Estimation of Credit Card Exposure at Default (EAD)

This version: July 2013

Mindy Leow & Jonathan Crook
Credit Research Centre
University of Edinburgh Business School

Abstract

Using a large portfolio of defaulted loans and their historical observations, we estimate EAD on the level of obligors by estimating the outstanding balance of an account, not only for the account at the time of default, but at any time over the entire loan period. We theorize that the outstanding balance on a credit card account at any time during the loan is not only a function of the spending and repayment amounts by the borrower, but is also subject to the credit limit imposed by the card issuer. Therefore, we predict for outstanding balance using a two-step mixture model. We first develop a discrete-time repeated events survival model to predict for the probability of an account having a balance greater than its limit. Next, we develop two panel models with random effects to estimate for balance and limit, where the final prediction for balance would be a product of the estimated balance and limit, depending on how likely the borrower is to have a balance greater than the limit. Using this mixture model, we find that we are able to get good predictions for outstanding balance, not only at the time of default, but at any time over the entire loan period, which would allow us to make predictions for outstanding balance and hence EAD before default occurs, for delinquent accounts. Overall r-square values achieved are 0.49 when looking over the entire loan period for all delinquent accounts, and 0.55 when only looking at default-time observations.

Keywords: Exposure At Default (EAD), panel models, survival models, macroeconomic variables, time-varying covariates

1. Introduction

The three loss components defined in the Basel Accords are Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD), and loss is then calculated to be a product of the three. Since the credit crisis of 2008, there has been increased awareness of the risk models for these components, and in particular, for retail loans. However, these have been mainly focused on PD and LGD models, and how they should and can be improved (see Thomas (2010) for a review). The analysis and modelling of EAD at account level has so far been neglected and assumed to be an easily tabulated deterministic variable. This might be the case for loans with fixed loan amounts over fixed terms and pre-agreed monthly repayment amounts, making it possible to try and get at least a reasonable range for EAD should the loan be expected to default in the following time horizon, e.g. in the next 12 months. However, in the case of revolving loans, i.e. loans with no fixed loan amount or term, debtors are given a line of credit, with a credit limit up to which they can draw upon at any time (as long as they have not gone into default), and this could make it difficult for financial institutions to predict account level outstanding balance should an account go into default, especially if accounts deteriorate into default quickly and draw heavily on the card just before default. As we are looking at retail loans here, the obvious type of product here would be credit cards.

Another issue associated with the analysis and modelling of EAD is the measurement of EAD. EAD is similar to LGD in that its value is only of interest in the event default occurs (although its value still needs to be estimated for the calculation and preparation of economic capital). However, unlike LGD, where loss is calculated to be at some time point after default, EAD is known the very instant the account goes into default. Therefore, although default-time variables could be used in the modelling of LGD, they cannot be used for EAD. As such, various indicators have been created to be estimated instead of EAD, taking into account the current balance and available limit. Common variables estimated in lieu of EAD in the literature are the Loan Equivalent Exposure (LEQ) Factor, the Credit Conversion Factor (CCF) and the Exposure At Default Factor (EADF), all of which have their advantages and limitations and are explored in more detail in Section 2.2.

Only a few papers have looked at EAD for corporate loans, and even fewer on retail loans, and it is common to assume some value of EAD at the portfolio level. The few EAD papers that examine corporate data (for example, see Araten and Jacobs (2001), Jimenez et al. (2009), Jacobs (2008)) in the literature focus mostly on the possible determinants of EAD, how EAD is affected by the economy and relationships between EAD and the behavioural of

delinquent firms. Jimenez and Mencia (2009) takes an overall view of credit risk and looks at the PD, EAD and LGD for many corporate sectors at the aggregate level. However, they did not explicitly model EAD and accounted for it on the portfolio level by matching a suitable distribution (either the Inverse Gaussian or the Gamma distribution) to the empirical distribution of EAD. In terms of retail loans, Qi (2009) looks at EAD for unsecured credit cards, trying to predict for LEQ by looking at the level of credit drawn at one year before default. However, no macroeconomic variables were included in the model. All come to the conclusion that EAD plays an important part in the calculation of provision of capital and should be more carefully incorporated into risk and loss calculations.

Using a large portfolio of defaulted loans and their historical observations, this paper estimates EAD at the level of the obligor by estimating the outstanding balance of an account, not only for the account at the time of default, but at any time over the entire loan period, up to the time of default. This way, we avoid the issues plaguing the measurement of EAD (as seen by the various indicators used to represent EAD), and because we are able to make a prediction for outstanding balance at any point in the life of the account, when used together with predictions for PD and LGD, we can predict the EAD should the loan go into default.

We theorize that the outstanding balance on a credit card account at any time during the loan is not only a function of the spending and repayment amounts by the borrower, but is also subject to the credit limit imposed by the card issuer. Once any borrower has an outstanding amount on the account equal to the credit limit, they should not be able to draw upon any more credit until they make some repayment towards their outstanding balance (although it is possible). This means that although a borrower could default for any amount between £0 and his credit limit, estimation of balance could be more efficient if we know how likely a borrower is to default near his credit limit. Therefore, we predict for outstanding balance using a two-step mixture model. We first develop a survival model to predict for the probability of an account having a balance greater than its limit. Next, we develop two sub-models estimating for balance and limit, where the final prediction for balance would be a product of the estimated balance and limit, depending on how likely the borrower is to have a balance greater than the limit. The covariates of all three models are drawn from three groups: application time variables, behavioural variables and macroeconomic variables.

The development and validation of this mixture model contributes to the literature in two ways. First, it is the first paper to predict for the outstanding balance at any time during the

life of a revolving loan. Second, we incorporate macroeconomic variables into the model and so provide a framework suitable for stress testing later.

The rest of this paper is structured with Section 2 detailing the data and variables, including some empirical analysis of EAD and common dependent variables used in lieu of EAD. Methodology and results are given in Sections 3 and 4 respectively, and Section 5 concludes.

2. Data and variables

The data is supplied by a major UK bank and consists of a large sample of credit card accounts, geographically representative of the UK market. The accounts were drawn from a single product, and opened between 2001 and 2010. Accounts were observed and tracked monthly up to March 2011 or until it was closed, whichever is earlier. A minimum repayment[1] is calculated in each month for each account. Accounts progress through states of arrears depending on whether they are able to make the minimum repayment amount. It is also possible for accounts to recover from states of arrears should the borrower make repayment amounts large enough to cover accumulated minimum repayment amounts that were previously missed. An account is then said to go into default if it goes into 3 months in arrears (not necessarily consecutive). For more details on the dataset and movement of accounts between states, see Leow and Crook (2012).

