Orth, Walter

Working Paper

The predictive accuracy of credit ratings: measurement and statistical inference

Discussion Papers in Statistics and Econometrics, No. 2/10

Provided in Cooperation with:
University of Cologne, Institute of Econometrics and Statistics

Suggested Citation: Orth, Walter (2010) : The predictive accuracy of credit ratings: measurement and statistical inference, Discussion Papers in Statistics and Econometrics, No. 2/10, University of Cologne, Seminar of Economic and Social Statistics, Cologne

This Version is available at:
https://hdl.handle.net/10419/45360

DISCUSSION PAPERS IN STATISTICS
AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 2/10

The Predictive Accuracy of Credit Ratings: Measurement and Statistical Inference

by

Walter Orth

This draft: February 16, 2011
First draft: March 22, 2010

DISKUSSIONSBEITRÄGE ZUR
STATISTIK UND ÖKONOMETRIE

SEMINAR FÜR WIRTSCHAFTS- UND SOZIALSTATISTIK
UNIVERSITÄT ZU KÖLN

Albertus-Magnus-Platz, D-50923 Köln, Deutschland

The Predictive Accuracy of Credit Ratings: Measurement and Statistical Inference

Walter Orth*

This draft: February 16, 2011
First draft: March 22, 2010

Abstract

Credit ratings are ordinal predictions for the default risk of an obligor. To evaluate the accuracy of such predictions commonly used measures are the Accuracy Ratio or, equivalently, the Area under the ROC curve. The disadvantage of these measures is that they treat default as a binary variable thereby neglecting the timing of the default events and also not using the full information from censored observations. We present an alternative measure that is related to the Accuracy Ratio but does not suffer from these drawbacks. As a second contribution, we study statistical inference for the Accuracy Ratio and the proposed measure in the case of multiple cohorts of obligors with overlapping lifetimes. We derive methods that use more sample information and lead to more powerful tests than alternatives that filter just the independent part of the dataset. All procedures are illustrated in the empirical section using a dataset of S&P Long Term Credit Ratings.

JEL classification: C41, C52, G17, G24, G32

Keywords: Ratings, predictive accuracy, Accuracy Ratio, Harrell's C, overlapping lifetimes

*Department of Statistics and Econometrics, University of Cologne; walter.orth@uni-koeln.de;
I would like to thank Karl Mosler, Gabriel Frahm, Stephan Popp and participants of the FFM Conference 2010 and the CEQURA Conference 2010 for their helpful commments.

1. Introduction

Ratings are ordinal predictions for the default risk of an obligor. Like in any prediction problem the evaluation of predictive accuracy is an essential constituent. The measure most commonly used by rating agencies, regulators and researchers is the Accuracy Ratio, which is the summary statistic of the so-called Cumulative Accuracy Profile.[1] For the calculation of the Accuracy Ratio it is necessary to choose a fixed time horizon and classify the obligors into two groups, those who defaulted within the chosen time span and those who did not default. However, reducing the data by this kind of classification leads to a loss of information. First, the timing of defaults is neglected. Second, certain right-censored observations have to be omitted. The latter concerns those obligors which are observed - without default - for only a fraction of the chosen prediction horizon. The share of these kind of right-censored observations and with it the loss of information grows with the prediction horizon. In this paper we show how we can extend the methodology of the Accuracy Ratio to include the full information contained in rating datasets. We do so by introducing a so-called concordance index named Harrell's C to the credit risk literature, a measure which has been proposed in the biostatistical literature for the purpose of evaluating predictive accuracy (Harrell et al., 1996). We further propose a modification of Harrell's C that takes the prediction horizon into account and is expected be more suitable for credit risk applications.

Measures of predictive accuracy like the Accuracy Ratio or Harrell's C are, of course, subject to sampling variability. Analyzing this variation is useful not least for confidence intervals and hypothesis tests. For approximately independent data, methods for statistical inference for the Accuracy Ratio and Harrell's C have been established (Bamber, 1975; DeLong et al., 1988; Newson, 2006). However, in typical rating datasets, ratings change over time, which means that we have a time series of default predictions for each obligor. If we want to evaluate the whole time series of default predictions and build cohorts of obligors in time intervals that are shorter than the prediction horizon we get a sample that consists of partially overlapping data which clearly exhibit dependence. To construct a summarizing index under such a multiple cohort sampling scheme, one may take a weighted mean of the indices of the individual cohorts or calculate the index for the pooled cohort by simply aggregating all the individual cohorts to one large cohort (Cantor & Mann, 2003).

[1] Some authors focus instead on the ROC curve and its summary statistic, the Area under the ROC curve. However, the Accuracy Ratio and the Area under the ROC curve contain exactly the same information, since there is a simple linear relationship between the two (Engelmann et al., 2003). The next section describes this relation explicitly.

However, statistical inference for such summarizing indices clearly has to take the dependence of the data into account. We show how this can be done by deriving asymptotic formulae for the weighted mean case and describe how resampling procedures can be used for both the weighted mean and the pooled version. With these procedures, more information is used than by using just a subsample of the data that consists of approximately independent observations. In statistical terms, the benefits are narrower confidence intervals and more powerful tests.

