A New Mixture Model for the Estimation of Credit Card Exposure at Default

A New Mixture Model for the Estimation of Credit Card Exposure at Default

Mindy Leow[1,2] & Jonathan Crook[1,3]

[1] Credit Research Centre, University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, Scotland, United Kingdom
[2] Corresponding author: +44 131 650 9850; mindy.leow@ed.ac.uk
[3] j.crook@ed.ac.uk

A New Mixture Model for the Estimation of Credit Card Exposure at Default

Abstract

Using a large portfolio of historical observations on defaulted loans, we estimate Exposure at Default (EAD) at the level of the obligor by estimating the outstanding balance of an account, not only 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 a function of the spending and repayment amounts by the borrower and is also subject to the credit limit imposed by the card issuer. The predicted value is a weighted average of the estimated balance and limit, with weights depending on how likely the borrower is to have a balance greater than the limit. The weights are estimated using a discrete-time repeated events survival model to predict the probability of an account having a balance greater than its limit. The expected balance and expected limit are estimated using two panel models with random effects. We are able to get considerably more accurate predictions for outstanding balance, not only at the time of default, but at any time over the entire default loan period, than other techniques in the literature and that are used in practice.

Keywords: risk management, forecasting, 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 expected 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 estimate 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). 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. The data we have here are of retail loans, where the product is 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 predicted 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 models. 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. However, note that these terms are not universal and could be defined differently or have different

acronyms by different institutions or papers. Basel II refers to a Credit Conversion Factor, "CCF", but does not define it except to state that it is a factor of any further undrawn limit (see Basel Committee on Banking Supervision (2004), Paragraph 316, 474-478), so it is not clear that there is a standard industry practice towards EAD modelling. There is some reference in the literature to current industry practices (e.g. see Taplin et al. (2007), Risk Management Association (2004)), which indicates that portfolios are segmented, aggregated and an average LEQ is calculated.

Only a few papers have looked at EAD for corporate loans, and even fewer on retail loans, and it is common in practice 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), Jiménez et al. (2009), Jacobs Jr. (2008)) looked at the possible determinants of EAD, how EAD is affected by the economy and relationships between EAD and the behaviour of delinquent firms. Jiménez and Mencía (2009) take an overall view of credit risk and look at the time series of PD, EAD and LGD for consumer loans and mortgages as well as for many corporate sectors, but all at the aggregate level. They did not explicitly model EAD at account level and accounted for it at 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) looked at EAD for unsecured credit cards, trying to predict LEQ by looking at the level of credit drawn at one year before default. However, no application or 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. Taplin et al. (2007), working with data from business credit cards, criticized the use of LEQ (referred to as "CCF" in their paper) and attempted to breakdown the influences of balance and limit on balance at default at observation time. They proposed regression models that estimate EAD as a function of balance and limit, but did not give any indication of covariates used or any performance measures.

Using a large portfolio of defaulted loans and their historical observations, in this paper we directly estimate 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 default account, when used together with predictions for PD and LGD, we can predict the EAD when the loan goes 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 the probability of an account having a balance greater than its limit. Next, we develop two sub-models estimating for balance and limit, and the final prediction for balance is a product of the estimated balance and limit, depending on how likely the borrower is to have a balance greater than the limit. To make comparisons with current practices in the industry, we also estimate regression models for two alternative expressions for EAD, i.e. CCF and LEQ. The covariates of all 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, this is the first paper to predict the outstanding balance for defaulted loans 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 amount is calculated in each month for each account and accounts progress through states of arrears depending on whether they are able to make the minimum repayment amount. 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. 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 movement of accounts between states, see Leow and Crook (2014), but note that the percentage used here is different.

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 have £0 balance. 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

From the data, we see that some accounts go into default with an outstanding balance greater than their credit limit, 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.