Accounts that have a credit limit of £0 at any point in the loan are removed, based on the assumption that these accounts would have been singled out as problem loans by the bank. It is possible for accounts to be in credit, such that balance is negative, so balance is constrained such that observations that have negative balance are £0. Due to the 6 month lag imposed on time-dependent covariates and the minimum time required for accounts to go into default, we also remove accounts that have been on the books less than 9 months.

2.1. Empirical analysis of EAD

[1] This minimum repayment amount is 2.5% of the previous month's outstanding balance or £5, whichever is higher, unless the account is in credit, in which case the minimum repayment amount is £0, or the account has an outstanding balance of less than £5, in which case the minimum repayment amount would be the full outstanding amount. Note that this percentage is different to that used in LEOW, M. & CROOK, J. N. 2012. Intensity Models and Transition Probabilities for Credit Card Loan Delinquencies.

From the data, we see that some accounts go into default with an outstanding balance greater than their credit limits, in which case we would be able to get a good estimate for EAD given that the credit limit is known (or can be predicted) before default (plus an estimated percentage to account for cumulated interest and fees, although this was not adjusted for in this work). Other borrowers might go into default with a significantly smaller balance compared to their available credit limit, in which case a prediction for balance itself is required. This is illustrated in Figure 1, which gives the distribution of the ratio of balance over limit at the time of default (only for ratios less than 3 for a clearer picture of the distribution). The peak in the graph corresponds to borrowers defaulting with a balance equal to their credit limit, but we also do see a sizeable proportion of borrowers who default with balances on either side of their credit limits.

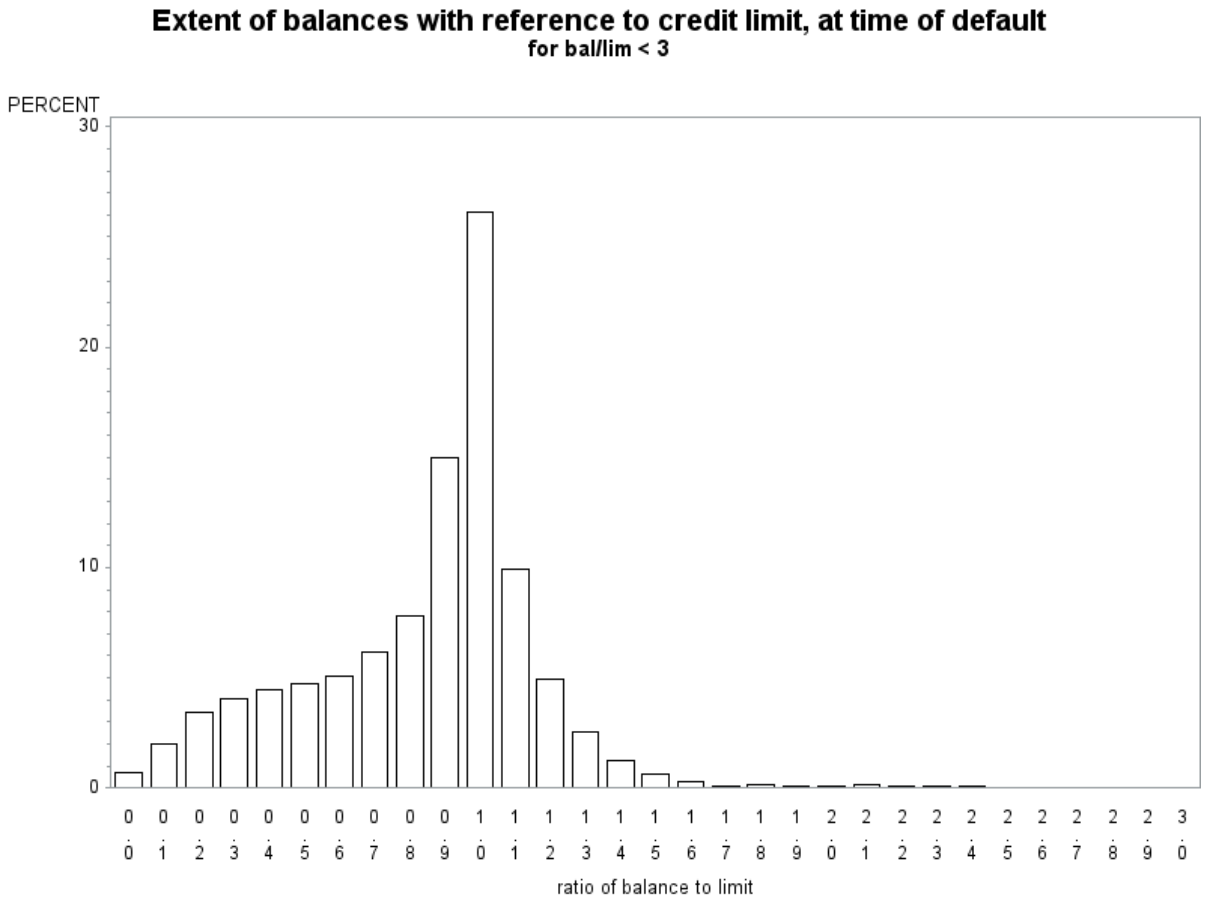

Figure 1: Distribution of ratio of balance over limit at time of default (for ratios less than 3)

2.2. Dependent variables

As mentioned earlier, variables estimated in lieu of EAD in the literature are the Exposure At Default Factor (EADF), the Credit Conversion Factor (CCF) and the Loan Equivalent Exposure (LEQ) Factor (a more comprehensive review can be found in Moral (2006)). All three variables are ratios created using the outstanding balance at default over some indication of limit or balance at an observation time before default, and are described in more detail below. Because each account would only have a single time of default and correspondingly, a single value of the dependent variable, they would be estimated via cross-sectional models. However, it is possible to modify the variables slightly such that balance at default is replaced by balance at some duration time *t*, and where the observation time is lagged *l* months, and definitions of the variables below are thus given in this most general form. This would allow for a more flexible definition of the dependent variables, and increase the number of observations per account, allowing for other methods of modelling, for example, panel modelling.

From here on, outstanding balance of account *i* at time *t* is represented by B_{it}, and limit of account *i* at time *t* is represented by L_{it}. To simplify the notation, the subscript *i* representing account *i* is dropped for the equations in this sub-section.

The EADF_t is the ratio of the balance at time *t* (at default), over the limit at observation time $t-l$ (an observation time before default), given in Equation 1:

$$\text{EADF}_t = \begin{cases} \dfrac{B_t}{L_{t-l}} & \text{for } L_{t-l} \neq 0 \end{cases} \tag{1}$$

The limit is usually the limit at the time of application as that would be a known quantity once the account is opened. However, as it is possible for the credit limit to change during the lifetime of the account, it is also possible to take the limit at some observation time before default. It is unlikely that limit will be £0 at any time during the loan as these accounts would already be flagged up as delinquent accounts or closed. Although we expect EADF_t to range between 0 and 1, it is possible and quite common to see outstanding balances greater than the assigned limits, perhaps due to accumulated interest or banks allowing borrowers to go over their limits, giving values much greater than 1.

