares in total errors of type II.

[334] Each such section represents one discrete rating class. Linearity of these sections implies, that within the respective sections (rating classes) an identical error-I-II-trade-off has to be made, which means that corporations *within* rating classes are homogenous with respect to default probabilities.

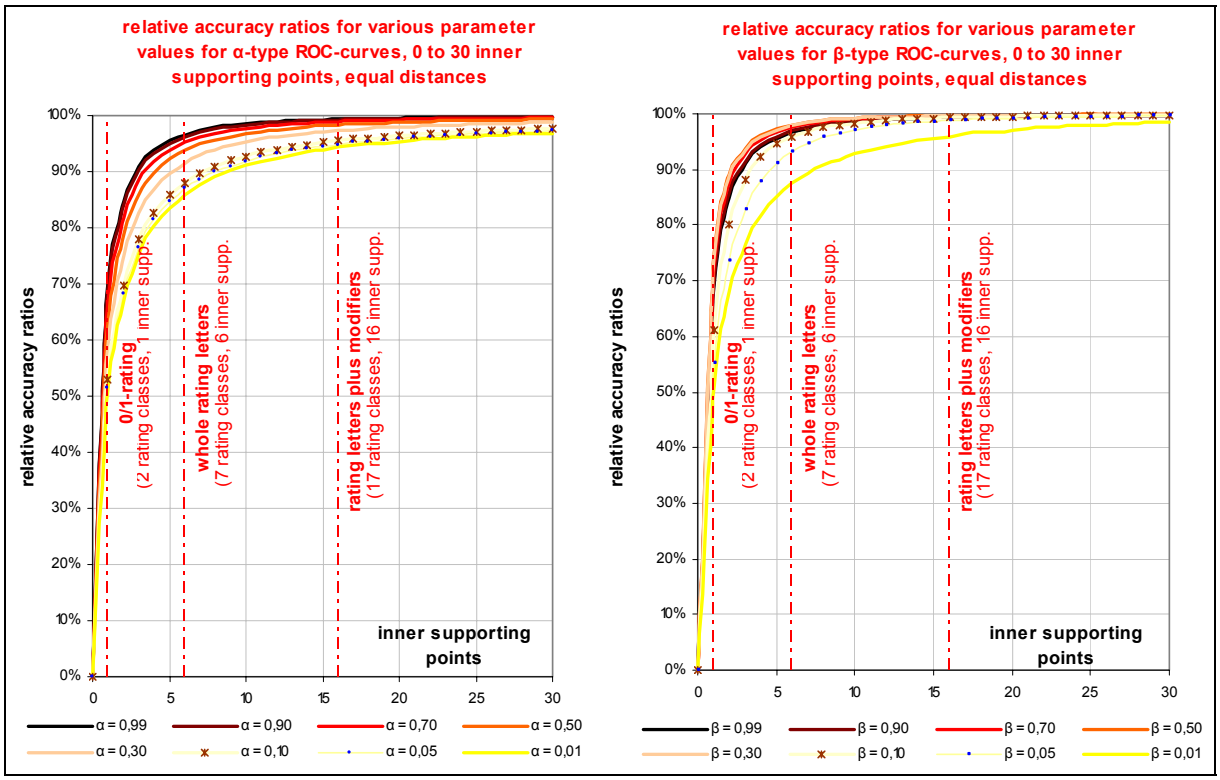

Figure 43: relative accuracy ratios for various parameter values for α- (left hand side) and β-type ROC-curves (right hand side) and for 0 to 30 inner supporting points (1..31 discrete rating classes), equal shares in total errors of type II per rating class

For 17 rating classes, for instance, following relative accuracy ratio values originate: 95.5% (α=0.1) vs. 99.3% (β=0.1) or 98.3% (α=0.5) / 99.7% (β=0.5). Therefore, based on a 17ary rating scale, the accuracy ratio of a very bad rating model (α=β=0.5, AR_α=AR_β=33%) would produce accuracy ratio values that are about 0.3% - 1.7% (0.1 – 0.6 percentage points) below the "true accuracy" ratio of the continuous score model, while the respective values for a good rating model (α=0.1, β=0.1, AR_α=AR_β=82%) would be up to 0.7%-4.5% (0.5 – 3.5 percentage points) too low.

These results, however, are rather too pessimistic estimations for the information losses that can be attributed to discretizing rating scales – in particular for α-type curves.

Additional examinations revealed, that information losses in case of equally wide rating classes (in respect with shares in errors of type II) for α-type curves are nearly exclusively ascribable to information losses that accrue in the first rating class (see Figure 44 left hand side, for an illustration see also Figure 42 left hand side).

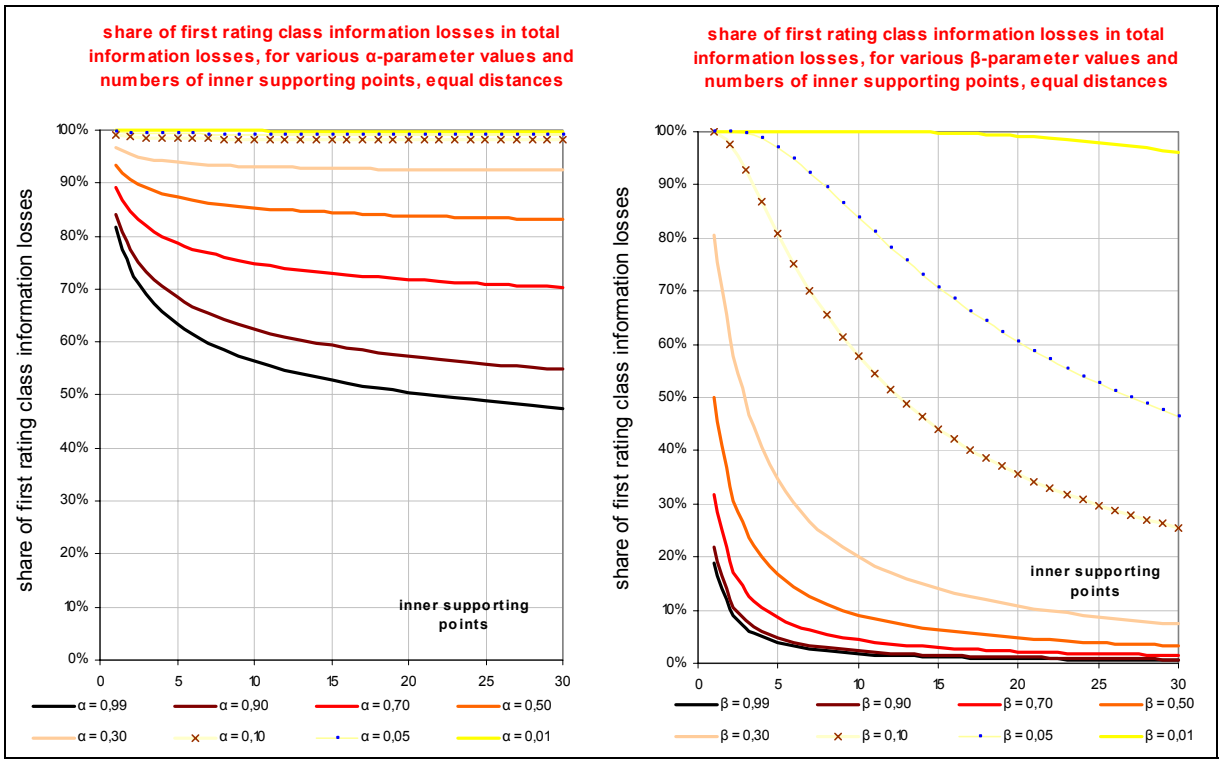

Figure 44: share of information losses in total information losses that accrue in the first rating class, for α- (left hand side) and β-type curves (right hand side) and 0..30 inner supporting points (1..31 rating classes), equal shares in total errors of type II per rating class

For α=0.1 and 30 rating classes, for instance 98.1% (!) of all information losses that totally result by the linearly approximation of the ROC-curve accrue in the first rating class (for β=0.1 and 30 rating classes it still are over 25%) – while the residual 1.9% of information losses are shared among the 29 remaining rating classes.