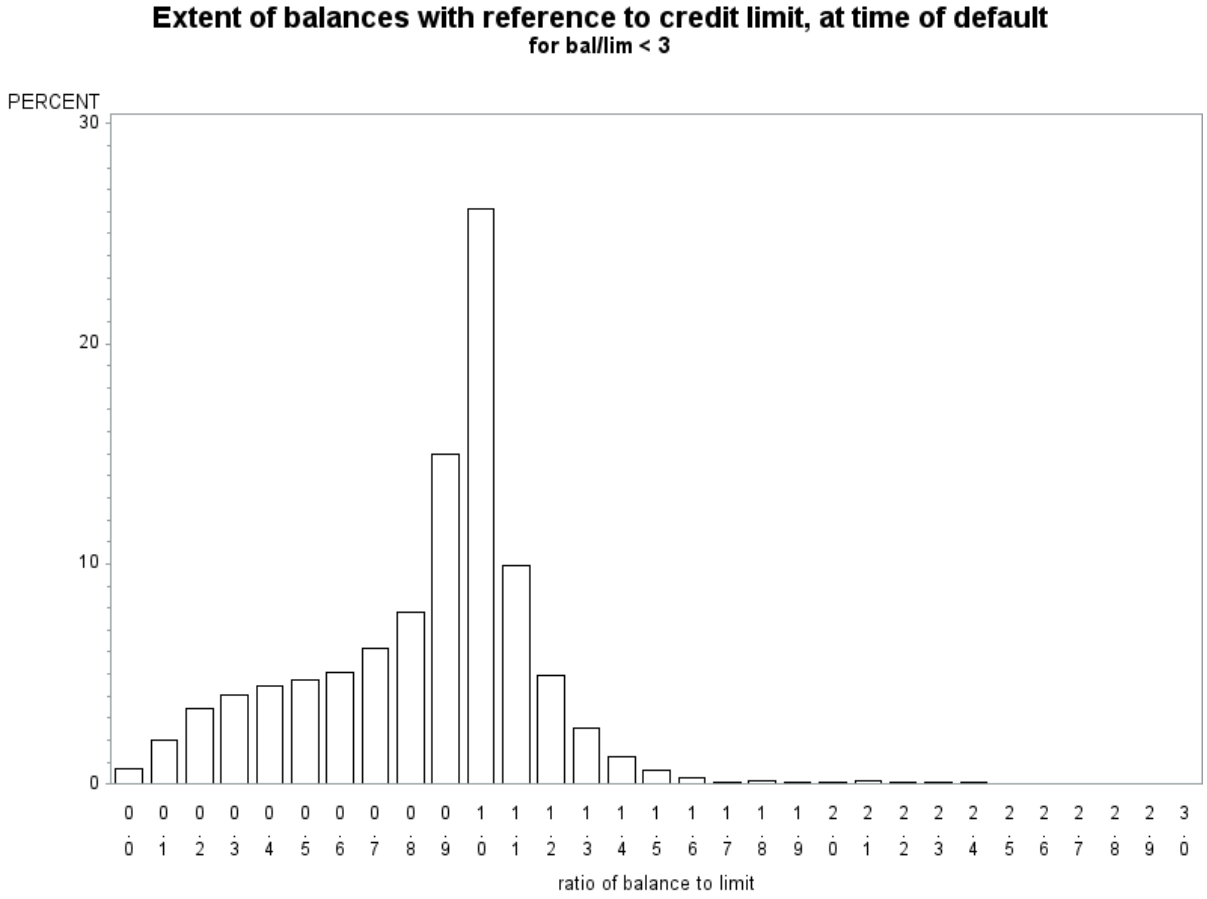

Figure 1: Distribution of ratio of balance over limit at time of default (for ratios less than 3)

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 *D* is replaced by balance at some duration time τ, and where the observation time is lagged *l* months, which 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.

We note that expressions for different factors in EAD estimation are not universal and could be defined differently in different papers. EADF, CCF and LEQ defined here are in keeping with the notations used in Jacobs Jr. (2008) and Qi (2009). From here on, outstanding balance of account *i* at duration time τ is represented by $B_{i\tau}$, and limit of account *i* at duration time τ is represented by $L_{i\tau}$. We also construct a binary variable $d_{i\tau}$ that takes on the value 1 if account *i* defaults at time τ and d_i that takes on the value 1 if account *i* defaults at some time in the future. To simplify the notation, the subscript *i* representing account *i* is dropped for the equations in this sub-section.