In addition to our theoretical considerations we provide an empirical illustration of the proposed methods using a dataset of Standard & Poor's Long Term Credit Ratings for North American firms. We analyze prediction horizons ranging from 6 to 60 months. As one main finding, we observe illusively high values for the Accuracy Ratio at long horizons which are a direct result from the omission of censored observations. This may lead to an overestimation of the long-run accuracy of ratings by investors and risk managers who use the Accuracy Ratio. In another part of our empirical study, we provide an example where the aforementioned enhanced testing power leads to a difference in the test decision.

The methods proposed in this paper are relevant for the development and validation of default prediction models and rating systems. They are useful for all cases in which the prediction horizon covers more than one sample period, and they are especially beneficial for the evaluation of multi-period default predictions. On the one hand, multi-period predictions are necessary if longer horizons, say multiple years, are of interest - like it is, for instance, in the case of S&P's Long Term Credit Ratings. The Basel Committee for Banking Supervision has also emphasized the importance of a multi-year perspective claiming that "banks are expected to use a longer time horizon [than one year] in assigning ratings" (Basel Committee on Banking Supervision, 2006, § 414). On the other hand, a multi-period set-up is also useful for one-year predictions by allowing the use of all the information in, say, monthly or quarterly data. For instance, it is well known that the analysis of ratings on a yearly basis omits valuable information about intra-year rating transitions (Lando & Skodeberg, 2002).

The remaining part of the paper is organized as follows. In the next section, we give a brief review of the Accuracy Ratio and its theoretical background. Then, we introduce Harrell's C and present an adjusted version in section 3 before we go on with the part on statistical inference for the Accuracy Ratio and (adjusted) Harrell's C in section 4. Section 5 contains the empirical investigation while section 6 concludes.

2. The Accuracy Ratio

To measure predictive accuracy, the approach most popular in the rating industry as well as in the academic world is based either on the Cumulative Accuracy Profile (CAP) and its summary statistic, the Accuracy Ratio (AR), or the Receiver Operating Characteristic (ROC) curve and its summary index, the area under the ROC curve (AUROC).[2] While the construction of the underlying curves differs, the Accuracy Ratio and the AUROC curve contain the same information since there is a simple linear relation between the two (Engelmann et al., 2003):

$$AR = 2 \cdot AUROC - 1 \tag{1}$$

A comprehensive description of the graphical interpretation of these indices and an overview over further measures can be found in Thomas et al. (2002, chap. 7). The Accuracy Ratio and AUROC are designed to measure the discriminative power of a rating system. If default probabilities are assigned to ratings further dimensions of predictive accuracy arise (Krämer & Güttler, 2008). However, we will not pursue this case further here.

In the following, we will focus on the Accuracy Ratio but of course everything extends to the AUROC. Besides its graphical derivation, there is another simple method to calculate the Accuracy Ratio that provides a good intuition about what this index measures. Denote the numerical rating (high values indicate low risk) of the ith defaulting obligor and the jth non-defaulting obligor by X_i^D and X_j^{ND}, respectively. The number of defaulting and non-defaulting obligors in the sample are referred to as n_1 and n_2. Define

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i^D < X_j^{ND}\,, \\ -1 & \text{if} \quad X_i^D > X_j^{ND}\,, \\ 0 & \text{if} \quad X_i^D = X_j^{ND}\,. \end{cases} \tag{2}$$

Then, the Accuracy Ratio is given by

$$AR = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij}\,. \tag{3}$$

We will call c_{ij} the concordance score of the pair of the ith defaulting and the jth non-defaulting obligor. Concordance is given if the rating of the defaulting obligor was worse than the rating of the non-defaulting obligor, while we have discordance

[2] Sometimes, the Accuracy Ratio is also referred to as the Gini coefficient.

in the opposite case. The case of identical ratings is captured by a concordance score of zero. The concordance score is evaluated for every pair of a defaulting and a non-defaulting obligor. Thus, the Accuracy Ratio is the fraction of pairs where the rating was concordant with the outcome minus the fraction of discordant pairs. In line with this interpretation, the corresponding population value - for which the Accuracy Ratio is an unbiased estimator - is

$$P(X_i^D < X_j^{ND}) - P(X_i^D > X_j^{ND})\,, \tag{4}$$

for a randomly selected pair i, j of the population (DeLong et al., 1988).