Obviously, α-type ROC-curves are highly non-linear in the range of the first rating class (see the pronounced concave curve progression in Figure 42, left hand side). Therefore it was to be expected, that not only the share of information losses of the first rating class, but also the total amount of information losses could be reduced, if the class size of the first rating class was scaled down in order to improve the quality of the linear interpolation. One way of achieving this, is to apply the above mentioned procedure of linearly increasing class widths (see Figure 45 for the resulting relative accuracy ratios depending on the quality of the rating (α- and β parameter values and Figure 46 for the share of the first rating class in all information losses).[335,336]

[335] Whether this procedure does *minimize* information losses given a ROC-curve of type α (or β or both) is unknown and irrelevant. It is irrelevant, because a rater cannot perfectly control this feature of his rating systems anyway - and he might pursue other targets than minimizing information losses, such as making sure sufficient class width for enabling meaningful statistical tests. Further, it is not likely that a rating class distribution that is *optimal* (in what ever sense) for a particular ROC-curve function, is also optimal for any other ROC-curve function. The above mentioned procedure was chosen, because it enables a better differentiation in the relevant fore part of the ROC-curve and because it is easy to implement and to communicate.

[336] For further approaches for formally and numerically modeling rating class widths ("plausible but suboptimal boundaries", p. 17) see JANKOWITSCH, PICHLER, SCHWAIGER (2003, p. 15f. and 21f.).

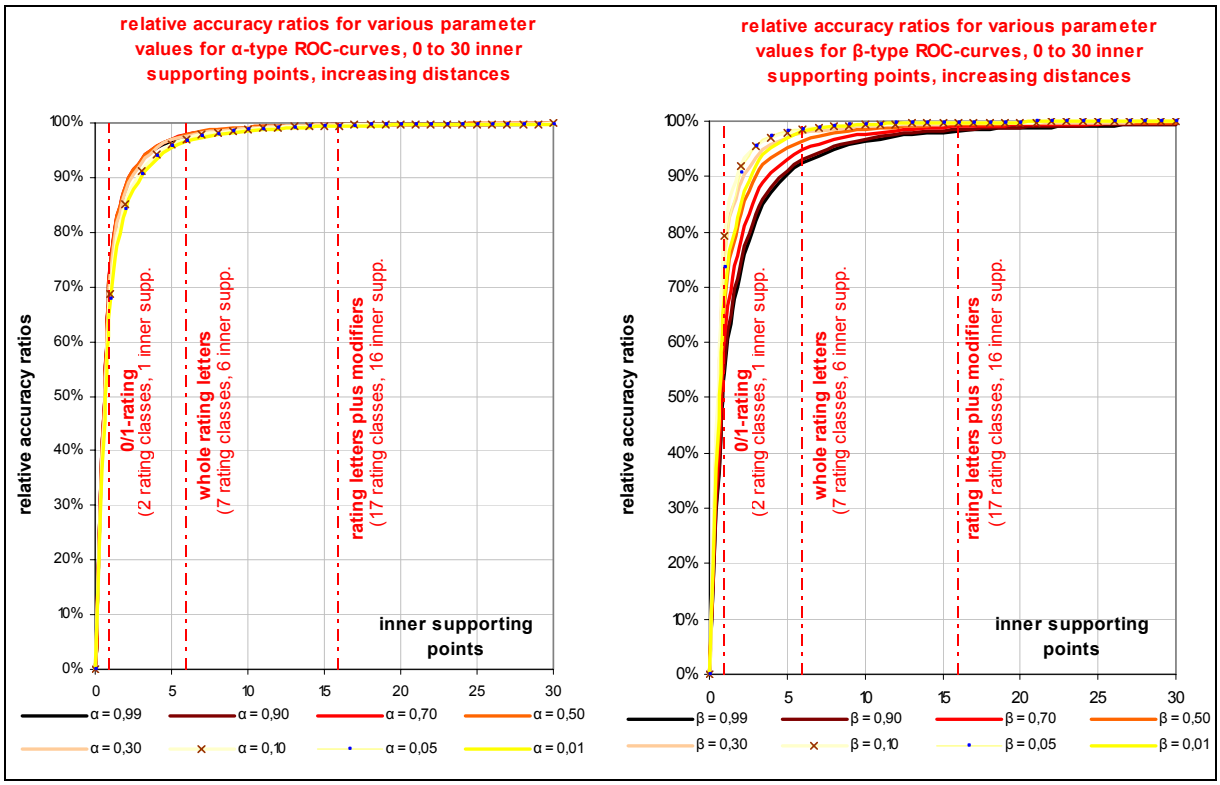

Figure 45: relative accuracy ratios for various parameter values for α- (left hand side) and β-type ROC-curves (right hand side) and for 0 to 30 inner supporting points (1..31 discrete rating classes), linearly increasing rating class widths

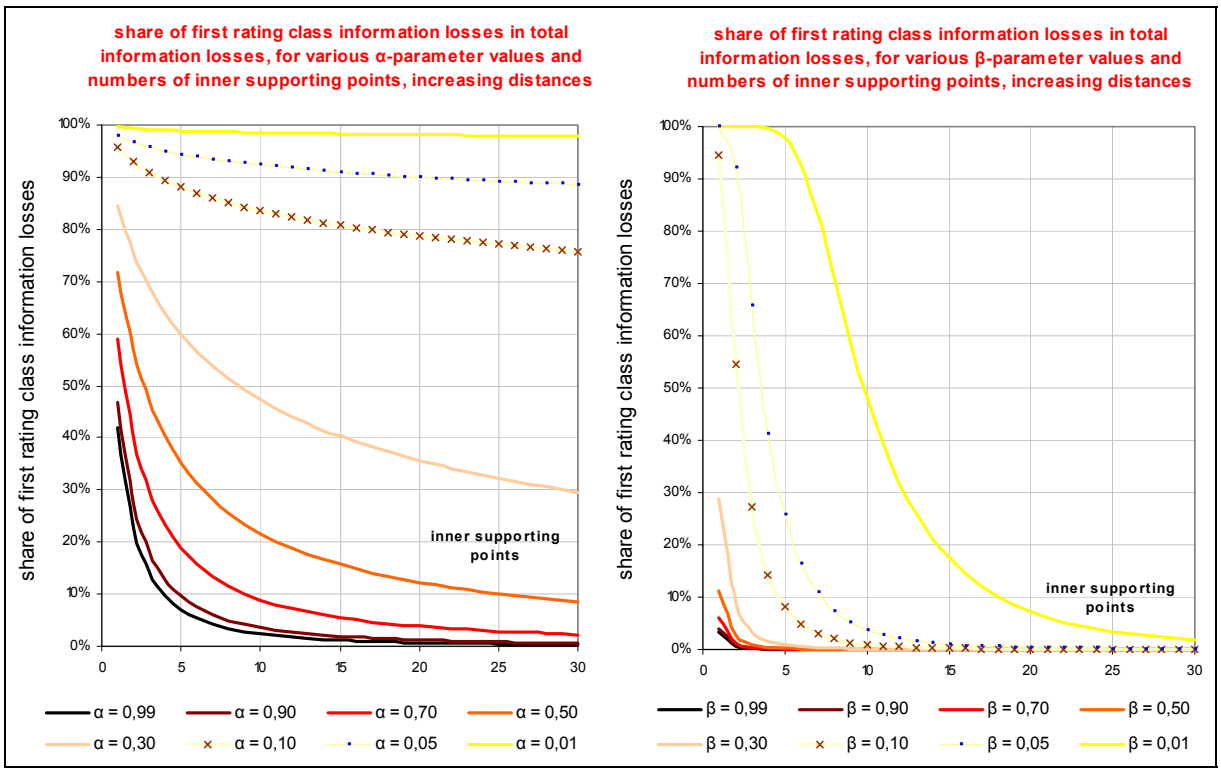

Figure 46: share of information losses that accrue in the first rating class in total information losses, for α- (left hand side) and β-type curves (right hand side) and 0..30 inner supporting points (1..31 rating classes), linearly increasing rating class widths

For the procedure with linearly increasing rating class widths and 17 rating classes, for instance relative accuracy ratios of 99.5% (α=0.1) vs. 99.7% (β=0.1) or 99.6% (α=0.5) vs. 99.3% (β=0.1) result. Therefore, based on a 17ary rating scale, the accuracy ratio of a very bad rating model (α=β=0.5, AR_α=AR_β=33%) would produce accuracy ratio values that are about 0.4% - 0.7% (ca. 0.2 percentage points) below the "true accuracy" ratio of the continuous score model, while the respective values for a good rating model (α=0.1, β=0.1, AR_α=AR_β=82%) would be 0.3%-0.5% (ca. 0.3 percentage points) too low.

