Original Article

# A simulation study of Basel II expected loss distributions for a portfolio of credit cards

Received (in revised form): 20th October 2009

**Tony Bellotti**

gained a BSc honours degree in Computer Science with Mathematics from the University of Bristol in 1987 and spent several years in industry as a business analyst/programmer. He received his PhD in computational learning from Royal Holloway, University of London in 2006. He is currently a research fellow working on statistical and machine learning approaches to consumer credit risk in the Credit Research Centre at the University of Edinburgh.

**ABSTRACT** **Credit scoring models have been used traditionally as the basis of decisions to reject or accept credit applications. They are also used to categorize applicants or existing accounts into risk groups. Based on estimates of probability of default (PD), the risk groups may seem well separated. However, by considering distributions on risk elements such as model estimation uncertainty, exposure at default and loss given default, a simulation approach is used to compute Basel II expected loss distributions for a portfolio of credit cards. These show that discrimination between risk groups is not as clear as is immediately suggested simply by PD estimates. Based on these distributions, we also show that measuring extreme credit risk with Value at Risk can lead to considerable underestimation if distributions on these risk elements are not entered into the computation.**

*Journal of Financial Services Marketing* (2010) **14,** 268–277. doi:10.1057/fsm.2009.25

**Keywords:** consumer credit; Basel II; simulation; expected loss

## INTRODUCTION

One of the problems that has been identified as a cause of the current credit crunch is the financial industry's over-reliance on complex quantitative models of financial risk and the wide and complacent use of risk management tools such as Value at Risk (VaR). It is claimed that VaR has been treated as reliable even when it was evident that not all uncertainties in the risk model had been sufficiently accounted for, and that this has led to an underestimation of extreme risk.[1] Even simple quantitative models may lead to dangerous underestimation of risk if not all elements of uncertainty are properly considered. In this article the sensitivity of consumer credit portfolios to this problem is investigated.

Consumer credit scoring models have been used since the 1940s to determine risk of individual applicants for credit and to support the decision to accept or reject an application. From a historic credit database, obligors are classified according to whether or not they default on their loans. The scorecard model is then built to explain default with respect to application

**Correspondence:** Tony Bellotti
Credit Research Centre, University of Edinburgh Business School, William Robertson Building, 50 George Square, Edinburgh EH8 9JY, UK
E-mail: tony.bellotti@ed.ac.uk

 

information. Typically, variables such as income, age, housing and employment status are significant variables. Although exotic models such as neural networks and support vector machine have been used to build scorecards,[2] still the simple logistic regression remains a popular choice for score card modelling.[3] Credit scoring models can be used to categorize obligors into risk groups and a probability of default (PD) can be assigned to each. These often suggest good separation between groups with increasing PD for groups with higher risk.[4,5] However, as these PDs are point estimates at account level, they do not show how differentiated the groups are in terms of distribution of expected risk at the group level. Further, PD is only one component in the calculation of expected loss (EL) on an account given by

$$EL = PD \times EAD \times LGD \quad (1)$$

where EAD is exposure at default and LGD is loss given default, as specified by the Basel II Accord.[6,7] LGD is the fraction of EAD that is not recovered by the bank within a given period after the default event. Since these risk elements are omitted, assessing risk groups by mean PD gives poor insight into the true risk given by the credit model.

Little is known about the distribution of loss for consumer credit portfolios once model uncertainties are considered along with EAD and LGD. Rodriguez and Trucharte make this observation for loss distributions on mortgage portfolios in particular.[5] In this article we investigate EL distributions on a portfolio of credit cards using a simulation approach and review how these estimates change with the inclusion of model uncertainties, EAD and LGD. We show that the credit scoring model gives a more uncertain picture of credit loss than we might initially believe. Financial practitioners and regulators are interested in assessing ‘worst case’ risk and this is often measured using VaR. For a given confidence level, $\alpha$ per cent, VaR is defined as the worst loss we can expect to incur with probability $\alpha$ per cent. For example, if 99 per cent VaR is calculated, there is a 1 per cent chance that losses will exceed this value. It follows that VaR is calculated as the $\alpha$-percentile of the loss distribution. Consequently, if the relevant risk elements are not taken into account, the loss distribution may be too conservative and VaR for a credit portfolio will be underestimated.