The Accuracy Ratio is a special case of a more generally defined measure of predictive accuracy called Somers' D (Somers, 1962). The connection is important in our context since the index which will be introduced in the next section, Harrell's C, is also based on Somers' D. Consider predicting a variable Y with a predictor X. The sample size is denoted by n. For ease of exposition, sort the values of Y in ascending order so that $Y_i \leq Y_j$ for $i < j$.[3] Let

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i < X_j\,,\, Y_i < Y_j\,, \\ -1 & \text{if} \quad X_i > X_j\,,\, Y_i < Y_j\,, \\ 0 & \text{else}\,. \end{cases} \tag{5}$$

Then Somers' D is defined as follows:

$$D_{XY} = \frac{1}{n_u} \sum_{i=1}^{n} \sum_{j>i} c_{ij} \tag{6}$$

$$n_u = \sum_{i=1}^{n} \sum_{j>i} 1_{[Y_i \neq Y_j]} \tag{7}$$

The denominator of D_{XY} excludes any ties in Y since in these cases it is not possible to assess a "correct" or "incorrect" order of the predictors. In contrast, ties on X represent a case of mediocre prediction and are subsumed under "else". The Accuracy Ratio is simply the special case with Y being a binary variable representing default. In rating datasets, classification of obligors into defaulters and non-defaulters and thus constructing a binary variable Y leads to the aforementioned information loss. How this loss of information can be avoided will be the topic of the next section.

[3] It does not matter how ties on Y are ordered since pairs with equal values of Y are not usable for Somers' D.

3. Harrell's C

Consider first the following motivating example. At time t, two firms have ratings AA and B, respectively. When the prediction horizon is five years and the AA firm defaults at $t+1$, while the B firm's lifetime is censored at $t+4$,[4] this pair is dropped for the calculation of the Accuracy Ratio although for this pair ratings and outcomes are clearly discordant. In fact, the firm that was rated B at time t has to be dropped completely in the case of the Accuracy Ratio since it can not be classified in either the defaulting or the non-defaulting group. In contrast, Harrell's C uses every observation. In the example given above the corresponding pair would - in line with intuition - receive a concordance score of -1 (with the analogous meaning as explained in section 2).

We will now give the formal definition of Harrell's C and then discuss the various individual cases. Again, X_i is the numerical rating (high values correspond to low risk) of obligor i, $i = 1, \ldots, n$. After being rated X_i, obligor i is observed not to default for a time denoted by Y_i. We will refer to Y_i as the lifetime of obligor i. If the observation is then ended by a default event, the censoring indicator variable C_i is set to zero. If obligor i is no longer observed due to right censoring, the value of C_i is one. Again, it is convenient to sort the lifetimes in ascending order so that $Y_i \leq Y_j$ for $i < j$. As a natural extension of Somers' D to censored data, we then define the concordance score as[5]

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i < X_j \,,\, Y_i < Y_j \,,\, C_i = 0 \,, \\ -1 & \text{if} \quad X_i > X_j \,,\, Y_i < Y_j \,,\, C_i = 0 \,, \\ 0 & \text{else} \,. \end{cases} \tag{8}$$

Then Harrell's C is given by:

$$C = \frac{1}{n_u} \sum_{i=1}^{n} \sum_{j>i} c_{ij} \tag{9}$$

$$n_u = \sum_{i=1}^{n} \sum_{j>i} 1_{[Y_i \neq Y_j \,,\, C_i = 0]} \tag{10}$$

n_u is the number of usable pairs. In words, a pair of observations is usable if a) the obligors' observed lifetimes are not equal and b) the obligor with the shorter

[4] This means that the B firm does not default until period $t+4$, but is no longer observed thereafter.

[5] Harrell et al. (1996) actually normalize the measure between zero and one by assigning concordance scores of 1, $\frac{1}{2}$ and 0 instead of 1,0 and -1 as we do. We stick to the latter version to ensure comparibility with the Accuracy Ratio.

observed lifetime experiences a default event, i.e. the lifetime is not censored. These conditions ensure that for every pair one obligor has indeed "outlived" the other obligor thereby enabling a sensible comparison of both. Given a usable pair, we can distinguish two cases. The first one consists of two obligors, both defaulting but after different time spans. Concordance is achieved if the rating of the obligor with the earlier default event was worse than the rating of the obligor defaulting later, while discordance is given in the opposite case and a concordance score of zero is assigned in the case of equal ratings. In the second case, one obligor defaults after a certain time span and the other obligor's lifetime is censored at a later point in time. For concordance, we require that the defaulting obligor was lower rated. Accordingly, we assign a concordance score of -1 in the opposite case and a score of zero for equal ratings.

Similar to Pencina & D'Agostino (2004) we can derive the population value of Harrell's C as

$$P(X_i < X_j | Y_i < Y_j, C_i = 0) - P(X_i > X_j | Y_i < Y_j, C_i = 0)\,, \tag{11}$$

for two randomly selected individuals i and j from the population. That is, given a pair is found to be usable, Harrell's C estimates the probability of concordance minus the probability of discordance.

A potential source of criticism may be the fact that Harrell's C theoretically covers an unlimited prediction horizon.[6] This is due to the origin of Harrell's C in biostatistics but may not be suitable in credit risk applications since the maturity of most credits is limited. For this reason, we propose the following modification of Harrell's C. Denote the maximum prediction horizon that is of practical interest as H. Similarly to Equation (8) we define

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i < X_j\,,\, Y_i < Y_j\,,\, C_i = 0\,,\, Y_i < H\,, \\ -1 & \text{if} \quad X_i > X_j\,,\, Y_i < Y_j\,,\, C_i = 0\,,\, Y_i < H\,, \\ 0 & \text{else}\,. \end{cases} \tag{12}$$

The adjusted index is then calculated as follows:

$$C_{adj} = \frac{1}{n_u} \sum_{i=1}^{n} \sum_{j>i} c_{ij} \tag{13}$$

$$n_u = \sum_{i=1}^{n} \sum_{j>i} 1_{[Y_i \neq Y_j\,,\, C_i = 0\,,\, Y_i < H]} \tag{14}$$

[6] Practically, the prediction horizon is limited by the sample period.