Conclusion Procedure II:

In Table P numerical results for the relative accuracy ratios of discrete scale rating models in relation to continuous scale rating models are given for four different numbers of rating classes, two different ROC-curve functions, three different rating quality levels and two different methods for determining rating class distributions.

Relative accuracy ratio values that are considered especially relevant for the respective numbers of rating classes have been shaded (for further explanations see below).

number of rating classes	quality of the rating model	equal class widths		linearly increasing class widths		mean value of α-β-functions	
		α-function	β-function	α-function	β-function	equal class widths	linearly increasing class widths
2	good (AR=82%)	0.529	0.610	0.688	0.794	0.570	**0.741**
	mediocre (AR=54%)	0.580	0.744	0.717	0.757	0.662	**0.737**
	bad (AR=33%)	0.621	0.750	0.732	0.667	0.686	**0.699**
7	good (AR=82%)	0.880	0.959	0.971	0.985	0.920	**0.978**
	mediocre (AR=54%)	0.915	0.979	0.977	0.978	0.947	**0.977**
	bad (AR=33%)	0.938	0.980	0.978	0.964	0.959	**0.971**
10	good (AR=82%)	0.919	0.980	0.986	0.992	**0.950**	0.989
	mediocre (AR=54%)	0.946	0.990	0.989	0.989	**0.968**	0.989
	bad (AR=33%)	0.963	0.990	0.990	0.982	**0.977**	0.986
17	good (AR=82%)	0.955	0.993	0.995	0.997	0.974	**0.996**
	mediocre (AR=54%)	0.973	0.996	0.996	0.996	0.985	**0.996**
	bad (AR=33%)	0.983	0.997	0.996	0.993	0.990	**0.995**
30	good (AR=82%)	0.976	0.998	0.998	0.999	0.987	**0.999**
	mediocre (AR=54%)	0.987	0.999	0.999	0.999	0.993	**0.999**
	bad (AR=33%)	0.993	0.999	0.999	0.998	0.996	**0.998**

Table P: relative accuracy ratios for four different numbers of rating classes, two different ROC-curve functions, three different rating quality levels and two different methods for determining rating class distributions

The examinations show, that there is no unequivocal interrelationship between the number of rating classes and the information losses compared with a continuous scale rating model, but that additional influencing variables have to be considered.

Besides the functional forms of the continuous ROC-curves, which empirically may well be described by a mixture of ROC_α- and ROC_β-curves[337], see the examinations in Appendix I, results are strongly influenced by the assumed distributions of corporations among the different rating classes – for which, however, practically no empirical information is available, except for rating systems with either 7 or 17 classes.

In the following it is assumed, that the "worst rating classes" are notedly below average sized,[338] so that the results of the method with increasing class sizes should reveal more realistic results than the method with equally sized rating classes (see the respective shadings in Table P). The results of the method with equal class sizes, though, are relevant in all those cases, where the quality of rating models has to be determined based on groups that are equally sized by *definition.* Often each such group embraces exactly 10% of all corporations ("decile"), see the respective shadings in Table P.[339]

As a matter of principle, if information losses have to be estimated for a *concrete* rating model, it is advisable to implement the distribution (of corporations among the various rating classes) actually observed for this rating model instead of above mentioned methods for partitioning corporations among the various classes.

As a rule of thumb, it may be stated that by using a 7ary scale information losses of about 3% have to be expected in comparison with a continuous rating scale and by using a 17ary scale rating losses of about only 0.5%.

As was already found by procedure I (see beginning of Appendix IV), disproportionate information losses occur in particular in the worst rating class.

---

[337] For that reason, in Table P mean values that were determined according to the α- and β-methods are especially shaded and highlighted.

[338] With shares of 3.1% resp. 6.2% in all corporations, rating classes CCC/C (S&P) resp. Caa/C (MOODY'S) are clearly underrepresented based on 7ary scales (average class coverage = 1/7 = 14%). On 17ary scales, however, their shares are roughly equivalent to the average class coverage (1/17 = 5.9%). But if it is accounted for, that classes CCC/C resp. Caa/C are actually aggregates of five modified-rating sub classes (CCC+, CCC, CCC-, CC, C resp. Caa1, Caa2, Caa3, Ca, C), see S&P (2003b, p.8) and MOODY'S (2004b, p.6), the average share of these groups is only 0,6% (S&P) resp. 1,2% (MOODY'S) – which is clearly below average size, too.
In the major agencies' default studies by 2004 CCC/C resp. Caa/C classes were disclosed only in aggregated form (see i.e. S&P (2004, p.13f), S&P (2005, p. 18, 31) or MOODY'S (2004, p. 22, 26)).
It may be speculated, whether an aggregated disclosure was chosen in order to conceal a rather low discriminative power of the agencies' ratings *within* the CCC/C resp. Caa/C group. See on this for instance HAMILTON's (2004, p. 18) data: the average 3-years-default rates for *1996*-2003 [according to ibid, p. 4, alphanumeric modifiers for the Caa rating category were introduced in June *1997*] for Caa1 (33.4%) is bigger (!) than for Caa2 (31.3%), which in turn is bigger (!) than the default rate for Caa3 (24.4%).
Only in its most recent default study, see MOODY'S (2005, p. 18), MOODY'S also revealed historical default rates for the Caa-subgroups (Caa1, Caa2, and Caa3) and for the aggregated group Ca/C for the 1998 (!)-2004 period. According to the data given there, the realized default rates behaved very plausibly: $PD_{Caa1} < PD_{Caa2} < PD_{Caa3} < PD_{Ca/C}$ both for 1- and 3-years default rates. [But still questions remain why Ca and C default performance were not revealed separately.]

[339] See on this e.g. MCQUOWN (1993, p. 17), KEALHOFER (2003, p. 35), SHUMWAY (2001, p. 118ff.)

Appendix V: The effect of preselection of portfolios on the accuracy of insolvency predictions

Intention: Impacts of "positive preselections" on the accuracy of rating models shall be quantified, which occur when banks reject a certain share of potential customers due to their bad solvencies, or if they transfer low solvency (but non-defaulted) existing customers to "special asset groups" - but measure the performance of their rating models based on the default behavior of the remaining pool of accepted and non-transferred costumers.
Respective impacts of "negative preselections" shall be determined as well.[340]

Further proceedings: Based on the formally analyzable, parametrical ROC-curve functions that were developed and empirically tested in Appendices I and IV, it is both formally and numerically tested, how *preselection* (as defined above) affects accuracy measures of rating models. Empirical crosschecks with data of real rating models are carried out, too.

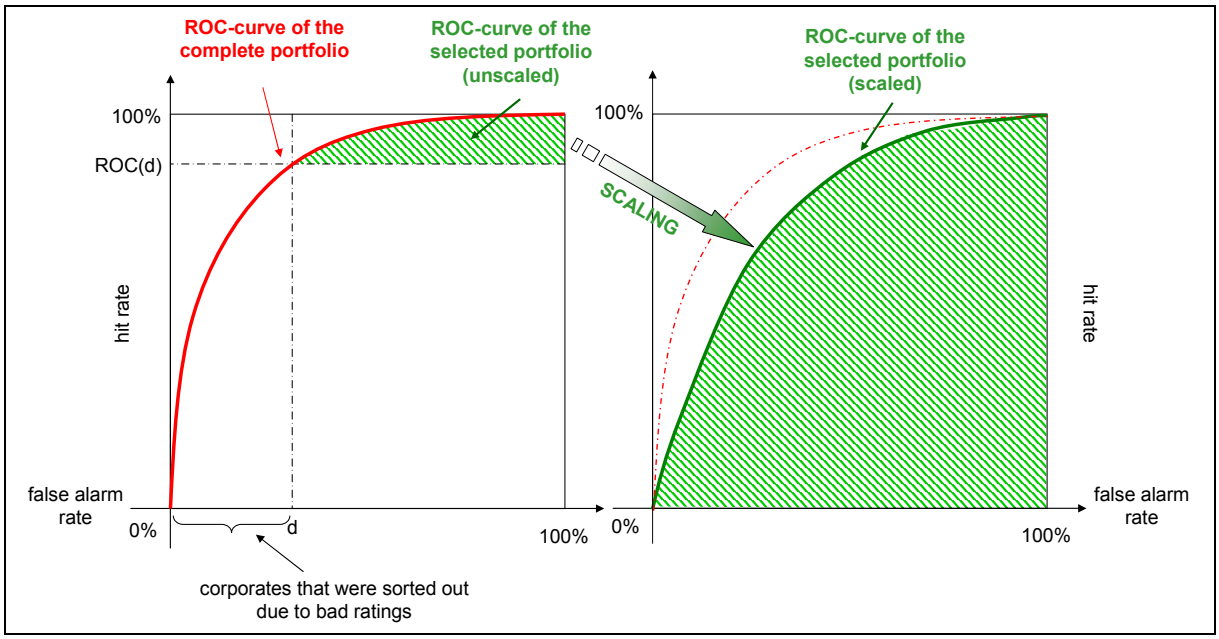

Figure 47: comparison of a total sample's and a preselected sample's ROC-curve, conceptual sketch

