“Testing the consistency of internal credit ratings”

AUTHORS	Edward H.K. Ng
ARTICLE INFO	Edward H.K. Ng (2012). Testing the consistency of internal credit ratings. *Banks and Bank Systems*, *7*(3)
RELEASED ON	Friday, 19 October 2012
JOURNAL	"Banks and Bank Systems"
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES **0**

NUMBER OF FIGURES **0**

NUMBER OF TABLES **0**

Edward H.K. Ng (Singapore)

Testing the consistency of internal credit ratings

Abstract

The push for banks to develop their own internal rating based systems under the New Basel Capital Accord highlights the need for credit ratings consistency. As pointed out in Carey (2001), ratings inconsistency across banks can lead to a host of problems and banks are aware of the need for such consistency as found in a survey by Treacy and Carey (2000). Unlike testing for accuracy (see Lopez and Saidenberg, 2000), validating consistency usually requires cross-sectional data that are rarely available to individual banks. In many economic regimes, there is no corporate credit rating or credit bureau to speak of which leaves a bank to depend on its own internal data to verify consistency. Business strategy or other choices can lead a bank to an internal database composition significantly different from that of the underlying population or its peers. This may result in ratings for an obligor inconsistent with those of other credit institutions. This paper proposes a consistency test that does not require cross-sectional data. When applied to logistic regression modeling of corporate credit, the results suggest that ratings consistency is sensitive to the default rate in the data sample employed which reaffirms the need for attention in this area.

Keywords: banking, credit risk, credit rating, internal rating based system, consistency.
JEL Classifications: G21, G33.

Introduction

The internal rating based (IRB) approach strongly advocated for the New Basel Capital Accord (or Basel II as it is widely known) brings to the fore the issue of credit ratings consistency. As Carey (2001) explains, ratings inconsistency across banks can lead to a host of problems that can diminish the intended effects of the Accord and he has found such inconsistencies using a proprietary dataset. In many economic regimes, however, it is not even possible to conduct the type of analysis that Carey has done. Outside of the OECD, few countries have created or maintained any national-level databases, whether public or private, for credit risk modeling. With few exceptions, corporates are not rated and financial data on unlisted firms are not available unless purchased at a high unit cost from the official business registry. Credit bureaus are non-existent or are just regulatory agencies created to maintain records of defaults and little else. What these mean is that unlike OECD banks which can validate their ratings against that done by professional agencies (for corporate credit at least), individual banks in these countries are left with only their own data to test the consistency of their internal credit ratings.

Treacy and Carey (2000) find that banks are aware of the need for rating consistency but as pointed out in Altman and Saunders (2001), testing for it is a challenge. While Lopez and Saidenberg (2001) have proposed an approach to comparing the accuracy of rating models, there is yet no test for consistency. This paper aims to help fill the void with a proposed consistency test that does not require the use of external data for validation. While not comprehensive in purpose, the test does allow a bank to ascertain if it has an internal data sample that can lead to credit ratings that are unstable and potentially consistent with that of its peers. Following this introduction, section 1 briefly discusses issues of consistency pertaining to a bank's internal credit rating. Section 2 explains the test proposed and section 3 reports some results from applying this test to eleven years of North American corporate financial ratios. The results suggest that ratings may be less inconsistent using data samples with low default rates. The final section concludes the paper.

1. Data sample composition and ratings consistency

Compared with accuracy, consistency is not given much attention in statistical methods. There is no consensus on the definition of the construct and it is usually discussed in conjunction with robustness.

With the push for banks to have IRB systems, the need for consistency comes into a much sharper focus. The probability of default (PD) is a good metric for comparing credit ratings between banks as reasoned in Carey (2001) and perfect consistency would mean that every bank derives exactly the same PD for each credit entity. Such perfection is an ideal rather than a reality but it provides a point of reference for measuring inconsistency. A wide PD estimates dispersion will suggest ratings inconsistency among banks but there is still a need in empirical analysis to reject the null of no inconsistency with known confidence.

Poor modeling skill is a potential cause of inconsistency but this can be quite easily rectified by the engagement of qualified professionals. Another potential cause that is more insidious and not well understood is how data sample composition may affect rating consistency.

The paper has benefited from comments at seminars in the National University of Singapore and Singapore Management University.