The rationale of the adjustment is simple. Everything what happens after H is ignored. For instance, with H equal to 3 years, pairs of observations that do not include a default within the first 3 years are now not usable. This corresponds to the fact that we now require for a usable pair that the shorter observed lifetime is ended by a default event that occurs before H. The modification is easy to implement by simply conducting an artificial censoring at H for lifetimes that last longer than H. To be specific, values of Y larger than H are then set to H and their censoring indicator is set to one. Then, the standard formula (8) can be applied.

While the number of unusable pairs grows with this adjustment it is important to note that still no observations have been completely removed. Thus, the amount of information – as measured by the number of usable pairs – included in C_{adj} is still distinctively higher than in the case of the Accuracy Ratio. We can distinguish two kind of pairs that are used for C_{adj} but not for the Accuracy Ratio. The first type covers obligors defaulting at different points in time before H. The second type refers to cases where one obligor defaults at a certain point in time and another obligor whose lifetime is censored at a later point in time but before H.

The interpretation of C_{adj} is still in line with Harrell's C and the Accuracy Ratio. All these measures are bounded between -1 and 1 and yield the proportion of concordant pairs minus the proportion of discordant pairs among all usable pairs. Moreover, Harrell's C has been implemented in various software packages. For instance, it is available in STATA through the user-written *somersd* program by Roger B. Newson and in R it is part of the *Hmisc* package (function *rcorr.cens*).

Harrell's C has the advantages of using the timing of the default events and the information in censored observations. Banasik et al. (1999) provide a detailed discussion why the timing of defaults is relevant from an economic point of view. The authors further advocate a survival analysis approach to default prediction and such models have indeed become very popular over the recent years. Harrell's C as a measure that originates from survival analysis perfectly fits into this framework.

4. Statistical inference

For a single cohort of obligors or more generally for approximately independent observations, a framework for statistical inference with respect to the Accuracy Ratio and Harrell's C has already been established (Bamber, 1975; DeLong et al., 1988; Newson, 2006). In the latter study, the author combines a computationally efficient jackknife procedure with the Delta method. This approach is also applicable

to the adjusted version of Harrell's C presented in section 3 since all our indices are ratios of U-statistics for which the jackknife has been shown to be generally appropriate (Arvesen, 1969). In an application to the credit risk area, Engelmann et al. (2003) present and apply the methods of Bamber (1975) and DeLong et al. (1988) to the Accuracy Ratio and find them to be adequate. However, to the best of our knowledge, there is no study so far that deals with the problem of statistical inference in the multiple cohort case and takes into account the dependence structure of such datasets. The multiple cohort case is relevant because it allows to extract the maximum amount of information out of the dataset. To see this, let us first briefly clarify what is meant by a multiple cohort sampling scheme.[7]

A cohort consists of all obligors that have a rating at a given point in time t. For the members of the cohort, the rating at t and the lifetimes beginning at t (together with the censoring indicators) are recorded. As an example, consider a firm that was rated, say, BB at the beginning of 2009 and defaulted in october 2010. The firm thus enters the cohort that was built at the start of 2009 with its BB rating and a lifetime of 21 months (and a censoring indicator of 0). Now assume that in the beginning of 2010 the same firm was rated CCC. In the cohort built at the start of 2010 the same firm is included again with its CCC rating and a lifetime of 9 months (and again a censoring indicator of 0). The reason why the same firm is included in both cohorts is that we want to evaluate both the performance of the BB rating in the beginning of 2009 and the CCC rating in the beginning of 2010. Note also that the firm would be included in the cohort of 2010 even if the rating would not have changed. Obviously, if we build an aggregated or pooled cohort out of all the individual cohorts the pooled observations are dependent because of partially overlapping lifetimes. In our example, the overlapping period consists of the 9 months in 2010. The overlapping sample problem gets more pronounced if we build cohorts at a higher frequency and if we consider longer prediction horizons. For instance, in the empirical section we will build cohorts on a monthly basis which leads to even larger overlappings than in our example.[8] Due to the strong dependencies in the pooled cohort methods for statistical inference designed for approximately independent samples as the ones mentioned in the beginning of this section are not directly applicable in our setting.

Returning to a more general setup, let us assume that there is a sequence of points in time $t, t = 1, \ldots, T$, and a cohort is built at each t with I_t denoting the chosen index

[7]Sometimes the term "static pool" instead of "cohort" is used.