Among the risk elements given in equation (1), PD is commonly modelled and discussed. However, models of LGD for consumer credit are rarely discussed in the academic literature[8,9] and models of EAD are even rarer, drawing little academic interest.[10] Often LGD or EAD or both are taken as fixed values in estimates of credit loss. A conservative approach can be taken whereby EAD is estimated as the credit limit on an account.[11] However, it is possible for an obligor to default well below or possibly above that limit and it is not clear why an obligor should default at the credit limit or how common this would be. Altman *et al*[12] refer to commercial models that take LGD as fixed, Rodriguez and Trucharte[5] use a fixed value of LGD when computing loss rates for mortgages, Jokivuolle *et al*[13] use empirical EAD but fix LGD for corporate loans, and Rösch[14] only reports default rate (DR) distributions and does not consider the effect of EAD and LGD on loss distributions. Finally, Basel II acknowledges the difficulty of estimation by allowing banks to use given supervisory values for LGD and EAD if they choose the foundational internal ratings-based (IRB) approach to capital requirements computation.[7] However, this article explores how these restrictions on risk elements affect the measure of credit risk. In particular, four risk elements in the computation of EL are considered.

1. Uncertainty at the account level: no model will exactly explain default for each

individual account, therefore the deviation of each obligor's tendency to default from that estimated by the model represents an uncertainty when assessing risk.
2. Model estimation uncertainty.
3. The distribution of EAD.
4. The distribution of LGD along with correlation with PD: there is considerable evidence that LGD is positively correlated with PD and this could impact EL.[12]

These risk elements are discussed in detail in the following section, along with the modelling and simulation approaches and details of the credit card data we use. In the subsequent section empirical results are given and conclusions are given in the final section.

## METHODS AND DATA

We use a logistic regression model of default given by the latent model structure

$$\begin{aligned} d_i^* &= \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \\ d_i &= \mathrm{I}(d_i^* > 0) \end{aligned} \quad (2)$$

where $d_i = 1$ if account $i$ defaults and is 0 otherwise, $\mathbf{x}_i$ is a vector of application variables for account $i$, $d_i^*$ is a latent variable, $\boldsymbol{\beta}$ is a vector of coefficients to be estimated, $\varepsilon_i$ is a residual drawn independently from a standard logistic distribution and represent the uncertainty at the account level, I(.) is the indicator function and $T$ is matrix transpose.[15] It follows that PD is given by

$$p_i = 1/(1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})) \quad (3)$$

A point estimate of DR across a test set of $n$ cases is then given by

$$\frac{1}{n} \sum_{i=1}^{n} p_i$$

We use a data set of 50 000 Brazilian credit card accounts that were made available for academic research for the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09) competition by Neurotech Ltd (Brazil). The data represent a sample drawn from a period of 1 year between 2003 and 2004 and includes application information such as age, income, occupation, residential status and telephone details, along with an indicator of default. Approximately one-fifth of accounts were recorded as defaults within the sample. Since we only have data on accepted credit card applications, we do not consider rejected applications. For application approval models this is a problem as modelling on just accepted data creates a bias and may require reject inference.[16] However, because we are considering credit loss on a portfolio of active accounts, the rejects are not relevant to the calculation and so reject inference is not required in this case.

We split the data randomly into a training data set and a test data set of 25 000 records each. A logistic regression model was constructed on the training data and was statistically significant with a log-likelihood ratio of 1720 and *P*-value less than 0.0001. The full list of covariates and coefficient estimates are given in Appendix A.

The coefficient estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ form a scorecard and the term $s_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ is a credit score for account $i$.[17] We categorize the cases into three approximately equal sized risk groups according to credit scores given in the training set. Risk groups 1 and 3 consist of the one-third least and most risky applications, respectively. In practice, more risk groups would usually be considered, but for this exercise we consider the simple case with just three to allow for straightforward reporting and analysis of results.

Estimation uncertainty can be measured by the standard errors on the coefficient estimates reported in Appendix A taking into account their covariance structure. However, a simpler alternative is to remodel default with the single covariate $s_i$. Clearly this model will have a coefficient estimate of 1 but, nevertheless, this estimate will have a standard error $\sigma_s$, fixed for all cases, that provides a measure of estimation uncertainty for the credit score model.