In a typical modeling process, sample data are carefully selected to reflect the true underlying distribution. For most banks, two factors can lead to unintended deviation from this principle. First, a bank is likely to have chosen a market niche which naturally leads to a limited range of credit customer profiles. A bank focusing on industry *K* will have a higher than average concentration of its database from that industry. In a business cycle unfavorable to industry *K*, a typical defaulting obligor will have the profile of an industry *K* firm. Unless a bank can supplement its internal data with external ones to constitute a sample reflective of the population, all estimated PDs will be influenced by determinants from industry *K*.

Second, the proportions of corporate default or default rates are generally low in most developed economies. This rate for North American listed non-financial firms is less than 1% each year in the 1990's. In credit risk modeling, default serves as the target outcome for deriving a model and for some statistical methods, the paucity of observations poseses a problem. Logistic regression, for instance, starts with the sample odds of a default to arrive at the credit risk model. A low default rate in the modeling sample can limit the power of this method. It is not uncommon then for bank credit rating modelers to select a sample "loaded" with defaults to overcome the problem. Frydman, Altman and Kao (1985), for instance, employed a sample of 200 New York Stock Exchange listed firms comprising 58 (29%) that have gone bankrupt to demonstrate the effectiveness of decision trees in predicting corporate bankruptcy. While that paper does not address credit rating, it is a short step to use the same approach to derive PDs and map them to a rating structure[1]. Aside from accuracy in bankruptcy prediction, would credit ratings derived by a sample using 29% default be consistent with that using the population default rate of less than 1%? Looking ahead, simulation results reported in section 4 suggest that it is unlikely to be so.

2. A consistent credit rating order

Many borrowers or obligors are common across banks. Internal credit ratings consistency would generally mean that ratings for such an obligor should be similar across lending banks. The question then is how dissimilar must ratings be for one to reject consistency.

Consider two obligors *X* and *Y*. If estimates of their PDs are exactly the same everywhere, it is can be concluded that credit ratings are consistent, even if consistently wrong. Estimated PDs for a common obligor are, however, unlikely to be the same across banks unless the same data values and methodology are used. Differences can be random and trivial so there is a need to separate such from more systematic disparity which is evidence of inconsistency.

Let us denote the PD for *X* as PD_X. If $PDdiff = (PD_X - PD_Y)$ is always of the same sign with a narrow dispersion, rating consistency is unlikely to be a concern. It is when the *PDdiffs* are large that matters and Carey (2001) has found evidence of some being so. Still, it is not possible to conclude inconsistency using *PDdiff* alone.

Two *PDdiff*s of different signs from two sources of estimations would be stronger evidence. This means that one source has concluded that *X* is more credit risky than *Y* and the other source the reverse. If such occurrences are few, it could be attributed to chance but pervasiveness in such a pattern will be evidence of inconsistency. Finding the credit risk order of *X* and *Y* being reversed may not amount to much but if one bank rates *W*, *X*, *Y* and *Z* in descending order of PD and another arrives at the opposite order, the consistency of ratings across the two banks should be a cause for concern. It is this pervasive difference, represented by the number of reversals, that the proposed distribution-free test measures.

2.1. Counting rank order reversals. A pairwise comparison of credit ratings by each bank for each common obligor can be performed as in Carey (2001) to evaluate rating consistency. For any bank, this requires the availability of another bank's rating of the same obligor for the same credit facility at about the same time. This is an unrealistic possibility for most banks but the underlying principle of comparing credit ratings can still be applied to internal data alone. As it is probably the most uniform metric for credit rating, I will use PD as the measure for the rest of this paper.

As explained above, a persistent change in the sign of *PDdiff* indicates a lack of consistency in one obligor being rated as more (or less) credit risky than another. Determining if a change is persistent or statistically significant can be done with a sufficiently large dataset.

A bank can draw a first sample from its internal credit database and develop a credit rating model using a chosen statistical method. With replacement, it can then draw a second independent sample and derive the credit ratings again. If credit ratings so derived are consistent, one would expect that *PDdiff* for any pair of obligors common to both samples to have the same sign. This null hypothesis can be rejected if the sign changes for a significant number of pairs. The number of sign changes can be enumerated using rank orders.

[1] The use of decision trees to develop an IRB system is not uncommon.