The basic concept of the following analyses is outlined in Figure 47. The ROC-curve of a portfolio from which a share *d* of the corporations with the worst ratings was removed[341] results from scaling the respective ROC-curve-section (see the hatched area in the left hand illustration) with a vertical scale factor of (1-d)-1 and a horizontal scale factor of (1-ROC(d))-1. See also Figure 48 for some numerically obtained examples of ROC-curves for selected portfolios, if the complete portfolio's ROC-curve is modeled by a ROC_α-curve with α=0.3.

[340] See the remarks on positive and negative preselections of bank portfolios in chapter 3.1.

[341] Strictly speaking, d (see the figure above) states the share of excluded *non-defaulters* (in all non-defaulters) and not the share of excluded *corporations* (in all corporations). For small average default rates PD, which are characteristic for insolvency prediction studies - at least for one-year prediction horizons, not much precision is lost thereby. The exact share of excluded corporations is given by $d \cdot (1-PD) + ROC(d) \cdot PD$.

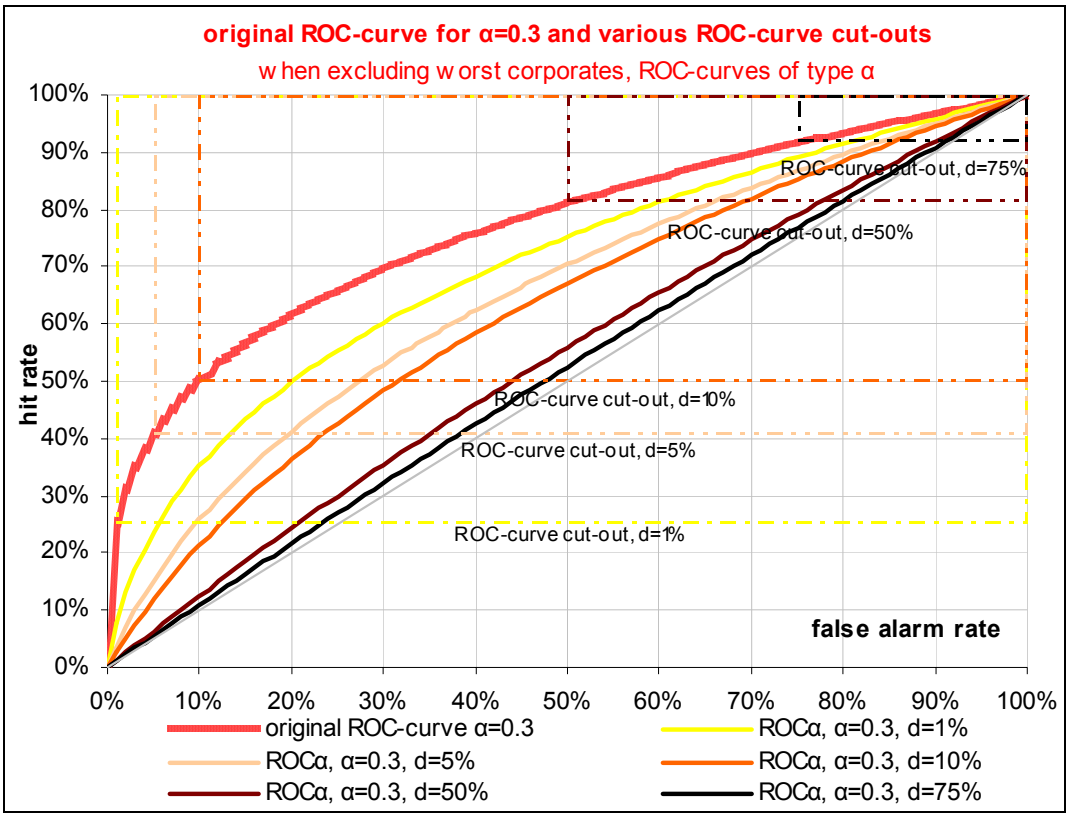

Figure 48: comparison of a total sample's ROC_α-curve (with $\alpha=0.3$) and preselected samples' ROC-curves for various d (percentages of excluded worst corporations)

In general, the area under the scaled ROC-curve of a selected portfolio $AUC_{ROC,d}$, is given in dependence of the ROC-curve function of the complete portfolio, ROC(x), by:

F 241) $$AUC_{ROC,d} = \frac{\int_d^1 ROC(x)dx - (1-d)\cdot ROC(d)}{(1-d)\cdot(1-ROC(d))}$$

In the following, accuracy ratios are calculated based on $AUC_{ROC,d}$ and are accounted for in relation to the accuracy ratio of the complete portfolios. ROC-curves are modeled by the formally analyzable ROC-curve-functions $ROC_\alpha(x)=x^\alpha$ and $ROC_\beta(x)=1-(1-x)^{1/\beta}$ that were introduced in Appendices I and IV. As was shown in Appendix I, both functions (or better: function *families*) represent rather extreme examples of possible ROC-curves. Empirical ROC-curves can be well described as mixtures of ROC_α- and ROC_β-curves, whereby some empirical ROC-curves are rather complying with α-type and other with β-type ROC-curves.

While the accuracy of ROC-curves is determined by the parameter α or β (see formulas F 224 and F 235), α-type ROC-curves are comparatively discriminative in the range of low solvency corporations and indiscriminative in the range of good solvency corporations (see Appendix I), it is just the opposite way around in case of β-type ROC-curves.

Therefore, it may be reckoned, that the relative performance of both ROC-curve functions are struck unequally when *low solvency* corporations are sorted out. A more adverse performance may be expected for ROC_α-curves, because just those corporations are removed from the sample, where the respective rating models have "comparative advantages" in comparison with rating models, whose ROC-curves are better described by β-type ROC-curves. In case of sorting out *good solvency* corporations, ROC_α-curves are expected to be less negatively affected than ROC_β-curves.

For α-type ROC-curves following equation holds:

F 242) $$\int_d^1 ROC_\alpha(x)dx = \int_d^1 x^\alpha dx = \left[\frac{x^{\alpha+1}}{\alpha+1}\right]_d^1$$

F 243) $$\int_d^1 x^\alpha dx = \frac{1-d^{\alpha+1}}{\alpha+1}$$ and an insertion to formula F 241 yields:

F 244) $$AUC_{ROC,d,\alpha} = \frac{\frac{1-d^{\alpha+1}}{\alpha+1} - (1-d)\cdot d^\alpha}{(1-d)\cdot(1-d^\alpha)}$$

F 245) $$AUC_{ROC,d,\alpha} = \frac{1-d^{\alpha+1} - (d^\alpha - d^{\alpha+1})\cdot(\alpha+1)}{(\alpha+1)\cdot(1-d)\cdot(1-d^\alpha)}$$

F 246) $$AUC_{ROC,d,\alpha} = \frac{1-d^\alpha + \alpha d^\alpha \cdot (d-1)}{(\alpha+1)\cdot(1-d)\cdot(1-d^\alpha)}$$

F 247) $$AUC_{ROC,d,\alpha} = \frac{1-d^\alpha + \alpha d^{\alpha+1} - \alpha d^\alpha}{\alpha - \alpha d - \alpha d^\alpha + \alpha d^{\alpha+1} + 1 - d - d^\alpha + d^{\alpha+1}}$$

