Validation methodologies for default risk models

The Basle Committee has identified credit model validation as one of the most challenging issues in quantitative credit model development. **Jorge Sobehart**, **Sean Keenan** and **Roger Stein** of Moody's Investors Service address issues of data sparseness and the sensitivity of models to changing economic conditions along the Basle guidelines.

A MAJOR CHALLENGE in developing models that can effectively assess the credit risk of individual obligors is the limited availability of high-frequency objective information to use as model inputs. Most models estimate creditworthiness over a period of one year or more, which often implies the need for several years of historical financial data for each borrower[1].

While reliable and timely financial data can usually be obtained for the largest corporate borrowers, they are difficult to obtain for smaller borrowers, and are particularly difficult to obtain for companies in financial distress or default, which are key to the construction of accurate credit risk models. The scarcity of reliable data required for building credit risk models also stems from the highly infrequent nature of default events.

In addition to the difficulties associated with developing models, the limited availability of data presents challenges in assessing the accuracy and reliability of credit risk models.

In its recent report on credit risk modelling, the Basle Committee on Banking Supervision highlighted the relatively informal nature of the credit model validation approaches at many financial institutions. In particular, the Committee emphasised data sufficiency and model sensitivity analysis as significant challenges to validation. The Committee has identified validation as a key issue in the use of quantitative default models and concluded that "*…the area of validation will prove to be a key challenge for banking institutions in the foreseeable future*."[2]

This article describes several of the techniques that Moody's has found valuable for quantitative default model validation and benchmarking. More precisely, we focus on (a) robust segmentation of data for model validation and testing, and (b) several robust measures of model performance and inter-model comparison that we have found informative and currently use. These performance measures can be used to complement standard statistical measures.

We address the two fundamental issues that arise in validating and determining the accuracy of a credit risk model under: what is measured, or the metrics by which model 'goodness' can be defined; and how it is measured, or the framework that ensures that the observed performance can reasonably be expected to represent the behavior of the model in practice.

Model accuracy

When used as classification tools, default risk models can err in one of two ways[3]. First, the model can indicate low risk when, in fact, the risk is high. This Type I error corresponds to the assignment of high credit quality to issuers who nevertheless default or come close to defaulting in their obligations. The cost to the investor can be the loss of principal and interest, or a loss in the market value of the obligation.

Second, the model can assign a low credit quality when, in fact, the quality is high. Potential losses resulting from this Type II error include the loss of return and origination fees when loans are either turned down or lost through non-competitive bidding. These accuracy and cost scenarios are described schematically in Figures 1 and 2. Unfortunately, minimising one type of error usually comes at the expense of increasing the other. The trade-off between these errors is a complex and important issue. It is often the case, for example, that a particular model will outperform another under one set of cost assumptions, but can be disadvantaged under a different set of assumptions.

Since different institutions have different cost and pay-off structures, it is difficult to present a single cost function that is appropriate across all firms. For this reason, here we use cost functions related only to the information content of the models.

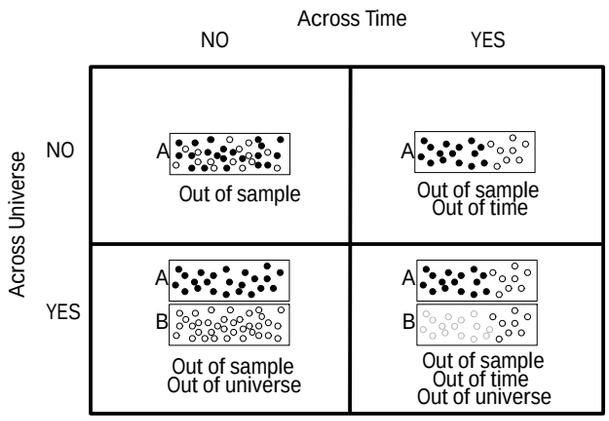
A validation framework

Performance statistics for credit risk models can be highly sensitive to the data sample used for validation. To avoid embedding unwanted sample dependency, quantitative models should be developed and validated using some type of out-of-sample[4], out-of-universe and out-of-time testing approach on panel or cross-sectional data sets.

