

# EVIDENCE ON THE INCOMPLETENESS OF MERTON-TYPE STRUCTURAL MODELS FOR DEFAULT PREDICTION

TECHNICAL PAPER 1-2-1-2000

## ABSTRACT

### AUTHOR

Roger M. Stein

One may argue that structural models of default incorporating equity market information, such as the Merton (1974) model, are complete in the sense that univariate (single-factor) models are thought to be sufficient to capture all significant aspects of the future prospects for a firm. This assertion is testable empirically. In this short paper we provide some evidence that unmodified, Merton-type models are not, in fact, complete in the sense that additional information provides better discrimination between defaulters and non-defaulters even when conditioned on Merton-based variables. Using Moody's extensive database of corporate defaults, we first show heuristically that partitioning a standard Merton model by a second variable provides more information about default. We then show that econometric tests of significance refute the assertion that additional information does not help explain default. Finally, we show that even a simple regression-based multi-factor model appears to outperform its single-factor (basic Merton-only) counterpart in rigorous (out-of-sample and out-of-time) validation. This suggests merit to exploring enhancements to the Merton framework such as, for example, those introduced by the Vasicek-Kealhofer model.

© 2005 Moody's KMV Company. All rights reserved. Credit Monitor®, EDFCalc®, Private Firm Model®, KMV®, CreditEdge, Portfolio Manager, Portfolio Preprocessor, GCorr, DealAnalyzer, CreditMark, the KMV logo, Moody's RiskCalc, Moody's Financial Analyst, Moody's Risk Advisor, LossCalc, Expected Default Frequency, and EDF are trademarks of MIS Quality Management Corp.

#### **NOTE ON REVISIONS**

This paper was originally released by Moody's Risk Management Services in 2000, prior to the acquisition of KMV by Moody's. Since that time, Moody's KMV has conducted additional research in this area and we have revised the content of the article to reflect this research. We have also edited some of the language in the paper to better reflect differences between generic Merton-type models of the sort tested in this paper and the VK model to which Moody's did not have access at the time of the original publication; nonetheless, the main results are the same as in previous versions of the paper.

Published by:  
Moody's KMV Company

#### To Learn More

Please contact your Moody's KMV client representative, visit us online at [www.moodyskmv.com](http://www.moodyskmv.com), contact Moody's KMV via e-mail at [info@mkmv.com](mailto:info@mkmv.com), or call us at:

FROM NORTH AND SOUTH AMERICA CALL:  
1 866 321 MKMV (6568) or 415 296 9669

FROM EUROPE, THE MIDDLE EAST, AND AFRICA CALL:  
44 20 7778 7400

FROM ASIA, NEW ZEALAND, AUSTRALIA AND INDIA CALL:  
813 3218 1160

# 1 INTRODUCTION

The problem of default prediction has been an active one in the finance literature for many years. Most academic accounts credit Beaver (Beaver 1966) with developing the first (univariate) statistical model of financial distress, although Altman's (Altman 1968) Z-score discriminant model has become, for most, the prototypical statistical model in the field.

Several years after Altman published his research, an alternative approach to default prediction was suggested by Black and Scholes (Black and Scholes 1973) in their seminal paper on options pricing theory. This work was then extended by Merton (Merton 1974). These authors made the insightful observation that a firm's equity could be viewed simultaneously as an option on the underlying value of the firm, with an expiration time equal to the default horizon. Under this framework, equity holders would only be incented to repay the maturing debt obligations in cases where the option was "in the money" with respect to the market value of the firm. This would only be the case when the market value of a firm's assets exceeded the par value of the maturing liabilities. Otherwise the equity holders would elect to let the option "expire," thus defaulting on the liabilities.<sup>1</sup> The probability that this "option" is out of the money is theoretically related to the probability that the firm will default.

The structural model of Merton represents a framework for determining a default probability for a leveraged firm. While theoretically elegant, it turns out to produce probabilities that are unrealistic in practice as well as implying debt spreads that are at significant variance from those observed empirically (see: (Kim, Ramaswamy and Sunderasan 1993)). This is due, in part to the breakdown of certain strong assumptions that underlie the model.

Proponents of the Merton-based approach sometimes contend that the Merton model, in its original form, is both straightforward to implement and sufficient to explain all of the variability in the default process.