The CCF_t is the ratio of the balance at time *t* (at default) over the balance at some observation time $t-l$, given in Equation 2:

$$\text{CCF}_t = \begin{cases} \dfrac{B_t}{B_{t-l}} & \text{if } B_{t-l} \neq 0 \\ 0 & \text{if } B_{t-l} = 0 \end{cases} \tag{2}$$

The CCF_t tries to get better predictions for balance at default by taking into account the outstanding balance of an account at some observation time before default. However, it is possible that the outstanding balance at the selected observation time is £0, or even negative (the account is in credit), which would give $\text{CCF}_t = 0$, and this raises the issue of the treatment of these accounts. It is possible that these accounts are up-to-date and thus not likely to be delinquent, but it is also possible that these accounts could deteriorate quickly into delinquency and default, and it could be difficult to differentiate between these two groups. Although on one hand, it is likely that accounts that go into default have large balances on their account prior to default (debtors who default due to behavioural issues), it is also possible that accounts go from a low or zero balance to default within a short period of time (debtors who default due to unexpected circumstances), which could then imply a different set of predictors for each group.

The LEQ_t factor tries to make a more sophisticated prediction for balance by not only taking into account balance at some observation time before default, but also the remaining amount of credit the debtor is able to draw upon, given in Equation 3:

$$\text{LEQ}_t = \begin{cases} \dfrac{B_t - B_{t-l}}{L_{t-l} - B_{t-l}} & \text{if } L_{t-l} \neq B_{t-l} \\ 0 & \text{if } L_{t-l} = B_{t-l} \end{cases} \tag{3}$$

The different values the LEQ_t can take could arise due to a number of different situations and which would give different implications. Should the outstanding balance be equal to limit at the time of observation, we get an LEQ_t value of 0. This is a group of debtors who have used their maximum available limit and are likely to default, but would be difficult to include and handle in the modelling because the LEQ_t value computed does not have the same implications as the other LEQ_t values computed for when balance and limit are not equal.

The majority of accounts would have a positive LEQ_t, which could be due to one of two situations: (a) when balance at default is greater than balance at observation, and balance at observation is below the credit limit at observation, which would be the most common progression into default; or (b) when balance at observation is greater than balance at

default, and balance at observation is already greater than the limit at observation, which would represent debtors who are actually recovering from a large balance (and where perhaps extending the credit without putting the account into default might give lower loss). Although these two groups of debtors would have LEQ_t in the same range, we expect their characteristics and circumstances to be quite different. It is also possible to have negative LEQ_t: (a) when balance at observation is larger than limit at observation and balance at default is larger than balance at observation, which would represent debtors who are spiralling further into debt and default; or (b) when balance at observation is larger than balance at default, but both are below the limit at observation. Again, we have two groups of debtors with negative LEQ_t values but where they have arrived via different circumstances. The possible range of LEQ_t, coupled with the fact that different types of borrowers and circumstances could give LEQ_t in the same range, would make it difficult to estimate and model LEQ_t.

Although these dependent variables have tried to predict EAD using a combination of limit, balance and available remaining credit at some observation time before default, each have weaknesses. Qi (2009) and Jacobs (2008) looked at these dependent variables but do not draw any conclusions about which would be the most appropriate. Also, predictive results from papers in the literature using these dependent variables generally been poor. We therefore decided to focus on the estimation of outstanding balance itself, as we explain in Section 3.

2.3. Explanatory and macroeconomic variables

Common application variables are available, including age, time at address, time with bank, income, presence of landline and employment type. Behavioural variables are also available on a monthly basis, including repayment amount, credit limit, outstanding balance and number and value of cash withdrawals or card transactions. From these, further behavioural indicators could be derived, for example, the number of times an account oscillates between states of arrears and being up-to-date, the proportion of time the account has been in arrears and the average card transaction value. Any behavioural variables used in the model are lagged by 6 months.

The macroeconomic variables considered here are listed in Table 1. The main source of macroeconomic variables is the Office of National Statistics (ONS), supplemented by data from Bank of England (BOE), Nationwide and the European Commission (EC) where

appropriate. The non-seasonally adjusted series is selected unless unavailable. Any macroeconomic variables used in the model are lagged by 6 months. This lag was determined by the arbitrary time horizon chosen for the predictions.

Table 1: Table of macroeconomic variables

Variable	Source	Description
AWEN	ONS	Average earnings index, including bonus, including arrears, whole economy, not seasonally adjusted
CIRN	BOE	Monthly weighted average of UK financial institutions' interest rate for credit card loans to households, not seasonally adjusted
CLMN	ONS	Claimant count rate, UK, percentage, not seasonally adjusted
CONS	EC	Total consumer confidence indicator, UK, seasonally adjusted
HPIS	Nationwide	All houses, seasonally adjusted
IOPN	ONS	Index of production, all production industries, not seasonally adjusted
IRMA	BOE	Monthly average of Bank of England's base rate
LAMN	ONS	Log (base e) of total consumer credit, amounts outstanding, not seasonally adjusted
LFTN	ONS	Log (base e) of FTSE all share price index, month end, not seasonally adjusted
MIRN	BOE	Monthly weighted average of UK financial institutions' interest rate for loans secured on dwellings to households, not seasonally adjusted
RPIN	ONS	All items retail price index, not seasonally adjusted
UERS	ONS	Labour Force Survey unemployment rate, UK, all, ages 16 and over, percentages, seasonally adjusted

2.4. Training and test set split

Although we are interested in the prediction of outstanding balance of an account in each time step, these predictions of balance only become EAD values if and when accounts go into default. As such, we only use accounts that do (eventually) go into default. Because we only use observations from accounts that do go into default for the development of the EAD model, we do not need to be concerned with accounts that are inactive, e.g. have zero

transactions and zero balance on the card for an extended period of time, but remain in the portfolio.

The dataset is split such that the training set consists of all accounts that do go into default and were opened on or before 31 December 2008, giving about 86,000 unique accounts. An out-of-sample test set (Test set I) is created using the remaining default accounts, consisting of all observations of all accounts opened on or after 01 January 2009. Test set I consists of about 7,000 unique accounts, giving more than 100,000 month-account observations. A second test set is created (Test set II), a subset of Test set I, where only default-time observations are included. Test set I would give an indication of how well the model is able to predict for balance for accounts that are likely to be delinquent but have not yet gone into default, whilst Test set II would be an indication of how well the model is able to predict at default-time.