The EADF_D is the ratio of the balance at default time *D*, over the limit at observation time $D-l$, given in Equation 1:

$$\text{EADF}_D = \begin{cases} \dfrac{B_D}{L_{D-l}} & \text{for } L_{D-l} \neq 0 \end{cases} \tag{1}$$

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_D to range between 0 and 1, as shown in Figure 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. A further problem is that as an account moves towards default and its balance increases, lenders may respond different between accounts; in some cases increasing the limit, in others reducing the limit. This may introduce unexplained heterogeneity in a cross sectional model of EADF.

The CCF_D is the ratio of the balance at default time *D* over the balance at some observation time $D-l$, given in Equation 2:

$$\text{CCF}_D = \begin{cases} \dfrac{B_D}{B_{D-l}} & \text{if } B_{D-l} \neq 0 \\ 0 & \text{if } B_{D-l} = 0 \end{cases} \tag{2}$$

The CCF_D tries to get better predictions for balance 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 $\text{CCF}_D = 0$, 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. Should the account have a very low balance during observation time and defaults with a large balance, CCF could become an extremely large value, causing difficulties with data analysis and model estimation. Although on the one hand, it is likely that accounts that go into default have large balances on their account prior to default (for example, 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 (for example, debtors who default due to unexpected circumstances), which could then imply a different set of predictors for each group. From the point of prediction, a value of 0 for CCF does not make any sense as this would mean a prediction of £0 for balance at some time in the future, and possibly at default.

The LEQ_D 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 undrawn limit, i.e. the remaining amount of credit the debtor is able to draw upon, as given in Equation 3:

$$\text{LEQ}_D = \begin{cases} \dfrac{B_D - B_{D-l}}{L_{D-l} - B_{D-l}} & \text{if } L_{D-l} \neq B_{D-l} \\ 0 & \text{if } L_{D-l} = B_{D-l} \end{cases} \tag{3}$$

The different values that the LEQ_D can take could arise due to a number of different situations and which would give different implications. Should the account have zero undrawn limit, i.e. outstanding balance equal to limit, at the time of observation, we get an LEQ_D 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_D value computed does not have the same implications as the other LEQ_D values computed for when balance and limit are not equal. Also, when there is small undrawn limit, LEQ_D could become excessively large and unstable.

The majority of accounts would have a positive LEQ_D, 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. The latter 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_D in the same range, we expect their characteristics and circumstances to be quite different. It is also possible to have negative LEQ_D: (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_D values but where they have arrived via different

circumstances. The possible range of LEQ_D, coupled with the fact that different types of borrowers and circumstances could give LEQ_D in the same range, would make it difficult to estimate and model LEQ_D.

One weakness of several of the above methods is that according to how they are defined, these variables could become unstable if the denominator is very small. Many authors imposed restrictions on the range of values. Qi (2009) included only accounts at default time where undrawn limit is greater than 50 USD; Jacobs Jr. (2008) restricted the values of LEQ to between 0 and 1 and replaced outliers with the maximum and minimum values of his selected range. In his CCF model, he restricted the range of CCF to between 1 and 99 percentile, and replaced outliers with these maximum and minimum values. Both authors effectively ignored accounts that go from up-to-date to default suddenly or within a short time period, but this was the only way to get plausible results. Taplin et al. (2007) did not attempt to estimate LEQ (referred to "CCF" in their paper) as they would have to exclude about 50% of their observations. Notice also that predictive results from papers in the literature using these dependent variables have generally been poor. We therefore decided to focus on the estimation of outstanding balance (given default) itself, as we explain in Section 3. We also comment on using LEQ and CCF as alternatives in EAD modelling there.

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 oscillate 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.

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

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. We use the non-seasonally adjusted series unless unavailable because the balance and limit data are also not seasonally adjusted. Any macroeconomic variables used in the model are also lagged by 6 months.