According to formula F 8, the accuracy ratio can be derived from AUC as follows:

F 248) $$AR_{d,\alpha} = 2\cdot AUC_{ROC,d,\alpha} - 1$$ and thus

F 249) $$AR_{d,\alpha} = \frac{2 - 2d^\alpha + 2\alpha d^{\alpha+1} - 2\alpha d^\alpha - \alpha + \alpha d + \alpha d^\alpha - \alpha d^{\alpha+1} - 1 + d + d^\alpha - d^{\alpha+1}}{\alpha - \alpha d - \alpha d^\alpha + \alpha d^{\alpha+1} + 1 - d - d^\alpha + d^{\alpha+1}}$$

F 250) $$AR_{d,\alpha} = \frac{1 - \alpha - d^\alpha + \alpha d^{\alpha+1} - \alpha d^\alpha + \alpha d + d - d^{\alpha+1}}{(\alpha+1)\cdot(1-d)\cdot(1-d^\alpha)}$$

The *relative* accuracy ratio, $AR_{rel,\alpha,d}$, which states the accuracy ratio of the preselected portfolio in relation to the accuracy ratio of the complete portfolio, is defined as following:

F 251) $$AR_{rel,\alpha,d} = \frac{AR_{\alpha,d}}{AR_\alpha}$$ and thus, in connection with formula F 224, it follows that

F 252) $$AR_{rel,\alpha,d} = \frac{1 - \alpha - d^\alpha + \alpha d^{\alpha+1} - \alpha d^\alpha + \alpha d + d - d^{\alpha+1}}{(1-\alpha)\cdot(1-d)\cdot(1-d^\alpha)}$$

This term can not be substantially simplified and its derivations with respect to d and α are quite intricate. In the following, it is therefore examined only numerically for various combinations of d and α, see Figure 49.

Thereby it was shown, that relative accuracy ratios of preselected portfolios are the worse (i.e. smaller), the more corporations are excluded, which was expected. Relative losses in accuracy ratios are the bigger, the more discriminative the respective rating models are, i.e. the smaller α is. However, within an interval of realistic α-values ($0.1 < \alpha < 0.5$, see on this Appendix I and IV) results are not differing materially.

The results are mainly driven by d, the share of excluded low solvency corporations.

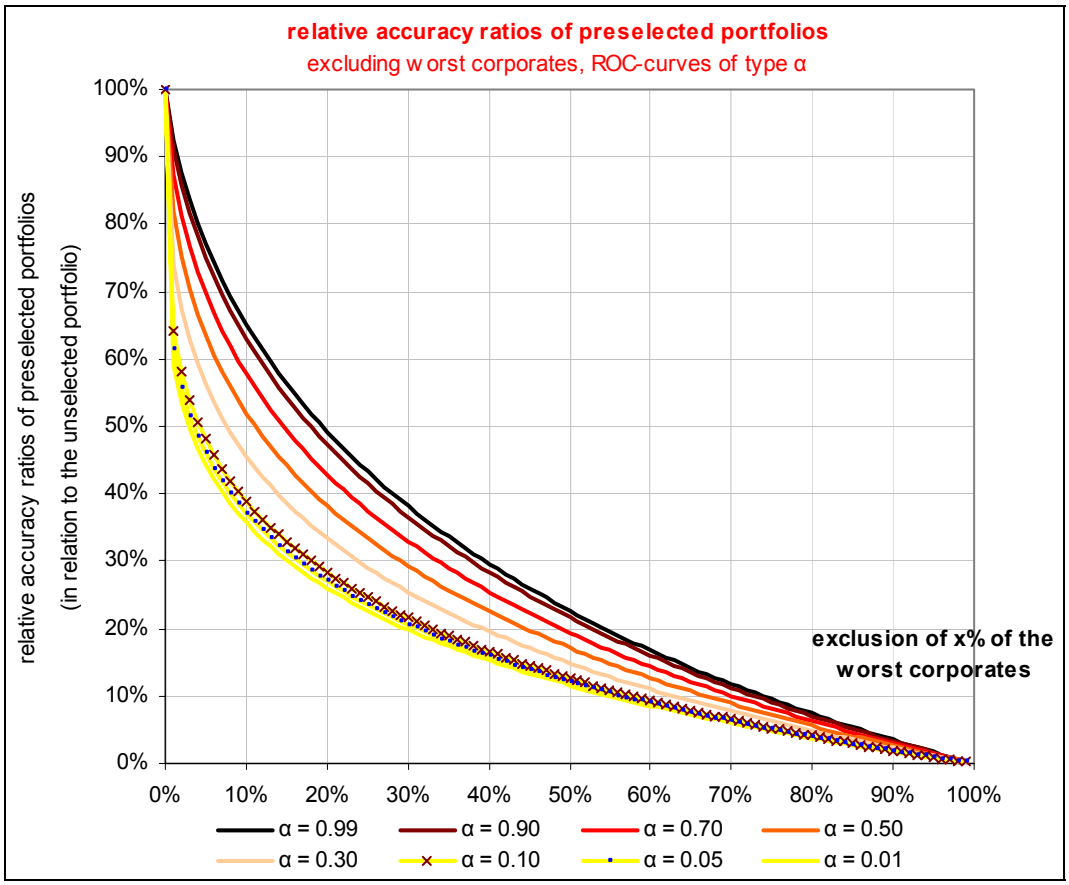

Figure 49: relative accuracy ratios of preselected portfolios for α-type ROC-curves, subject to various parameter values for α and d (share of the excluded "worst" corporations)

If only 10% of all corporations are excluded, the accuracy ratio of the rating model – measured on the remaining portfolio – is reduced by more than a half! If 50% of the corporations (with the lowest solvencies) are excluded, the measured accuracy of the preselected portfolio even reduces to less than the sixth part of the original value.

In the following, an analog analysis is carried out for β-type ROC-curves.

For β-type ROC-curves following equation holds:

F 253) $$\int_d^1 ROC_\beta(x)dx = \int_d^1 1-(1-x)^{1/\beta}dx = \left[x + \frac{\beta}{1+\beta}\cdot(1-x)^{(1+\beta)/\beta}\right]_d^1$$ and thus

F 254) $$\int_d^1 1-(1-x)^{1/\beta}dx = 1-\left(d + \frac{\beta}{1+\beta}\cdot(1-d)^{(1+\beta)/\beta}\right)$$ simplified to

F 255) $$\int_d^1 1-(1-x)^{1/\beta}dx = 1-d-\frac{(1-d)\cdot\beta}{1+\beta}\cdot(1-d)^{1/\beta}$$ and finally

F 256) $$\int_d^1 1-(1-x)^{1/\beta}dx = \frac{(1-d)\cdot\left(1+\beta-\beta\cdot(1-d)^{1/\beta}\right)}{1+\beta}$$

By insertion to formula F 241 for determining $AUC_{ROC,d,\beta}$ it follows that:

F 257) $$AUC_{ROC,d,\beta} = \frac{\int_d^1 ROC_\beta(x)dx - (1-d)\cdot ROC_\beta(d)}{(1-d)\cdot(1-ROC_\beta(d))}$$

F 258) $$AUC_{ROC,d,\beta} = \frac{\frac{(1-d)\cdot\left(1+\beta-\beta\cdot(1-d)^{1/\beta}\right)}{1+\beta} - (1-d)\cdot\left(1-(1-d)^{1/\beta}\right)}{(1-d)\cdot\left(1-\left(1-(1-d)^{1/\beta}\right)\right)}$$

F 259) $$AUC_{ROC,d,\beta} = \frac{\frac{1+\beta-\beta\cdot(1-d)^{1/\beta}}{1+\beta} - 1 + (1-d)^{1/\beta}}{(1-d)^{1/\beta}}$$

F 260) $$AUC_{ROC,d,\beta} = \frac{1+\beta-\beta\cdot(1-d)^{1/\beta} - (1+\beta) + (1+\beta)\cdot(1-d)^{1/\beta}}{(1+\beta)\cdot(1-d)^{1/\beta}}$$

F 261) $$AUC_{ROC,d,\beta} = \frac{(1-d)^{1/\beta}}{(1+\beta)\cdot(1-d)^{1/\beta}}$$

F 262) $$AUC_{ROC,d,\beta} = \frac{1}{1+\beta} = AUC_{ROC,\beta}$$ for all d, cf. formula F 233

Therefore, according to the formal analysis, β-type ROC-curves are completely unaffected by excluding discretionarily large shares of low solvency corporations, be it 1%, 50% or even 99% (graphical presentations of the results are set aside for this reason)[342] – while rating models that can be described by α-type ROC-curves are incurring information losses of over 50% when excluding only 10% of the worst corporations.
The analyses showed, that for preselected portfolios (whereby low solvency corporations were sorted out) rating model performance values may either decline materially – or may stay completely unaffected. The extent of information losses is mainly determined by whether the rating model is either relatively more discriminative in the range of low solvency corporations (which means, that its ROC-curve rather conforms to an α-type ROC-curve) or in the range of good solvency corporations (which means, that its ROC-curve rather conforms to a β-type ROC-curve).