3. Methodology

We predict for outstanding balance using a mixture model. We assume the random variable, balance of account *i* at time *t* could be the account limit or less than this. Therefore, the expected balance for account *i* at time *t* is given in Equation 4:

$$E(B_{it}) = \left(P(B_{it} = L_{it}) \times E(L_{it} | B_{it} = L_{it})\right) + \left(P(B_{it} < L_{it}) \times E(B_{it} | B_{it} < L_{it})\right) \qquad (4)$$

Since some accounts can have a value of B_{it} that is greater than the credit limit and we assume such accounts have an expected valance equal to the expected limit, we replace the first probability condition in Equation 4 by $P(B_{it} \geq L_{it})$. We therefore parameterise three models. First, a model of the probability that the outstanding balance of an account is larger than the credit limit; second, a model to predict the outstanding balance; and third, a model to predict the credit limit, where we allow parameters to predict balance and limit to differ.

From the training dataset, we first estimate the probability that the outstanding balance at any duration time *t* is equal or greater than the limit at time *t*. This is done by defining the event 'overstretched', S_{it}, for account *i* at time *t* which takes the value 1 if outstanding balance is greater than the limit at time *t*; and 0 otherwise, given in Equation 5.

$$S_{it} = \begin{cases} 1 & \text{if } B_{it} \geq L_{it} \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

Given this definition, it is possible for an account to experience the event more than once (at different times of the loan), so a discrete-time repeated events survival model, with clustered standard errors, is then used to estimate this model in SAS, given in Equation 6 (see Allison (2010), Chapter 8 for details).

$$\log\left(\frac{S_{it}}{1-S_{it}}\right) = \nu + \beta_1 X_i + \beta_2 Y_{i,t-6} + \beta_3 Z_{t-6} \quad (6)$$

where ν is the intercept term; X_i are account-dependent, time-independent covariates, i.e. application variables; $Y_{i,t-6}$ are account-dependent, time-dependent covariates, lagged 6 months, i.e. behavioural variables; Z_{t-6} are account-independent, time-dependent covariates, lagged 6 months, i.e. macroeconomic variables; and $\beta_1, \beta_2, \beta_3$ are unknown vectors of parameters to be estimated.

Next, we develop two sub-models, to predict either balance or limit, using two separate training datasets. For all the accounts in the training set, we look at the entire history of each account, and subset the training dataset further by segmenting between accounts which ever had balance exceeding limit (but not necessarily in default) at any point in the loan; and accounts that never had balance exceeding limit throughout the life of the loan. The subset consisting of accounts where balance exceeded credit limit at some point during the loan is the limit training set, which is used to estimate the limit at time *t*. By structuring the sample in this way, we parameterise the distribution of B_{it} given $B_{it} \geq L_{it}$. We could not include only observations where $B_{it} > L_{it}$ because we wish to use the panel aspect of the data and such a condition is rare anyway. The other subset consists of accounts where balance never exceeded limit throughout the observation time of the loan is the balance training set, which is used to estimate for the balance at time *t*. Hence, we parameterise the B_{it} given $B_{it} < L_{it}$ distribution. By segmenting the accounts in this way (see Table 2), we are able to use the full history of each account in the estimation of either balance or limit as it changes over time and over the course of the loan period. This methodology and the training and test sets created are represented in Figure 2.

Table 2: Number of observations for balance and limit subsets

Model	Number of accounts	Number of observations	Minimum observations per account	Maximum observations per account	Average observations per account
Balance	40,428	862,695	4	115	21.3
Limit	45,361	977,306	3	113	21.5

