Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Banking & Finance 30 (2006) 797–810

Journal of BANKING & FINANCE
Applied Section

www.elsevier.com/locate/jbf

Inferring the default rate in a population by comparing two incomplete default databases

Douglas W. Dwyer [a], Roger M. Stein [b,*]

[a] Moody's KMV, New York, United States
[b] Moody's Investors Service, 99 Church Street, New York, NY 10007, United States

Received 14 October 2005; accepted 2 November 2005

Abstract

It is often the case in default modeling that the need arises to calibrate a model to some prior probability of default. In many situations, a researcher may not know the true prior default rate for the population because the data set at hand is itself incomplete, either with respect to default identification (hidden defaults) or default under reporting. In situations where a researcher has access to two incomplete default data sets, for example in the case of two banks that have merged, it is possible to infer the number of "missing" defaults, which we demonstrate in this short note. We discuss an approach to estimating this quantity and show an example in which we infer the number of missing defaults in the combined legacy databases of the former Moody's Risk Management Services and the former KMV Corporation. While calibration is one application of this approach, the method is a general one that can be applied in other settings as well.

JEL classification: C40; C43; C52; C81; G33

Keywords: Bankruptcy modeling; Credit modeling; Model validation and calibration; Capture–recapture; Database merging

[*] Corresponding author. Tel.: +1 212 553 4928; fax: +1 212 298 7024.
E-mail address: roger.stein@moodys.com (R.M. Stein).

0378-4266/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbankfin.2005.11.001

1. Introduction

Default databases play a key role in the development, validation and application of credit models. Nevertheless, it has often been difficult to ascertain the extent to which these databases accurately capture *all* of the default events that have occurred over a particular time period or market segment. While it is generally understood that not all default events are captured in any one dataset, estimates of the magnitude of missed defaults are previously non-existent (to our knowledge) even though such information is extremely valuable for credit risk management.

For example, a crucial component in calibrating theoretical default models to real-world applications is the assessment of the relationship between a model's output and a real-world probability of default. This is true irrespective of whether the model output is the result of a probit function as might be the case with a statistical model; a theoretical distance to default as might be the case with a structural model; a risk neutral implied probability of default as might be the case with a reduced form model; or a historical default rate, as might be the case with an internal rating system (see, for example, Duffie and Singleton, 2003 or Lando, 2004 for a review of various approaches to estimating probabilities of default).

Recently, issues of calibration and estimation of reliable long-run PDs has attracted the attention of both regulators and practitioners due to its central role in determining capital adequacy and other risk measures. For example, the New Basel Accord (BIS, 2004) stipulates that a bank may use data on internal default experience for the estimation of PD. A bank must demonstrate in its analysis that the estimates are reflective of underwriting standards and of any differences in the rating system that generated the data and the current rating system. Where only limited data are available, or where underwriting standards or rating systems have changed, the bank must add a greater margin of conservatism in its estimate of PD (cf., paragraph 451 of 'A Revised Framework').

Calibration is a process by which a model's output is converted into actual default rates. Calibration typically involves two steps. The first requires mapping a model score to an empirical probability of default using historical data. The second step, which sometimes receives less attention in the literature, entails adjusting for the difference between the default rate in the historical database and the actual default rate[1,2] (cf., Dwyer and Stein, 2005).

In addition to confounding calibration, an incomplete default database can also adversely impact the validation of default probability models. In such validation exercises, it is typically the case that users focus on issues both of model power (the model's ability to distinguish between defaulting and non-defaulting firms) and on the *accuracy of a model's calibration* (the appropriateness of the probability levels that the model produces). While much of the literature on default model validation focuses on aspects of power through the use of power curves and their associated statistics, a powerful model can turn out to be poorly calibrated and the "most accurate" probability model may not be the most powerful (cf., Stein, 2002). Validation of probabilities can be confounded when an incomplete data set is the reference database and this can have direct practical consequences as the

[1] It is not uncommon for these rates to be different due to data gathering and data processing constraints.

[2] In technical terms, it is necessary to adjust the model prediction for the true prior probability or base rate.

precision of the probabilities produced by a default model are particularly important for capital allocation and risk management.

Differences between sample default rates and actual population default rates may arise through the design of the modeling task. For example, in a matched sample setting a modeler might artificially create a data set containing an equal number of defaulting and solvent firms. More often, however, it arises as a natural consequence of the limited ability of a researcher to gather complete data. In such situations, a researcher may not know the true prior default rate for the population because the data set at hand is itself incomplete either with respect to default identification (hidden defaults) or default under reporting.