Such assertions typically rely on several implicit assumptions, the most prominent of which are that (a) equity markets (on average) contain complete information about the credit quality of the firm and do not contain non-credit related information<sup>2</sup>; and (b) the original Merton framework is the correct one with which to completely decode the market information and translate it into credit evaluations. These are strong assumptions.

However, the purpose of this paper is not to test whether these assumptions hold empirically. Here we set the much more modest goal of examining the single-factor Merton-based default model to determine whether it is a complete model or whether additional variables increase the predictive power in the same way that the Fama-French multifactor model explained more equity return variability than the CAPM.

It is important to note that here we use an implementation of the original Merton model. More complex versions of the model, such as the Vasicek-Kealhofer model implemented by Moody's KMV can be shown to be better behaved both with respect to default prediction and pricing (See Arora, Bohn, and Zhu (2005)). These modified versions of the original Merton framework are implemented in ways designed to counteract informational problems (e.g. transitory noise, periodic lack of trading, etc.) sometimes found in the equity markets. In this research we are examining whether the "plain-vanilla" implementation of Merton's model can be improved by adding additional information.

---

<sup>1</sup> In this context, the parameters for valuing the "option to default" are a strike price, given by the par value of the liabilities maturing at time  $T$ , and the current value and volatility of the underlying market value of the firm's assets. This later valuation (market value of assets), cannot be determined directly and is typically inferred numerically, based on the market value and volatility of a firm's equity and information about the capital structure of the firm.

<sup>2</sup> Note that this does not necessarily require market efficiency, although it does imply that investors would need to be relatively rational and motivated by similar objectives, investment horizons, and liquidity needs. Further, this would require that investors have, on average, access to the same information and interpret it similarly.

In the remainder of this paper, we examine this question using three approaches of increasing rigor: heuristic, econometric, and out-of-sample validation of predictive power. The paper proceeds as follows. In section 2, we show visually that a single financial ratio (ROA<sup>3</sup>) provides additional information about default probabilities when conditioning is done using a Merton variable. In section 3 we extend this analysis more formally using regression techniques and show that ROA reduces variance significantly, even in the presence of the Merton variable. In section 4 we show that the improvement we describe in section 3 is not just descriptive, but can be used predictively: the relationship is not only significant statistically, but has practical significance as well. We show this through a walk-forward out-of-sample analysis. We present conclusions in the final section, 5.

## 2 GRAPHICAL ANALYSIS (EDA)<sup>4</sup>

We begin by performing exploratory data analysis. While not rigorous, this provides some intuitive insight into the relationships under study.

In each year, we partition the firms in the data set into the bottom  $k\%$  of the population with respect to the Merton measure. Thus the bottom quantile contains the worst  $k\%$  of the firms as measured by the Merton variable and the top quantile contains the  $1-k\%$  best firms by that measure. Having so segmented the population, we further segment the lowest quantile into three sets, including the best and worst 20% of that quantile as measured by ROA (in this case Adjusted Net Income/Total Assets). Thus the bottom 20<sup>th</sup> percentile contains those firms that are both in the lowest  $k\%$  of the population as measured by the Merton variable *and* in the lowest 20% of that segment's ROA.

We then examined the default rate for each segment one year later. A typical example is shown graphically in, Figure 1 below, in which  $k=20\%$ .

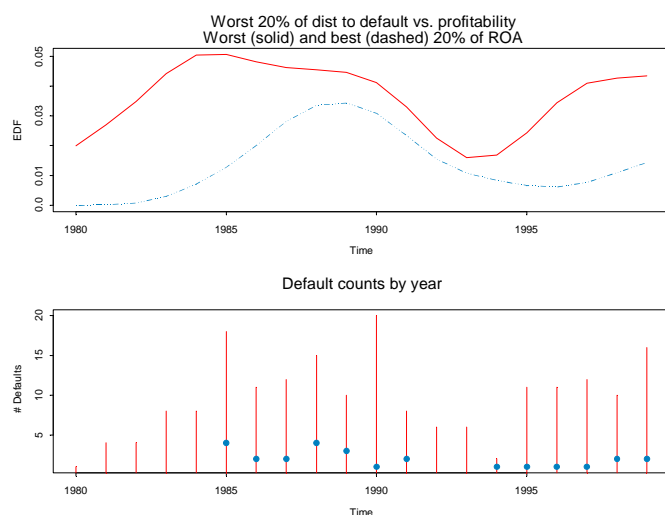


FIGURE 1 The impact of conditioning the Merton variable on another financial variable (ROA)  
 This figure demonstrates the impact of conditioning the Merton variable further on an additional financial variable. In each year there is a clear differential in default probability for those firms with high ROA and those with low ROA, after controlling for the bottom 20% of the Merton variable. The relationship inverts with low ROA firms defaulting at a higher rate. (Smoothed curves shown.)

