WORKING PAPER SERIES

Measures of Predictive Success for Rating Functions

Sebastian Ostrowski/Peter Reichling

Working Paper No. 18/2010

FACULTY OF ECONOMICS
AND MANAGEMENT

Impressum (§ 5 TMG)

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Die Dekanin

Verantwortlich für diese Ausgabe:
Sebastian Ostrowski
Otto-von-Guericke-Universität Magdeburg
Fakultät für Wirtschaftswissenschaft
Postfach 4120
39016 Magdeburg
Germany

http://www.ww.uni-magdeburg.de/fwwdeka/femm/

Bezug über den Herausgeber
ISSN 1615-4274

Measures of Predictive Success for Rating Functions*

Sebastian Ostrowski and Peter Reichling**

August, 2010

Abstract

Aim of our paper is to develop an adequate measure of predictive success and accuracy of rating functions. At first, we show that the common measures of rating accuracy, i.e. area under curve and accuracy ratio, respectively, lack of informative value of single rating classes. Selten (1991) builds up an axiomatic framework for measures of predictive success. Therefore, we introduce a measure for rating functions that fulfills the axioms proposed by Selten (1991). Furthermore, an empirical investigation analyzes predictive power and accuracy of Standard & Poor's and Moody's ratings, and compares the rankings according to area under curve and our measure.

Keywords Accuracy Measure · Rating Functions · Predictive Success · Discriminative Power

J.E.L. Classification C52 · G21 · G33

*The authors are very grateful to Stefan Hlawatsch and Bodo Vogt for their helpful comments and suggestions

**Both authors are from Otto-von-Guericke-University Magdeburg, Faculty of Economics and Management, Department of Banking and Finance, ✉: Postfach 4120, 39016 Magdeburg, Germany, ☎: +49 391 67-12256, 🖷: sebastian.ostrowski@ovgu.de

1 Introduction

A rating should provide information about default risk in terms of quantifying a probability that the debtor will not meet its payment obligations. Therefore, a single rating class should correspond to a certain default rate, which means that a specified proportion of the debtors within that rating class is assumed to default. In general, there are only two states occurring in each rating class, namely default or non-default[1] of a company or a credit. An important factor for the goodness of a rating function is its discriminative power, which means a distinction between defaulted and non-defaulted companies. Measures for this feature are adopted from signalling theory where a similar structure of information retrieval exists so that certain signals have to be separated from noise.

One of the first who introduced such measures was Bamber (1975). He analyzed the receiver operating characteristic (ROC) graph and the area below this graph, which is today known as the area under curve (AUC). A linear transformation of the AUC is called accuracy ratio (AR), which corresponds to the cumulative accuracy profile (CAP) and was described in the context of rating functions by Keenan and Sobehart (1999). They propose that "one of the most useful properties of CAPs is that they reveal information about the predictive accuracy of the model over its entire range of risk scores". In other words, the focus is put on the ordering of debtors according to their risk scores. This may be satisfactory for credit risk models without using different classifications like a Z- or Zeta-score model proposed by Altman (1968); Altman et al. (1977); Altman and Saunders (1998). However, when using classification models like Standard & Poor's (S&P) and Moody's it is not sufficient to just validate the ordering of debtors but additionally the accuracy within each rating class. This is not ensured by the common and frequently applied measures AUC and AR.

This shortcoming will be resolved by introducing a new measure of predictive success based on axioms proposed by Selten (1991). He analyzed measures of predictive success for area theories for characteristic function experiments.[2] The adaption of area theoretical

[1] We interpret "default" in terms of occurrence of a credit event.

[2] See Selten and Krischker (1983).

measures for ratings is reasonable since an area theory predicts subsets within an observation space. In the rating context, the observation space is described by the number of all rated companies and the defaulted ones are the predicted subset.

The paper is organized as follows: Section 2 introduces the common measures AUC and AR in more detail and shows how these measures may lead to false conclusions. In Section 3 the axiomatic approach of Selten (1991) is presented and applied to construct a new measure for predictive success of rating functions. In Section 4 our new measure is compared to AUC by an empirical analysis of S&P and Moody's ratings for the period from 1982 to 2001. Section 5 concludes.

2 Common Accuracy Measures

To explain the subsequent derivation we present some general notations used throughout the paper. Assume a rating function with k rating classes and the two observations default (D) and non-default (ND) within each class. Furthermore, the number of debtors in each rating class is denoted by $n_i, i = 1, \ldots, k$. Given the distribution of n companies over the rating classes, a contingency table can be constructed as presented in Table 1.

Table 1: Exemplarily Contingency Table

The table shows the contingency table of a rating function with k rating classes.