However, even this seemingly rigorous approach can generate false impressions about a model's reliability if done incorrectly. Hold out testing can easily miss important model problems, particularly when processes vary over time, as credit risk does.

In the following section, we describe a validation framework that accounts for variations across both time and across the population of obligors[5].

Figure 3: Schematic of out of sample validation techniques



Source: Moody's Risk Management Services

Moody's fits a model using a sample of historical data on firms and tests the model using both data on those firms one year later, and using data on new firms one year later (upper portion of exhibit). Dark circles represent model estimation data and white circles represent validation data.
We do "walk-forward testing" (bottom left) by fitting the parameters of a model using data through a particular year, and testing on data from the following year, and then inching the whole process forward one year.
The results of the testing for each validation year are aggregated and then resampled (lower left) to calculate particular statistics of interest.

A schematic of the framework is shown in Figure 3.

The figure breaks up the model testing procedure along two dimensions: (a) time (along the horizontal axis); and (b) the population of obligors (along the vertical axis). The least restrictive validation procedure is represented by the upper-left quadrant, and the most stringent by the lower-right quadrant. The other two quadrants represent procedures that are more stringent with respect to one dimension than another.

The upper-left quadrant describes validation data chosen completely at random from the full data set. This approach assumes that the properties of the data stay stable over time (stationary process). Because the data are drawn at random, this approach validates the model across the population of obligors preserving its original distribution.

The upper-right quadrant describes one of the most common testing procedures. In this case, data for building a model are chosen from any time period prior to a certain date and validation data are selected from time periods only after that date. A model constructed with data from 1990 to 1995 and tested on data from 1996 through 1999 is a simple example of this out-of-time procedure.

Because model validation is performed with out-of-time samples, time dependence can be detected using different validation sub-samples. However, since the sample of obligors is drawn from the population at random, this approach also validates the model preserving its original distribution.

The lower-left quadrant represents the case in which the data are segmented into a model estimation set and a validation (out-of-sample) set containing no firms in common. If the population of the validation set is different from that of the model estimation set, the data set is out-of-universe. An example of out-of-universe validation would be a model that was trained on manufacturing firms but tested on other industry sectors. This approach validates the model homogeneously in time and will not identify time dependence in the data.

Finally, the most flexible procedure is shown in the lower-right quadrant. In addition to being segmented in time, the data are also segmented across the population of obligors. An example of this approach[6] would be a model constructed with data for all rated manufacturing firms from 1980 to 1989 and tested on a sample of all retail firms rated Ba1 or lower for 1990 to 1999.

Because default events are infrequent and default model outputs for consecutive years are highly correlated, it is often impractical to create a model using one data set and then test it on a separate 'hold-out' data set composed of completely independent cross-sectional data. While such out-of-sample and out-of-time tests would unquestionably be the best way to compare models' performance, default data are rarely available. As a result, most institutions face the following dilemma:

◆ If too many defaulters are left out of the in-sample data set, estimation of the model parameters will be seriously impaired and over-fitting becomes likely.

◆ If too many defaulters are left out of the hold-out sample, it becomes exceedingly difficult to evaluate the true model performance due to severe reductions in statistical power.

In light of these problems, an effective approach is to rationalise the default experience of the sample at hand by combin-

Figure 1: Types of errors

	ACTUAL Low Credit Quality	High Credit Quality
MODEL Low Credit Quality	Correct Prediction	Type II Error
MODEL High Credit Quality	Type I Error	Correct Prediction

Figure 2: Cost of errors

	ACTUAL Low Credit Quality	High Credit Quality
MODEL Low Credit Quality	Correct Assessment	Opportunity costs, and lost potential profits. Lost interest income and origination fees. Premature selling at disadvantageous prices.
MODEL High Credit Quality	Lost interest and principle through defaults. Recovery costs. Loss in market value.	Correct Assessment

Source: Moody's Risk Management Services

ing out-of-time and out-of-sample tests.