Fortunately, in some situations it may be possible to infer the true default rate based on the data in hand if a second data set exists that has also been collected in the same credit market, but independently (e.g., by a different organization or through a different process). This turns out not to be uncommon. For example, it is often the case that a bank may have access to both its internal data on defaults as well as data from a commercial vendor such as Dun and Bradstreet in the US or BvD in Europe. These data sets may be pooled to arrive at a more complete set of defaulting and non-defaulting firms.

A particularly interesting case in which this inference might be applied is the case in which two banks merge, particularly if they previously competed in the same market segments. In such cases, the banks often have overlapping client-bases and thus may have the ability to apply the techniques we discuss here to at least a portion of the new combined portfolio in order to better characterize the default rates in the sub-populations within them.

In situations where a researcher has access to two incomplete default data sets, it is possible to infer the number of "missing" defaults using a procedure developed by Sekar and Deming (1949).[3] We demonstrate this approach in this short note and show an application to defaults for North American public firms with less than $1 mm in sales.

The remainder of this document proceeds as follows: In Section 2 we review the relevant statistical theory. In Section 3 we present two applications of this methodology to two independently compiled public default databases to estimate the overall number of missing defaults (and thus the adjusted default rate). Section 4 examines how sensitive our results are to assumptions regarding correlation between default capture probabilities in the two data sets. Section 5 and 6 provide discussion and conclusion, respectively.

2. Statistical background

Fig. 1 provides a graphical depiction of the quantities that we can observe in two databases and the unknown quantity (number of missing defaults) we wish to estimate. In the figure, the left circle shows the defaults in database 1 and the right circle shows defaults that were in database 2. The overlap represents defaults that were captured in both databases. The light gray area in the larger circle represents the quantity that we are trying to estimate. It shows defaults that were captured in neither database 1 nor database 2.

[3] The application that Sekar and Deming addressed was of comparing the birth and death records maintained by the registry to those obtained by a house-to-house canvas in India. The estimator that they provide is identical to the so-called Petersen estimate common in the capture–recapture literature. Estimating populations through capture–recapture sampling has a rich history dating back to the 17th century (e.g., see Mammo, 1998, or ABA/CEELI, 2000).

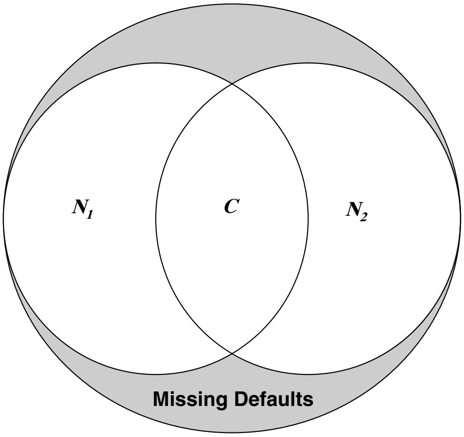

Fig. 1. Missing defaults. The left circle shows the defaults in one database, and the right circle shows defaults that were in the other. The overlap represents default in both databases and the light grey area in the larger circle shows defaults that were in neither.

Following the notation of Sekar and Deming (1949) let

C	denote the number of defaults detected in both databases
N_1	denote the number of defaults detected in database 1 but not database 2
N_2	denote the number of defaults detected in database 2 but not database 1
N	denote the total number of defaults
M_1	denote the number of defaults detected in database 1 ($N_1 + C$)
M_2	denote the number of defaults detected in database 2 ($N_2 + C$)
D_1	denote a random variable with 1 indicating that a given default is captured in database 1 and 0 otherwise
D_2	denote a random variable with 1 indicating that a given default is captured in database 2 and 0 otherwise

The random variables D_1 and D_2 have Bernoulli distributions. To estimate the size of the missed defaults, let $p_1 = E(D_1)$ and $p_2 = E(D_2)$, i.e., the probability that a given default event will be captured in database 1 and database 2, respectively. For now, we assume that capture events are independent (we later partially relax this assumption), so

$$P((D_1 = 1) \cap (D_2 = 1)) = p_1 p_2.$$

We have:

$$\mathrm{p}\lim_{N\to\infty} \frac{N_1 + C}{N} = p_1, \tag{1}$$

$$\mathrm{p}\lim_{N\to\infty} \frac{N_2 + C}{N} = p_2, \tag{2}$$

$$\mathrm{p}\lim_{N\to\infty} \frac{C}{N} = p_1 p_2. \tag{3}$$

Inserting (1) and (2) into (3) and rearranging yields the following estimate[4] of N:

$$\widehat{N} = N + \widehat{N}_m = C + N_1 + N_2 + N_1N_2/C, \tag{4}$$

$\widehat{N}$ represents the estimate of the total number of defaults in the universe, based on our observations of the defaults in each database, $\widehat{N}_m = N_1N_2/C$ is an estimate of the number of missing defaults.