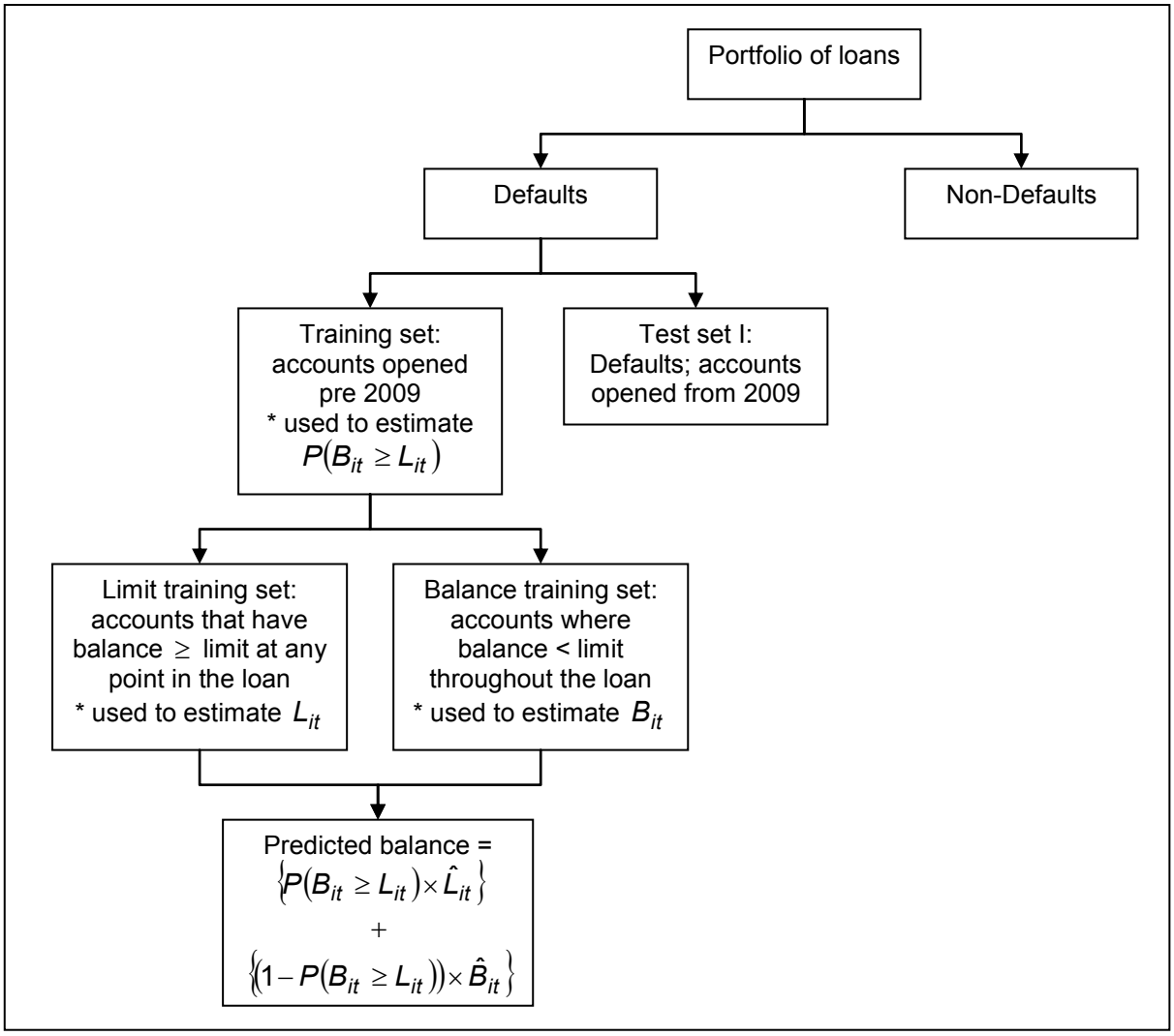

Figure 2: Flowchart of methodology and training and test set splits

The limit, L_{it}, and balance, B_{it} for each account *i* at time *t* are estimated using panel models with random effects. As these are standard specifications of panel models (see Equation 7), we do not include the technical equations here and instead refer to Cameron and Trivedi (2005), Gujarati (2003) and Verbeek (2004) for details. Since each account has multiple observations (month-account observations), we adjust for serial correlation by using a clustered sandwich estimator (on account ID) to estimate variance and standard errors (Drukker (2003)).

$$y_{it} = \mu + \gamma_1 X_i + \gamma_2 Y_{it} + \gamma_3 Z_t + \alpha_i + \varepsilon_{it} \quad (7)$$

where μ is the intercept term, X_i are account-dependent, time-independent covariates, i.e. application variables; Y_{it} are account-dependent, time-dependent covariates, i.e.

behavioural variables; Z_t are account-independent, time-dependent covariates, i.e. macroeconomic variables; $\gamma_1, \gamma_2, \gamma_3$ are unknown vectors of parameters to be estimated; and $\alpha_i + \varepsilon_{it}$ is the error term, with $\alpha_i \sim IID\left(0, \sigma_\alpha{}^2\right)$ and $\varepsilon_{it} \sim IID\left(0, \sigma_\varepsilon{}^2\right)$.

Both models were estimated using Generalised Least Squares (GLS) estimators. Covariates include application variables, behavioural variables, lagged 6 months and macroeconomic variables, lagged 6 months, defined in Equations 8 and 9. We note that different sets of parameters are used in each model, depending on the relevance of the variables to balance or limit, as well as their statistical significance. Variations of these models include lags of various periods between 3 and 9 months.

$$\hat{L}_{it} = f\left(X_i^L; Y_{i,t-6}^L; Z_{t-6}^L\right) \tag{8}$$

$$\hat{B}_{it} = f\left(X_i^B; Y_{i,t-6}^B; Z_{t-6}^B\right) \tag{9}$$

The mixture model could then be used to predict for balance at any given time during the loan. We apply this model on the out-of-sample test set I (of defaults, for all observations) by first applying the survival model to all accounts to predict the probability of being overstretched at each time *t*. Then, regardless of the estimated probability, we apply the balance panel model and the limit panel model onto all observations of all accounts to get an estimated balance and estimated limit, again at each time *t*. Because the models were estimated for the subsets described above, these predicted values for B_{it} and L_{it} are the values of B_{it} given $B_{it} < L_{it}$ and L_{it} given $B_{it} \geq L_{it}$ respectively. The final predicted value for balance of an account *i* at time *t*, $\tilde{B}_{it}$, is then a combination of the repeated events survival model estimating the probability of balance exceeding limit at time *t*, and the panel models estimating either balance or limit at time *t*. This is the expected value of balance and limit, given the probabilities of the balance exceeding the limit at time *t*, defined in Equation 10.

$$\tilde{B}_{it} = \left\{P(S_{it}) \times \hat{L}_{it}\right\} + \left\{\left(1 - P(S_{it})\right) \times \hat{B}_{it}\right\} \tag{10}$$

where $P(S_{it}) = P(B_{it} \geq L_{it})$ and is the estimated probability of account *i* is overstretched at time *t*, i.e. that the balance for account *i* at time *t* exceeds the limit for account *i* at time *t*; and $\hat{L}_{it}$ and $\hat{B}_{it}$ are the estimated values for limit and balance respectively, from their respective panel models.

In order to validate this mixture model and assess how well it predicts, using observed balances over time, B_{it}, and predicted balances over time, $\tilde{B}_{it}$, we are able to calculate r-square values. These are then calculated for the two test sets: Test set I, for all accounts, for all observations; and Test set II, for all accounts, only at time of default.

4. Results

4.1. Survival model for being overstretched

The parameter estimates for the discrete-time repeated events survival model predicting for the event overstretched is given in Table 3. We find that the signs of the parameter estimates are intuitive: for example, the probability of being overstretched decreases with age as well as with higher income. In terms of behavioural variables, we find that the probability of being overstretched reflects how well borrowers manage their accounts, so borrowers who move in and out of arrears frequently (see rate of total jumps) or are frequently in arrears (see proportion of months in arrears) tend to have a higher probability of being overstretched. In terms of macroeconomic variables, an increase in general wealth, for example, an increase in the HPI, or the FTSE would decrease the probability of being overstretched; but easier access to credit (via an increase in credit amount outstanding) increases the probability of being overstretched.

Table 3: Parameter estimates of survival model for event overstretched and panel models for balance and limit