<sup>3</sup> (Net Income – Extraordinary Items)/Total Assets.

<sup>4</sup> Note that the data set used in these analyses and those that follow were derived from the pre-acquisition data set. The combination of the Moody's and KMV default databases has resulted in a larger data set (cf. ,Dwyer and Stein (2003)).

Examining the figure, we observe three things. First there is a clear differential in default rates between those firms that have low ROA and those that have high ROA. This, in and of itself, is not surprising. However, the fact that this relationship obtains even for firms that have been already segmented by the Merton variable is more interesting. It implies that the predictive power of the Merton-based measure can be improved by including additional information.

We examined the relationship for varying values of  $k$ , the Merton variable conditioning percentile, from  $k=20\%$  to  $k=5\%$  and found that in general this relationship holds irrespective of the value of  $k$ . As  $k$  gets small, however, the number of data points in each year decreases drastically and in some years there are only two or three defaults in the segment, causing the results to be unreliable. We find, in general, that regardless of the bin size, adding a second predictor appears to increase the ability of the model to predict default.

While such analysis provides useful insights, it does not directly answer the question of whether the inclusion of another variable provides significant additional information beyond that already contained in the Merton variable.

In the next section, we examine this question in a traditional econometric framework.

### 3 REGRESSION TESTS

A common econometric method for testing whether additional variables add information is to use regression techniques of various sorts. A typical approach is to examine the results of a regression that includes the variables of interest to determine whether the additional variable provides a significant reduction of variance over those that do not include the variable.

For example, under some very general assumptions, those variables that have high t-values in the regression are said to be significant, and those with low values are said to not contribute to explaining the variable of interest. By convention, variables with t-statistics whose probabilities are below 5% are considered to be significant.

It should be noted that the t-test can be misleading under certain correlation structures. An F-test between a two models, one containing only the Merton variable and one containing both the Merton variable and ROA, provides an overall estimate of the joint significance of the variables in the regression. To the extent that the two variables do not jointly increase the explanatory power of the model over the single (Merton) variable, the F statistic will not be significant. Because the t-statistic tends to be more intuitive, we discuss both.

We conducted a series of regressions to evaluate the significance of the ROA measure in the presence of the Merton variable. First, we performed a cross-sectional analysis using the entire 20 years of data. Then we looked at cohorts of the data to ensure that any behavior we observed was not due to one specific historical period.

We took 5-year cohorts<sup>5</sup> of data starting in 1980 and estimated a series of logistic regressions in which the Merton variable and ROA<sup>6</sup> were regressed against a one year default flag. (The flag was set to 1 if a firm defaulted within one year of the observed variables and 0 if it did not.) This produced a series of 16 cohorts on which we estimated regression equations, one for each year from 1985 through 1999.

For each regression, we calculated t-statistics for the coefficient on ROA. To the extent that these statistics were significant, the regression rejected the assertion that the Merton variable contained all of the information necessary to explain default, (beyond random variations).

For the regression over the full 20-year period, the t-statistic on the ROA coefficient was -15.9, and the F statistic between the regression without ROA and with ROA was 2461, producing p values in both cases well beyond the 0.01% (0.0001) level and providing strong evidence that ROA adds significantly to the explanation of default, even in the presence of the Merton variable.

---

<sup>5</sup> Due to the small number of defaults, we needed to use more than one year to ensure sufficient statistical power. Note that the correlation in the data from period to period would make direct comparisons between two cohorts difficult, and we do not attempt to do this here.

<sup>6</sup> A simple variance reducing transform (trimming) was used here to control for outliers.

The results of the 5-year cohorts, shown in Table 1, also overwhelmingly refute the assertion that the Merton variable measure contains all information relevant to predicting default. In no year was the t-statistic on the ROA coefficient insignificant.