	Rating class in $t = 0$	Observation in $t = 1$ D	ND
Prediction	1	A_1	B_1
	$\vdots$	$\vdots$	$\vdots$
	k	A_k	B_k

The hit rate hr_i for rating class i is defined as the proportion of defaulted debtors in rating class i over all defaulted debtors. To construct the ROC the cumulated hit rate HR_s is needed, which equals the sum of all hit rates hr_i up to a predefined rating class

s. According to the notation used in Table 1, hr_i and HR_s result as:

$$hr_i = \frac{A_i}{\sum_{j=1}^{k} A_j} \tag{1}$$

$$HR_s = \sum_{j=1}^{s} hr_j = \sum_{j=1}^{s} \frac{A_j}{\sum_{l=1}^{k} A_l}. \tag{2}$$

Correspondingly, the false alarm rate far_i for rating class i is defined as the ratio of non-defaulted obligors in class i to the overall number of non-defaulted obligors. Formally written, far_i and the cumulated false alarm rate FAR_s result as:

$$far_i = \frac{B_i}{\sum_{j=1}^{k} B_j} \tag{3}$$

$$FAR_s = \sum_{j=1}^{s} far_j = \sum_{j=1}^{s} \frac{B_j}{\sum_{l=1}^{k} B_l}. \tag{4}$$

The ROC curve results from plotting the cumulated hit rates on the ordinate against the cumulated false alarm rates on the abscissa for each rating class including the points (0,0) and (1,1). AUC is, as the name suggests, the area below this curve and can be calculated as:

$$\text{AUC} = \sum_{i=1}^{k} (FAR_i - FAR_{i-1}) \cdot \frac{HR_i + HR_{i-1}}{2} \tag{5}$$

$$\text{where} \qquad FAR_0 = HR_0 = 0 \text{ and } FAR_k = HR_k = 1.$$

Alternatively, AUC can also be calculated when using an integral structure:

$$\text{AUC} = \sum_{i=1}^{k} \int_{0}^{far_i} \left(\frac{hr_i}{far_i} \cdot x + HR_{i-1} \right) \mathrm{d}x \tag{6}$$

Engelmann et al. (2003) proved the linear relationship between AUC and AR: AR $= 2 \cdot \text{AUC} - 1$. Therefore, we limit our further analysis just on the AUC.

In the following, we show that the interpretation of AUC may be doubtful or may result in false conclusions. To have a better insight into the arising problems, Table 1 is less meaningful for our further analysis and will be restructured in more detail in Table 2.

Table 2: Extended Contingency Table

The table shows the extended contingency table of a rating function with k rating classes.

	Rating class in $t = 0$		Observation in $t = 1$ D	ND
Prediction	1	D	A_1	B_1
		ND	C_1	D_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	k	D	A_k	B_k
		ND	C_k	D_k

To explain the new structure, we take a look at the rows of rating class 1. The first row presents the number of predicted defaults within this rating class, which equal the sum of A_1 and B_1. Correspondingly, the second row, indicated with ND, presents the number of predicted non-defaults within this rating class, where the number results from the sum of C_1 and D_1. When now looking at the column denoted with D, the true number of defaulted debtors in rating class 1 results as the sum of A_1 and C_1. Analogously, the same results for the non-defaulted debtors in the last column.

Since only the numbers of defaulted and non-defaulted debtors is predicted and there is no company-specific prediction, only three cases may appear: we predict either more, less or exactly the number of defaults. When predicting more defaults than actually occur, the value of C_1 is zero and the excess of predicted defaults will appear in cell B_1. In case of predicting less defaults than actually occur, B_1 equals zero and the shortfall appears in C_1. When exactly predicting the number of actual defaults, B_1 and C_1 will both be zero.

A contingency table of two exemplarily rating functions with three rating classes, 30 debtors with ten defaulters is presented in Table 3. The AUC value for rating function I is 1. This suggests that rating function I is a perfect one. Indeed, all defaulters are in rating class 1 but every single debtor in this rating class was not recognized as a defaulter. Furthermore, the accuracy in the other two rating classes is low and not existent,

Table 3: Contingency Table for Two Exemplarily Rating Functions

The table shows the extended contingency table of two exemplarily rating functions with three rating classes.

	Rating class in $t=0$		Rating Function I Observation in $t=1$		Rating Function II Observation in $t=1$	
			D	ND	D	ND
Prediction	1	D	0	0	5	0
		ND	10	0	0	10
	2	D	0	10	3	0
		ND	0	0	0	7
	3	D	0	5	2	0
		ND	0	5	0	3

respectively. Rating class 3 has a merely random prediction process and the classes 1 and 2 do not show any correct prediction.

As a contrary example, we look at rating function II. Here, the rating function does perfectly predict the number of defaulters and non-defaulters in every single rating class. However, AUC amounts to only 0.4875, which is below the AUC value of a random rating function (0.5). Thus according to AUC, rating function I is preferred to rating function II. Obviously, rating function I exhibits a better ordering property of the defaulters than rating function II since all defaulters are concentrated in rating class 1. However, the higher (true) predictive power seems to be inherent in rating function II since the defaulters and non-defaulters in each rating class are predicted correctly.

The general aim of a rating function is a precise estimation of the default rate for each rating class. In case a rating function cannot fulfill this crucial requirement, a wrong risk premium will be added to the other credit costs. AUC and AR are just measures of successful ranking but not of successful prediction in each rating class. Therefore, we will present a new measure of predictive success in the following section.