In applications, this formula can be applied to groups that are ‘‘homogeneous enough’’ to justify the assumption of independence and then sum up across all the groups to get an estimate of total default events. Sekar and Deming (1949) show that if there is positive (negative) correlation between D_1 and D_2 then the estimate in Eq. (4) is asymptotically biased downward (upward).[5] Section 5 examines the sensitivity of estimates to the magnitude of correlation between the capture probabilities.

2.1. A short example

As an example, assume that we have two default databases, with the following distributions of defaults:

Database 1	Database 2	Both 1 and 2	Total unique
50	90	20	120

In terms of the notation defined above, we have:

$$M_1 = 50, \quad M_2 = 90, \quad C = 20, \quad N_1 = 30 \quad \text{and} \quad N_2 = 70.$$

By (4) we get:

$$\widehat{N} = C + N_1 + N_2 + N_1N_2/C = 20 + 30 + 70 + \frac{70 * 30}{20} = 225,$$

where 225 is equal to the total number of defaults in the two databases (120) plus the estimated number of missed defaults (105).

[4] Since the last term involves dividing by a random variable the estimate is not unbiased in small samples. Further, if $C = 0$ then the estimate is undefined. The so-called Chapman estimate adds 1 to C to address this issue. Rivest and Levesque (2001) find that the Chapman estimate has more desirable small sample properties. We have experimented with the Chapman estimate and found it to have a minor impact on our results, perhaps due to the reasonably large samples we are examining.

[5] For some intuition behind this result consider the following example: Suppose that there were some class of defaults for which the probability of detection were zero for both organizations and that for the remainder of the defaults, the probability of detection is one for both organizations. Across the universe of defaults, the detection probability is positively correlated between the two organizations. Both organizations will observe exactly the same defaults and $N_1 = N_2 = 0$. Without controlling for the correlation between the two capture rates, we would estimate the population of defaults as the numbers of defaults observed by both organizations, C. This estimate is clearly less than the true value, because, by construction in our example, there are some defaults that go unobserved by both organizations.

How much of a difference would this make in our estimation of default probabilities? If we continue the example by assuming that the combined database has 20,000 firm years in it in total (a not unreasonable example), this implies that the prior probability of default should move from 0.6% to 1.1%. Note that in this case, we are assuming that we have full knowledge of the total number of non-defaulting firms.[6]

There are, of course, a variety of potential sources for bias in this estimate. First, the estimator is biased in small samples due to the non-linearity of the formulation. Second, D_1 and D_2 may not be completely independent. We discuss this in Section 5. Third, there is the potential for misclassification of defaults and for differences in the definition of defaults. Finally, there is likely to be a certain amount of merging error, i.e., both organizations identify the same default but fail to identify that it is the same default due to different identifiers in the database. We shall address each of these issues.

3. Sample application

Prior to the merger of KMV and Moody's Risk Management Services in 2002, both legacy organizations actively tracked default events for companies with public securities through a wide variety of public and private sources. In the case of legacy KMV (LKMV), the default database was used primarily for calibration and validation, and in the case of legacy Moody's Risk Management Services (LMRMS) the default database was used for model development as well as calibration and validation.

After the merger of the two firms, the combined Moody's KMV (MKMV) organization had the opportunity to extend the analyses of the legacy organizations to better characterize the true baseline default rates in various markets. Due to reporting issues and other data problems described above, it is typically not possible to capture all defaults with complete accuracy. As a result, both for modeling, calibration and validation purposes and for tracking purposes, it is useful to obtain an estimate of how many defaults are "missing" from the (now combined) database. Further, it is interesting to determine whether we might observe systematic patterns in the missed data.