Subsequently, a crosscheck with empirical data is carried out, that tests whether real rating models' *selection sensitivities* are actually as diverse as formal analyses imply. In the following, three real rating models are examined, which were already used in Appendix I.
Results are displayed for the ratings of S&P (1981-2003 pool data) and MOODY'S (1983-2003 pool data), whose ROC-curves are a mixture of α- and β-type ROC-curves but which are closer to β-type ROC-curves and for the CREDITREFORM Bonitätsindex Deutschland, whose ROC-curve nearly perfectly matches an α-type ROC-curve (see Appendix I).

---

342 Even rating models are conceivable, whose predictive qualities are *improving* when lowest solvency corporations are excluded. See for instance the ROC-curve of a rating model with minimal discriminative power at low solvency corporations and maximal discriminative power at good solvency corporations - which implies a linear progression of the ROC-curve for low errors of type II until intersecting the 100%-hit-rate-line. If a share *a*, with $0<a<1$, of the low solvency corporations is excluded, the remainder portfolio's ROC-curve still rises linearly until it intersects the 100%-hit-rate-line – the slope of this line, however, will be steeper by factor $(1-a)^{-1}$ – and thus its AUC_{ROC} and AR will increase.

In the first instance, accuracy ratios of the complete portfolios are calculated according to formula F 21, which requires rating class specific data concerning relative frequencies and historical default rates. Afterwards, the corporations of the worst (remaining) rating class are removed and accuracy ratio calculations for the remainder-portfolio are updated. These steps are repeated so long, until the remainder-portfolios are exclusively composed of (homogenous) corporations of the best respective rating classes.

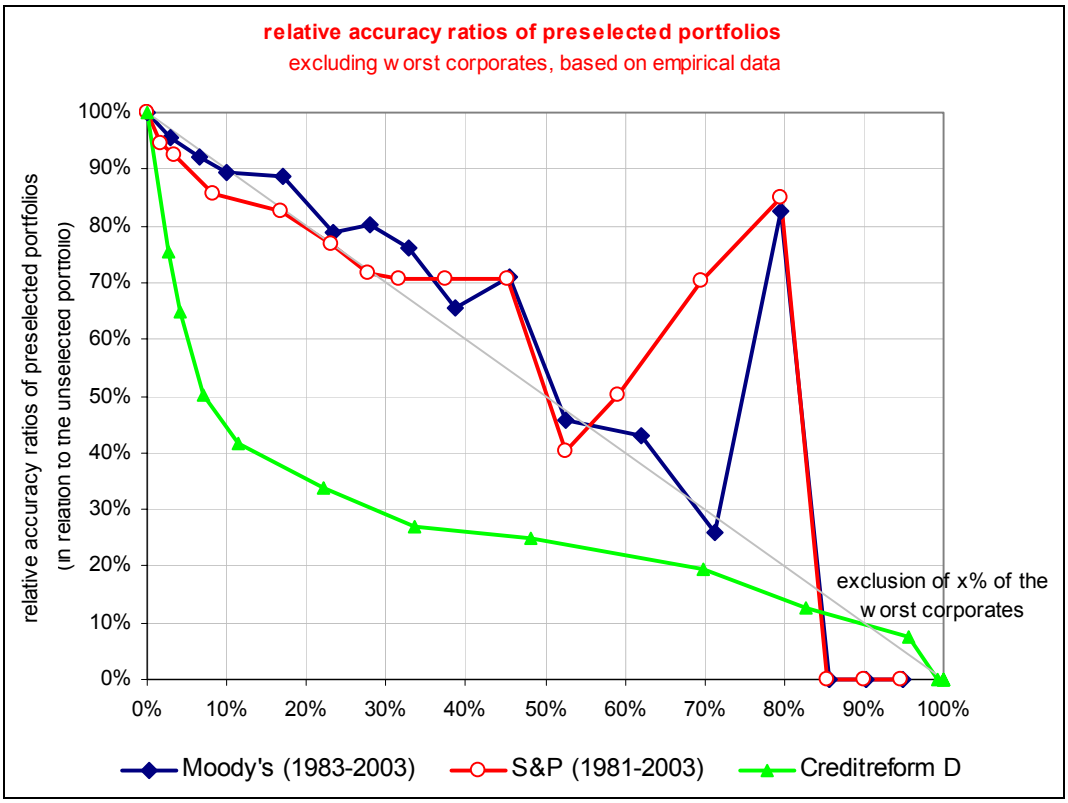

Figure 50: relative accuracy ratios of preselected portfolios; excluding corporations with worst ratings; based on empirical data of MOODY'S portfolios (1983-2003), S&P's portfolios (1981-2003) and CREDITREFORM Bonitätsindex Deutschland (1998-2000), source: own examinations

As could be expected, the progression of the relative-accuracy-ratio-graph for the CREDITREFORM-Bonitätsindex quite closely matches the progression of ROC_α-curves (cf. Figure 49 for α=0.3). The exclusion of only 10% of the corporations (with the lowest solvencies) reduces the accuracy ratio of the CREDITREFORM Bonitätsindex – which is already extremely low, see chapter 3.5 – for the preselected portfolio by 50%-60%!

According to the results for the MOODY'S and S&P's ratings, exclusions of low solvency corporations of up to 50% of the original sample[343] come along with nearly proportional accuracy losses for the remainder portfolio (see the progressions of the relative-accuracy-ratio-graphs along the secondary diagonal), i.e. an exclusion of 10% of the "worst" corporations induces an accuracy loss of (only) about 10%, an exclusion of 20% of the corporations induces an accuracy loss of 20%, etc.

The aft ranges of the curves are marked by discontinuities that may be evoked by statistical artifacts owing to the sparsity of defaults (within one-year prediction horizons) for investment grade, in particular for AA and AAA rated corporations.[344]

---

[343] The boundary between investment and speculative grade ratings is - based on both major agencies' pools data - marginally above 30%. At around 50% starts the domain of A/AA/AAA (resp. A/Aa/Aaa) ratings.

[344] High accuracy ratios based on the remaining portfolios are obtained for both agencies, if all corporations of the rating grades BBB+ to CCC/C (resp. Baa1 to Caa/C) are separated out, because historically both for MOODY'S and S&P's ratings only in rating class AA- (resp. Aa3) very few one-year defaults ever occurred,

The decline of relative accuracy ratio values for increasing exclusion rates down to 0% – in contrast to the formally derived total selection-insensitivity of "pure" ROC_α-curves – may also be traced back to the fact, that only *rating class specific* data was available, which implies that at the latest when only one rating class is left, the remaining portfolio's accuracy ratio must equal zero.

For the sake of completeness, in the following consequences on prediction accuracies (based on the remaining portfolios) in case of sorting out corporations with the *best* ratings are examined, too. These investigations are relevant for instance for investors, who intentionally specialize on low solvency corporations or for banks, if they don't succeed in acquiring costumers with excellent solvencies.