To compute EL on a credit card portfolio, estimates of EAD and LGD are needed. For this data set we do not have information about exposures or recovery and, as has been discussed, very little has been published on how to derive models for these. Nevertheless we can assume distributions about them. EAD has a value greater than zero and like other monetary values such as wealth or income it is reasonable to expect it to have a right skew. In particular, we expect a large minority of obligors to build up exceptionally larger arrears than the average. For these reasons we choose a log-normal distribution on EAD, as this distribution only allows values greater than zero and can model the right skew. Jiménez and Mencía also use an inverse Gaussian distribution for EAD, which is close to log-normal.[18] The standard deviation $\sigma_E$ of the distribution needs to be set and we consider two possibilities: 1.5 and 2. The first gives a conservative estimate of the right tail with the 99th percentile being only 34 times the median value of EAD, while $\sigma_E = 2$ is less conservative having a longer right tail with the 99th percentile being 108 times the median value. LGD distributions are better understood and are bimodal with many cases having a value of 0 or 1.[8,12,19] Since we are considering credit card data, we follow empirical evidence from Querci for LGD distributions for unsecured loans.[19] Thus LGD is modelled as a truncated normal distribution such that the percentage of 0s and 1s is 30 per cent and 42 per cent, respectively. Figure 1 shows a typical random draw from this distribution.

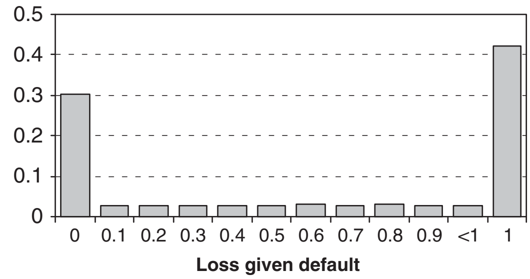


**Figure 1:** Simulated LGD distribution.

Monte Carlo simulation is used to construct an empirical distribution of EL on the test data set. We consider how the distribution changes as each risk element is included in the computation. Although PD is computed in equation (3), this does not allow us to consider random perturbations owing to the residual $\varepsilon_i$. For this reason we take the approach of simulating a default or non-default event, SD, for each test account based on equation (2) given a simulated value for $\varepsilon_i$, following Jokivuolle *et al.*[13] Therefore simulated EL on the whole data set of *n* cases is given by

$$L = \sum_{i=1}^{n} \mathbf{I}((1+m)\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \varepsilon_i > 0)\, e_i\, l_i \qquad (4)$$

where *m* is model estimation uncertainty drawn from $N(0, \sigma_s)$ or 0 if estimation uncertainty is not included in the simulation, $e_i$ and $l_i$ are EAD and LGD, respectively, drawn from the distributions discussed above, or set to 1 if either is not included in the simulation. Additionally the correlation between PD and LGD is managed using Cholesky decomposition before generation of $l_i$, given that PD is determined by equation (3). Details are given in Appendix B. The study by Miu and Ozdemir gives a guide to possible values for the PD/LGD correlation coefficient.[20] They measure a moderate level as 0.2 so we consider this value along with a higher value of 0.4.

Repeated simulations give different values of *L*, which together form the loss distribution. As is typical in the industry and following Basel, we evaluate worst case loss using VaR. The scale of EL from different computations is not directly comparable, therefore VaR is reported as a ratio of the median EL. In particular, we are interested in the difference between VaR and median EL and report relative VaR difference given by

$$\mathrm{VaR}_{\Delta} = \left( \frac{\mathrm{VaR}}{\mathrm{median\ EL}} - 1 \right) \times 100\%. \quad (5)$$

## RESULTS

Table 1 shows point estimates of DR on the out-of-sample test data set for each risk group. These results suggest good separation between groups with a distinct DR within each group, which are accurate when compared with the observed DR.