3 A New Measure of Predictive Success and its Axiomatic Foundation

Selten and Krischker (1983) were one of the first to analyze measures of predictive success

with respect to their general structure. They analyzed these measures in the context of experiments and their results. In general, such measures contain two parts. One part refers to the accuracy of the prediction, i.e. the relative frequency of correct predictions which is also called hit rate.[3] The second part refers to the precision of a prediction, i.e. the relative size of the predicted outcome to all outcomes. Selten (1991) refers to this part as the area of a theory. According to their properties, Selten (1991) analyzed different measures. These measures are either a difference measures, i.e. the difference between hit rate and corresponding area, or a ratio measure, i.e. the ratio of hit rate and area.

Selten (1991) argued with the help of a small numerical example that a difference measure is more favorable than a ratio measure. He considered two theories, the first exhibits a hit rate of 0.9 and an area of 0.1 and the second exhibits a hit rate of 0.01 and an area of 0.0001. A ratio measure would prefer the second theory since its value of 100 is greater than 9, which is the ratio of the first theory. However, this implication may not be true since a hit rate of 0.01 means that in 99 of 100 cases the prediction of this theory is wrong. In contrast to this, the first theory predicts 90 of 100 cases correctly and uses ten percent of the set of all outcomes. It is obvious that the trade-off between hit rate and predicted area may have unfavorable impacts on the decision between two theories. Transferred to the framework of rating functions, this means that it almost does not matter how good the prediction is if the area of a rating function is sufficiently small, so that the rating function outperforms other rating functions.

Reconsidering the structure of Equation (6), AUC is similar to a ratio measure. The numerator contains the hit rate, but with respect to the ordering of defaults and not as the relative size of the predicted outcome to all outcomes. Thus, once more the predictive power of a rating function cannot be addressed with the AUC measure. The denominator consists of far_i, which is kind of an area measure in a more abstract way.[4] If we consider the whole set of rated companies, a rating function partitions this set into subsets represented by rating classes. Each class can be partitioned into two subsets: defaulters

[3] This hit rate needs not be identical to the hit rate of the AUC framework.

[4] It may happen that $far_i = 0$ but this shall not influence the mentioned shortcomings of AUC.

and non-defaulters. The false alarm rate describes the relative size of wrongly assessed companies in that rating class given that hr_i is not zero. Again, just the ordering property is addressed. Taking a look again at rating class 1 of rating function I in Table 3, it is apparent that a small area, i.e. a false alarm rate near or equal to zero, leads to a high AUC value, independent of how accurate the prediction was.

Inspired by these considerations, we introduce a difference measure to overcome the drawbacks of AUC. Since hit rate and area are the two driving factors, we start looking at the hit rate. To determine the predictive power of a rating function, it is not only important to look at the predictive power with respect to the defaulters but also with respect to non-defaulters. Both numbers are related to each other in every rating class since the total number per class is partitioned into defaulters and non-defaulters. We define the hit rate as the relative deviation of the predicted number to the actual number of either defaulted or non-defaulted debtors within each rating class and denote the hit rate as $r_{\{.\}}$, where the subscript D denotes default and ND denotes non-default. The following equations for hit rates refer to rating class i in Table 2:

$$r_{D,i} = \begin{cases} 0 & , A_i = B_i = C_i = 0 \\ 1 - \frac{|(A_i+B_i)-(A_i+C_i)|}{\max\{(A_i+B_i),(A_i+C_i)\}} & , \text{otherwise.} \end{cases} \tag{7}$$

This can be rearranged to:

$$r_{D,i} = \begin{cases} 0 & , A_i = B_i = C_i = 0 \\ 1 - \frac{C_i}{A_i+C_i} & , B_i = 0 \wedge A_i \neq 0 \neq C_i \\ 1 - \frac{B_i}{A_i+B_i} & , C_i = 0 \wedge A_i \neq 0 \neq B_i. \end{cases} \tag{8}$$

Analogously, the non-default hit rate reads as:

$$r_{ND,i} = \begin{cases} 0 & , B_i = C_i = D_i = 0 \\ 1 - \frac{C_i}{C_i+D_i} & , B_i = 0 \wedge D_i \neq 0 \neq C_i \\ 1 - \frac{B_i}{B_i+D_i} & , C_i = 0 \wedge D_i \neq 0 \neq B_i. \end{cases} \tag{9}$$

The following interpretation of the hit rate is restricted to the default case as the non-default case can be interpreted analogously. Consider the first situation where the hit rate equals zero. In this case there are no defaulters in this class and no default was predicted. This implies that the corresponding hit rate for the non-defaulters in this rating class equals one. In the other situation we subtract the relative prediction error from one to get the hit rate. By this definition we assure that deviations in both directions, i.e. predicting more than or less than the actual number, are equally evaluated. Consider for example a rating class with ten defaulters but only one default was predicted. Then the relative deviation equals 0.9 and, thus, the hit rate amounts to 0.1. Consider another case with one defaulter but ten predicted defaults. Here, again the relative deviation equals 0.9 and so again the hit rate equals 0.1. In both situations the distance between predicted and actual defaults is the same, resulting in the same hit rate value. Thus, neither a too optimistic nor a too pessimistic prediction is advantageous.