Code	Parameter	Discrete-time repeated events survival model for P(B>=L)			Panel model with random effects for balance			Panel model with random effects for limit		
		Estimate	WaldChiSq	ProbChiSq	Estimate	z	P>|z|	Estimate	z	P>|z|
Intercept	Intercept	-7.4852	15.1526	0.0000	-1706.4380	-12.9200	0.0000	3413.3990	15.9200	0.0000
	Application variables									
ageapp_1	Age at application group 1	-	-	-	-	-	-	-	-	-
ageapp_2	Age at application group 2	-0.1468	64.6533	0.0000	56.1902	6.4000	0.0000	25.3712	8.7300	0.0000
ageapp_3	Age at application group 3	-0.2183	108.9908	0.0000	76.1836	6.8700	0.0000	43.7350	11.7000	0.0000
ageapp_4	Age at application group 4	-0.1550	42.6659	0.0000	146.6656	11.2700	0.0000	55.7777	11.9700	0.0000
ageapp_5	Age at application group 5	-0.1501	32.8285	0.0000	157.0375	11.2500	0.0000	61.4651	11.6800	0.0000
ageapp_6	Age at application group 6	-0.2160	56.7911	0.0000	164.6499	10.9000	0.0000	65.0601	10.9000	0.0000
ageapp_7	Age at application group 7	-0.2609	60.8063	0.0000	184.2070	11.0400	0.0000	72.7960	9.5600	0.0000
ageapp_8	Age at application group 8	-0.3779	79.2617	0.0000	178.1197	9.8700	0.0000	81.9746	8.6300	0.0000
ageapp_9	Age at application group 9	-0.4486	72.4532	0.0000	178.0454	8.1300	0.0000	82.3753	7.3500	0.0000
ageapp_10	Age at application group 10	-0.6343	127.3410	0.0000	73.0292	3.2400	0.0010	64.5486	5.5300	0.0000
ECode_A	Employment code, group A	-	-	-	-	-	-	-	-	-
ECode_B	Employment code, group B	-0.0180	0.7509	0.3862	8.2236	0.6800	0.4960	-26.1230	-6.1000	0.0000
ECode_C	Employment code, group C	0.0873	2.7354	0.0981	-93.1608	-4.2800	0.0000	-20.5453	-2.0900	0.0360
ECode_D	Employment code, group D	-0.1636	39.8432	0.0000	-89.8889	-8.3400	0.0000	41.5112	11.1700	0.0000
ECode_E	Employment code, group E	-0.1217	53.9036	0.0000	27.6582	2.8200	0.0050	73.3981	16.7500	0.0000
INC_L	Income, ln	-0.1581	232.4980	0.0000	130.6251	18.7400	0.0000	69.5979	20.5700	0.0000
INC_M0	Binary indicator for missing or 0 income	-1.5088	237.9265	0.0000	1011.0890	15.7900	0.0000	578.3791	19.0100	0.0000
LLine	Binary indicator for presence of landline	-0.0040	0.0615	0.8041	82.5093	9.4000	0.0000	-	-	-
NOCards	Number of cards	-0.0908	204.8178	0.0000	26.6488	8.3200	0.0000	10.5419	8.2200	0.0000
TAAdd	Time at address (years)	0.0002	0.0680	0.7943	0.0000	0.0000	0.0000	-	-	-
TWBank	Time with bank (years)	-0.0984	24.0764	0.0000	0.3232	8.1500	0.0000	0.3668	16.6200	0.0000

TWBank_MU	Binary indicator for missing or unknown time with bank	-0.0016	342.2373	0.0000	-	-	-	-	-	-
X_A	Variable X, group A	-	-	-	-	-	-	-	-	-
X_B	Variable X, group B	0.3315	304.5941	0.0000	-131.7043	-12.8200	0.0000	-75.3408	-16.8300	0.0000
X_C	Variable X, group C	0.4286	359.2445	0.0000	-162.8874	-13.9400	0.0000	-58.4784	-12.1900	0.0000
X_D	Variable X, group D	0.3243	225.0762	0.0000	-104.4849	-9.6300	0.0000	-51.5318	-11.8500	0.0000
X_E	Variable X, group E	0.5634	742.8391	0.0000	-169.3372	-14.3800	0.0000	-174.9060	-28.1100	0.0000
	Behavioural variables, lagged 6 months									
ATRV_lag6	Average transaction value	-0.0011	457.4898	0.0000	0.1647	19.1200	0.0000	0.0226	3.6400	0.0000
CASC_lag6	Number of cash withdrawals	0.1675	103.9683	0.0000	-	-	-	-	-	-
CASV_lag6	Amount of cash withdrawal	0.0001	3.4970	0.0615	-	-	-	-	-	-
CRLM_lag6	Credit limit	-	-	-	0.0889	20.1000	0.0000	0.8378	133.9900	0.0000
JUMP_lag6	Rate of total jumps	0.5324	129.2653	0.0000	90.9499	4.0500	0.0000	-	-	-
PARR_lag6	Proportion of months in arrears	0.5574	102.2871	0.0000	-413.1779	-15.1800	0.0000	-	-	-
PAYM_lag6	Repayment amount	-	-	-	-	-	-	0.0190	10.1900	0.0000
SCBA_lag6	Outstanding balance	-	-	-	0.2417	70.5100	0.0000	0.0354	17.7500	0.0000
	Macroeconomic variables, lagged 6 months									
AWEN_lag6	Average wage earnings	0.0002	0.0883	0.7663	-	-	-	-	-	-
CIRN_lag6	Credit card interest rate	0.0104	83.5315	0.0000	11.1188	3.0000	0.0030	-63.9783	-40.8000	0.0000
CONS_lag6	Consumer confidence	-	-	-	10.0082	23.5100	0.0000	-	-	-
HPIS_lag6	House Price Index	-0.0013	9.6272	0.0019	-2.6700	-19.8000	0.0000	0.8436	18.4900	0.0000
IOPN_lag6	Index of production	-0.0039	77.9723	0.0000	-	-	-	-	-	-
IRMA_lag6	Base interest rate	-0.0213	12.7909	0.0003	-	-	-	-	-	-
LAMN_lag6	Amount outstanding, ln	0.8031	22.8974	0.0000	-	-	-	-258.0192	-13.8000	0.0000
LFTN_lag6	FTSE Index, ln	-0.2386	27.7009	0.0000	-169.2306	-10.2800	0.0000	-	-	-
RPIN_lag6	Retail Price Index	0.0010	0.5418	0.4617	13.0255	20.1900	0.0000	-	-	-
UERS_lag6	Unemployment rate	-	-	-	61.8278	11.8600	0.0000	-5.7529	-3.1100	0.0020
	Model specific required variables									
duration	Survival time (months)	-0.0381	4484.3489	0.0000	-	-	-	-	-	-

	before next event									
period	Number of times event has happened	0.1932	1396.0967	0.0000	-	-	-	-	-	-
time_on_books	Time on books (months)	-	-	-	5.9166	22.7500	0.0000	3.2404	24.7800	0.0000

4.2. Panel models for balance and limit

The parameter estimates for both panel models are given in Table 3. We acknowledge that the balance from 6 months previous is included as a parameter in the balance model, and credit limit from 6 months previous is included as a parameter in the limit model. Although this would raise the issue of endogeneity in econometric interpretation, it is not an issue in this case as we are using the model solely for the purpose of prediction. Although the panel models are developed with random effects, these random effects are not known for accounts in the test set(s). The random effects associated with each account in the test set is assigned to be the mean values of α_i and ε_{it}, that is zero in both cases.

Table 4: Performance indicators for panel models, for training set

Model	Overall R^2 (trn)	σ_u	σ_e	ρ
Balance	0.2748	368.1667	636.6160	0.2506
Limit	0.9239	162.7045	281.6164	0.2502

The predictive statistics for panel models for balance and limit, based on the training set are given in Table 4. We expect it to be easier to predict for limit, as this would be based on a combination of application time and behavioural indicators. This is reflected in the impressive r-square value for the limit model, although the fact that credit limit from 6 months previous was also included in the model contributes substantially. The panel model for balance does not predict as well, as factors affecting outstanding balance of an account would include borrower circumstances which would be impossible to take into account given the information we have.