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. We also believe that balances of defaulted and non-defaulted accounts behave differently, and we see from Figure 2 that balances of non-default accounts are on average lower, and have more occurrences of 0 than balances of default accounts. 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.

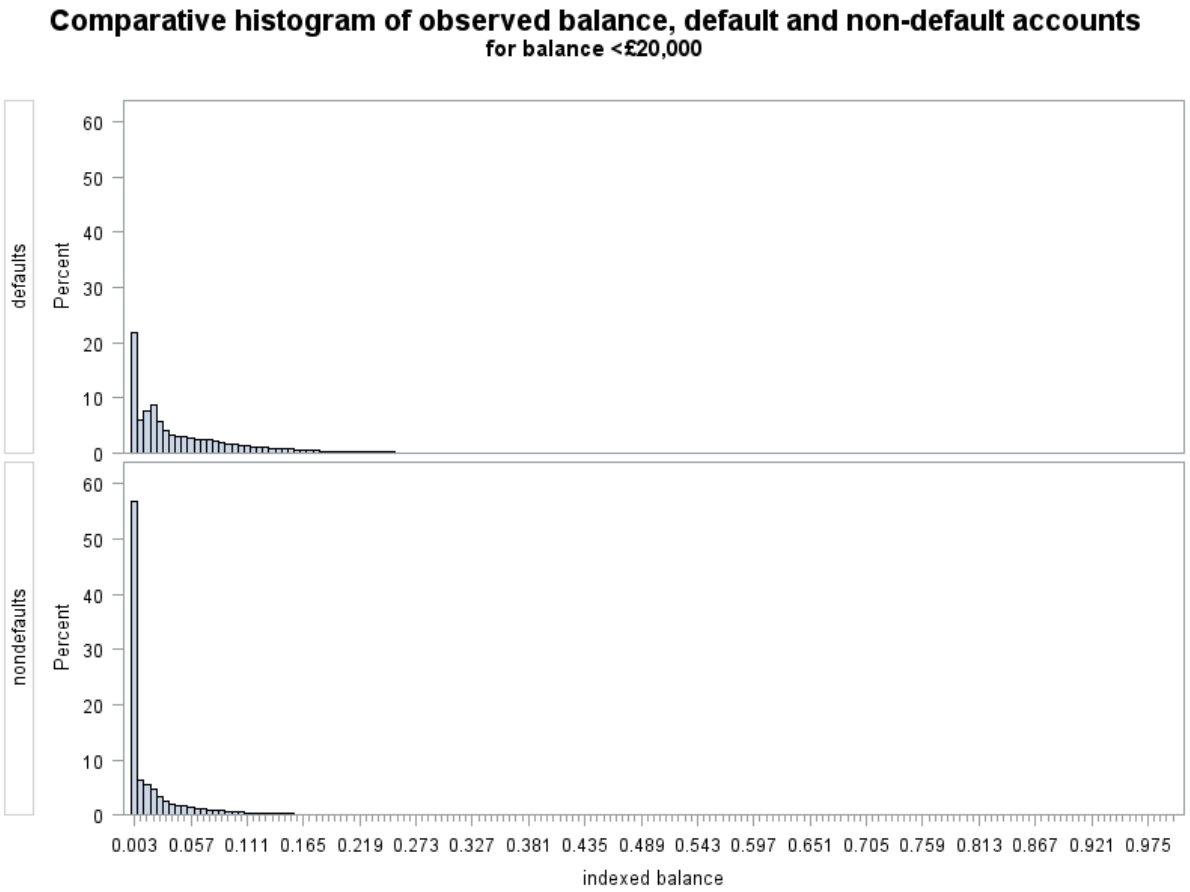

Figure 2: Distributions of observed balance, for default and non-default accounts, for balance less than £20,000.

The dataset is split such that the training set consists of all accounts that do go into default at some time and were opened on or before 31 December 2008, giving about 94,000 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 12,000 unique accounts, giving more than 66,000 month-account observations. A second test set is created (Test set II), a subset of Test set I, where only observations at the time of default 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

may not yet have gone into default, whilst Test set II would be an indication of how well the model is able to predict at default-time.