The second driving factor of our new measure is the area. As mentioned before, the area denotes the relative size of the outcome to all outcomes in this class. For example in the default case, the area is defined as the ratio of defaulters in this class to the number of all obligors within that class. Thus, the areas $a_{\{.,i\}}$ for the default and non-default case in rating class i read as:

$$a_{D,i} = \frac{A_i + C_i}{A_i + B_i + C_i + D_i} \quad \text{and} \quad a_{ND,i} = \frac{B_i + D_i}{A_i + B_i + C_i + D_i} \tag{10}$$

Eventually, the measures of predictive success for defaults and non-defaults result as the difference between hit rate and corresponding area for each rating class: $m_{D,i} = r_{D,i} - a_{D,i}$ for the default case of rating class i (the non-default case results analogously: $m_{ND,i} = r_{ND,i} - a_{ND,i}$). There are $2 \cdot k$ measures for each rating function, which is inconvenient for comparisons of rating functions. Thus, an aggregate measure may serve better for this purpose. We suggest to create a weighted sum of the two measures $m_{D,.}$ and $m_{ND,.}$ over all rating classes by weighting $m_{D,i}$ with hr_i and $m_{ND,i}$ with far_i. This is meaningful since the importance of each measure can be described by the proportion of the specific

outcome in a rating class to the total number of that outcome over all rating classes. Therefore, the final measure for a rating function can be described by:

$$m = \sum_{i=1}^{k} \left(hr_i \cdot m_{D,i} + far_i \cdot m_{ND,i} \right). \tag{11}$$

We apply this new measure to the rating functions presented in Table 3. The corresponding values for rating function I and II are $m^{\mathrm{I}} = -1.75$ and $m^{\mathrm{II}} = 0.995$ so that rating function II is better than rating function I, which was suggested by the numbers in the table. However, the resulting figures are not always directly comparable unless they have a different sign. The problem is due to different numbers of debtors that defaulted or non-defaulted in each rating class. Thus, the minimal and maximal values that the measure can take depends on the rating class and are different for each rating function. Even for one rating function these values may change over time. Therefore, it is necessary to create a standardized measure for each rating function and each time period considered. To derive the maximum value of a rating function we look at the general structure for a perfect prediction. A perfect prediction is characterized by a hit rate of one for both the number of defaults and the number of non-defaults in each rating class. This determines the area in each rating class since there is no wrong prediction and, therefore, according to the notation used before, B_i and C_i are zero. Hence, the maximum value $m_{\max}$ can be calculated as:

$$\begin{aligned} m_{\max} &= \sum_{i=1}^{k} \left(\underbrace{\frac{A_i}{\sum_{j=1}^{k} A_j}}_{=hr_i} \cdot \left(\underbrace{1}_{=r_{D,i}} - \underbrace{\frac{A_i}{A_i + D_i}}_{=a_{D,i}} \right) + \underbrace{\frac{D_i}{\sum_{j=1}^{k} D_j}}_{=far_i} \cdot \left(\underbrace{1}_{=r_{ND,i}} - \underbrace{\frac{D_i}{A_i + D_i}}_{=a_{ND,i}} \right) \right) \\ &= \left(\frac{1}{\sum_{j=1}^{k} A_j} + \frac{1}{\sum_{j=1}^{k} D_j} \right) \cdot \sum_{i=1}^{k} \frac{A_i \cdot D_i}{A_i + D_i} \end{aligned} \tag{12}$$

For instance, the maximum value for rating function II in Table 3 equals 0.995.

Rating function I exhibits a negative measure value and the question arises how bad is the rating function compared to the lowest possible m given the observations. The number of defaults and non-defaults within each rating class are observed values and the area is fixed. The worst prediction, and thus the lowest m, will be reached when predicting either only defaults or only non-defaults, depending on the specific values of hr_i and far_i. When predicting just one outcome, for instance just defaults D, the measure of this outcome will become zero and the corresponding measure for non-defaults will be $-a_{ND,i}$. Therefore, the minimum of $hr_i \cdot (-a_{D,i})$ and $far_i \cdot (-a_{ND,i})$ for each rating class i has to be considered and aggregated over all rating classes. The minimum value $m_{\min}$ can be calculated as:

$$\begin{aligned} m_{\min} &= \sum_{i=1}^{k} \min\left\{hr_i \cdot m_{D,i}, far_i \cdot m_{ND,i}\right\} \\ &= \sum_{i=1}^{k} \min\left\{ -\frac{(A_i + C_i)^2}{n_i \cdot \sum_{j=1}^{k}(A_j + C_j)}, -\frac{(B_i + D_i)^2}{n_i \cdot \sum_{j=1}^{k}(B_j + D_j)} \right\}, \end{aligned} \tag{13}$$

where n_i denotes the number of debtors in rating class i. Thus, the minimum value for rating function I in Table 3 equals -2 and it is evident by Equation (13) that this minimum value is always negative.

Now, we derive a standardized measure M for comparing different rating functions. The relative quality of a rating function is described by its distance to its minimum and maximum value. Therefore, a standardized measure that only takes values from the unit interval can be computed as:

$$M = \frac{m - m_{\min}}{m_{\max} - m_{\min}} \tag{14}$$

The corresponding values of rating functions I and II are $M^{\mathrm{I}} = 0.125$ and $M^{\mathrm{II}} = 1$, respectively, indicating that rating function II shows perfect predictive power.