The procedure we describe is often referred to in the trading model literature as "walk-forward" testing and works as follows[7]. Select a year, for example, 1989. Then, fit the model using all the data available on or before the selected year. Once the model form and parameters are established, generate the model outputs for all the firms available during the following year (in this example 1990).

Note that the predicted model outputs for 1990 are out-of-time for firms existing in previous years, and out-of-sample for all the firms whose data become available after 1989. Now move the window up one year, using all of the data through 1990 to fit the model and 1991 to validate it. The process is repeated using data for every year.

Finally, collect all the out-of-sample and out-of-time model predictions in a validation result set that can then be used to analyse the performance of the model in more detail. Note that this approach simulates the process by which the model will actually be used in practice. Each year, the model can be refitted and used to predict the credit quality of all known obligors, one year hence. The process is outlined in the lower left of Figure 4.

For example, for Moody's Public Firm Default Risk Model, we selected 1989 as the first year for which to construct the validation result set. Following the above procedure, we constructed a validation result data set containing over 54,000 observations (firm years), obtained from a sample representing about 9,000 different firms, and including over 530 default events from Moody's extensive database.

Once a result set of this type has been produced, a variety of performance measures of interest can be calculated. It is important to note that the validation set is itself a sub-sample of the population and, therefore, may yield spurious model performance differences based only on data anomalies. Several resampling techniques are available to leverage the available data and reduce the dependency on the particular sample at hand[8].

A typical resampling technique proceeds as follow[9]. From the result set, a sub-sample is selected at random. The selected performance measure (for example, the number of defaults correctly predicted) is calculated for this sub-sample and recorded.

Another sub-sample is then drawn, and the process is repeated. This continues for many repetitions until a distribution of the performance measure is established. A schematic of this validation process is shown in Figure 4.

Resampling approaches provide two related benefits. First, they give an estimate of the variability around the actual reported model performance. In those cases in which the distribution of means converges to a known distribution, this variability can be used to determine whether differences in model performance are statistically significant using familiar statistical tests. In cases where the distributional properties are unknown, non-parametric permutation type tests can be used instead.

Second, because of the low numbers of defaults, resampling approaches decrease the likelihood that individual default events (or non-defaults) will overly influence a particular model's chances of being ranked higher or lower than another model.

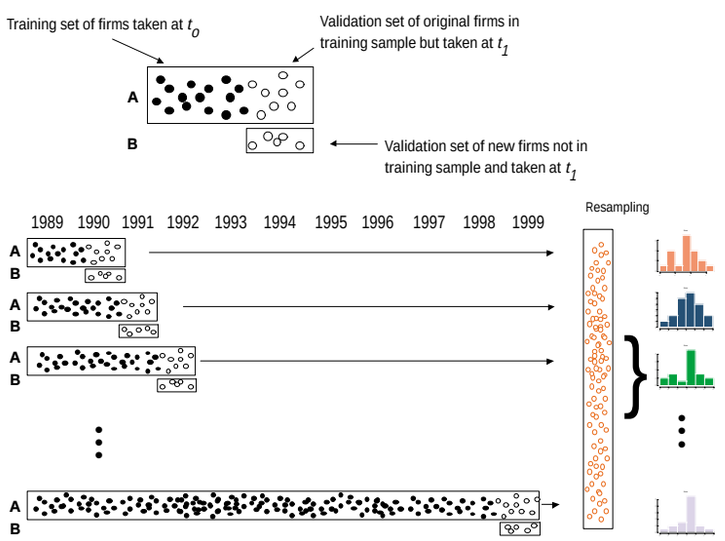
Model performance and benchmarking

We introduce four objective metrics for analysing information redundancy[10], and measuring and comparing the performance of credit risk models to predict default events: cumulative accuracy profiles, accuracy ratios, conditional information entropy ratios, and mutual information entropy. These techniques are quite general and can be used to compare different types of models.