[8]Moody's is also building cohorts each month in its calculation of Accuracy Ratios (Cantor & Mann, 2003).

of predictive accuracy (e.g., the Accuracy Ratio or C_{adj}) for the cohort built at time t. Given a prediction horizon H, this would correspond to a sample length of $T+H$. As a first issue, one has to decide how to combine the indices $I_t, t = 1, \ldots, T$, to one single measure of predictive accuracy. Cantor & Mann (2003) propose either using some type of weighted mean of I_t or simply calculating the index for the pooled cohort. As weighting schemes, the authors consider equal weights, the number of observations and the number of defaults while finally using the second alternative in their empirical part. We first analyze the weighted mean approach. Consider the following general weighted mean:

$$I = \sum_{t=1}^{T} w_t I_t \tag{15}$$

The weights are normalized to sum up to one. Due to the "overlapping lifetimes problem" sketched above we expect strong autocorrelation of the time series I_t.[9] We assume that $\mathrm{Corr}(I_t, I_{t+j}) = \rho_j$ depends only on j but not on t (assumption 1). This assumption seems reasonable since the main source of dependence between I_t and I_{t+j} is the overlapping fraction of the underlying lifetimes which is equal to $\min(0, 1 - j/H)$, regardless of t. In contrast, the variance of I_t, denoted by σ_t^2, is allowed to vary with t so that we do not assume stationarity. Further, we assume that the dependence of the indices vanishes if the time between the cohort building dates is equal to or larger than the prediction horizon, i.e. $\rho_j = 0$ for $j \geq H$ (assumption 2). In these cases, overlapping lifetimes do not occur anymore. Under these assumptions, the variance of I can be expressed as

$$V(I) = \sum_{t=1}^{T} w_t^2 \sigma_t^2 + 2 \sum_{j=1}^{H-1} \rho_j \sum_{t=1}^{T-j} w_t \sigma_t w_{t+j} \sigma_{t+j} \,. \tag{16}$$

For the derivation of this formula, we have also used the additional assumption that the weights are deterministic. Strictly speaking, from the three types of weights mentioned above, only equal weights are deterministic. However, bootstrap experiments with fixed and varying weights show that this source of variation is negligible. Estimators for σ_t are available for every t by the procedures of Bamber (1975) and Newson (2006). For ρ_j, a natural choice are the empirical autocorrelations, which are consistent estimators of the true autocorrelations and do not require the construction of a time series model. The formula used in the empirical section is

$$\widehat{\rho}_j = \max\left(0, \frac{1/(T-j)\sum_{t=1}^{T-j}(I_t - \overline{I})(I_{t+j} - \overline{I})}{1/T\sum_{t=1}^{T}(I_t - \overline{I})^2}\right) \tag{17}$$

[9] This is confirmed by the empirical analysis with first-order autocorrelations ranging from 0.539 to 0.946.

which cancels out the effect of occasionally occurring negative autocorrelation estimates.

We now turn to the sampling distribution of I. For both the Accuracy Ratio and Harrell's C, the asymptotic normality of I_t directly follows from the fact that both the Accuracy Ratio and Harrell's C represent ratios of U-statistics for which asymptotic normality has been generally established (Hoeffding, 1948). We then assume that the weighted average I converges to the population value $\mu(I)$ (assumption 3). This excludes any trending behaviour. Under these assumptions, we can apply Slutsky's theorem and the Central Limit Theorem for M-dependent random variables[10] to derive the following formula, which can be used for confidence intervals and hypothesis tests:

$$\frac{I - \mu(I)}{\sqrt{\widehat{V}(I)}} \xrightarrow{d} N(0,1) \tag{18}$$

Note that the asymptotics of formula (18) require both the cohort sizes and the number of cohorts approaching infinity. Further, the use of formulas (16) and (18) is not restricted to the Accuracy Ratio and Harrell's C since the only condition for the index I_t was the asymptotic normality in the derivation of (18). Of course, other candidates for I_t may also be asymptotically normal. In order to perform hypothesis tests regarding the difference in the predictive accuracy of two different rating systems, say A and B, we only have to substitute I_t by $(I_{t,A} - I_{t,B})$, σ_t^2 by $\sigma_{t,A-B}^2 = \sigma_{t,A}^2 - 2 \cdot Cov(I_{t,A}, I_{t,B}) + \sigma_{t,B}^2$ and ρ_t by $\rho_{t,A-B}$, the autocorrelation of the time series $(I_{t,A} - I_{t,B})$. The necessary covariances, $Cov(I_{t,A}, I_{t,B})$, can be computed with the methods of DeLong et al. (1988) and Newson (2006). Asymptotic normality of $(I_{t,A} - I_{t,B})$ follows from the joint asymptotic normality of $I_{t,A}$ and $I_{t,B}$.

Alternatively, resampling methods can be used for inference. They are an especially important alternative for datasets with just a few number of cohorts where it is not possible to estimate the autocorrelations for the time series of indices accurately. Jackknife and bootstrap approaches can be applied to both the weighted average and the pooled version of the indices. Clearly, the resampling procedures have to take the dependence structure of the data into account as well. If we interpret all the lifetimes of one obligor as one cluster and assume independence between clusters, i.e. between different obligors, we can apply the cluster versions of the jackknife and the bootstrap. This is achieved by resampling from the set of obligors instead

[10] Our time series of indices is M-dependent in the sense that indices separated by more that M periods are assumed to be independent (see assumption 2), i.e. in our case $M = H - 1$. For details about this kind of Central Limit Theorem see, for instance, Shumway & Stoffer (2006), appendix A.

of the set of all observations which contains several lifetimes for each obligor. Then, the usual jackknife and bootstrap formulas can be applied. For further details, see Shao & Tu (1996), Field & Welsh (2007) and Busing et al. (1999). In a similar application, Cantor et al. (2008) use this kind of bootstrap to calculate confidence intervals for default rates. Default rates are also usually calculated in a multiple cohort setting so that the same rationale applies.