4.3. Overall performance

Table 5: Overall r-square values based on mixture model, for test sets

Model, Test set	R^2
Mixture Model, Test I (defaults, all observations)	0.4989
Mixture Model, Test II (defaults, default time only)	0.5517

After applying the mixture model onto the test sets, we compute overall r-square, given in Table 5. We see that the model is able to achieve a modest r-square of 0.49 when predicting for balances for accounts that are likely to be delinquent. When looking

specifically at default-time observations, the model predicts even better, achieving an r-square of 0.55.

In comparison, the regression models developed for credit card LEQ by Qi (2009) achieved adjusted r-square values of between 0.06 to 0.37, depending on whether the accounts were current or delinquent, and whether outliers were excluded from the model development. Jacobs (2008), working on corporate data, achieved pseudo r-square values of 0.20, 0.23 and 0.16 for LEQ, CCF and EADF respectively. He also developed a model for multiple quantile LEQ regression and achieved a pseudo r-square of 0.85. However, we note that this model incorporates an estimate for loss using conditional PD and LGD and assumes that EAD would share the same risk drivers, which is not necessarily true, especially for retail loans.

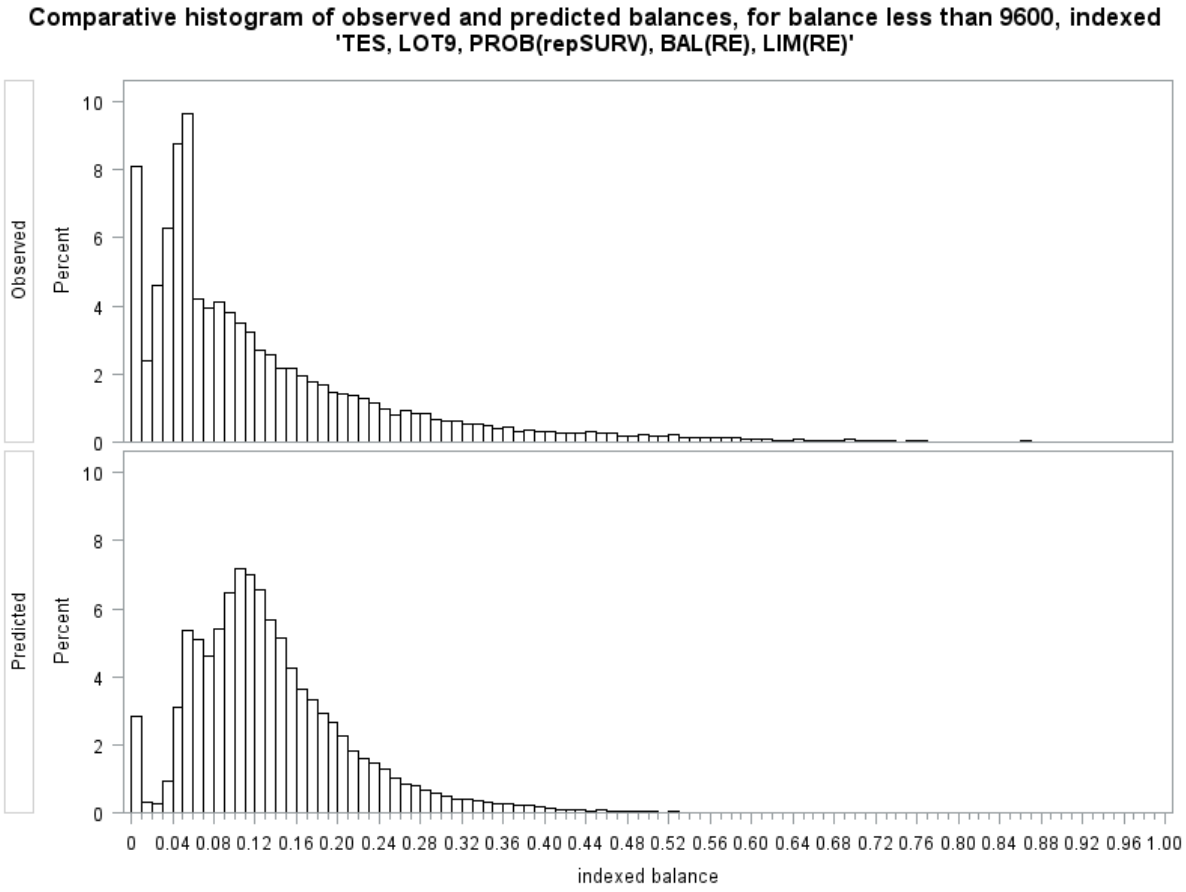

Figure 3: Comparative histogram of predicted and observed balances, for Test set I, all observations for all default accounts (where balance lies between £0 and £9,600). The top panel gives the distribution of observed balance; the bottom panel gives the distribution of predicted balance given by the mixture model.

We also look at the distributions of predicted balance, $\tilde{B}_{it}$ and compare them against the distributions of observed balance over time, B_{it}. Figure 3 compares the distributions of predicted and observed balance of all observations of all accounts where for the purpose of a clearer illustration, we look at values of balance between £0 and £9,600. The figure shows that the model is over-estimating balance most of the time. Figure 4 compares the distributions of predicted and observed balances for Test set II, only default time observations for all default accounts. Again, the values of balance are limited to between £0 and £9,600 for clearer representation of the distributions. A similar picture is seen here, where predicted balance is generally over-estimated except at very low values of balance.