**Table 1:** Point estimates of DR for each risk group in the test set compared with observed DR

| *Risk group* | *Expected DR* | *Observed DR* |
|---|---|---|
| All | 0.199 | 0.196 |
| 1 | 0.099 | 0.090 |
| 2 | 0.189 | 0.183 |
| 3 | 0.308 | 0.312 |

In our experiments, we found that 8000 simulations were sufficient to produce stable loss distributions. These are shown in Figure 2 for different specifications of credit loss computation. They show that when EAD and LGD are included in the computation, then the right tail of the distribution becomes more extended. Figure 2(a) shows a distribution generated with just random perturbations at the account level. The level of uncertainty captured by the residual is high, as the model cannot be expected to capture the direct reason for default, which may be a loss of job, negative equity or a sudden personal crisis such as illness or divorce. The credit scoring model can only model a *tendency* to default. For this reason it may be surprising that the distribution of estimated DR in Figure 2(a) is so narrow. However, this

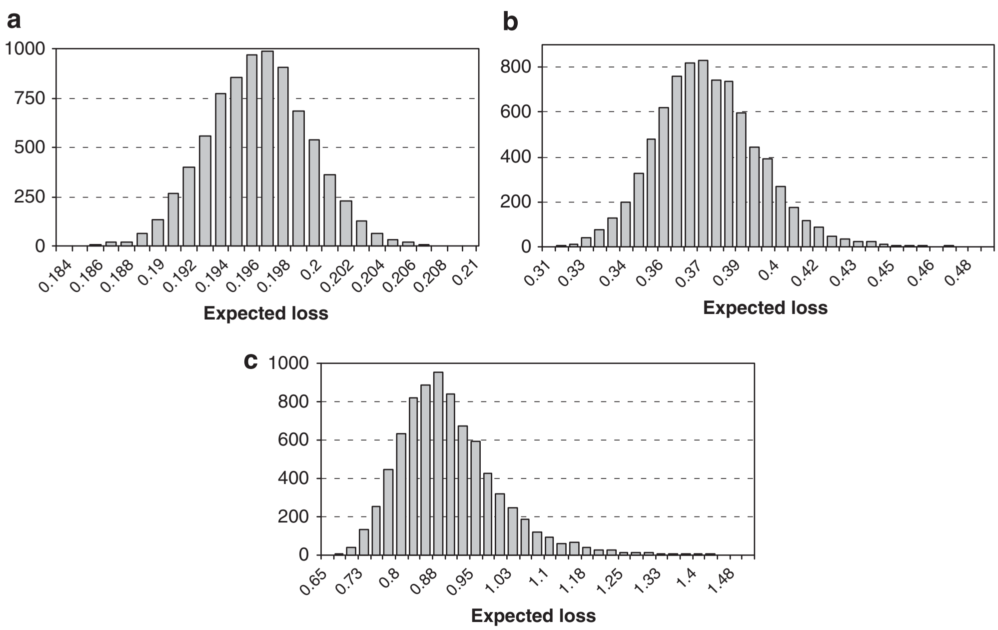


**Figure 2:** Loss distributions for different computations of EL: (**a**) SD×1×1; (**b**) SD×EAD (1.5)×LGD (0.2); (**c**) SD×EAD (2)×LGD (0.2).
*Notes*: The figure in brackets following EAD is the standard deviation of the EAD log-normal distribution $\sigma_E$ and the figure in brackets following LGD is the PD/LGD correlation coefficient. SD indicates simulation of default event.

  

**Table 2:** VaR for different loss calculations

| *Specification of loss calculation (EL)* | | | | | *Outcome* |
|---|---|---|---|---|---|
| *PD* | *×EAD* | *× LGD* | *LGD/PD correlation* | *Estimation uncertainty?* | *$VaR_{\Delta}$ at 99% level (%)* |
| PD | ×1 | ×1 | — | no | 1.4 |
| SD | ×1 | ×1 | — | no | 2.9 |
| SD | ×1 | ×1 | — | yes | 3.9 |
| SD | × EAD(1.5) | ×1 | — | yes | 12.0 |
| SD | × EAD(2) | ×1 | — | yes | 33.0 |
| SD | × EAD(1.5) | × LGD | 0 | yes | 13.8 |
| SD | × EAD(2) | × LGD | 0 | yes | 44.0 |
| SD | × EAD(1.5) | × LGD | 0.2 | yes | 13.7 |
| SD | × EAD(2) | × LGD | 0.2 | yes | 40.5 |
| SD | × EAD(1.5) | × LGD | 0.4 | yes | 14.5 |
| SD | × EAD(2) | × LGD | 0.4 | yes | 40.9 |

VaR is reported as difference between VaR and median EL as a ratio of median EL, as given in equation (5). The figure in brackets following EAD is the standard deviation of the EAD log-normal distribution $\sigma_E$. SD indicates simulation of default event.

distribution aggregates over the whole data set of accounts thus yielding low joint risk. In general, the larger the test data set size, the smaller the risk from individual account level uncertainty. Similarly, the larger the training data set size, the lower the estimation uncertainty.