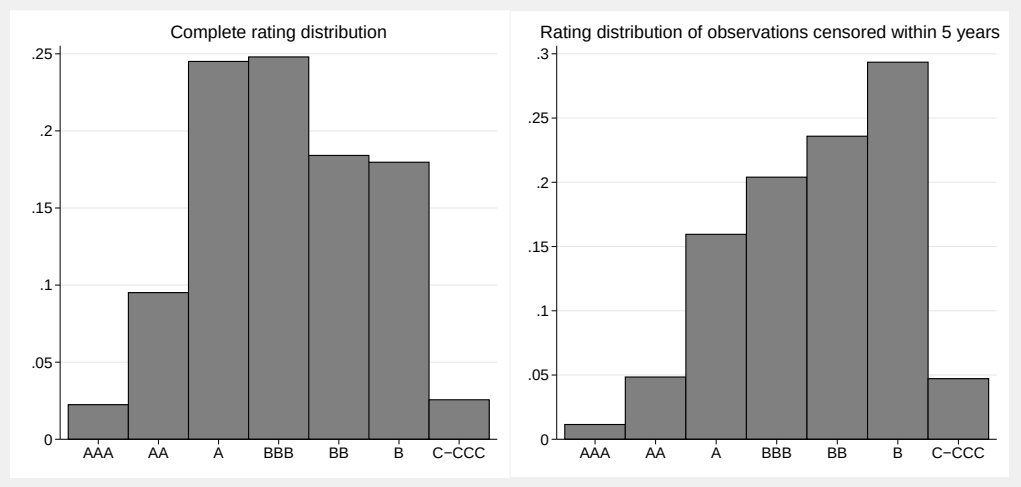
With respect to computational effort, the jackknife is much more efficient than the bootstrap for the pooled indices if the algorithm of Newson (2006) is applied while we observe no major differences for the weighted average indices. Thus, in the upcoming empirical section we will apply the asymptotic formulas derived above and the bootstrap to the weighted average indices while using the jackknife and the bootstrap for the pooled indices.

5. Empirical analysis

Our dataset consists of monthly Standard & Poor's Long Term Issuer Credit Ratings provided by Compustat. Long term ratings are particularly suitable in our context since the benefits from Harrell's C compared to the Accuracy Ratio get most visible in the evaluation of long term predictions. We consider prediction horizons from six months up to five years which is the maximum time horizon of S&P's Long Term Ratings (Standard & Poor's, 2010). After excluding missing observations we have 512 685 firm-months of 5151 North American public firms in the period from december 1985 to june 2009, including 609 defaults. Cohort building is performed on a monthly basis starting in december 1985 until june 2004. Thus, our time series of indices consists of 223 periods. Figure 1 shows the rating distribution of our sample. To investigate the censoring scheme in our data, Figure 1 also shows the rating distribution of the observations censored within five years. Clearly, lower rated firms have higher censoring rates so that the subsample of firms which were not censored within the first five years tends to contain primarily highly rated and defaulting firms. Recall that the five-year Accuracy Ratio uses only this uncensored subsample which obviously has to some degree different characteristics than the whole sample. Apart from this finding, the censoring problem leads to a substantial loss of information as 30.99% of all observations are omitted for the five-year Accuracy Ratio.

We can see the consequences of these problems among other things in Table 1 which gives Accuracy Ratios and adjusted Harrell's C's together with their standard errors and 95% confidence intervals. Looking at the levels of the indices, we see that for

Figure 1: Rating distribution of the full and censored sample



both the weighted average and the pooled versions the Accuracy Ratio declines distinctively less with the prediction horizon than Harrell's C. In particular the five-year Accuracy Ratio is almost as high as the three-year Accuracy Ratio. This is most likely due to the aforementioned fact that, as the prediction horizon grows, the subsample relevant for the Accuracy Ratio tends to consist of "very good" and "very bad" firms making discrimination obviously easier. It follows that the Accuracy Ratios at long horizons indicate a prognostic power of the rating system that is not really existent. Thus, investors and risk managers relying on the Accuracy Ratio are endangered to be too optimistic about the long-run predictive accuracy of ratings. Further, we see that the weighted average measures are generally higher than the pooled measures. This does not surprise as the weighted average indices aggregate measures for predictions made at certain points in time and do not compare ratings from different points of the business cycle as in the pooled cohort approach.