Assume that m unique obligors or observations are common to both credit risk modeling samples of sizes much larger than m. These m observations are assigned ranks 1 to m by their estimated PDs derived in the first sample. They are also ranked likewise in the second sample. Whether the ranking is in ascending or descending order is immaterial but the merit must be maintained in both samples.

Since each observation is unique and hence should have a different PD, there will be no tied ranks[1].

Let i_{xs} denote rank i of observation X in sample s. For each pair of observations X, Y where $i_{xs} < j_{ys}$, let

$$a_{xy} = 1, \text{ if } i_{xt} > j_{yt} \quad 0, \text{ otherwise.} \tag{1}$$

Equation (1) is essentially a count of the times the rank order between two observations in sample s is reversed in sample t. Given the design, this means that if X is estimated to be less credit risky than Y in sample s, it is the reverse in sample t. This satisfies a broad definition of inconsistency. Here, the magnitude of reversal is not taken into consideration. For instance, if X and Y were ranked 1 and 4 respectively in sample s and then 33 and 2 respectively in sample t, this rank order reversal would be counted as no different from observing rank 5 for X and 2 for Y in the latter sample. The magnitude of reversal can, however, be taken into consideration but deriving the distributional properties for this refinement is beyond the scope of this paper.

If we take the first-ranked observation and compare it to all observations of lower rank, there will be $m-1$ rank order comparisons. Repeating this for the second to the second-last ranked observation, there will be $\frac{m(m\text{-}1)}{2}$ comparisons for all m observations.

Let A_k be defined as

$$A_k = \sum_{j=i+1}^{m} a_{ij} \text{ for the } kth \text{ possible outcome.} \tag{2}$$

If the rank orders in the second sample turns out exactly as in the first sample, $A_k = 0$ since each $a_{ij} = 0$. Conversely, if all the ranks are exactly inverted with the last-ranked observation now being ranked first and so on, $A_k = \frac{m(m\text{-}1)}{2}$ as each of the $\frac{m(m\text{-}1)}{2}$ comparisons yield a value of 1. There are, hence, $\frac{m(m\text{-}1)}{2} + 1$ possible values for A_k ranging from 0 to $\frac{m(m\text{-}1)}{2}$.

If we define k' as the numerical value outcome of A_k and $c_{k'}$ as the counts for k', we can show that

$$\sum_{k'=0}^{m(m-1)/2} c_{k'} = m! \tag{3}$$

Alternatively, we can enumerate the number of different rankings possible for m observations, which again is $m!$. Since $m! > \frac{m(m+1)}{2}$, many values of k' will be repeated.

To illustrate the above process, a simple example of $m = 4$ is used and the results are displayed in Table 1.

Table 1. Illustration of the complete distribution of rank order reversal counts

	Sample	Panel A				Panel B						Panel C
	k	Observations				a_{ij}						$A_k = \sum_{j=i+1}^{4} a_{ij}$
		1	2	3	4	1,2	1,3	1,4	2,3	2,4	3,4	
Ranking in the initial sample		1	2	3	4							
Rankings in the alternative sample k	1	1	2	3	4	0	0	0	0	0	0	0
	2	1	2	4	3	0	0	0	0	0	1	1
	3	1	3	2	4	0	0	0	1	0	0	1
	4	1	3	4	2	0	0	0	0	1	1	2
	5	1	4	2	3	0	0	0	1	1	0	2
	6	1	4	3	2	0	0	0	1	1	1	3
	7	2	1	3	4	1	0	0	0	0	0	1
	8	2	1	4	3	1	0	0	0	0	1	2
	9	2	3	1	4	0	1	1	1	0	0	3
	10	2	3	4	1	0	0	0	0	1	1	2
	11	2	4	1	3	0	1	1	1	1	0	4
	12	2	4	3	1	0	0	0	1	1	1	3

[1] The precision of the PD can always be calibrated to ensure this unless two obligors have exactly the same variable values.