If u % of the "worst" corporations *remain* in the portfolio (in the following u is herein after referred to as "cut-off value"), the area under the scaled ROC-curve of the selected portfolio, $AUC_{ROC,u}$, subject to the ROC-curve of the complete portfolio is given by:

F 263) $$AUC_{ROC,u} = \frac{\int_0^u ROC(x)dx}{u \cdot ROC(u)}$$

For α-type curves it holds that:

F 264) $$\int_0^u ROC_\alpha(x)dx = \int_0^u x^\alpha dx = \left[\frac{x^{\alpha+1}}{\alpha+1}\right]_0^u \quad \text{and thus}$$

F 265) $$\int_0^u ROC_\alpha(x)dx = \frac{u^{\alpha+1}}{\alpha+1}$$

An insertion to formula F 263 yields:

F 266) $$AUC_{ROC,\alpha,u} = \frac{\frac{u^{\alpha+1}}{\alpha+1}}{u \cdot u^\alpha} \quad \text{and thus}$$

F 267) $$AUC_{\alpha,u} = \frac{1}{\alpha+1} = AUC_\alpha \quad \text{for all u (cf. formula F 222)!}$$

The results imply, that in case of α-type ROC-curves, excluding any fraction of the best solvency customers does not result in accuracy losses – irrespective of u and α (graphical presentations of the results are set aside for this reason).

For β-type curves it holds that:

F 268) $$\int_0^u ROC_\beta(x)dx = \int_0^u 1-(1-x)^{1/\beta}dx = \left[x + \frac{\beta}{1+\beta}\cdot(1-x)^{(1+\beta)/\beta}\right]_0^u$$

F 269) $$\int_0^u ROC_\beta(x)dx = u + \frac{\beta}{1+\beta}\cdot(1-u)^{(1+\beta)/\beta} - \frac{\beta}{1+\beta}$$

---

but *never* (at least not in 1981-2003 resp. 1983-2003) in the other rating classes AA, AA+, and AAA (resp. Aa2, Aa1, and Aaa). If in the next step corporations of rating class AA- (Aa3) are removed, prognosis accuracy sinks to zero, because all remaining rating classes are characterized by the same realized (one-year) default rate.

An insertion to formula F 263 yields:

$$\textbf{F 270)}\quad AUC_{ROC,\beta,u} = \frac{u + \frac{\beta}{1+\beta}\cdot(1-u)^{(1+\beta)/\beta} - \frac{\beta}{1+\beta}}{u\cdot\left(1-(1-u)^{1/\beta}\right)}$$

$$\textbf{F 271)}\quad AUC_{\beta,u} = \frac{u\cdot(1+\beta)+(1-u)\cdot\beta\cdot(1-u)^{1/\beta}-\beta}{(1+\beta)\cdot\left(u-u\cdot(1-u)^{1/ß}\right)}$$

$$\textbf{F 272)}\quad AUC_{ROC,\beta,u} = \frac{u+u\cdot\beta+\beta\cdot(1-u)^{1/\beta}-u\cdot\beta\cdot(1-u)^{1/\beta}-\beta}{u+u\cdot\beta-u\cdot(1-u)^{1/\beta}-u\cdot\beta\cdot(1-u)^{1/\beta}}$$

The *accuracy ratio* is given by (cf. formula F 248):

$$\textbf{F 273)}\quad AR_{\beta,u} = \frac{2u+2u\cdot\beta+2\beta\cdot(1-u)^{1/\beta}-2u\cdot\beta\cdot(1-u)^{1/\beta}-2\beta-u-u\cdot\beta+u\cdot(1-u)^{1/\beta}+u\cdot\beta\cdot(1-u)^{1/\beta}}{u+u\cdot\beta-u\cdot(1-u)^{1/\beta}-u\cdot\beta\cdot(1-u)^{1/\beta}}$$

$$\textbf{F 274)}\quad AR_{\beta,u} = \frac{u+u\cdot\beta-2\beta+(1-u)^{1/\beta}\cdot(2\beta-u\cdot\beta+u)}{(1+\beta)\cdot u\cdot\left(1-(1-u)^{1/ß}\right)}$$

For the relative accuracy ratio $AR_{rel,\beta,u}$, according to formulas F 251 and F 233 it follows that:

$$\textbf{F 275)}\quad AR_{rel,\beta,u} = \frac{AR_{\beta,u}}{AR_{\beta}} = \frac{AR_{\beta,u}}{\left(\frac{1-\beta}{1+\beta}\right)} \quad \text{and thus}$$

$$\textbf{F 276)}\quad AR_{rel,\beta,u} = \frac{u+u\cdot\beta-2\beta+(1-u)^{1/\beta}\cdot(2\beta-u\cdot\beta+u)}{(1-\beta)\cdot u\cdot\left(1-(1-u)^{1/ß}\right)}$$

Subsequently, this term is numerically determined and plotted for various u and β, see the following Figure 51. It appears, that relative accuracy ratios of the selected portfolios are the worse (lower) the more good-solvency corporations are excluded, i.e. the lower the cut-off value u is. Relative accuracy losses are the smaller, the more selective a rating model is, i.e. the smaller β is. In case of highly selective models ($\beta \geq 0.1$) relative accuracy starts decreasing materially only for very small u-values, i.e. when very many of the best corporations are excluded. In the relevant range (for large u and small α or β) it turns out, that selection sensitivity with respect to ROC-curve shapes is considerable lower than in case of excluding corporations with *worst* ratings.

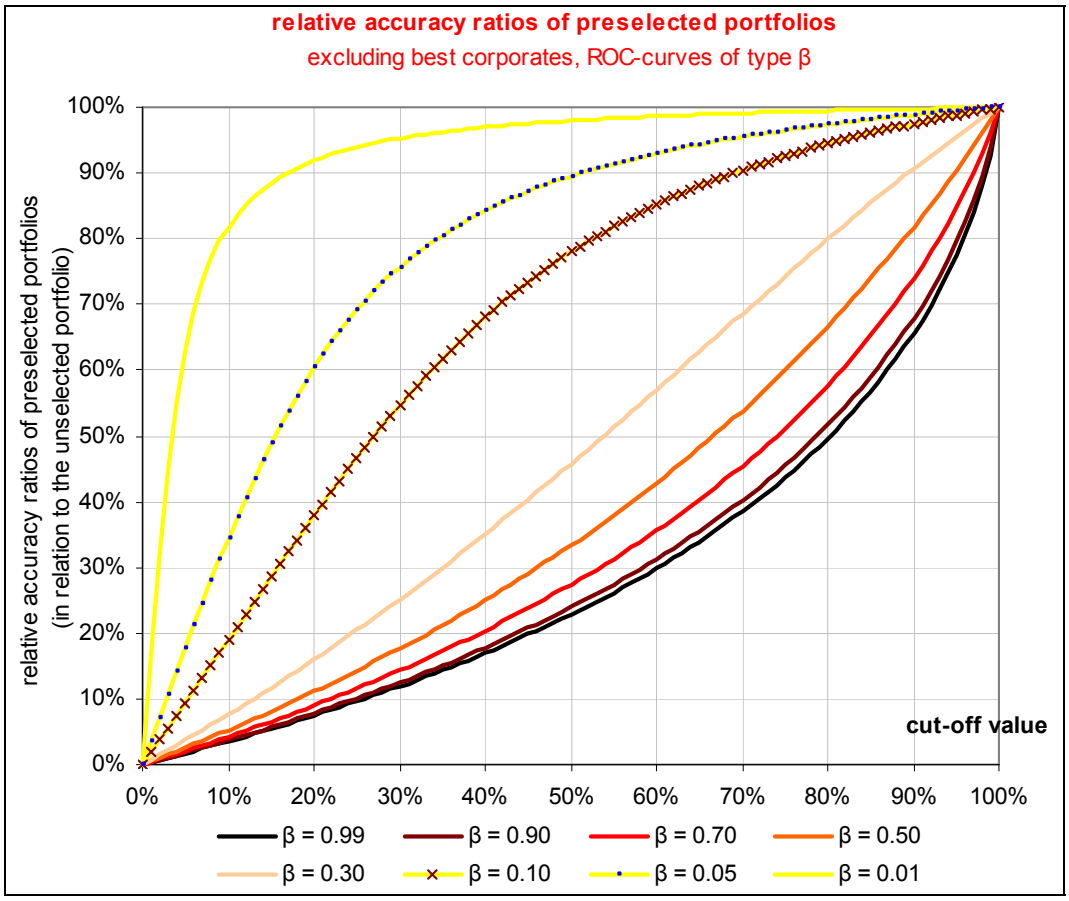

Figure 51: relative accuracy ratios of preselected portfolios for β-type ROC-curves, subject to various parameter values for β and u (cut-off value)

Again, selection sensitivity of three real rating models is determined based on empirical data, see Figure 52. Contrary to the proceedings described above, not the respective *worst* rating are removed step by step, but the respective *best* rating classes.