Over the 20-year period, the least significant regression result had a t-statistic of approximately 3.78 in absolute value ( $p < 0.00016$ ) and the median t-value was about 6 in absolute value ( $p \sim 2.0E-9$ ). The F statistics in each case were also highly significant, with probabilities well beyond the 0.0001 level.

TABLE 1 Results of Regression Tests

This table shows the results of the regression tests which examined the significance of the t-statistic on the ROA variable in the presence of the Merton variable. In all cases the additional variable was highly significant. Not shown are the results of the F tests which also had extremely small probabilities in all cases. These results provide strong empirical support for the assertion that the multifactor model explains more of the default behavior than the pure Merton model.

Cohort	t-stat	p( t )
UNIV	-15.90	7.6E-57
1984	-5.99	2.2E-09
1985	-7.68	1.7E-14
1986	-10.00	1.9E-23
1987	-8.82	1.3E-18
1988	-7.00	2.6E-12
1989	-6.33	2.5E-10
1990	-5.25	1.5E-07
1991	-5.32	1.0E-07
1992	-5.49	4.1E-08
1993	-4.96	7.3E-07
1994	-4.13	3.7E-05
1995	-3.78	1.6E-04
1996	-4.18	2.9E-05
1997	-5.26	1.5E-07
1998	-7.10	1.3E-12
1999	-8.35	6.9E-17

## 4 WALK-FORWARD TESTS

Although the specification tests in section 3 give some econometric evidence that the addition of a financial ratio (ROA) explained to a statistically significant degree more of the default process beyond the Merton model alone, they do not directly address the practical implications of the result.

It is reasonable to question whether the statistics relationships are merely providing a "rear view mirror" description of the process or whether the additional information contained in the non-Merton variables is useful in actually *predicting* default rather than just describing it.

In this section we attempt to answer this question directly using a rigorous validation approach developed for model validation and selection. The approach, termed *walk-forward testing* allows us to closely approximate the on-line performance of a model of interest over an extended timeframe. We feel that the methodology provides realistic estimates of model predictive performance under conditions of actual model usage.

We provide details of head-to-head comparisons of predictive accuracy of the pure Merton variable versus a simple binary regression that includes both the Merton variable and the financial ratio ROA. Using resampling methods, we perform significance tests and report the results.

## 4.1 Overview of Testing Methodology

We generate a set of predictions using a walk-forward analysis. The walk-forward approach, described more fully in (Stein 2002), proceeds as follows: The data is segmented into "in-time" and "out-of-time" samples. The in-time sample includes all data up through time  $t$ . The model parameters are estimated using the in-time sample. Predictions for the data in the out-of-time sample are generated for those records dated between time  $t$  and time  $t+I$ . The predictions are collected along with the actual observed outcomes. The in-time sample is then extended to include all records prior to time  $t+I$  and the process is repeated.

For example, a default model could be fit using all data prior to 1990. Predictions are made for all data in 1991. The predictions and actual outcomes are collected. The model would then be re-fit using data prior to 1991 and predictions made for 1992, and outcomes collected. The process would be repeated, walking forward through the entire data set.

Once the results have been collected, we calculate various comparative metrics of interest. Using resampling approaches we also perform significance tests on the results. We use data from the period 1990 through 1999 for these tests (over 50,000 firm years). The process is described more fully in Figure 2.

Note that this approach has two major benefits with respect to default modeling. First, it allows us to get a realistic view of how a particular model would perform over time. In most contexts, the parameters of models are adjusted as new data becomes available and as credit cycles evolve. This approach closely mimics that process while at the same time preventing overfitting in the modeling (no data used to fit the model should be used to test it). Thus it tests not only a particular model, but also the modeling methodology that is being used.

Secondly, it allows us to leverage to a higher degree the available (but rare) default event data. Since defaults are infrequent (typically between 1% and 2% per year, historically), a modeler is often faced with the hard choice of using more (fewer) defaults to parameterize a model and fewer (more) to test it. In the case that more are used for parameterization, testing typically suffers due to reduced power. In the case that more are used for testing, model quality typically suffers due to statistical instability.

The approach outlined above often permits a modeler to make adequate use of default data for model fitting, while also providing sufficient statistical power for testing.

## 4.2 Overview of Statistical Test of Predictive Power

We perform a non-parametric statistical test to compare the pure Merton model with the enhanced one. To avoid issues of mapping model scores to default probabilities, we chose tests based solely on discriminatory power. The metric of interest we measure is the *accuracy ratio*.