For example, we applied the method of Section 2 to the combined data set of all the public North American defaults that had less than $1 million in sales and were captured by either LKMV or LMRMS between January 1, 1990 and December 31, 2002.[7] Such firms represent between 5% and 10% of the defaults in the two databases. A priori, we expected there to be a larger number of missing defaults in this segment since smaller defaults tend not to receive as much press coverage as their larger counterparts and since we have observed that market participants tend to focus less on (and thus are less likely to recollect) smaller defaults. Following LMRMS default database convention, we only allow a company to enter the database as a default once; subsequent defaults by the same

[6] We view this assumption as a reasonable starting point. Banks are often able to track loans that are originated or loans on which principle and interest are being collected. It is typically more difficult to identify which firms ultimately default as these loans are often passed to separate work-out groups, the activities of which are not part of the operational data collection procedures of the bank. In the case of firms that must report publicly (such as listed firms in the US or other certain classes of enterprise in Europe and Asia) the issue of identifying the borrower population is typically minor.

[7] We use the term "publicly traded" to refer to a firm whose equity is publicly traded, i.e., as an indicator of whether or not it has a stock price.

Table 1
Estimates of total and missing defaults among firms with less than $1 mm in sales

	LKMV	LMRMS	Total
Defaults observed	237	93	251
Estimated total			279(10)
Estimated missed	42	186	28
Percent captured	85%	33%	90%

This table estimates the total number of defaults of firms with less than $1 mm in sales in the US and Canada from the beginning of 1990 until the end of 2002. The capture probabilities of both LKMV and LMRMS are assumed to be independent within this size segment. The standard error of the estimated total is in parenthesis.

firm are eliminated.[8] For the purpose of this research we defined a default on a publicly listed firm to be:

- Missed interest or principle payment (on either a bond or a bank loan).
- Chapter 7 or Chapter 11 bankruptcy.
- A distressed exchange.

Restricting the scope of the study to small firms eliminates a key source of heterogeneity in the likelihood that a default will be captured.[9]

Table 1 presents the counts of small firm defaults captured by both LKMV and LMRMS as well as the total observed by either firm. Further, it presents estimates of the number of missing defaults, the total number of defaults and an asymptotic standard error.[10] Finally, it presents capture rates for the two organizations, i.e., the percent of the total estimated defaults observed by each organization. This table assumes independence of default detection between the two organizations for this size category.

As expected, the estimates of missed defaults are proportionately higher for the simple cases than for the combined case, particularly within the LMRMS database. The low capture rates among small firms for LMRMS is likely to be a by-product of its default database being built from the Moody's Investors Service (MIS) default database which was designed primarily to track defaults on rated entities.[11]

By combining the two databases, we estimate that 90% of the total defaults among small companies were captured in the full sample. Said another way we estimated that there were an additional 28 defaulted firms, or 10% of the complete population, that

[8] LKMV traditionally allowed the same firm to default more than once provided the defaults were more than 30 months apart. Many firms have emerged from Chapter 11 only to go back into Chapter 11 at a later date. Therefore, the LKMV approach does have merit. Nevertheless, in order to facilitate an "apples to apples" comparison, we chose to adopt the LMRMS convention for the pooled data in this experiment.

[9] A possible extension of this analysis is to allow the probability of detection to be a continuous function of the size of the firm (cf., Borchers et al., 2002).

[10] We show the asymptotic standard error provided by Sekar and Dewing: $\sqrt{(1-p_1)(1-p_2)\widehat{N}/(p_1p_2)}$, assuming no correlation in capture rates. We explore the reliability of the estimate and the effect of correlation later in this paper.

[11] Moody's Investors Service (MIS) has published 18 *Annual Default Studies* documenting the default experiences of rated bonds since 1970 (Hamilton, 2005). In these studies, MIS is chiefly interested in "rigorously assess the performance of its ratings as predictors of default" (Hamilton, 2002). Consequently, the focus has been on rated defaults, which are not typically small companies. LMRMS built its default database from this one and added other defaults to it from other sources.

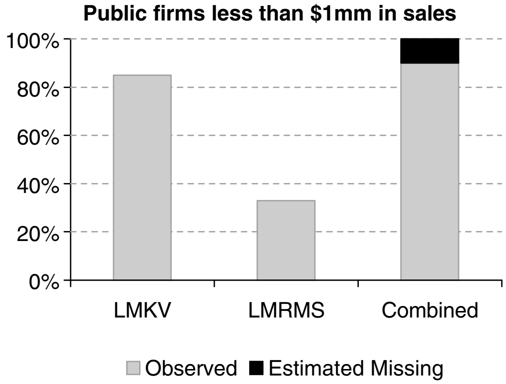

Fig. 2. Percent of total defaults observed and estimates of missing defaults. Percent of total defaults observed by each organization as well as for the combined organization. By combining the two default databases, one is able to obtain an estimate of the number of missing defaults.

neither database captured but that should be included in calculating a baseline default rate for this segment. Fig. 2 depicts this graphically.