Table 2 shows our main results with effect on VaR at a 99 per cent level for different specification of computation of EL. It shows that as risk elements are added, VaR increases, relative to the median estimate. The most pronounced increase occurs when the log-normal EAD distribution is included. The $VaR_{\Delta}$ rises from 3.9 per cent to 12 per cent or 33 per cent, depending on the standard deviation of the EAD distribution. Including the LGD distribution in the computation shows a modest increase in VaR, but this effect is more pronounced when linked with a less conservative specification of EAD ($\sigma_E = 2$). However, including the PD/LGD correlation does not generally increase VaR and in some cases VaR even decreases, regardless of the correlation coefficient used.

Figure 3 shows how well distributions for each risk group are separated. It is clear that the naïve computation without EAD or LGD, Figure 3(a), shows good separation between risk groups. It would imply room for the use of several more risk groups. However, when EAD and LGD are included in the computation, the spread of each distribution widens. With a conservative EAD distribution, Figure 3(b), the three risk groups are still quite distinct, but there is not much available room to expand to greater numbers of risk groups. With a less conservative specification of EAD, Figure 3(c), the overlap between the distributions is larger and the three groups are less distinct.

## CONCLUSION

Seemingly, point estimates from credit scoring models are reliable and allow us to separate applications and accounts meaningfully into distinct risk groups. However, once we consider distributions of EL given a variety of risk elements, we find these models give broad distributions and the boundaries between the different risk groups becomes fuzzy. This implies that a small number of risk groups should be used if we want to ensure they represent practically distinct categories of risk.

We have used a static data set taken from 1 year of credit card data such that DR is artificially set to 20 per cent: much higher than we would normally expect. These limitations would certainly affect the computed loss distributions. A lower DR would lead to a proportionate change in