Since these statistics are somewhat variable with respect to the samples on which they are calculated, in addition to reporting the results of the comparison, we also bootstrap (Efron and Tibshirani 1993) the result set samples. This permits us to determine the degree to which the observed difference in performance may be explained by chance. The bootstrap facilitates Wilcoxon's signed rank test, a simple and well-known non-parametric permutation-based test.<sup>7,8</sup>

---

<sup>7</sup> See, for example, (Maritz 1995) or (Sprent 1998) for a discussion of distribution free and resampling based statistical procedures.

<sup>8</sup> We chose distribution-free approaches in order to avoid requiring assumptions that may not be met in practice (e.g., those relating to distribution of defaults).

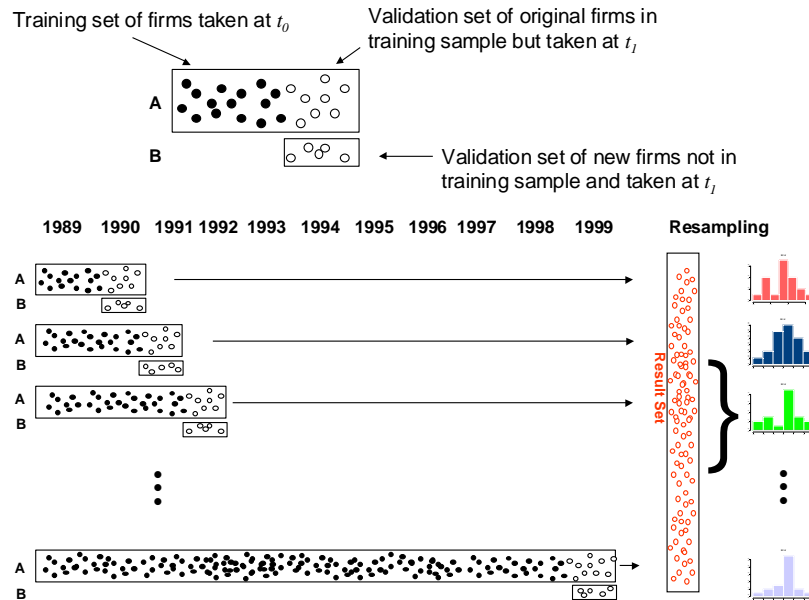


FIGURE 2 Moody's testing methodology: end-to-end (adapted from Sobehart, Keenan, and Stein (2000))  
 The model is fit using a sample of historical data on firms and tests the model using both data on those firms one year later, and data on new firms one year later (upper portion of Figure). Dark circles represent "in-time" data used to fit a model and white circles represent "out-of-time" 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. We do this to address issues of overfitting. The results of the testing for each validation year are aggregated and then resampled (lower left) to calculate particular statistics of interest.

We start by bootstrapping 1000 replications from the original sample. We then calculate a pair of accuracy ratios, one for each model, for each replication and calculate the difference in accuracy ratios<sup>9</sup> for each of the pairs.

For example, the first of the 1000 differences was calculated by drawing a bootstrap sample (with replacement) from the original set predictions (calculated using a walk-forward analysis) for both models. The predictions were drawn in pairs to ensure that both models were run on the same data set. The resulting set of paired predictions was then used to calculate an accuracy ratio for each model. The difference of these two accuracy ratios was recorded.

Wilcoxon's signed rank test is similar to a simple median test. The test involves ranking the *absolute values* of each of the differences in accuracy ratios. The sign of the original difference is then attached to these ranks, and the sum of the signed ranks is taken.

Under the assumption that the multi-factor model is no better at predicting default than the pure Merton model, there should be about as much of a chance of a positive difference (due to sampling error) as a negative difference in accuracy ratios, and the ranks should be distributed evenly with a sum of about zero. In the event that the sum were much greater than zero, it would imply that more of the "big" differences were positive (favoring the multi-factor model), providing evidence that the multi-factor model were more predictive.

To determine the significance of the statistic, we take advantage of a normal approximation, converting the statistic into a familiar Z statistic with mean zero and variance of unity (see, for example, (Sprent 1998)).

<sup>9</sup> Accuracy ratios measure the percentage of the theoretically perfect model that a particular model captures.