4. Small sample properties of the estimator

In order to provide some sensitivity for the accuracy of the estimates, we performed a simulation study in which we evaluated the distributions of estimate errors, some examples of which we show in Fig. 3. To create the figure, we performed simulations in which we generated random complete databases with a target default rate of 2%. We then generated two incomplete databases each with a target expected missingness rates. We then used Eq. (4) to estimate the number of missing defaults, based only on the two incomplete databases. Because we had an accurate estimate in each case of the complete database (the two databases were derived from the simulated complete database), we were able to compare the estimated results with the known number of defaults in the complete database. Fig. 3 shows the median of the absolute value of the errors, abs[(actual – estimated)/actual], for a variety of missingness rates. We performed 1000 simulations for each combination of missingness rates, assuming database sizes of 10,000 observations and, on average, 200 and 2000 defaults.

The graphics show chromatically how the degree of missingness affects the error rate $(N - \widehat{N})/N$. Higher error rates are darker. We assume an average default rate of 2% and a database size of 10,000 which is within the ranges seen in many banks. The largest errors shown by the darkest shades, are around 5% and are seen for the largest missingness values, in this case 50% in each database, and low numbers of defaults (~200). The size of the errors increase as the missing percentage increases.

We can gain perspective on these error rates by examining Fig. 3. The plots in the figure show how the median percentage error in the simulations changes with different combinations of missingness in the two databases. From the plots it is clear that the errors widen as the missingness rate increases for both databases. The figure shows a clear pattern of darkening as we move to the upper right, the direction of higher error rates for both databases.

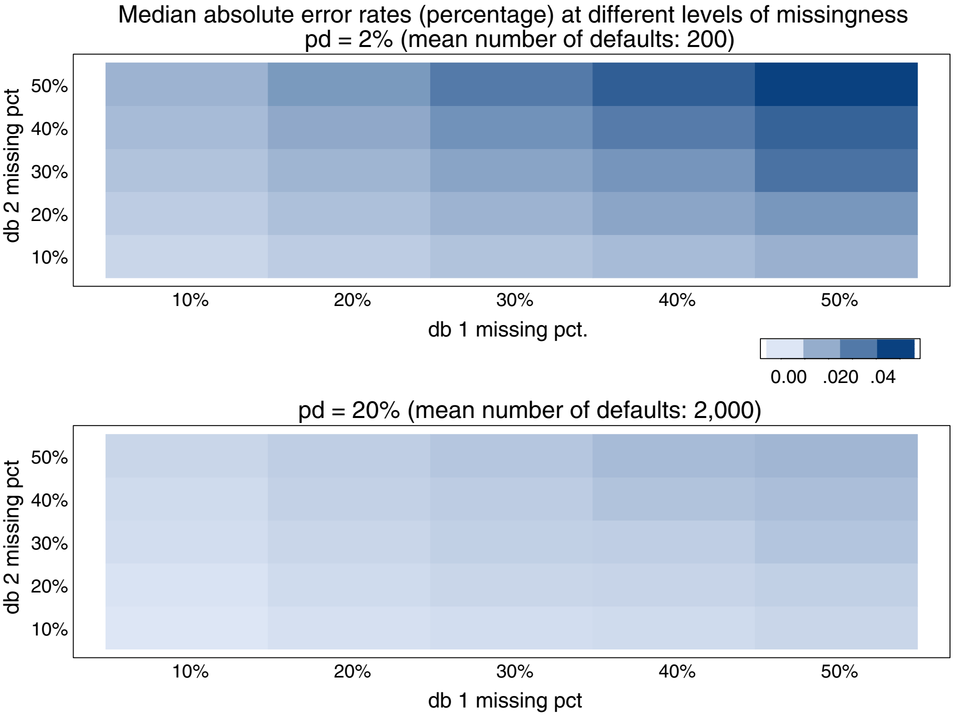

Fig. 3. Results of simulation showing error rate of estimated vs. actual missing observations.

This evidences the intuitive relationship between database errors and the accuracy of the estimates of $\widehat{N}$.

Given the results of Fig. 3, it is interesting to consider the reliability of the asymptotic standard error discussed in Footnote 10. We find that in general, estimates of the standard error based on the asymptotic value are overly optimistic. That is, the actual standard errors should be wider to accurately reflect the variability of the estimates.