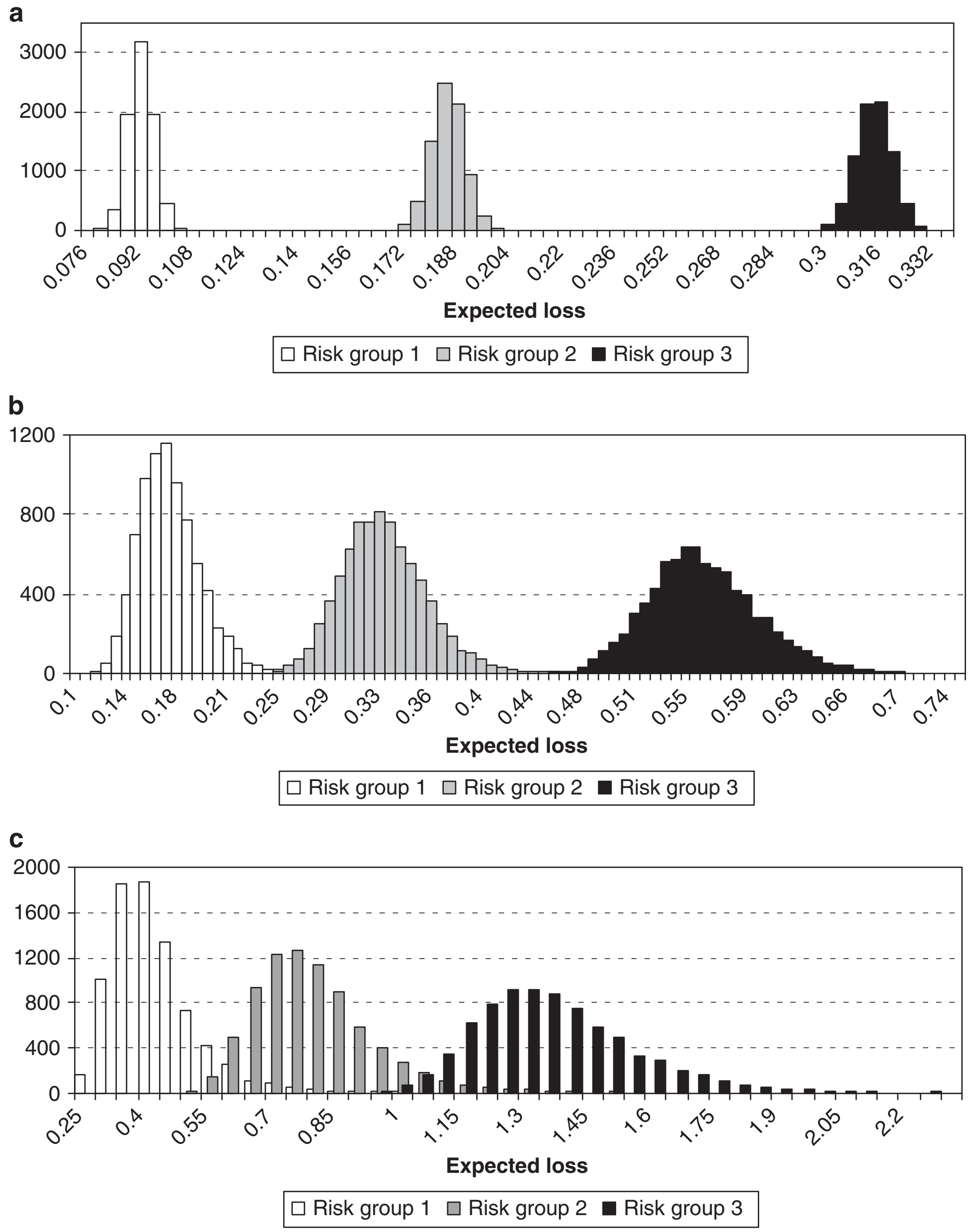


**Figure 3:** Loss distributions for risk groups for different specifications of EL: (**a**) SD×1×1; (**b**) SD×EAD (1.5)×LGD (0.2); (**c**) SD×EAD (2) ×LGD (0.2).
*Notes*: The figure in brackets following EAD is the standard deviation of the EAD log-normal distribution $\sigma_E$ and the figure in brackets following LGD is the PD/LGD correlation coefficient. SD indicates simulation of default event.

scale of EL, but otherwise we may expect approximately the same shape distributions. The lack of dynamics in the data does not take into consideration the possibility of population drift in distribution of defaults over time.[16] This could lead to even broader loss distribution if this uncertainty is taken into account. Three elements of population drift can be measured: vintage effect, asset correlation and systemic risk factors. Vintage effect is the level of risk inherent in groups of loans taken out at the same time, usually owing to lending policy at that time, and can lead to fluctuations in DR for a portfolio across time. Often credit models assume default events are independent. However, it is not clear this is a reasonable assumption. Including asset correlation and systemic risk factors in our model enables us to model dependency between defaults over time.[14,21] We expect macroeconomic conditions to contribute to the correlation of defaults over time and these can also be considered as risk factors in the model.[22] Including these dynamic elements in our model would allow us to produce accurate *forecast* loss distributions that are essential for operational use. Therefore further work is required using consumer credit data across an extended time period, where this is available, in order to build models that include dynamic risk factors. Additionally, inclusion of dynamic risk factors would enable stress testing and consideration of downturn conditions as required by the Basel II Accord.[6,22] We can expect that such work will lead to broader loss distributions. Indeed, a study for corporate loans reports a similar credit loss distribution shape to the one we report in Figure 2(c) but where inclusion of macroeconomic variables yields much longer right tails.[13]

Nevertheless, within the context of a static data set, our simulations show the extent to which inclusion of risk elements increases measured credit risk. We find that the inclusion of an EAD distribution had the greatest effect on the distribution of EL. We show that for a real data set, credit risk is underestimated without it. We have assumed a log-normal distribution for EAD and have considered two plausible parameterizations (conservative and non-conservative). However, little is published about the distribution of EAD or modelling EAD for consumer credit and further empirical work is required to understand this.[10] This was not possible for the data set we used as EAD was not available. Although including LGD also had an effect, we found that the LGD/PD correlation had only a minimal effect on EL in general.