In order to demonstrate the applicability of the methodology described here, we compared six univariate and multivariate models of credit risk using Moody's proprietary databases, including our default database and our credit modelling database. We compared the following models:

1) a simple univariate model based on return on assets (ROA)
2) reduced Z' score model[11] (1993)
3) Z' score model (1993)
4) a hazard model[12] (1998)
5) a variant of the Merton model based on distance to default[13],

Figure 4: Testing methodology: end-to-end

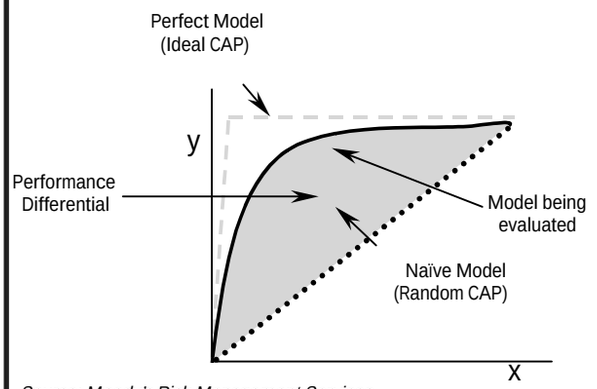


Source: Moody's Risk Management Services

Moody's fits a model using a sample of historical data on firms and tests the model using both data on those firms one year later, and using data on new firms one year later (upper portion of exhibit). Dark circles represent model estimation data and white circles represent validation data.

We do "walk-forward testing" (bottom left) by fitting the parameters of a model using data through a particular year, and testing on data from the following year, and then inching the whole process forward one year. The results of the testing for each validation year are aggregated and then resampled (lower left) to calculate particular statistics of interest.

Figure 5: Type I CAP curve



Source: Moody's Risk Management Services

The dark curved line shows the performance of the model being evaluated. It depicts the percentage of defaults captured by the model (vertical axis) vs. the model score (horizontal axis). The heavy dotted line represents the naïve case (which is equivalent to a random assignment of scores). The grey dashed line represents the case in which the model is able to discriminate perfectly and all defaults are caught at the lowest model output. The grey region represents the performance differential between the naïve model and the model being evaluated.

and
6) Moody's Public Firm model, a model based on ratings, market and financial information (2000).

These models represent a wide range of modelling approaches listed in order of complexity.

Cumulative Accuracy Profiles (CAPs)

Cumulative Accuracy Profiles (CAP) can be used to make visual qualitative assessments of model performance. While similar tools exist under a variety of different names (lift-curves, dubbed-curves, receiver-operator curves, power curves, etc) Moody's use of the term CAP refers specifically to the case where the curve represents the cumulative probability of default over the entire population, as opposed to the non-defaulting population only.

To plot a Type I Cumulative Accuracy Profiles, companies are first ordered by model score, from riskiest to safest. For a given fraction $x\%$ of the total number of companies, a CAP curve is constructed by calculating the percentage $y(x)$ of the defaulters whose risk score is equal to or lower than the one for fraction x. Figure 5 shows an example of a CAP plot.

A good model concentrates the defaulters at the riskiest scores and, therefore, the percentage of all defaulters identified (the y axis in the figure above) increases quickly as one moves up the sorted sample (along the x axis). If the model-assigned risk scores randomly, we would expect to capture a proportional fraction of the defaulters with about $x\%$ of the observations, generating a straight line or Random CAP (the dotted line in Figure 5).

A perfect model would produce the *Ideal CAP*, which is a straight line capturing 100% of the defaults within a fraction of the population equal to the fraction of defaulters in the sample. Because the fraction of defaulters is usually a small number, the ideal CAP is very steep.

One of the most useful properties of CAPs is that they reveal information about the predictive accuracy of the model over its entire range of risk scores for a particular time horizon.

Figure 6 shows the CAP curves for several models using the validation sample. Similar results are obtained for the in-sample tests[14]. Note that Moody's model appears to outperform all of the benchmark models consistently.

Accuracy ratios

It is often convenient to have a single measure that summarises the predictive accuracy of a model. To calculate one such summary statistic, we focus on the area that lies above the Random CAP and below the model CAP. The more area there is below the model CAP and above the Random CAP, the better the model is doing overall (*see Figure 5*).