We now turn to the analysis of standard errors and confidence intervals. Regarding the weighted average indices, the asymptotic formulas derived in section 4 tend to be more liberal than the bootstrap results which is a common finding in such comparisons (Horowitz, 2001). While the suggested finite-sample bias of the asymptotic formulas seems to be moderate, the computational effort of the bootstrap might be worthwhile for more precise inference. For the pooled indices, the differences between the cluster jackknife and the cluster bootstrap are very small. Since the jackknife is computationally more efficient, we recommend its use in this situation. Finally, looking at the standard errors and the length of the confidence intervals over time, it is obvious that the uncertainty about rating accuracy grows with the

Table 1: Indices of predictive accuracy, their standard errors and 95% confidence intervals

Panel A: Weighted average indices (number of firms per cohort as weights)

	Adjusted Harrell's C				Accuracy Ratio			
Prediction horizon (months)	6	12	36	60	6	12	36	60
Index	.8686	8340	.7475	.7168	.8725	.8422	.7768	.7679
Formulas (16),(17),(18)								
Standard error	.0087	.0114	.0133	.0144	.0086	.0112	.0130	.0163
CI lower bound	.8515	.8116	.7214	.6886	.8557	.8202	.7514	.7360
CI upper bound	.8856	.8563	.7736	.7450	.8893	.8643	.8022	.7999
Cluster bootstrap								
Standard error	.0105	.0116	.0175	.0204	.0103	.0114	.0173	.0204
CI lower bound	.8436	.8074	.7111	.6715	.8477	.8165	.7406	.7237
CI upper bound	.8831	.8512	.7790	.7509	.8869	.8594	.8077	.8025

Panel B: Pooled indices

	Adjusted Harrell's C				Accuracy Ratio			
Prediction horizon (months)	6	12	36	60	6	12	36	60
Index	.8562	.8116	.7368	.7135	.8599	.8200	.7682	.7660
Cluster jackknife								
Standard error	.0106	.0115	.0141	.0155	.0106	.0114	.0141	.0157
CI lower bound	.8354	.7891	.7091	.6831	.8391	.7977	.7406	.7353
CI upper bound	.8769	.8340	.7645	.7440	.8806	.8424	.7959	.7967
Cluster bootstrap								
Standard error	.0111	.0114	.0144	.0152	.0107	.0113	.0140	.0157
CI lower bound	.8340	.7886	.7077	.6837	.8386	.7971	.7397	.7332
CI upper bound	.8773	.8325	.7640	.7424	.8798	.8413	.7958	.7985

The number of bootstrap replications is 1000. Bootstrap confidence intervals are calculated via the percentile method. Jackknife confidence intervals are calculated using jackknife standard errors and asymptotic normality.

prediction horizon. Note that for a single cohort, the standard errors do not rise with the prediction horizon. However, for the aggregated indices they do, since the overlapping lifetimes problem is more pronounced in this case leading to higher dependencies in the data for longer horizons.

In section 4, we have argued that inference based on multiple cohorts including overlapping lifetimes extracts the maximum amount of information out of the dataset. From a statistical point of view, this leads to smaller standard errors, narrower confidence intervals and more powerful tests. We now demonstrate the latter by example. The following test is motivated by the observation that the information that a firm reached its rating by a downgrade may be useful in predicting future defaults (Lando & Skodeberg, 2002; Guettler & Raupach, 2010). Thus, we created a rating scale that includes new additional grades for downgraded firms. For instance, we classify a firm that reached a BBB− rating by a downgrade between the firms

Table 2: Tests for significant differences in adjusted Harrell's C

	Overlapping lifetimes included				Overlapping lifetimes excluded			
Prediction horizon (months)	6	12	36	60	6	12	36	60
Index	.8686	.8340	.7475	.7168	.8678	.8318	.7466	.6780
Index+	.8695	.8350	.7482	.7173	.8686	.8322	.7467	.6781
Difference	9.92e-4	1.05e-3	7.52e-4	5.59e-4	8.24e-4	4.14e-4	1.04e-4	9.82e-5
St. error of diff.	1.71e-4	1.79e-4	1.29e-4	1.27e-4	3.39e-4	2.38e-4	2.54e-4	2.91e-4
p value	6.06e-9	4.09e-9	5.64e-9	1.06e-5	.0152	.0808	.6834	.7355

The columns under "Overlapping lifetimes included" refer to monthly cohort building. "Overlapping lifetimes excluded" columns use only data from cohorts which are separated by $H-1$ months where H is the prediction horizon. Index refers to Harrell's C in the weighted average version for the S&P fine-grained rating scale as in Table 1. Index+ augments the rating scale by an additional grade for firms who reached their rating grade by a downgrade. The equality of the population indices is tested against the two-sided alternative. Standard errors and p values are based on formulas (16), (17) and (18).

that did not reach BBB− by a downgrade and the firms which are one grade lower, in this case BB+. The null hypothesis of the test presented in Table 2 is that this augmented rating scale has the same predictive power as the original rating scale which is tested against the two-sided alternative. We use Harrell's C in the weighted average version as our measure of predictive accuracy. On the one hand, in the first four columns of Table 2, we perform the test using again monthly cohort building and apply the asymptotic formulas as described in section 4.[11] On the other hand, we do the same test using only cohorts where the time between the cohort building dates is $H-1$ months (H being the prediction horizon) so that no overlapping lifetimes occur.[12] The latter case refers to inference which avoids to deal with the dependence induced by the overlapping lifetimes problem. The standard errors are in this case computed with the elementary part of formula (16) that does not include the terms which involve autocorrelations.