The results found empirically may quite well be explicated by the theoretical analyses. In case of CREDITREFORM Bonitätsindex Deutschland, whose ROC-curve nearly perfectly matches an α-type ROC curve, the (considerably low) accuracy virtually stays constant when measured on the remaining portfolio – even when up to 80% of the corporations with the best solvencies are sorted out! Only if more than 80% are sorted out, relative accuracy decreases noticeably – which is also due to the fact, that the empirical analyses are employing *rating class specific* data (concerning relative frequencies and default rates) and not continuous or quasi-continuous individual scores. If only one rating class remains, relative accuracy measured on the remaining portfolios must sink to zero.

The curve progression of relative accuracy ratios for S&P's and MOODY'S ratings, whose ROC-curves rather match β-type ROC-curves, can be explained quite well by a curve progression as those that are given in Figure 51 (for β=0.1).

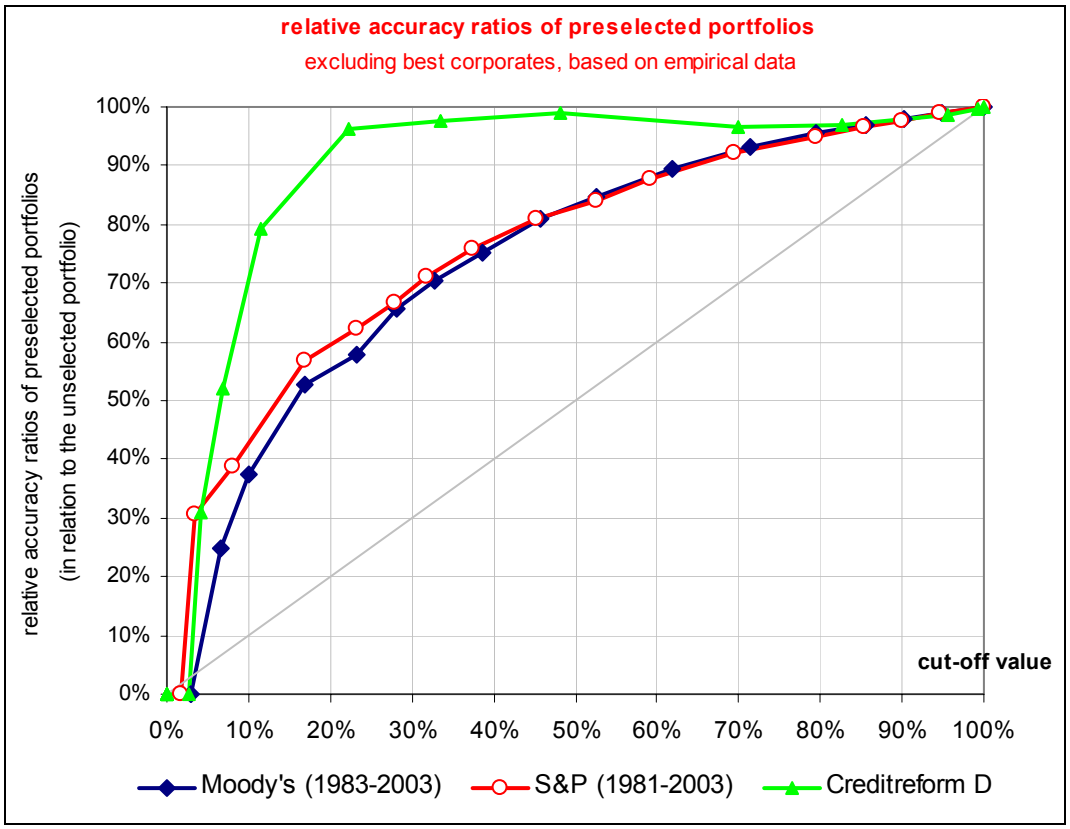

Figure 52: relative accuracy ratios of preselected portfolios; excluding corporations with best ratings; based on empirical data of MOODY'S portfolios (1983-2003), S&P's portfolios (1981-2003) and CREDITREFORM Bonitätsindex Deutschland (1998-2000), source: own examinations

Overall, empirically and theoretically derived results permit the conclusion, that sorting out corporations with good solvencies induces clearly less dramatic and under-proportional accuracy losses (when measuring predictive quality of rating models based on the remaining portfolios) than in case of sorting out corporations with the worst solvencies.

As a rough rule of thumb it may be stated, that each percent of excluded corporations with good solvencies comes along with accuracy losses (measured on the remaining portfolio) of about ¼ percent. That is, excluding 20% of the best rated corporations from a portfolio will reduced measured accuracy by (only) 5% and excluding 40% of all corporations will reduce measured accuracy by (only) 10% or less.

Dresden Discussion Paper Series in Economics

17/03 **Lehmann-Waffenschmidt, Marco / Reina, Livia:** Coalition formation in multilateral negotiations with a potential for logrolling: an experimental analysis of negotiators' cognition processes

18/03 **Lehmann-Waffenschmidt, Marco / Böhmer, Robert:** Mentality Matters – Thorstein Veblens ‚Regime of Status' und Max Webers ‚Protestantische Ethik' aus der Sicht des (radikalen) Konstruktivismus. Eine Anwendung auf die ökonomischen Probleme des deutschen Wiedervereinigungsprozesses

19/03 **Eisenschmidt, Jens / Wälde, Klaus:** International Trade, Hedging and the Demand for Forward Contracts

20/03 **Broll, Udo / Wong, Kit Pong:** Capital Structure and the Firm under Uncertainty

01/04 **Lehmann-Waffenschmidt, Marco:** A Note on Continuously Decomposed Evolving Exchange Economies

02/04 **Friedrich, B. Cornelia:** Competition and the Evolution of Market Structure in the E-conomy.

03/04 **Berlemann, Michael / Dittrich, Marcus / Markwardt, Gunther:** The Value of Non-Binding Announcements in Public Goods Experiments. Some Theory and Experimental Evidence

04/04 **Blum, Ulrich / Schaller, Armin / Veltins, Michael:** The East German Cement Cartel: An Inquiry into Comparable Markets, Industry Structure, and Antitrust Policy

05/04 **Schlegel, Christoph:** Analytical and Numerical Solution of a Poisson RBC model

06/04 **Lehmann-Waffenschmidt, Marco:** Die ökonomische Botschaft in Goethes „Faust"

07/04 **Fuchs, Michaela / Thum, Marcel:** EU Enlargement: Challenges for Germany's New Laender

08/04 **Seitz, Helmut:** Implikationen der demographischen Veränderungen für die öffentlichen Haushalte und Verwaltungen

09/04 **Sülzle, Kai:** Duopolistic Competition between Independent and Collaborative Business-to-Business Marketplaces

10/04 **Broll, Udo / Eckwert, Bernhard:** Transparency in the Interbank Market and the Volume of Bank Intermediated Loans

11/04 **Thum, Marcel:** Korruption

12/04 **Broll, Udo / Hansen, Sabine / Marjit, Sugata:** Domestic labor, foreign capital and national welfare

13/04 **Nyamtseren, Lhamsuren:** Challenges and Opportunities of Small Countries for Integration into the Global Economy, as a Case of Mongolia

01/05 **Schubert, Stefan / Broll, Udo:** Dynamic Hedging of Real Wealth Risk

02/05 **Günther, Edeltraud / Lehmann-Waffenschmidt, Marco:** Deceleration - Revealed Preference in Society and Win-Win-Strategy for Sustainable Management. Concept and Exprimental Evidence

03/05 **Sennewald, Ken:** Controlled Stochastic Differential Equations under Poisson Uncertainty and with Unbounded Utility

04/05 **Sennewald, Ken / Wälde, Klaus:** "Itô's Lemma" and the Bellman equation: An applied view

05/05 **Neumann, Anne / Siliverstovs, Boriss:** Convergence of European Spot Market Prices for Natural Gas?

06/05 **Hirschhausen, Christian von / Cullmann, Astrid:** Efficiency Analysis of German Electricity Distribution Utilities

07/05 **Seitz, Helmut / Freigang, Dirk / Kempkes, Gerhard:** Demographic Change and Federal Systems: Some Preliminary Results for Germany

08/05 **Bemmann, Martin:** Verbesserung der Vergleichbarkeit von Schätzgüteergebnissen von Insolvenzprognosestudien