Table 1 (cont.). Illustration of the complete distribution of rank order reversal counts

	Sample	Panel A				Panel B						Panel C
	k	Observations				a_{ij}						$A_k = \sum_{j=i+1}^{4} a_{ij}$
	13	3	1	2	4	1	1	1	0	0	0	3
	14	3	1	4	2	1	0	0	0	0	1	2
	15	3	2	1	4	1	1	1	1	0	0	4
Rankings in the alternative sample k	16	3	2	4	1	1	0	0	0	1	1	3
	17	3	4	1	2	0	1	1	1	1	0	4
	18	3	4	2	1	0	1	1	1	1	1	5
	19	4	1	2	3	1	1	1	0	0	0	3
	20	4	1	3	2	1	1	1	0	0	1	4
	21	4	2	1	3	1	1	1	1	0	0	4
	22	4	2	3	1	1	1	1	0	1	1	5
	23	4	3	1	2	1	1	1	1	1	0	5
	24	4	3	2	1	1	1	1	1	1	1	6

Notes: For four unique observations 1, 2, 3, 4, there are 4! or 24 different possible rank orders. Panel A shows the 24 ways the 4 observations can be ranked in Sample 2 compared to their ranks in Sample 1. In Panel B, $a_{ij} = 1$ if $q_{j2} < p_{i2}$ and $q_{j1} > p_{i1}$ and 0 otherwise, where p_{ik} is the rank p of observation i in sample k. Panel C shows the sum of the binary outcomes across the six pairs of comparisons.

As shown in the table, we will have $\frac{4(4-1)}{2} = 6$ rank order comparisons with four unique observations labeled 1, 2, 3 and 4. In Panel A, we see that there are 4! = 24 possible rankings in the alternate sample. Panel B shows the binary outcomes 0 or 1 for each of the 6 rank order comparisons. In the alternative sample $k = 3$, for instance, the rank orders of observations 2 and 3 are reversed. Observation 2 is now ranked number 3 and observation 3 is ranked number 2. Regardless of the statistical method employed, this reversal can be regarded as rating inconsistency.

In Panel C, the outcomes are summed across the 6 comparisons. With $k = 1$, this sum $A_1 = 0$ since all rank orders remain unchanged. With $k = 24$, the sum $A_{24} = 6$ as all rank orders are reversed. Between these two samples, A_k ranges from 1 to 5 with repetitions.

It is a short step from here to devise a test against the null hypothesis of no inconsistency. A large A_k indicates more rank order reversals which means that relative credit riskiness between two obligors has not been maintained. From Table 1, we can count the number of times k' is 0, 1, 2, 3, 4, 5 or 6. The number of occurrences is 1, 3, 5, 6, 5, 3, and 1 respectively which forms a symmetric distribution.

To complete the illustration, we can use the number of occurrences to form a cumulative probability distribution as is done in most nonparametric tests. The probability of observing $k' = 6$ is $\frac{1}{24}$ or 0.0417. If we use a one-tail test at 95% confidence, we will need to detect 6 rank order reversals for a sample of 4 observations to reject the null hypothesis of consistency.

2.2. Tabulation of cumulative probabilities. The number of possible rank order outcomes for an alternative sample increases dramatically with the sample size. With 4 observations, it is 24 possible outcomes. With 50 observations, the number jumps to 50! or 3.04141 x 10^{64}. Even with 20 observations, the number is 2.4329 x 10^{18}. Such large numbers are unwieldy as even the counts for each k' or $c_{k'}$ can exceed ten digits. For practical purposes, therefore, dealing with percentages is more convenient.

For any value Z between 0 and $\frac{m(m\text{-}1)}{2}$, we can derive the percentage F such that

$$F = \frac{\sum_{k'=0}^{Z} c_{k'}}{m!}. \quad (4)$$

Equation (4) re-expresses the cumulative counts of all possible outcomes from 0 to Z into a proportion which is equivalent to a cumulative probability. We can test for the null hypothesis of consistency at F% confidence level and reject it if A_k falls within the $(1-F)$% critical region.

I have derived the values Z and F for 3 to 50 observations. As the sample size increases, the probability of observation $Z > 0.5(\frac{m(m\text{-}1)}{2} + 1)$ approaches zero. For a sample of 50 observations, there are 1226 possible values for Z from 0 to 1225. For the sake of brevity, the number of rank order reversals necessary to reject the null of consistency at 90%, 95% and 99% confidence levels for sample sizes 6 to 50[1] are shown in Table 2.

[1] The complete distribution is available from the author on request.