To examine this, we used an approach suggested by Diebold et al. (1998) in which we test the appropriateness of the standard errors by examining the precision of the coverage that confidence intervals based on these standard errors would imply. If the standard errors are accurate, then, under the normal assumption, 95% of the values should be within ±1.96 standard errors, 99% should be within ±3 standard errors, etc. (see, Diebold et al. (1998) for more detail on the approach). Since each estimated standard error is applied to a single point estimate, we can only evaluate this coverage in aggregate by looking at a large number of point-estimate/standard error combinations. We did this through simulation.

In experiments, not reported in detail here, we simulated values of p_1 and p_2 in {0.05, 0.1 and 0.2} with probabilities of default in the range {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.15 and 0.2}. We found that, depending on the level of p_1 and p_2, for low numbers of defaults (typically 1000 or 2000, depending on the level of missingness) the tests suggested by Diebold et al. (1998) tend to reject the appropriateness of the Sekar and Deming (1949) asymptotic estimate suggesting that the asymptotic assumptions do not hold in these sparser data contexts.

5. Sensitivity analysis: The impact of correlation

In the previous analysis, we assumed that the capture probabilities of the two organizations were independent, at least within the group that was being analyzed. As shown by Sekar and Deming (1949), if the correlation is positive (negative) our results are biased downward (upward). In this section, we examine the degree to which our estimates might change if we assumed varying levels of correlation between the capture probabilities.

Admitting correlation into the formulation results in an additional term in (3):

$$\begin{aligned}\mathrm{p}\lim_{N\to\infty}\frac{\mathrm{C}}{\mathrm{N}} &= p((D_1=1)\cap(D_2=1)) = E(D_1D_2)\\ &= p_1p_2+\rho\sqrt{p_1p_2(1-p_1)(1-p_2)},\end{aligned} \tag{5}$$

where $p((D_1=1)\cap(D_2=1))$ denotes the probability that the default is captured in both data sets 1 and 2, respectively, and ρ is the correlation coefficient.[12] The second equality is straightforward to derive.[13] Substituting in for p_1 and p_2 and multiplying both sides by $\frac{\widehat{N}^2}{C}$ yields:

$$\begin{aligned}\widehat{N} &= C+N_1+N_2+\frac{N_1N_2}{C}+\frac{\rho}{C}\sqrt{M_1M_2(\widehat{N}-M_1)(\widehat{N}-M_2)}\\ &= \frac{M_1M_2}{C}+\frac{\rho}{C}\sqrt{M_1M_2(\widehat{N}-M_1)(\widehat{N}-M_2)}\end{aligned}$$

recall that $M_i = N_i + C$, which is the total number of defaults captured in database i. Some algebraic manipulation shows that in the case of non-zero correlation, an estimate of $\widehat{N}$ is given by the following application of the quadratic formula:

$$\widehat{N} = \frac{-b\pm\sqrt{b^2-4ac}}{2a}, \tag{6}$$

where

$$\begin{aligned}a &= 1-\frac{\rho^2M_1M_2}{C^2},\\ b &= -\frac{2M_1M_2}{C}+\frac{\rho^2M_1M_2}{C^2}(M_1+M_2),\\ c &= \frac{M_1^2M_2^2}{C^2}(1-\rho^2).\end{aligned}$$

When the correlation coefficient is positive (negative) the estimate is given by the positive (negative) root.[14] It can be shown that as $\widehat{N}$ goes to infinity as $\rho\to\frac{C}{\sqrt{M_1M_2}}$ from the left (see

[12] Note that Lucas (1995) used a similar formulation to specify default correlation based on joint default probabilities.

[13] In the case of two binary variables, the correlation coefficient is defined as: $\rho\equiv\frac{\mathrm{Cov}(D_1,D_2)}{\sqrt{\mathrm{Var}(D_1)\mathrm{Var}(D_2)}}$ which reduces to $\frac{E(D_1D_2)-p_1p_2}{\sqrt{p_1,p_2(1-p_1)(1-p_2)}}$.

[14] As discussed in Footnote 4, positive (negative) correlation increases (decreases) the number of missing defaults relative to the case of zero correlation. Note that the two solutions to (6) do not depend on the sign of the correlation coefficient because ρ is always squared in the solution. Therefore, the positive root must correspond to positive correlation and the negative root must correspond to negative correlation. Note also that there are cases where one of the solutions will be inadmissible (e.g., for ρ greater than the limit discussed, below).