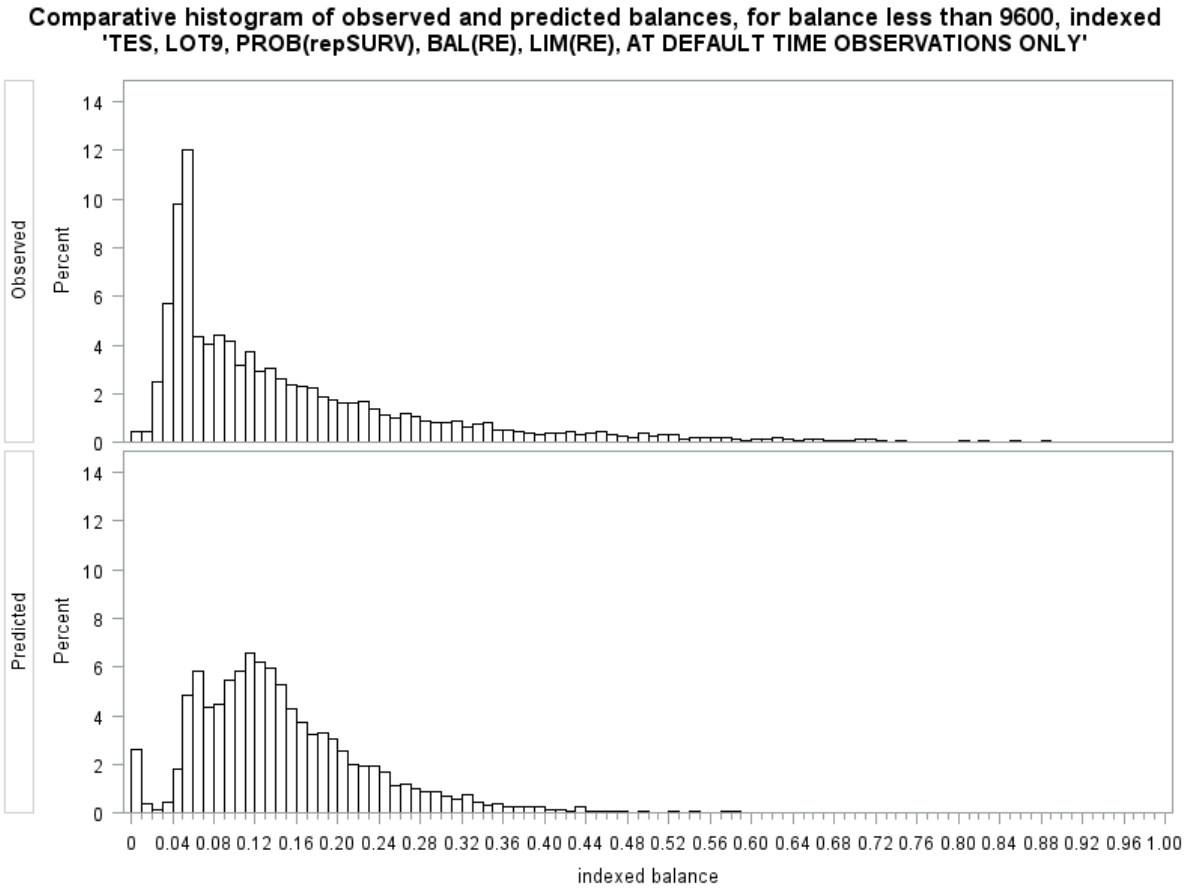

Figure 4: Comparative histogram of predicted and observed balances, for Test set II, at default time only for all default accounts (where balance lies between £0 and £9,600). The top panel gives the distribution of observed balance; the bottom panel gives the distribution of predicted balance given by the mixture model.

However, we note that what we are trying to predict here is the outstanding balance for individual accounts, of which many factors play a part. What we have done here is to include application time variables, behavioural indicators and macroeconomic variables, but

this is by no means a comprehensive list of covariates that define balance. We are unable to capture certain individual borrower circumstances that affect an account holder's spending and repayment habits, which undoubtedly play a part in the observed outstanding balances.

5. Concluding Remarks

Using a large portfolio of defaulted loans and their historical observations, we developed a mixture model to predict for balance at any time *t*. First, a discrete-time repeated events survival model was developed to estimate the probability of an account being overstretched, i.e. having a balance greater than its limit, at any time *t*. This model incorporated time-dependent variables and gave intuitive parameter estimates. Next, two panel models with random effects were developed to estimate balance and limit separately, at any time *t*. The final prediction for balance at time *t* is then said to be either the estimated limit or balance, depending on how likely the borrower is to be overstretched at that time *t*. This is the sum of two products: the probability of being overstretched multiplied by the estimation for limit; and the probability of not being overstretched multiplied by the estimation for balance (c.f. Equation 10).

Using this mixture model, we find that we are able to gain good predictions for outstanding balance, not only at the time of default, but at any time over the entire loan period, which would allow us to make predictions for outstanding balance and hence EAD before default occurs, for delinquent accounts. Overall r-square values achieved are 0.49 when looking over the entire loan period for all delinquent accounts, and 0.55 when only looking at accounts at default time. Although our dataset was able to provide detailed information on behavioural indicators for accounts over the course of the loan, and we were able to match macroeconomic indicators to the accounts at the relevant times, these indicators are ultimately unable to accurately reflect the individual borrower circumstances. Given the difficulties and individual intricacies involved in predicting outstanding balance and EAD for individual accounts, we believe the r-square values achieved here are commendable.

Following this work, we plan to incorporate stress testing into our risk models. We plan to combine PD, LGD and EAD models, and to stress test each component model independently. The obvious covariates to stress test within the models would be the macroeconomic variables; however, we would also like to consider methods which would allow us to stress the behavioural variables as well. It is not always clear how behavioural variables are affected by the economy, especially in the case of retail loans where the economy is expected to affect individuals differently and to varying degrees. The different

combinations of PD_{it}, LGD_{it} and EAD_{it} computed would enable us to get a distribution for $loss_{it}$, from which we expect to be able to predict for expected and unexpected losses better.

References

ALLISON, P. D. 2010. *Survival Analysis Using SAS: A Practical Guide* Cary, NC, SAS Institute Inc.

ARATEN, M. & JACOBS, M. J. 2001. Loan Equivalents for Revolving Credits and Advised Lines. *The RMA Journal*.

CAMERON, A. C. & TRIVEDI, P. K. 2005. *Microeconometrics: Methods and Applications*, Cambridge University Press.

DRUKKER, D. M. 2003. Testing for Serial Correlation in Linear Panel-Data Models. *The Stata Journal,* 3**,** 168-177.

GUJARATI, D. N. 2003. *Basic Econometrics*, McGraw Hill.

JACOBS, M. J. 2008. An Empirical Study of Exposure at Default. *Office of the Comptroller of the Currency Working Paper*.

JIMÉNEZ, G., LOPEZ, J. A. & SAURINA, J. 2009. Empirical Analysis of Corporate Credit Lines. *Review of Financial Studies,* 22**,** 5069-5098.

JIMÉNEZ, G. & MENCÍA, J. 2009. Modelling the distribution of credit losses with observable and latent factors. *Journal of Empirical Finance,* 16**,** 235-253.

LEOW, M. & CROOK, J. N. 2012. Intensity Models and Transition Probabilities for Credit Card Loan Delinquencies.

MORAL, G. 2006. EAD Estimates for Facilities with Explicit Limits. *In:* ENGELMANN, B. & RAUHMEIER, R. (eds.) *The Basel II Risk Parameters.* Springer Berlin Heidelberg.

QI, M. 2009. Exposure at Default of Unsecured Credit Cards. *Office of the Comptroller of the Currency Working Paper*.

THOMAS, L. C. 2010. Consumer Finance: Challenges for Operational Research. *Journal of the Operational Research Society,* 61**,** 41-52.

VERBEEK, M. 2004. *A Guide to Modern Econometrics*, John Wiley & Sons.