Table 2. Minimum counts of rank reversals needed to reject the null of no inconsistency in credit ratings between two equal samples of 6 to 50 observations at 90%, 95% and 99% confidence levels

No. of observations	90% level	95% level	99% level
6	11	12	13
7	15	16	18
8	19	21	23
9	24	26	29
10	30	32	35
11	36	38	42
12	42	45	50
13	50	52	58
14	57	61	66
15	66	69	76
16	74	78	85
17	84	88	96
18	94	98	107
19	104	109	118
20	115	120	130
21	126	132	143
22	138	145	156
23	151	158	170
24	164	171	184
25	178	185	199
26	192	200	215
27	206	215	231
28	222	231	247
29	237	247	264
30	254	264	282
31	270	281	300
32	288	299	319
33	305	317	339
34	324	336	358
35	343	355	379
36	362	375	400
37	382	396	421
38	403	417	443
39	424	439	466
40	445	461	489
41	467	483	513
42	490	506	537
43	513	530	562
44	537	554	587
45	561	579	613
46	585	604	640
47	611	630	677
48	636	657	694
49	663	683	722
50	689	711	751

Note: For a sample of 50 observations, it takes 711 or slightly less than 58% of rank orders to be reversed to reject the null at 5% significance level.

3. Sample composition and rating consistency

The test proposed can be applied to evaluate rating consistency, whether as a diagnosis of statistical method employed or the composition of the data sample used. A bank can easily compare the ratings derived using different methodologies to determine if the results obtained are consistent. Any modeling inadequacy, however, is not a real concern as expertise can be engaged to overcome the problem. The possibility of data sample composition affecting rating consistency on the other hand can pose a challenge that has yet to be assessed.

Can credit ratings in one bank be inconsistent with that of another simply because of differences in their credit databases or the selection of a sample to derive a credit rating model? It is this question that the rest of the section tries to provide an answer to and the findings suggest that sample composition affects consistency.

According to Quantity Impact Study or QIS 3 published by the Basel Committee on Banking Supervision in May 2003, about 3% of participating banks' corporate portfolio is in default. Though economically significant, this fraction is quite low from a statistical modeling perspective. More importantly, this percentage may still be higher than some corporate default rates. Using the Compustat database, I computed the fraction of non-financial firms filing for Chapter 11 each year from 1991 to 2000. The highest fraction is 0.80% which occurred in 1997. The average for the decade was 0.46%. For a bank with a corporate loan portfolio resembling the composition of the Compustat database, there will only be about 5 bankruptcies for every 1,000 customers. Can such a low proportion of target outcomes affect the consistency of PD estimation? An application of the consistency test proposed indicates that it does.

3.1. Estimating PD. Following the process described, the first step in formulating the test is to estimate PDs for each firm. A common approach is to apply a logistic regression or logit on financial ratios. Standard and Poor's propose a Criteria of Financial Soundness list comprising twenty-five such ratios. Martin (1977) advocates logit for modeling credit risk as it is designed for binary outcomes and produces a PD that can be directly derived from the model. This statistical method has been subsequently applied in several published studies on credit risk modeling.

Using 1991 to 2001 corporate financial data from the Compustat Industrial database, I selected a sample of unique non-financial firms listed in North America. For a firm that has survived the entire sample period, only the financial data in the first year, that is 1991, are used. This is to maximize power in the statistical model since consecutive year financial ratios of the same firm may be not very different.

Firms that have filed for Chapter 11 or gone bankrupt in each year are identified[1]. Financial ratios are derived according to the Standard and Poor's Criteria of Fi-

[1] While firms are also delisted for other reasons, Chapter 11 is the only code in the Compustat database that is clearly associated with bankruptcy.

nancial Soundness to minimize any potential bias arising from partiality to the choice of model variables. Since bankruptcy has to be a result of a failure to repay debts, it is reasonable to equate it with default.

For a defaulting firm, financial ratios of the year prior to bankruptcy are used. As with most other studies like Frydman et al. (1985), for instance, the objective is to derive a model of financial ratios that can predict bankruptcy at least one year ahead.

After filtering those with incomplete data, the final sample comprises 7,413 unique firms of which 254 have gone bankrupt. This offers a bankruptcy proportion of nearly 3.43% which is close to average corporate default rate of 3.00% reported in QIS 3. But this proportion masks the underlying annual default rate of about 0.50% throughout the sample period[1].