## 4.3 Results

The results of the resampling tests are shown in Table 2, below.

TABLE 2 Results of resampling tests

This table shows the results of the resampling tests for the Wilcoxon's Signed Rank test (converted to Z score) for different numbers of replications. In all cases the probabilities are extremely small providing strong support for the assertion that the multi-factor model outperforms the pure Merton variant.

Replications	Z	p(Z)
100	5.17	<< 0.0001
500	10.70	<< 0.0001
1000	15.40	<< 0.0001

In all cases the probabilities are extremely small. The probabilities describe the statistical likelihood that differences in model performance were due to chance, i.e., that the multifactor model performs no differently than the pure Merton variant. The very small probabilities therefore provide strong support for the assertion that the multifactor model outperforms the pure Merton variant<sup>10</sup>.

It is interesting to note that although the power of the tests increased as the sample size grew, the relative proportion of positive differences, about 70%, did not fluctuate much from sample size to sample size.

Note that although we have presented very strong evidence that the multi-factor model outperforms the pure Merton variant at levels well beyond those attributable to chance, there were a non-trivial number of individual cases in which it did not.

## 5 CONCLUSIONS

An interesting practical question is whether a traditional Merton model is complete in its information content and the representations it uses to transform that content into default prediction.

We provided several forms of empirical analysis in an effort to examine whether a naïve alternative multi-factor model that includes only a single additional variable, ROA, provides additional predictive and explanatory power.

Even with a very crude alternative multi-factor model, the evidence is strong that the inclusion of additional information can improve significantly both the explanatory and predictive power of a model. Note that this is consistent with the theoretical work of Duffie and Lando (2001). Also note however, that this does not imply that the additional information must enter the model in an additive form, as was done here. In fact, it is possible to incorporate this information within the structural framework of a model to produce enhanced performance while maintaining the elegance of the structural framework. Again, the implementation of the Vasicek-Kealhofer model at Moody's KMV is an example of how modifications are made to the classical Merton framework to improve the model's performance. For many applications, this integration may be preferable to an ex post introduction of the information.

---

<sup>10</sup> We note that for completeness we also tested a *univariate regression* approach (a regression containing only the Merton variable) against the multifactor regression alternative in the walk-forward context. However, the pure form of the model, which we used here (and which conforms to current usage conventions), gave superior predictive power to the univariate regression. Nonetheless, neither the regression nor the pure version of the single-factor model outperformed the multi-factor approach.



# REFERENCES

---

---

1. Altman, E. I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *Journal of Finance* (September): 589-609.
2. Arora, N., J. Bohn, and F. Zhu. (2005). "Reduced Form vs. Structural Models of Credit Risk: A Case Study of Three Models." MKMV Working Paper.
3. Beaver, W. H. (1966). "Financial Ratios as Predictors of Failure." *Journal of Accounting Research* .
4. Beck, J. R. and E. K. Shultz (1986). "The use of relative operating characteristic (ROC) curves in test performance evaluation." *Archives of Pathology & Laboratory Medicine* 110(January).
5. Black, F., and Scholes, M. (1973). "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* (81): 637-659.
6. Crosbie, P. and J. Bohn (2003). Modeling Default Risk, KMV Corporation.
7. Duffie, D. and D. Lando (2001), Term Structures of Credit Spreads with Incomplete Accounting Information, *Econometrica*, 69, 633—664..
8. Dwyer, Douglas W. and Roger M. Stein "Inferring the Default Rate in a Population by Comparing Two Incomplete Default Databases", Moody's KMV, New York, 2003.
9. Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York, Chapman & Hall.
10. Hanley, J. A. and B. J. McNeil (1982). "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve." *Radiology* (April): 29-36.
11. Kim, J., Ramaswamy, K., and Sunderasan, S. (1993). "Does Default Risk in Coupons Affect the Valuation of Corporate Bonds? A Contingent Claims Model." *Financial Management* : 117-131.
12. Maritz, J. S. (1995). *Distribution Free Statistical Methods*. New York, Chapman & Hall.
13. Merton, R. C. (1974). "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance* (29): 449-470.
14. Miller, R. (1998). "Refining Ratings." *RISK* (August).
15. Sprent, P. (1998). *Data Driven Statistical Methods*. London, Chapman & Hall.
16. Stein, Roger M. (2002). "Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation". Technical Report. New York: Moody's KMV.