The portfolio of non-default accounts is not used in either the modelling or the testing as we estimate balance given default. Applying the mixture model to observations of non-default accounts would give us the predicted balance should the account go into default, which is different to the observed balance, as seen in Figure 2, which would mean that we will not be able to score how well the model is predicting.

3. Methodology

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

$$\begin{aligned} E\left(B_{i\tau} \mid d_i = 1\right) = &\left(P\left(B_{i\tau} = L_{i\tau} \mid d_i = 1\right) \times E\left(L_{i\tau} \mid B_{i\tau} = L_{i\tau}, d_i = 1\right)\right) \\ &+ \left(P\left(B_{i\tau} < L_{i\tau} \mid d_i = 1\right) \times E\left(B_{i\tau} \mid B_{i\tau} < L_{i\tau}, d_i = 1\right)\right) \end{aligned} \tag{4}$$

Since some accounts can have a value of $B_{i\tau}$ that is greater than the credit limit and we assume such accounts have an expected value equal to the expected limit, we replace the first probability condition in Equation 4 by $P\left(B_{i\tau} \geq L_{i\tau} \mid d_i = 1\right)$. We therefore parameterise three models. First, a model of the probability that the outstanding balance of an account is larger than the credit limit, conditional on default; second, a model to predict the outstanding balance, conditional on default; and third, a model to predict the credit limit conditional on default, where we allow parameters to predict balance and limit to differ.

From the training dataset based on only default accounts, i.e. accounts that eventually go into default, we first estimate the probability that the outstanding balance at any duration time τ is equal or greater than the limit at duration time τ. This is done by defining the event 'overstretched', $S_{i\tau}$, for account *i* at time τ which takes the value 1 if outstanding balance is greater than the limit at time τ; and 0 otherwise, given in Equation 5:

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

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).

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

where ν is the intercept term; X_i are account-dependent, time-independent covariates, i.e. application variables; $Y_{i,\tau-6}$ are account-dependent, time-dependent covariates, lagged 6 months, i.e. behavioural variables; $Z_{\tau-6}$ 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.

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	44,893	998,302	4	115	22.2
Limit	48,706	1,094,801	3	113	22.5
CCF	79,338	79,338	1	1	1
LEQ	80,552	80,552	1	1	1

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, i.e. that defaulted at some time, 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,* conditional on default. By structuring the sample in this way, we parameterise the distribution of $B_{i\tau}$ given $B_{i\tau} \geq L_{i\tau}$ and given default. We could not include only observations where $B_{i\tau} > L_{i\tau}$ given default because we wish to use the panel aspect of the data and such a condition is rare. The other subset consisting 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 $B_{i\tau} < L_{i\tau}$ 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 3.

The limit, $L_{i\tau}$, and balance, $B_{i\tau}$ for each account *i* at time τ are estimated using panel models with random effects. The specification is given in Equation 7.

$$y_{i\tau} = \mu + \gamma_1 X_i + \gamma_2 Y_{i\tau} + \gamma_3 Z_\tau + \alpha_i + \varepsilon_{i\tau} \tag{7}$$

where μ is the intercept term, X_i are account-dependent, time-independent covariates, i.e. application variables; $Y_{i\tau}$ are account-dependent, time-dependent covariates, i.e. behavioural variables; Z_τ 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 $\alpha_i + \varepsilon_{i\tau}$ is the error term, with $\alpha_i \sim IID\left(0, \sigma_\alpha{}^2\right)$ and $\varepsilon_{i\tau} \sim IID\left(0, \sigma_\varepsilon{}^2\right)$.