Bankrupt and surviving firms are divided into two subpopulations from which random draws are made to form samples of 1,000 firms. For the first iteration, a sample is created with exactly 3% bankruptcy. This is done by randomly selecting 30 bankrupt firms from the pool of 254 available and 970 surviving firms from the 7,413 available. The process is repeated for 100,000 samples. For easy reference, sample p is denoted $S_{0.03,p}$ where the first subscript corresponds to 3% bankruptcy and p ranges from 1 to 100,000.

A second iteration creates another 100,000 samples but each with 20% bankruptcy. Such a proportion may seem unrealistically high relative to actual default rates but it is still lower than 29% in the data sample used by Frydman et al. (1985) for the pioneering work on employing decision trees to derive credit rating models. The samples are denoted $S_{0.20,p}$ here and p is again from 1 to 100,000.

Using the twenty-five Criteria of Financial Soundness financial ratios[2], logit is applied to each sample to derive a model predicting bankruptcy. There is no attempt to obtain a parsimonious model for each sample but rather to allow the models complete variability in the coefficient estimates of the twenty-five ratios. The objective is to hold constant the modeling approach and statistical method and examine the effect of data sample composition on the resultant *PD* estimates.

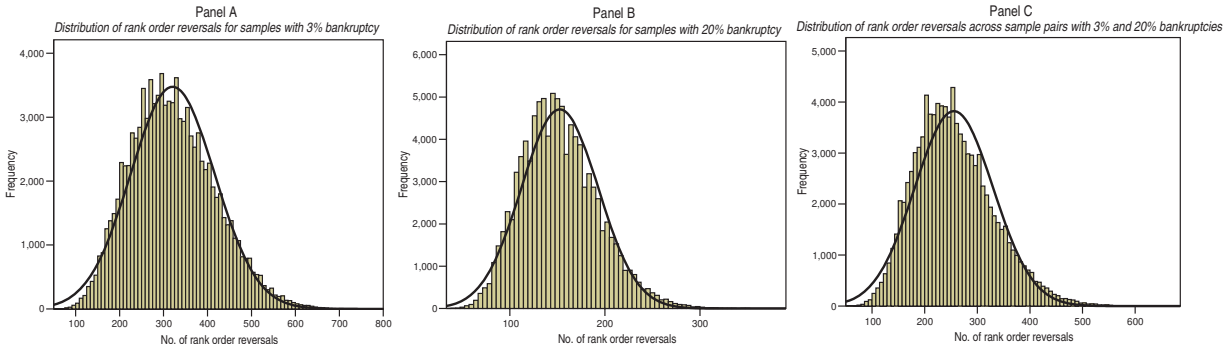
Logit allows us to compute a probability of the target event occurring which is bankruptcy in this case. This probability or *PD* can be expressed as

$$PD = [1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_k X_k)]^{-1}, \quad (5)$$

where the expression in the parentheses is the negative of the regression equation. For each sample, the *PD* estimates for each firm in the sample will be generated by the same logit regression coefficients. A higher PD should imply greater credit risk but for any firm that appears in more than one sample, *PDs* generated may differ across samples. Firm *X* having a larger *PD* in one sample but a smaller *PD* in another sample than firm *Y* would amount to inconsistency as defined in the test.

3.2. Ranking consistency. To test for inconsistency, 50 firms common to sample pairs are randomly selected. For the 100,000 samples with 3% bankruptcy, samples are simply paired in sequence with $S_{0.03,p}$ and $S_{0.03,p+1}$ making a pair. This provides 99,999 pairs. The same is done for the 100,000 samples with 20% bankruptcies.

The 50 firms are ranked in descending order of their PDs in both samples in a pair. According to test specifications, firm *X* ranking above firm *Y* in one sample but below it in the other is a rank order reversal which is evidence of inconsistency in credit ratings across the two samples in the pair. Rank order reversals are exhaustively enumerated for all 50 firms which require 50! comparisons.