The maximum area that can be enclosed above the Random CAP is identified by the Ideal CAP. Therefore, the ratio of the area between a model's CAP and the random CAP to the area between the ideal CAP and the random CAP summarises the predictive power over the entire range of possible risk values. We refer to this measure as the Accuracy Ratio (AR), which is a fraction between 0 and 1. AR values close to 0 display little advantage over a random assignment of risk scores, while those with AR values near 1 display almost perfect predictive power. Mathematically, the AR value is defined as

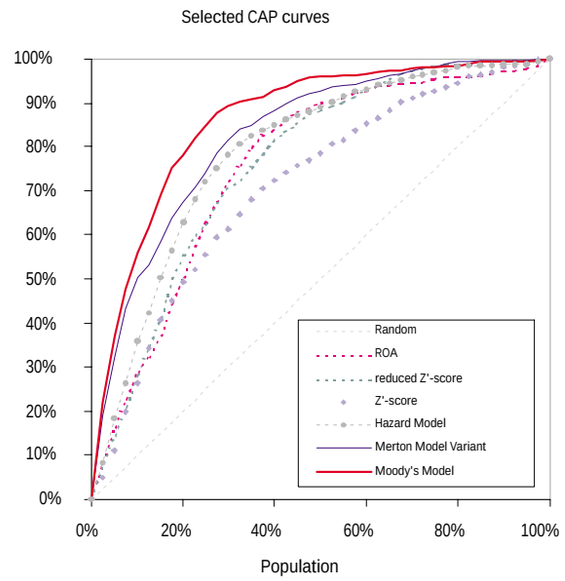
$$AR = \frac{2\int_0^1 y(x)dx - 1}{1-f} = \frac{1 - 2\int_0^1 z(x)dx}{f}$$

Here $y(x)$ and $z(x)$ are the Type I and Type II CAP curves for a population x of ordered risk scores, and $f = D/(N+D)$ is the fraction of defaults, where D is the total number of defaulting obligors and N is the total number of non-defaulting obligors. Note that our definition of AR provides the same performance measure for Type I and Type II CAP curves.

In a loose sense, AR is similar to the Kolmogorov-Smirnov (KS) test designed to determine if the model is better than a random assignment of credit quality. However, AR is a global measure of the discrepancy between the CAPs while the KS test focuses only on the maximum discrepancy and can be misleading in cases where two models behave quite differently, as they cover more of the data space from low risk to high risk model outputs. Also notice that, because the comparison of ARs is relative to a data set, our definition of the AR is not restricted to having completely independent samples as in the KS test[15].

Most of the models we tested had ARs in the range of 50% to 75% for (out-of-sample and out-of-time) validation tests. The results we report here are the product of the resampling approach described in the previous section. Thus, in addition to the reported value, we are also able to estimate an error bound for the statistic through resampling. We found that the maximum absolute deviation of the AR is of the order of 0.02 for

Figure 6: CAP curves for the tested models



Source: Moody's Risk Management Services

This composite figure shows the CAP curves for six models. All models were tested on the same data set. The 45° dashed grey line represents the naïve (which is equivalent to a random assignment of scores). Note that Moody's model performs better than the Merton model variant at discriminating defaults in the middle ranges of credits.

most models[16].

Table 1 shows AR values for the tested models for in-sample[17] and validation tests. To confirm the validity of the AR figures, we also checked whether a particular model differed significantly from the one ranked immediately above it by calculating KS statistics, using about 9,000 independent observations selected from the validation set. More precisely, KS tests showed that only the reduced Z'-score and ROA were not significantly different.

Conditional information entropy ratio

A different performance measure is based on the information about defaults contained in the distribution of model scores, or information entropy (IE). Intuitively, the information entropy measures the overall "amount of uncertainty" represented by a probability distribution. In the same way we reduced the CAP plot to a single AR statistic, we can reduce the information entropy measures into another useful summary statistic: the Conditional Information Entropy Ratio[18] (CIER).