As Equation 7 is a standard specification of a panel model, 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)). 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.

$$\left[\hat{L}_{i\tau} \mid d_i = 1\right] = f\left(X_i^L; Y_{i,\tau-6}^L; Z_{\tau-6}^L\right) \tag{8}$$

$$\left[\hat{B}_{i\tau} \mid d_i = 1\right] = f\left(X_i^B; Y_{i,\tau-6}^B; Z_{\tau-6}^B\right) \tag{9}$$

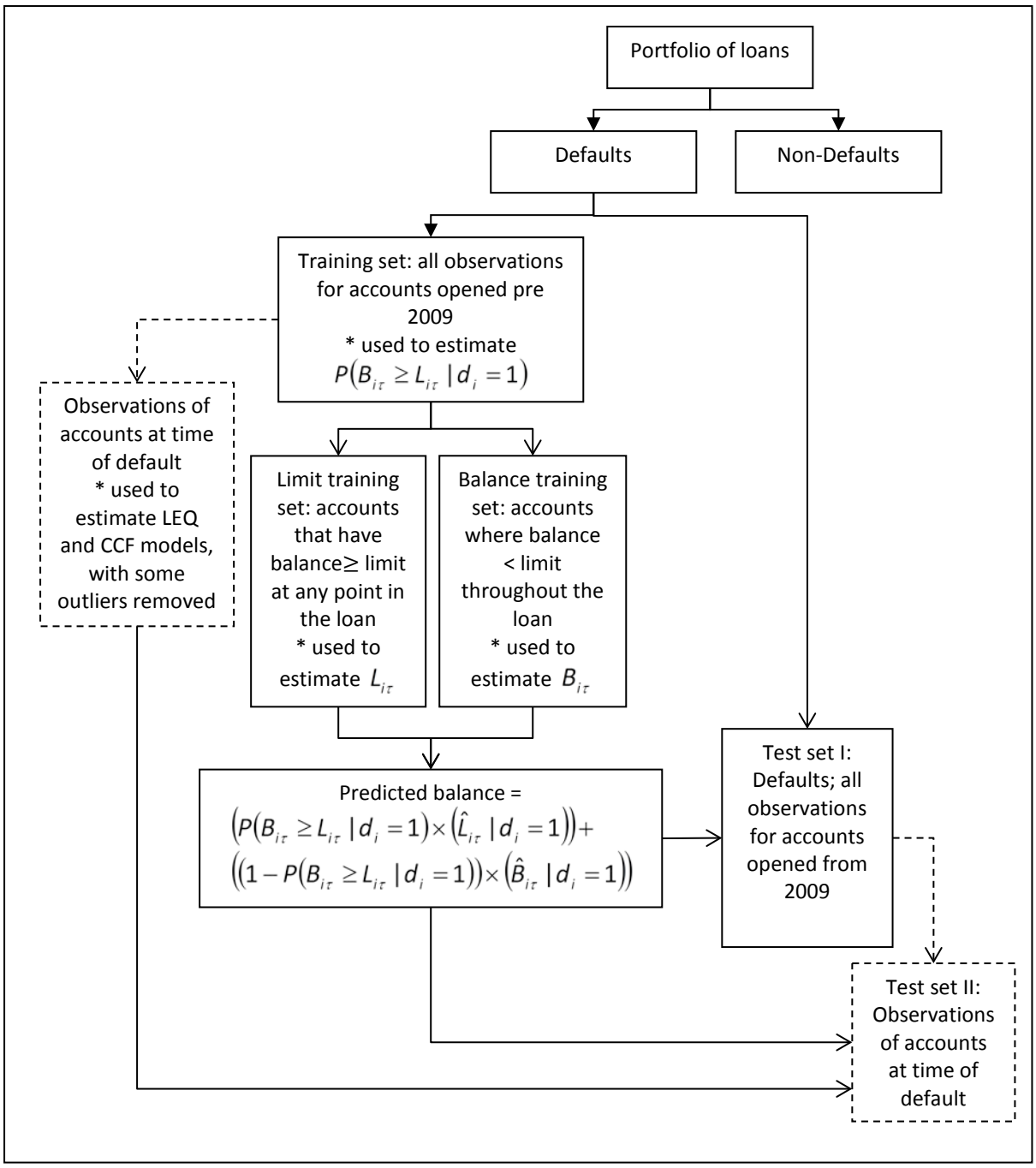

Figure 3: Flowchart of methodology and training and test set splits, where dotted lines represent subsets of the test and training sets that only consist of observations at default time.