Finally in this section, we check the axioms proposed by Selten (1991) that present desirable properties of measures of predicted success. These axioms are based on measures that are functions of hit rate-area combinations $(r, a) \in [0, 1] \times [0, 1]$, which is fulfilled for our definitions. The first two of Selten's axioms are propositions on monotonicity with

respect to hit rate and area, respectively:

$$\forall a \in [0,1], 0 \leq r_2 < r_1 \leq 1 : m(r_1, a) > m(r_2, a) \tag{15}$$

and

$$\forall r \in [0,1], 0 \leq a_1 < a_2 \leq 1 : m(r, a_1) > m(r, a_2). \tag{16}$$

The third of Selten's axioms states that the resulting measure should be continuous everywhere on the unit square. The fourth axiom states that there exists some function that allows for a cost-benefit evaluation. This means that one can decide whether a theory is better than another one by just comparing the differences in hit rates and areas. The fifth axiom refers to the equivalence of trivial theories, namely $m(0,0) = m(1,1)$. Finally, the sixth axiom refers to linearity of the measure:

$$\forall \alpha \in [0,1] : m\left(\alpha r_1 + (1-\alpha) r_2, \alpha a_1 + (1-\alpha) a_2\right) = \alpha m(r_1, a_1) + (1-\alpha) m(r_2, a_2) \tag{17}$$

Since our measure is a difference measure for each rating class (with a cardinal characterization), Theorem 2 in Selten (1991) shows that the axioms 1, 2, 5 and 6 are fulfilled. Axiom 3 is obviously fulfilled using the so called $\epsilon - \delta$ definition of continuity and choosing $\delta = \epsilon$, whereas axiom 4 is fulfilled by defining the canonical cost-benefit function

$$\Delta(r_1 - r_2, a_1 - a_2) = \begin{cases} 1 & , r_1 - r_2 > a_1 - a_2 \\ 0 & , r_1 - r_2 = a_1 - a_2 \\ -1 & , r_1 - r_2 < a_1 - a_2. \end{cases}$$

Our standardized measure M results as a linear transformation of the specific measure m_i, whereas hr_i, far_i, $m_{\max}$ and $m_{\min}$ are independent of the predictive power of the rating function under consideration and, thus, act as constants. Therefore, all desired properties of a measure of predictive success are fulfilled our standardized measure M.

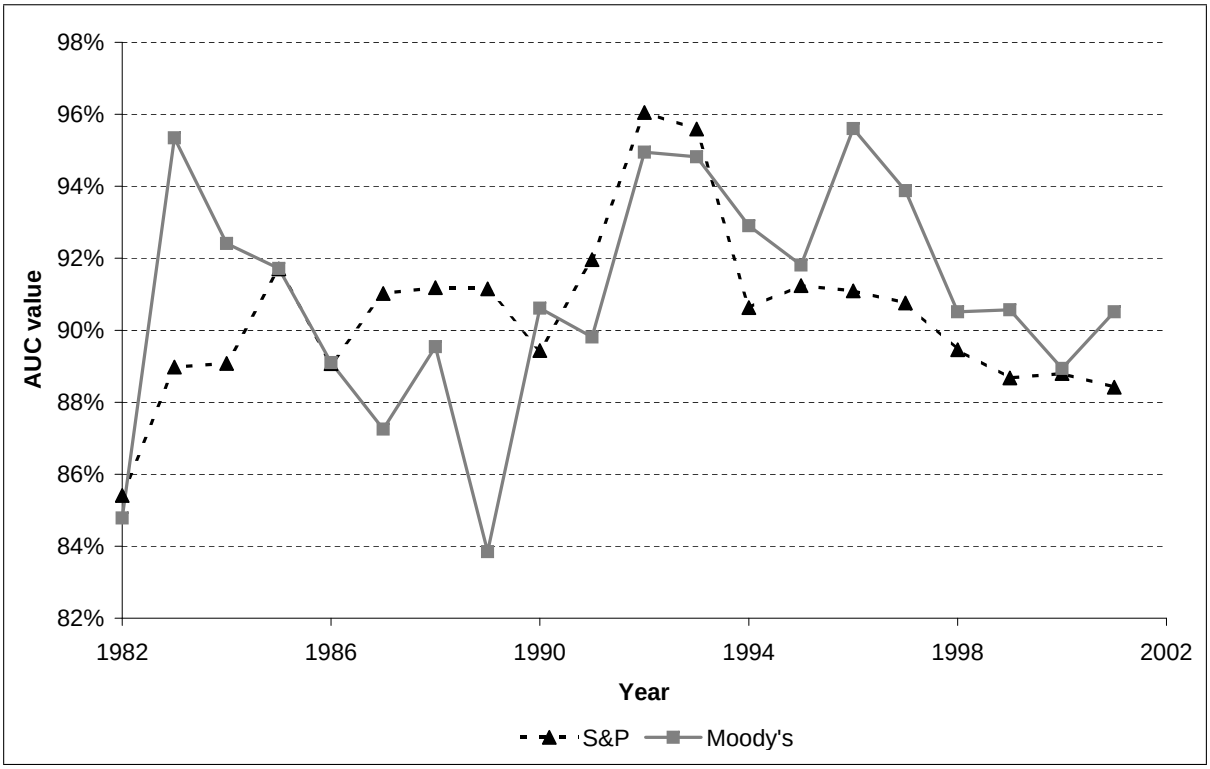
4 Empirical Analysis

In this section, we analyze the predictive power of the rating functions of Standard & Poor's and Moody's for the time period from 1982 to 2001 using contingency tables and seven rating classes for each company. The data set for S&P is obtained from Standard & Poor's (2002) comprising the rating classes AAA, AA, A, BBB, BB, B and CCC. The corresponding data for Moody's is obtained from Moody's Investors Service (2002) for rating classes Aaa, Aa, A, Baa, Ba, B and Caa.