To calculate the CIER, we first calculate the information entropy $H_0 = H_1(p)$ without attempting to control for any knowledge that we might have about credit quality. Here P is the aggregate default rate of the sample and H_1 is the information entropy defined in the Appendix [19]. This entropy reflects knowledge common to all models: the likelihood of the event given by the probability of default. We then calculate the information entropy $H_1(R)$ after having taken into account the risk scores $R = \{R_1,\ldots,R_N\}$ of the selected model. The CIER is defined as[20]

$$CIER(R) = \frac{H_0 - H_1(R)}{H_0}$$

If the model held no predictive power, the CIER would be 0. In this case the model provides no additional information on the likelihood of default that is not already known. If it were perfectly predictive, the CIER would be 1. In this case, there would be no uncertainty about the default event and, therefore, perfect default prediction. Because CIER measures the reduction of uncertainty, a higher value indicates a better model. Table 2 shows the CIER results. CIER errors are of the order of 0.02 and are obtained with a resampling scheme similar to the one described for the AR statistic.

Mutual information entropy

To this point we have been describing methods of comparing models on the assumption that the best performing model would be adopted. However, it is not unreasonable to question whether a combination of models might perform better than any individual one. Two models may both predict 10 out of 20 defaulters in a sample of 1,000 obligors.

Unfortunately, this information does not provide guidance on which model to choose. If each model predicted a different set of 10 defaulters, then using both models would have double the predictive accuracy of either model individually[21]. In practice, there is considerable overlap, or dependence, in what two models will predict for any given data sample.

To quantify the dependence between any two models A and B, we use a measure called the mutual information entropy (MIE). The mutual information entropy is a measure of how much information can be predicted about model B given the

Table 1: Selected Accuracy Ratios

	In-sample AR	Validation AR
ROA	0.53	0.53
Reduced Z'-Score	0.56	0.53
Z' -Score	0.48	0.43
Hazard model	0.59	0.58
Merton Model Variant	0.67	0.67
Moody's Model	0.76	0.73

Source: Moody's Risk Management Services

Table 2: Selected Entropy Ratios

	In-sample CIER	Validation CIER
ROA	0.06	0.06
Reduced Z'-Score	0.10	0.09
Z' -Score	0.07	0.06
Hazard model	0.11	0.11
Merton Model Variant	0.14	0.14
Moody's Model	0.21	0.19

Source: Moody's Risk Management Services

output of model A. MIE is defined as

$$MIE(r,R) = \frac{1}{H_0}\left(H_1(r) + H_1(R) - H_2(r,R)\right)$$

where r and R are the risk score sets of models A and B respectively, and $H_2(r,R)$ is the joint entropy defined in the Appendix. Because the MIE is calculated with the joint conditional distribution of models A and B, this measure requires a large number of defaults to be accurate. When default data are not widely available, this requirement can be relaxed by including reliable degrees of credit quality, such as agency ratings, instead of defaults only.

If models A and B are independent, the mutual information entropy is zero, while if model B is completely dependent on model A then MIE = 1-CIER(A). The additional uncertainty generated by model B can be estimated by comparing with the uncertainty generated by model A alone. In this context, the statistic serves much the same function as a correlation coefficient in a classic regression sense. However, the MIE statistic is based on the information content of the models.

Table 3 shows the difference D = MIE(A,B) – MIE(A,A), where A is Moody's model and B is any of the other selected models. In this example, we have compared all the benchmark models to Moody's model to determine if they contain redundant information.

Table 3: Difference of Mutual Information Entropy

	In-sample MIE	In-sample D	Validation MIE	Validation D
ROA	0.96	0.17	0.97	0.16
Reduced Z'-Score	0.93	0.14	0.96	0.15
Z' -Score	0.95	0.16	0.98	0.17
Hazard model	0.91	0.12	0.92	0.11
Merton Model Variant	0.87	0.08	0.87	0.06
Moody's Model	0.79	0	0.81	0

Source: Moody's Risk Management Services

The additional uncertainty generated by a model can be estimated by comparing it with the uncertainty generated by Moody's model alone. Table 3 shows the difference D = MIE(A,B) - MIE(A,A), where A is Moody's model and B is any of the other selected models.