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 accounts that default at some time in the observation window) by first applying the survival model to all accounts to predict the probability of being overstretched at each duration time τ. 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 τ. Because the models were estimated for the subsets described above, these predicted values for $B_{i\tau}$ and $L_{i\tau}$ are the values of $B_{i\tau}$ given $B_{i\tau} < L_{i\tau}$ and $L_{i\tau}$ given $B_{i\tau} \geq L_{i\tau}$ respectively, in both cases given default. The final predicted value for balance of an account *i* at duration time τ, given default, $\tilde{B}_{i\tau} \mid d_i = 1$, is then a combination of the repeated events survival model estimating the probability of balance exceeding limit at time τ, and the panel models estimating either balance or limit at time τ. This is the expected value of balance and limit, given the probabilities of the balance exceeding the limit at time τ, defined in Equation 10.

$$\left[\tilde{B}_{i\tau} \mid d_i = 1\right] = \left(P(S_{i\tau}) \times \left(\hat{L}_{i\tau} \mid d_i = 1\right)\right) + \left(\left(1 - P(S_{i\tau})\right) \times \left(\hat{B}_{i\tau} \mid d_i = 1\right)\right) \quad (10)$$

where $P(S_{i\tau}) = P(B_{i\tau} \geq L_{i\tau} \mid d_i = 1)$ and is the estimated probability that account *i* is overstretched at time τ, i.e. that the balance for account *i* at time τ exceeds the limit for account *i* at time τ; and $\hat{L}_{i\tau}$ and $\hat{B}_{i\tau}$ 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, $B_{i\tau}$, for accounts that default and predicted balances given default over time, $\tilde{B}_{i\tau}$, we are able to calculate a few performance measures, including r-square values, 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.

From the training set, observations only at time of default are used to estimate the LEQ and CCF cross-sectional regression models (represented by the dotted square from the training set in Figure 3). For all observations at time of default, CCF and LEQ are calculated based on observations 6 months before default, given in equations 2 and 3 respectively. Similar to

the EAD papers mentioned in the literature, some observations are then further excluded from this subset due to some very extreme observations of CCF and LEQ: default-time observations that have 0 or negative values of CCF, i.e. $CCF_D \leq 0$, and observations above the 95th percentile, i.e. $CCF_D > 9.6$ for the CCF model; default-time observations that have undrawn amount of less than £5 at observation time, i.e. $L_{D-6} - B_{D-6} < 5$, for the LEQ model. This is far from ideal as the observations being left out are genuine observations but which are many hundreds of standard deviations away from the mean. We note that these observations were not an issue in the mixture model. The final number of accounts and observations used in each training set is given in Table 2.

LEQ and CCF were regressed on the same covariates as those used in the survival model. To deal with outliers, Jacobs Jr. (2008) used a beta link generalised linear model for his estimation of LEQ and CCF. We find this is unnecessary in our data because the distribution of LEQ, once a sufficient number of outliers were excluded was almost normal, and the distribution of CCF could be made normal using a $\log_e$ transformation. These models are then applied onto the test sets (Test set I and Test set II) and performance measures are calculated. These two OLS regression models are not further documented in this paper.

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 4. 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 House Price Index (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.

4.2. Panel models for balance and limit

The parameter estimates for both panel models are given in Table 4. We acknowledge that the balance from 6 months previous is included as a variable in the balance model, and credit limit from 6 months previous is included as a variable 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.

The predictive statistics for panel models for balance and limit, based on the training set are given in Table 3. 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.