Note: Distribution of PD rank order reversals across two data samples. A rank order reversal occurs when the PD rank of firm *X* is above *Y* in one sample but below it in a paired sample. Panel A shows the distribution for all samples with 3% bankruptcy. Panel B shows the distribution for all samples with 20% bankruptcy. Panel C shows the distribution for reversals from one sample with 3% bankruptcy to another with 20% bankruptcy.

Fig. 1. Histogram of the number of rank order reversals

[1] The upward bias in ratio is natural since the bankrupt ones are pooled across the years resulting in a number that is far more than the bankruptcies occurring each year.

[2] This number is actually slightly fewer than the complete set of ratios listed by Standard and Poor's as some of the ratios differ only in terms of the definition of a financial measure.

Panel A of Figure 1 shows the histogram of the number of rank order reversals for the 3% bankruptcy samples. Panel B shows the same for the 20% bankruptcy samples.

Summary statistics of the distributions in number of rank order reversals are shown in Table 3.

Table 3. Summary statistics of PD rank order reversal distributions

Sample pair	Mean	Std. dev.	Minimum	Maximum	C.V.*
Both 3% bankruptcy	320.69	95.66	51	797	29.83%
Both 20% bankruptcy	152.51	40.36	36	388	26.46%
3% vs 20% bankruptcy	255.50	74.59	57	681	29.19%

Note: * Coefficient of variation. A rank order reversal occurs when the PD rank of firm *X* is above *Y* in one sample but below it in a paired sample. Samples with 3% bankruptcy are paired in sequence, that is sample *p* is paired with sample *p*+1 for *p*=1 to 99,999. This is repeated for samples with 20% bankruptcy. A final pairing is made for sample *p* with 3% bankruptcy vs sample *p* with 20% bankruptcy.

For the 3% bankruptcy samples the mean number of rank order reversals is 320.69 and the standard deviation is 95.66. The mean and standard deviation for the 20% bankruptcy samples are 152.51 and 40.36 respectively. Looking at Figure 1, it is evident that the distribution in Panel A lies to the right of that in Panel B which indicates that rank order reversals among the 3% bankruptcy samples are more than that in the 20% bankruptcy samples. The number of reversals for the 20% bankruptcy samples ranges for 36 to 388. For the 3% bankruptcy samples, the range is from 51 to 797.

To verify if the two distributions differ significantly, a simple test of difference in means in performed. The t-statistic obtained is 512.02 which is significant at 0.01% level. This is strong evidence of a difference in the distribution of rank order reversals between 3% and 20% bankruptcy samples.

The coefficient of variation for the 3% bankruptcy samples is 29.83% compared to the 26.46% for the 20% bankruptcy samples. Together with the significantly higher number of rank order reversals found earlier, this result suggests that credit ratings derived using a data sample with a low percentage of bankruptcies are less likely to be consistent. Ratings generated would also be of a wider dispersion. This finding has important implications on the development of IRB systems. If consistency is low when the credit rating model is derived from a data sample with 3% bankruptcy, it is likely to be even lower if the sample used is stratified to reflect more realistic default rates of 1% to 2%. Within even the same bank, therefore, the credit rating generated for an obligor may be dependent on the composition of the sample selected for modeling. Across banks, inconsistency will only be accentuated with different internal databases to begin with.

Since rank order reversals are significantly less for samples with 20% bankruptcy, would using a "loaded" sample approach reduce inconsistency? The answer is yes only if all credit rating models are derived the same way.

Panel C of Figure 1 shows the distribution of rank order reversals for sample pairs where one has 3% bankruptcy and the other 20% bankruptcy. Here sample *p* with 3% bankruptcy is paired with sample *p* with 20% bankruptcy. This is done for 100,000 samples of both bankruptcy rates. Summary statistics for the distribution or rank order reversals for this pairing are shown in the last row of Table 3.

As seen from Figure 1, this last distribution lies between those for 3% and 20% bankruptcies. The distribution mean is 255.50 and the standard deviation is 74.59. Rank order reversals range from 57 to 681. The coefficient of variation is 29.19%. While all the values fall between those for 3% and 20% bankruptcies, they are closer to the former than the latter. A test of difference in means lends some support to this conclusion. Against the 3% and 20% bankruptcy distributions, the t-statistics obtained are 206.60 and 386.19 respectively, both significant at 0.01% level. As the latter t-statistic is larger, it suggests that this distribution is closer to that for the 3% than the 20% bankruptcy.