We have also shown that when these risk elements are not included, the VaR can be extremely underestimated. The basic PD estimate (that is with EAD and LGD fixed) gives an implausibly small $\text{VaR}_{\Delta}$ of 1.4 per cent. Whereas once estimation uncertainty, LGD and EAD with a conservative standard deviation are included this rises to 13.8 per cent. This rises further to 33 per cent with a less conservative estimate of EAD. This gives some support to the critics of VaR. Even when applied to relatively simple underlying models, it is a risk measure that is highly sensitive to the inclusion or exclusion of risk elements and to how they are modelled. As Nocera concludes, VaR is not a bad methodology; it is simply in its use, or misuse, that there can be a problem.[1] If not all the risks are included correctly in the model, then VaR will not give a true picture of risk of extreme events.

## ACKNOWLEDGEMENTS

We thank the organizers of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-09) competition and Neurotech Ltd (Brazil) for making their credit card database available for academic research.

## REFERENCES AND NOTES

1 Nocera, J. (2009) Risk mismanagement. *The New York Times*. 4 January.
2 Baesens, B., van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54: 1082–1088.
3 Thomas, L.C., Edelman, D.B. and Crook, J.N. (2002) *Credit Scoring and its Applications. SIAM Monographs on mathematical Modeling and Computation*. Philadelphia, USA: SIAM.
4 Marrison, C. (2002) *Fundamentals of Risk Measurement*. New York: McGraw-Hill.
5 Rodriguez, A. and Trucharte, C. (2007) Loss coverage and stress testing mortgage portfolios: A non-parametric approach. *Journal of Financial Stability* 3: 342–367.

6 Basel Committee on Banking Supervision BCBS. (2005) Basel II: International convergence of capital measurement and capital standards: A revised framework. BIS at http://www.bis.org/publ/bcbsca.htm.
7 Hills, S. and Barrett, R. (2007) Explaining the credit risk elements in Basel II. In: M. Ong (ed.) *The Basel Handbook*, 2nd edn. London: Riskbooks.
8 Dermine, D. and Neto de Carvalho, C. (2005) Bank loan losses-given-default: A case study. *Journal of Banking and Finance* 30(4): 1219–1243.
9 Bellotti, T. and Crook, J. (2007) Modelling and Predicting Loss Given Default for Credit Cards. Working Paper, Quantitative Financial Risk Management Centre.
10 Financial Services Authority FSA. (2007) EAD expectations note. 10 December 2007.
11 Nankervis, R. and Nelson, M. (2007) IRB-compliant models in retail banking. In: M. Ong (ed.) *The Basel Handbook*, 2nd edn. London: Riskbooks.
12 Altman, E.I., Resti, A. and Sironi, A. (eds.) (2005) Loss given default: A review of the literature. *Recovery Risk*. London: Riskbooks.
13 Jokivuolle, E., Virolainen, K. and Vähämaa, O. (2008) Macro-model-based Stress Testing of Basel II Capital Requirements. Bank of Finland research discussion papers 17/2008.
14 Rösch, D. (2003) Correlations and business cycles of credit risk: Evidence from bankruptcies in Germany. *Swiss Society for Financial Market Research* 17(3): 309–331.
15 Verbeek, M. (2004) *A Guide to Modern Econometrics*, 2nd edn. Chichester: Wiley.
16 Hand, D. (2006) Classifier technology and the illusion of progress. *Statistical Science* 21(1): 1–14.
17 Except that usually in the industry, high credit scores indicate low risk, while the reverse is true of $s_i$.
18 Jiménez, G. and Mencía, J. (2007) Modelling the Distribution of Credit Losses with Observable and Latent Factors. Banco de Espana Research Paper.
19 Querci, F. (2005) Loss Given Default on a medium-sized Italian bank's loans: An empirical exercise. European Financial Management Association conference 2005 (Milan, Italy).
20 Miu, P. and Ozdemir, P. (2006) Basel requirement of downturn LGD: Modeling and estimating PD & LGD correlations. Financial Management Association European conference 2006.
21 Crook, J. and Bellotti, T. (2009) Asset Correlations for Credit Card Defaults. Working paper, Credit Research Centre, University of Edinburgh Business School (March 2009).
22 Bellotti, T. and Crook, J. (2009) Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society* 60: 1699–1707.