Two approaches are taken to compute the predicted default probability. The first approach uses a five-year moving average of the default rates of the preceding years as a proxy for the predicted default probability in the current year. Therefore, the first M value can be computed for the year 1987. The second, more intuitive, approach uses idealized default probabilities for each rating class as a proxy for the predicted default probability. Idealized default probabilities are computed in various ways often using Monte Carlo techniques, long-term historical data and agency-specific assessments for future developments regarding the specified rating class. We use rating agency-specific idealized one year default probabilities published by Johnston (2009) for the S&P data and obtained from Moody's Investors Service (2006) for the data of Moody's. These idealized default probabilities are fixed for each rating class and each year. This is reasonable because rating classes should be fixed over time so that a yearly comparison of different ratings is possible. When using these idealized default probabilities, the whole time period can be used to measure predictive power.

We start with a presentation of the corresponding AUC values for the two rating functions. These values present just the ranking ability of the rating classes as described in Section 2. Figure 1 presents AUC values for the rating functions of S&P and Moody's for the time period from 1982 to 2001. The AUC values in each year are indicated with a black triangle for S&P and with a grey square for Moody's. Single values are connected with grey and black lines to indicate the development over time. All AUC values are above 80% for both rating functions and all considered points in time, indicating a high ranking

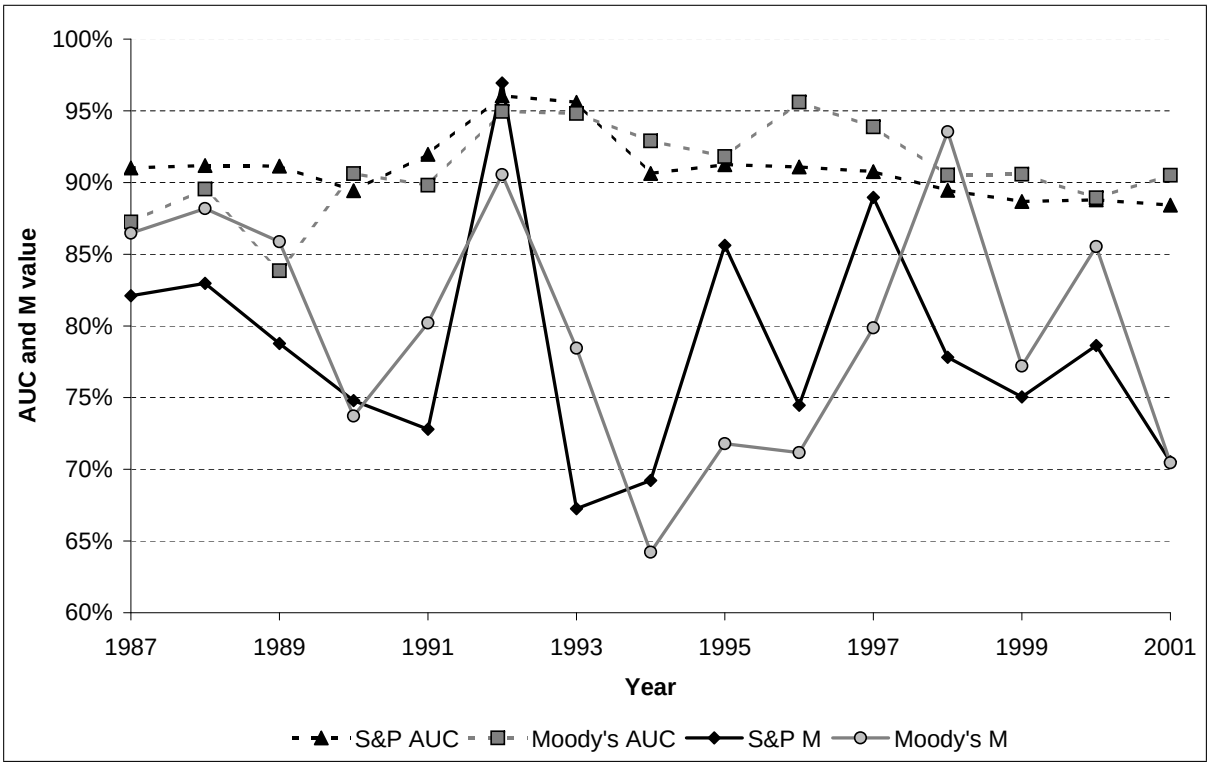
Figure 1: AUC values for S&P and Moody's



power for both agencies. Furthermore, Moody's rating accuracy seems to be more volatile over time and indicates a better ranking ability than S&P when looking at the descriptive statistics below. The average AUC values are 91.03% and 91.04% with corresponding standard deviations of 2.23% and 3.14% for S&P and Moody's, respectively. The maxima of the fluctuation intervals appear small with a span of about 10.7 percentage points for S&P and about 11.8 percentage points for Moody's.