The coefficient of variation provides additional insights. 29.19% differs only slightly from the 29.83% for the 3% bankruptcy samples. In terms of consistency, this means that variability of credit ratings derived across samples with 3% and 20% bankruptcies is no better than variability across all 3% bankruptcy samples. Across two banks with exactly the same credit database, consistency in credit ratings is not improved if one bank uses a 3% bankruptcy sample and the other uses a 20% bankruptcy to model credit risk over the use of 3% bankruptcy by both banks.

Overall, the results indicate that a low proportion of target outcomes can pose a problem to consistency. Default rates are generally low for bank loans. Deliberate "loading" by selecting a higher proportion of target outcomes in a data sample may alleviate this problem but it can lead to another which is that of accuracy. With a statistical method like logit where the odds prior is determined by the proportion of target outcomes in the sample, it is uncertain if a risk model derived using a sample with 20% bankrupt firms will be accurate in forecasting bankruptcy occurring at an actual rate of less than 3%.

A relatively simple way to reduce the number of rank order reversals which means an improvement in consistency is to use rating bands instead of point

PD estimates. The PD rank order of firms *X* and *Y* may be reversed across two data samples but if they fall into the same rating band, the reversal is nullified. How such banding and how many bands are optimal, however, beyond the scope of this paper?

Conclusion

With Basel II strongly advocating the development of IRB systems by individual banks, the issue of credit ratings consistency has come to the fore. Disparity in ratings for an obligor can result in differences in economic capital allocations and depreciate the intent of the New Accord. Unfortunately most banks do not have the universe of obligor data or access to external data to validate the consistency of their internal rating models. In this paper, I propose a test of inconsistency that can be employed even with a bank's own internal database. It is premised on a reasonable requirement of consistency, which is the order of estimated credit risk should be stable. A pervasive reversal of this order where *X* is adjudged more risky than *Y* in one sample but the reverse in another would be evidence of inconsistency. Using the concept of ranks, the distributions for up to 50 observations in a sample are derived and counts of rank reversals needed to reject the null of no inconsistency at 90%, 95% and 99% confidence levels are reported.

The test is then applied to a common corporate credit risk rating approach that employs financial ratios in a logistic regression. Using PD as the metric for credit rating, the results obtained suggest that sample composition in terms of the proportion of the modeling target, which is usually defaults or bankruptcies, affects consistency. Consistency is especially sensitive to PDs estimated using data samples with low default rates. The rank order of PDs derived using samples with 3% bankrupt firms are reversed significantly more times from one sample to another than those derived using 20% bankrupt firms. Across two samples, one with 3% and another 20% bankruptcies, the average number of reversals is less than those with 3% bankruptcy alone but the coefficient of variation for the reversals remain largely unchanged. This suggests that not only the consistency of credit ratings is obtained from data samples with low default rates a concern, even the common use of "loading" to create a higher default rate may not improve consistency. The use of rating bands to replace point PD estimates could, however, be one approach to improving consistency. All the results could, of course, be a function of employing logistic regression but that would not diminish the concern over credit ratings consistency.

References

1. Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance*, Vol. 23, pp. 589-608.
2. Altman, E.I. and A. Saunders (2001). An analysis and critique on the BIS proposal on capital adequacy and ratings, *Journal of Banking and Finance*, 25, pp. 25-46.
3. Carey (1998). Credit risk in private debt portfolio, *Journal of Finance*, Vol. 53, Issue 4, pp. 1363-1387.
4. Carey (2001). Some evidence on the consistency of banks' internal credit ratings, Federal Reserve Board, Working paper.
5. Frydman, Halina, E.I. Altman and D.L. Kao (1985). Introducing recursive partitioning for financial classification: the case of financial distress, *Journal of Finance*, Vol. 40, Issue 1, pp. 269-291.
6. Lopez, J.A. and M.R. Saidenberg (2000). Evaluating credit risk models, *Journal of Banking and Finance*, 24, pp. 151-165.
7. Martin, D. (1977). Early warning of bank failure, *Journal of Banking and Finance*, 1, pp. 249-276.
8. Treacy, W.F. and Mark S. Carey (2000). Credit risk rating systems at large US banks, *Journal of Banking and Finance*, 24, pp. 167-201.