Table 2
Sensitivity of estimates to correlation in default capture rates

ρ	Percent missed (%)
0.00	10
0.20	26
0.40	59
0.53	100

This table examines the sensitivity of the estimate of the percentage of defaults that are missed to the assumption regarding the correlation in the default detection probability. Table 1 assumes zero correlation.

Appendix A). Of course, the total number of defaults cannot exceed the number of firms in the population, which if known, places an upper bound on the extent of possible correlation when there are defaults captured in each data set that were not captured in the other (i.e., both N_1 and N_2 do not equal 0).

Table 2 expands the analysis of Table 1 by showing how the estimate of percent of defaults that were missing changes for different assumptions regarding the degree of correlation in the capture probabilities.

The analysis shows that, perhaps unsurprisingly, a moderate amount of correlation does lead to a substantial increase in the estimates of missing defaults. Specifically, changing the assumption regarding the correlation from 0 to 0.2 increases the percent of defaults that were missed from 10% to 26%. The estimate of the percent of defaults that were missed goes to 100% as the correlation coefficient goes to 0.53. This value is the value at which $\widehat{N}$ goes to infinity.

While instructive, this result is unfortunately not that informative in this setting. As ρ approaches $C/\sqrt{M_1M_2}$, its limiting value, the fact that there were *any* defaults in one database that were not in the other implies that the number of missed defaults goes to infinity.

With two databases, ρ cannot be directly estimated: We only have three pieces of information $\{N_1, N_2, C\}$ and are estimating three parameters $\{p_1, p_2, N\}$ while assuming a specific value of ρ. If we had three or more databases, the correlation coefficient between the detection rates could potentially be estimated as an additional parameter (cf., Borchers et al., 2002).

6. Conclusion

We have presented an approach to sizing the approximate size of the population of defaults based on samples from two different databases. The approach relies on using the overlap between the two databases to size the likely number of missing defaults.

We demonstrated the approach on a subset of the universe of defaults on publicly traded firms as captured independently by the legacy MRMS and legacy KMV firms. The combined database appears to contain capturing 90% of the total defaults among small firms, which is a larger number of defaults than either of the legacy databases taken individually (Fig. 2). Further, we are able to determine that, assuming no correlation in collection of defaults, the combined database appears to undercount defaults among smaller firms by 10%, but that we can adjust our prior estimate to correct for this undercount using the techniques described in this note.

In interpreting the results of this example, it is instructive to highlight aspects of the analysis where caution is warranted. The estimate of 28 missed small firm defaults may be an underestimate for a number of reasons. To the extent that there is positive correlation in default capture rates, we have shown that this will cause the rates to be understated. Furthermore, to the extent that there are classes of default that are generally not captured by either organization, this too will serve to understate the default counts. In fact, it seems plausible that there are distressed exchanges in this sub-population that are kept confidential so that the probability of capture by either organization is zero. A secondary source of error is that of merging error which, while we have attempted to address, is difficult to fully eradicate. Finally, differences in the definition of a default may result in one organization *choosing* not to capture a default that another *chooses* to capture. This issue is particularly pronounced for distressed exchanges.

Despite these caveats, we do think that in this example it would be prudent to adjust upwards the prior default rate by at least the inverse of the estimated capture rate. Doing so would increase the prior default rate by 11% for small firms. By combining two databases that were developed independently, we were able to obtain estimates of this adjustment factor.

We also applied this approach to rated companies. We initially found what appeared to be missing defaults in both databases. As it turned out, however, the differences were primarily the result of how the actual default events were assigned to different entities within a corporate family rather than the default events being missed. The differences in how the default events were recorded reflected the different purposes for which the two organizations used default databases.[15] Interestingly, however, though the underlying explanation for the differences in the default databases was not related to actual missing defaults, the approach did serve to highlight the potential issues that could arise in attempting to combine databases that use different conventions for identifying defaults.

More generally, we feel the approach outlined herein provides a reasonable method for adjusting default rates to reflect the sampling bias that may arise as a natural consequence of the difficulty in capturing historical data on defaulted firms. To the extent that researchers do not correct for such sampling effects, it is likely that PDs for certain elements of the borrower population will be systematically biased downward. Similarly, validation tests of default probabilities that rely on such samples may also be affected.