The results show that we can reject the null hypothesis at any horizon and at any conventional significance level if we include overlapping lifetimes. We conclude that the consideration of the downgrade effect indeed yields incremental predictive accuracy. However, such a conclusion is hardly possible without the use of overlapping lifetimes. In this case, we observe only marginally significant improvements at short horizons and no significant differences for longer horizons. One reason for this caused by random is that the point estimates for the difference of the indices are lower throughout all horizons. The other and systematic reason is that the stan-

[11] The results are robust to the alternative use of the cluster bootstrap.

[12] For instance, for five-year Harrell's C, we use the cohorts build in june of 2004, 1999, 1994 and 1989.

dard errors of the differences are higher – on average by 89% – reflecting the higher variability that is caused by the reduction of the dataset. To conclude, we see that there are realistic examples where the greater power of tests based on overlapping lifetimes results in different decisions.

6. Conclusions

In analyzing measures for the predictive accuracy of rating systems, this paper contributes to the existing literature mainly in two aspects. First, we propose a measure from the biostatistical literature, Harrell's C, as an improvement to the Accuracy Ratio and show how it can be modified for limited prediction horizons that are typical in credit risk applications. The main advantage of Harrell's C is that it uses the full information of the sample. Instead, the Accuracy Ratio omits certain censored observations which can, additionally, lead to a remaining subsample that is not fully representative. The empirical part shows that this is not just a theoretical obstacle and may result in an overestimation of rating accuracy in the long run. Second, we analyze the problem of statistical inference in rating datasets where overlapping lifetimes lead to dependence in the data that can not be ignored. For this purpose, we derive asymptotic formulas and describe how resampling procedures can be used appropriately. These multiple cohort procedures deliver narrower confidence intervals and more powerful tests than naive alternatives that use approximately independent data only. We show in the empirical analysis that the enhanced power is substantial and leads to different decisions in a simple but practically relevant example.

References

Arvesen, J. N. (1969). Jackknifing U-statistics. *Annals of Mathematical Statistics*, 40 (6), 2076–2100.

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.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but When Borrowers Default. *Journal of the Operational Research Society*, 50, 12, 1185–1190.

Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards – A Revised Framework*.

Busing, F. M. T. A., Meijer, E., & van der Leeden, R. (1999). Delete-m Jackknife for Unequal m. *Statistics and Computing*, 9, 3–8.

Cantor, R., Hamilton, D. T., & Tennant, J. (2008). Confidence Intervals for Corporate Default Rates. *Risk*, March 2008, 93–98.

Cantor, R. & Mann, C. (2003). *Measuring the Performance of Corporate Bond Ratings*. Special comment, Moody's Investors Service.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3), 837–845.

Engelmann, B., Hayden, E., & Tasche, D. (2003). Testing Rating Accuracy. *Risk*, January 2003, 82–86.

Field, C. A. & Welsh, A. H. (2007). Bootstrapping Clustered Data. *Journal of the Royal Statistical Society Series B*, 69, 369–390.

Guettler, A. & Raupach, P. (2010). The Impact of Downward Rating Momentum. *Journal of Financial Services Research*, 37, 1–23.

Harrell, F. E. J., Lee, K. L., & Mark, D. B. (1996). Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 15, 361–387.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics*, 19, 293–325.

Horowitz, J. L. (2001). The Bootstrap. *Handbook of Econometrics*, 5, 3159–3228.

Krämer, W. & Güttler, A. (2008). On Comparing the Accuracy of Default Predictions in the Rating Industry. *Empirical Economics*, 34, 343–356.

Lando, D. & Skodeberg, T. M. (2002). Analyzing Rating Transitions and Rating Drift with Continous Observations. *Journal of Banking & Finance*, 26, 423–444.

Newson, R. (2006). Confidence Intervals for Rank Statistics: Somers' D and Extensions. *The Stata Journal*, 6(3), 309–334.

Pencina, M. J. & D'Agostino, R. B. (2004). Overall C as a Measure of Discrimination in Survival Analysis: Model Specific Population Value and Confidence Intervals. *Statistics in Medicine*, 23, 2109–2123.

Shao, J. & Tu, D. (1996). *The Jackknife and Bootstrap*. Springer.

Shumway, R. H. & Stoffer, D. S. (2006). *Time Series Analysis and Its Applications*. Springer, 2nd edition.

Somers, R. H. (1962). A New Asymmetric Measure of Association for Ordinal Variables. *American Sociological Review*, 27(6), 799–811.

Standard & Poor's (2010). *The Time Dimension of Standard & Poor's Credit Ratings*. Global Credit Portal, September 22, 2010.

Thomas, L. C., Edelman, D. B., & Crook, J. B. (2002). *Credit Scoring and its Applications*. SIAM.