# APPENDIX A

See Table A1.

**Table A1:** List of covariates and logistic regression model coefficient estimates

| *Covariate* | *Estimate* | *Standard Error* | *Wald chi-square* |
|---|---|---|---|
| *Intercept* | 1.7262 | 0.2846 | 36.7813*** |
| *Pay date in month (+)*: | | | |
| From 1st to 5th | 0.1177 | 0.0764 | 2.3770 |
| From 6th to 9th | 0.1880 | 0.0561 | 11.2489*** |
| From 10th to 15th | −0.0507 | 0.0534 | 0.9001 |
| From 16th to 18th | 0 | | |
| From 19th to 21st | 0.0210 | 0.0623 | 0.1131 |
| From 22nd to 27th | 0.2194 | 0.0639 | 11.7773*** |
| After 27th | 0.3206 | 0.0649 | 24.3936*** |
| *Marital status (+)*: | | | |
| Single | 0 | | |
| Married | −0.2920 | 0.0425 | 47.1897*** |
| Divorced | 0.0883 | 0.0933 | 0.8950 |
| Widow | −0.0125 | 0.1036 | 0.0147 |
| Other | 0.0814 | 0.0700 | 1.3548 |
| *Age* | −0.0234 | 0.00193 | 146.7387*** |
| *Shop rank* | 0.0647 | 0.0713 | 0.8226 |
| *Applicant has a home telephone?* | −0.3751 | 0.0460 | 66.5224*** |
| *Residential type (+)*: | | | |
| Owned | 0 | | |
| Rented | 0.2576 | 0.0505 | 26.0207*** |
| Parent's house | −0.0453 | 0.0591 | 0.5872 |
| Other | −0.0541 | 0.0867 | 0.3899 |

**Table A1:** *continued*

| *Covariate* | *Estimate* | *Standard Error* | *Wald chi-square* |
|---|---|---|---|
| *Months in residence* | −0.00006 | 0.000141 | 0.1899 |
| *Does applicant live in same town as he/she lives?* | −0.1432 | 0.0350 | 16.7041*** |
| *Does applicant live in same state as he/she lives?* | 0.0355 | 0.1780 | 0.0397 |
| *Months in current job* | −0.00142 | 0.000305 | 21.6503*** |
| *Residential address is given* | −0.0246 | 0.1147 | 0.0462 |
| *Income (log)* | 0.0388 | 0.0275 | 1.9844 |
| *No income is indicated* | 0.1960 | 0.2060 | 0.9044 |
| *Quantity of additional cards in application* | 0.0434 | 0.0491 | 0.7812 |
| *Shop ID (weights of evidence)* | 0.7929 | 0.0646 | 150.7411*** |
| *Profession code (weights of evidence)* | 0.7976 | 0.0467 | 291.3696*** |
| *Residential code (+)*: | | | |
| 5 | −0.0818 | 0.0871 | 0.8821 |
| 23 | −0.0649 | 0.1455 | 0.1992 |
| 50 | 0.0818 | 0.0443 | 3.4057 |
| Other | −0.4923 | 0.2429 | 4.1099* |

Significance levels are shown at less than 0.05 level (*), and 0.001 level (***). Category variables that are separated into dummy indicator variables are shown by a (+). Excluded category has a coefficient estimate of 0.

## APPENDIX B

LGD is simulated using a truncated normal distribution: for each account $i$, a latent variable $l_i'$ is drawn from a given distribution $N(\mu_L, \sigma_L)$, then LGD is computed as

$$l_i = \begin{cases} l_i' & \text{if } 0 < l_i' < 1 \\ 0 & \text{if } l_i' \leqslant 0 \\ 1 & \text{if } l_i' \geqslant 1 \end{cases}.$$

Correlation with PD $p_i$ is introduced using correlation coefficient $\rho$, drawing a random number $z_i$ from $N(0, 1)$ and applying Cholesky decomposition (see, for example, Marrison, Chapter 6):[4]

$$l'_i = \mu_L + \sigma_L \left( \frac{p_i + \mu_P}{\sigma_P} \rho + z_i \sqrt{1 - \rho^2} \right)$$

where $\mu_p$ and $\sigma_p$ are constants introduced to standardize the distribution of PDs.