Acknowledgments

The authors wish to thank Irina Korablev, Matt Kurbat and Joachim Skor, Richard Cantor and David Hamilton for their generous comments. We are also grateful to two

[15] One of the key differences is in the treatment of subsidiary defaults. LMRMS drew heavily on the experience of Moody's Investors Service (MIS) in tracking the default events of companies that it had rated (cf., Hamilton, 2005). MIS assigns defaults to the specific entity that defaults unless there are also defaults on explicit guarantees from other members of the corporate family. LKMV in contrast, assigns defaults at subsidiaries to corporate parents if the subsidiary meets a certain size threshold. The different practices reflect the different purposes for which the default databases were being used. MIS uses its default database to track the performance of ratings that are assigned to the legal entity issuing the debt. LKMV used the database to develop and test its public firm model, which uses equity market information to assess default risk. For example, MIS tracks both Trump Atlantic City Associates and Trump's Castle Funding (the issuers) as having had grace period defaults in October of 2001, but not Trump Hotel and Casino Resort (the publicly traded holding company). LKMV, in contrast, treated Trump Hotel and Casino Resort as having experienced a default, albeit at one of its subsidiaries.

anonymous referees for very useful comments. All remaining errors are of course those of the authors.

Appendix A

Proposition. *$\widehat{N}$ goes to infinity as $\rho \to \frac{C}{\sqrt{M_1 M_2}}$ from the left.*

Proof. As $\rho \to \frac{C}{\sqrt{M_1 M_2}}$ from the left,

$$a \to 0$$

$$b \to -\frac{2M_1M_2}{C} + M_1 + M_2 \quad \text{and}$$

$$c \to \frac{M_1^2M_2^2}{C^2} - M_1M_2.$$

To show this result we need to demonstrate that the limit of this fraction goes to infinity as the denominator goes to zero. To do so, it is sufficient to show that the limit of b is *negative* since: we are using only the positive root of the quadratic formula. The limit of $-b + \sqrt{b^2 - 4ac}$ must be positive if the limit of b is negative (since $\lim_{\rho \to \frac{C}{\sqrt{M_1M_2}}} 4ac = 0$); and the denominator of the quadratic formula ($2a$) is going to zero (from the right).

To show that b is negative, substitute $N_i + C$ for M_i. Simplifying yields:

$$\lim_{\rho \to \frac{C}{\sqrt{M_1M_2}}} b = -2C - 2N_1 - 2N_2 - \frac{2N_1N_2}{C} + N_1 + N_2 + 2C$$

or

$$\lim_{\rho \to \frac{C}{\sqrt{M_1M_2}}} b = -N_1 - N_2 - \frac{2N_1N_2}{C}. \tag{7}$$

Since N_1, N_2 and C are always positive, the limit in (7) is clearly negative. □

References

American Bar Association Central and East European Law Initiative, 2000. Political Killings in Kosova/Kosovo March–June 1999, Washington, DC.

Bank for International Settlements, 2004. International Convergence of Capital Measurement and Capital Standards (A Revised Framework).

Borchers, D.L., Buckland, S.T., Zucchini, W., 2002. Estimating Animal Abundance: Closed Populations. Springer, London.

Diebold, F.X., Gunther, T.A., Tay, A.S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39 (4), 863–883.

Duffie, D., Singleton, K.J., 2003. Credit risk: Pricing, Measurement, and Management. Princeton Series in Finance. Princeton University Press, Princeton, NJ.

Dwyer, D., Stein, R.M., 2005. Moody's KMV RiskCalc™ v3.1 Technical Document. Moody's KMV, New York.

Hamilton, D., 2002. Default and Recovery Rates of Corporate Bond Issuers: A Statistical Review of Moody's Ratings Performance 1970–2001. Moody's Investors Service, New York.

Hamilton, D., 2005. Default and Recovery Rates of Corporate Bond Issuers: A Statistical Review of Moody's Ratings Performance 1970–2004. Moody's Investors Service, New York.

Lando, D., 2004. Credit Risk Modeling: Theory and Applications. Princeton Series in Finance. Princeton University Press, Princeton.

Lucas, D.J., 1995. Default correlation and credit analysis. Journal of Fixed Income (March), 76–87.

Mammo, A., 1998. Estimating the Extent of Illicit Drug Abuse in New Jersey Using Capture–Recapture Analysis. New Jersey Department of Health and Senior Services, Rockville, Maryland.

Rivest, L., Levesque, T., 2001. Improved log-linear model of estimators of abundance in capture–recapture experiments. The Canadian Journal of Statistics 29, 555–572.

Sekar, C.C., Deming, W.E., 1949. On a method of estimating birth and death rates and the extent of registration. American Statistical Association Journal 44, 101–115.

Stein, R.M., 2002. Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation. Moody's KMV, New York.