Summary

The benefits of implementing and using quantitative risk models cannot be fully realised without an understanding of how accurately any given model represents the dynamics of credit risk. This makes reliable validation techniques crucial for both commercial and regulatory purposes.

In the course of our research into quantitative credit modelling, we have found that simple statistics[22] (such as the number of defaults correctly predicted) are often inappropriate in the domain of credit models. As a result, we have developed several useful metrics that give a sense of the value added by a quantitative risk model.

The four such measures presented here permit analysts to assess the amount of additional predictive information contained in one credit model versus another. In situations where a specific model contains no additional information relative to another, the less informative should be discarded in favor of the more informative. In the special case where both models contribute information to each other, users may wish to combine the two to garner additional insight. ■

Jorge Sobehart is vice president, senior analyst, risk management services at Moody's Investors Service

e-mail: sobeharj@moodys.com

Sean Keenan is vice president, senior analyst, risk management services at Moody's. Roger Stein is vice president, senior credit officer, and director of quantitative modelling analytics

A full version of this paper, including a mathematical description of information entropy and a full list of references, is available from the authors. Contact +44 (0)20 7772 5454.

FOOTNOTES

1 See, for example, Herrity, Keenan, Sobehart, Carty and Falkenstein (1999).
2 Basel, op. cit., p. 50.
3 Accuracy may be only one of many measures of model quality. See Dhar and Stein (1997).
4 In-sample refers to observations used to build a model. Out-of-sample refers to observations that are not included in the in-sample set. Out-of-universe refers to observations whose distribution differs from the in-sample population. Out-of-time refers to observations that are not contemporary with the in-sample set.
5 This presentation follows closely that of Dhar and Stein (1998), Stein (1999), and Keenan and Sobehart (1999), with additional clarifications from Sobehart, Keenan and Stein (2000).
6 This case is particularly important when one type of error is more serious than another. To illustrate, an error of two notches for an Aa-rated credit is generally less costly than a similar error for a B-rated credit.
7 See Sobehart, Keenan and Stein (2000).
8 The bootstrap (e.g., Efron, B. and R. J. Tibshirani (1993)), randomisation testing (e.g., Sprent, P. (1998)), and cross-validation (ibid.) are all examples of resampling tests.
9 See, for example, Herrity, Keenan, Sobehart, Carty and Falkenstein (1999).
10 See Keenan and Sobehart (1999).
11 For the definition of the original Z score and its various revisions Z' see Caouette, Altman, Narayanan (1998).
12 For simplicity we selected the model based on Zmijweski's variables described in Shumway (1998).
13 For this research, Moody's has adapted the Merton model (1974) in a similar fashion to which KMV has modified it to produce their public firm model. More specifically, we calculate a Distance to Default based on equity prices and firm's liabilities. See also Vasicek (1984) and McQuown (1993). For an exact definition of Moody's distance to default measure see Sobehart, Stein, Mikityanskaya and Li (2000).
14 Here in-sample refers to the data set used to build Moody's model.
15 In fact, AR based on panel data sets will provide aggregated information about the time correlation of the risk scores.
16 Due to the high levels of correlation in the resampling, the maximum absolute deviation gives a more robust estimate of an error range than a corrected standard error.
17 Here in-sample refers to the data set used to build Moody's model.
18 This is similar to measures such as gain ratios used in the information theory and time series analysis literature (see, for example, Prichard and Theiler (1995)). However, our definition measures explicitly the uncertainty to predict defaults instead of the overall uncertainty in the distribution of model outputs.
19 For additional details see Keenan and Sobehart (1999).
20 CIER = 1- IER, where IER is the information entropy ratio defined in Herrity, Keenan, Sobehart, Carty and Falkenstein (1999). Here we introduce CIER for consistency with the concept of conditional entropy in Information Theory and Communication Theory.
21 Of course combining the models could also create ancillary trade-offs with respect to increased Type II error.
22 For an example of a more standard approach to validation see: Caouette, Altman and Narayanan (1998).