When calculating the corresponding M values using the five-years moving average of the default rates as an estimator for the predicted default probability, a different picture results. In Figure 2 both the AUC values and the M values of S&P's and Moody's rating functions are presented for the time period from 1987 to 2001. Note that the predicted default probabilities may yield predicted numbers of default and non-default that are not integer. However, we used the exact values and did not round when calculating the corresponding hit rates according to Equations (8) and (9). Triangles and squares connected with dotted lines in Figure 2 are AUC values for S&P and Moody's, whereas circles and diamonds connected with solid lines represent M values for Moody's and S&P, respec-

Figure 2: AUC and M values for S&P and Moody's using a five-years moving average of default rates



tively, for each year starting in 1987. M values fluctuate more over time than AUC values. When looking at the descriptive statistics average M values are 78.39% and 79.81% for S&P and Moody's, respectively, indicating lesser predictive power than AUC values do. Standard deviation rises to 7.95% and 8.47% for S&P and Moody's, respectively. Thus, predictive power is more volatile than suggested by AUC. Maxima of fluctuation intervals are larger with a span of 29.3 percentage points for Moody's and 29.7 percentage points for S&P, so the span almost tripled. Another interesting result is the different ranking between the two rating agencies, which is often inverted in comparison to the ranking suggested by AUC values. In ten out of 15 years the ranking of the agencies is different for the M measure compared with the AUC-based ranking.

Maybe more important, we will now evaluate the performance of the rating functions when using the idealized default probabilities (IPD). The results will yield high validity since the predictive power is directly referred to the prediction of number of defaults in each rating class stated by the rating agency. The idealized default probabilities for the rating classes of our data set are presented in Table 4. The IPDs of S&P are smaller than the

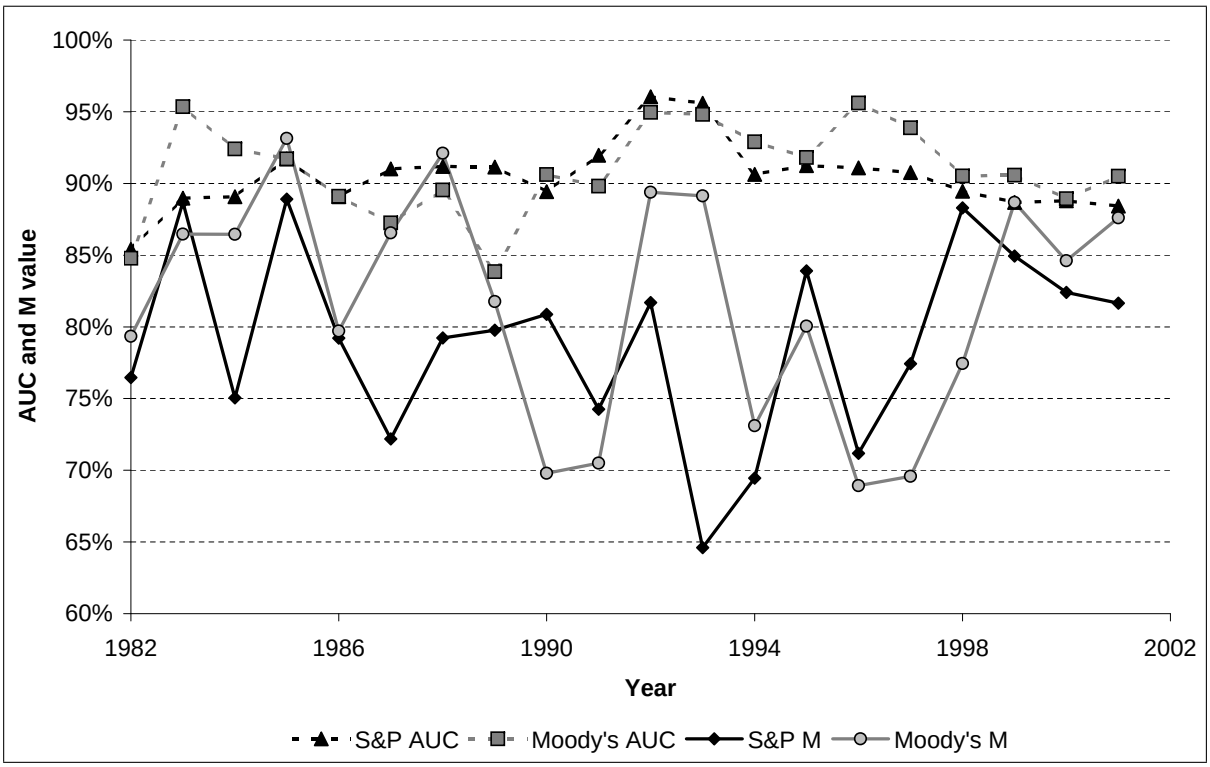
Table 4: Idealized One-Year Default Probabilities for S&P and Moody's

The table shows the idealized one-year default probabilities for the rating agencies S&P and Moody's and the seven rating classes described in the text in percent.

Rating classes of S&P	IPD	Rating classes of Moody's	IPD
AAA	0.0002	Aaa	0.0001
AA	0.0075	Aa	0.0014
A	0.0217	A	0.0109
BBB	0.2943	Baa	0.1700
BB	2.2956	Ba	1.5600
B	5.2946	B	7.1600
CCC	45.5600	Caa	26.0000

ones of Moody's for each rating class except for rating class B. Figure 3 presents the AUC and M values of S&P's and Moody's rating functions for the total time period from 1982 to 2001 using IPDs as predictors. Again, triangles and squares connected by dotted lines

Figure 3: AUC and M values of S&P and Moody's using IPDs



present the AUC values for S&P and Moody's, respectively, whereas circles and diamonds connected by solid lines represent the M values for Moody's and S&P, respectively, for each year. There is still a high fluctuation of M values in comparison to AUC values but the results improve somehow - compared to the case when five-years moving averages of

default rates are used - since the averages increase to 79.01% and 81.72% whereas the standard deviations decrease to 6.57% and 7.93% for S&P and Moody's, respectively. As expected, the maxima of the fluctuation intervals decreases to 24.3 and 24.2 percentage points for S&P and Moody's, respectively, but they are still more than twice as high as the range of the AUC values. Furthermore, the ranking of the two agencies according to M values differs from the ranking according to AUC values. In addition, the relative distance of predictive power between both companies is occasionally larger than implied by the AUC ranking (e.g. in 1993). Overall, the ranking according to the M measure differs in 13 out of 20 years from the AUC-based ranking.

5 Conclusion

The idea of our paper was to develop an adequate validation method for rating functions using the axiomatic approach proposed by Selten (1991). It was exemplarily shown that the common and frequently applied measures AUC and AR fail to measure predictive power within rating classes since they are just ranking measures. To overcome this shortcoming we introduced a new standardized measure that explicitly focuses on predictive success and neglects ranking ability of rating functions. Rating should be an indicator of the default risk a company or credit is exposed to. Our measure takes both defaults and non-defaults into consideration since these realizations are in a dual structure so that the prediction of one dimension has a direct influence on the predictive success of the other dimension.

The axiomatic framework settled for our derivation was introduced by Selten (1991) in a general manner. We adapted his ideas to rating functions and examined the performance of ratings by S&P and Moody's for the period from 1982 to 2001. Observed AUC values are high ($> 90\%$) and nearly stable over the whole period, indicating a good ranking ability. In contrast, our measure indicates a volatile accuracy regarding the predictive power of both rating functions and both prediction approaches using a moving average and idealized default probabilities. The main result obtained is a difference in the ranking

between both agencies over time. Here, AUC and our standardized measure often imply contrary results regarding which of the rating functions is more favorable.

References

Altman, E. I. (1968), 'Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy', *Journal of Finance* **23**, 589–609.

Altman, E. I., Haldemann, R. G. and Narayanan, P. (1977), 'ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations', *Journal of Banking and Finance* **1**, 29–51.

Altman, E. I. and Saunders, A. (1998), 'Credit Risk Measurement: Developments over the Last 20 Years', *Journal of Banking and Finance* **21**, 1721–1742.

Bamber, D. (1975), 'The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph', *Journal of Mathematical Psychology* **12**, 387–415.

Engelmann, B., Hayden, E. and Tasche, D. (2003), 'Measuring the Discriminative Power of Rating Systems'. Deutsche Bundesbank, Working Paper.

Johnston, M. (2009), 'Extending the Basel II Approach to Estimate Capital Requirements for Equity Investments', *Journal of Banking and Finance* **33**, 1177–1185.

Keenan, S. C. and Sobehart, J. R. (1999), 'Performance Measures for Credit Risk Models', *Moody's Risk Management Services* .

Moody's Investors Service (2002), 'Default and Recovery Rates of Corporate Bond Issuers', *Special Comment* .

Moody's Investors Service (2006), 'Mapping of Moody's U.S. Municipal Bond Rating Scale to Moody's Corporate Rating Scale and Assignment of Corporate Equivalent Ratings to Municipal Obligations', *Special Comment* .

Selten, R. (1991), 'Properties of a Measure of Predictive Success', *Mathematical Social Sciences* **21**, 153–167.

Selten, R. and Krischker, S. (1983), Comparison of Two Theories for Characteristic Function Experiments, *in* R. Tietz, ed., 'Aspiration Levels in Bargaining and Economic Decision Making', Springer, Berlin, pp. 259–264.

Standard & Poor's (2002), 'Ratings Performance 2001', *Special Report* .

Otto von Guericke University Magdeburg
Faculty of Economics and Management
P.O. Box 4120 | 39016 Magdeburg | Germany

Tel.: +49 (0) 3 91 / 67-1 85 84
Fax: +49 (0) 3 91 / 67-1 21 20

www.ww.uni-magdeburg.de

ISSN 1615-4274