Table 3: Performance indicators for panel models, for training set

Model	**Overall R-square (train)**	σ_u	σ_e	ρ
Balance	0.3853	543.6539	936.3293	0.2521
Limit	0.9585	203.0972	384.1876	0.2184

Table 4: 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	-8.6457	21.4697	<.0001	-3062.8870	-13.95	<.0001	3902.6650	12.85	<.0001
Application variables										
ageapp_1	Age at application group 1	-	-	-	-	-	-	-	-	-
ageapp_2	Age at application group 2	-0.1528	71.2035	<.0001	-10.7200	-1.05	0.2930	9.3943	2.93	0.0030
ageapp_3	Age at application group 3	-0.2234	118.0295	<.0001	-2.3682	-0.18	0.8590	25.5336	5.83	<.0001
ageapp_4	Age at application group 4	-0.1696	53.477	<.0001	49.6194	3.12	0.0020	37.0790	7.06	<.0001
ageapp_5	Age at application group 5	-0.1762	48.6003	<.0001	99.4107	5.54	<.0001	59.9693	9.36	<.0001
ageapp_6	Age at application group 6	-0.2349	72.2019	<.0001	117.2014	5.96	<.0001	66.8007	9.37	<.0001
ageapp_7	Age at application group 7	-0.2753	75.8899	<.0001	124.4876	5.85	<.0001	79.3083	8.84	<.0001
ageapp_8	Age at application group 8	-0.3795	92.5247	<.0001	138.8549	6.12	<.0001	77.0292	7.08	<.0001
ageapp_9	Age at application group 9	-0.4559	83.465	<.0001	152.2019	5.44	<.0001	76.5170	5.69	<.0001
ageapp_10	Age at application group 10	-0.621	128.2749	<.0001	32.2786	1.18	0.2390	40.0365	2.92	0.0030
ECode_A	Employment code, group A	-	-	-	-	-	-	-	-	-
ECode_B	Employment code, group B	-0.0136	0.4701	0.4929	45.6885	2.72	0.0070	-10.7177	-1.77	0.0770
ECode_C	Employment code, group C	0.0847	2.6442	0.1039	-78.9320	-3.07	0.0020	-6.4185	-0.56	0.5740
ECode_D	Employment code, group D	-0.1658	41.4698	<.0001	-38.0135	-3.11	0.0020	89.9695	18.68	<.0001
ECode_E	Employment code, group E	-0.1139	51.9118	<.0001	2.1046	0.17	0.8660	87.3885	16.19	<.0001
INC_L	Income, ln	-0.1776	340.6791	<.0001	197.7804	8.7	<.0001	136.3684	18.56	<.0001
INC_M0	Binary indicator for missing or 0 income	-1.6829	340.0634	<.0001	1740.4150	8.57	<.0001	1200.9230	18.01	<.0001

LLine	Binary indicator for presence of landline	-0.0102	0.4131	0.5204	78.0352	7.37	<.0001	-	-	-
NOCards	Number of cards	-0.0883	215.4338	<.0001	29.4743	5.83	<.0001	19.9496	9.97	<.0001
TAAdd	Time at address (years)	0.00021	0.0574	0.8106	-	-	-	-	-	-
TWBank_MU	Binary indicator for missing or unknown time with bank	-0.1004	26.2442	<.0001	-	-	-	-	-	-
TWBank	Time with bank (years)	-0.0017	408.0276	<.0001	0.0548	0.87	0.3820	0.4464	15.56	<.0001
X_A	Variable X, group A	-	-	-	-	-	-	-	-	-
X_B	Variable X, group B	0.3385	353.3018	<.0001	-71.9259	-4.51	<.0001	-93.5658	-16.08	<.0001
X_C	Variable X, group C	0.4371	398.3749	<.0001	-110.8362	-6.48	<.0001	-65.1864	-10.49	<.0001
X_D	Variable X, group D	0.3234	242.1324	<.0001	-78.8786	-5.84	<.0001	-58.4376	-11.07	<.0001
X_E	Variable X, group E	0.5607	810.4988	<.0001	-95.8007	-5.21	<.0001	-199.8428	-24.36	<.0001
Behavioural variables, lagged 6 months										
ATRV_lag6	Average transaction value	-0.0008	279.5528	<.0001	0.1581	9.09	<.0001	0.0284	3.72	<.0001
CASC_lag6	Number of cash withdrawals	0.1397	68.0088	<.0001	-	-	-	-	-	-
CASV_lag6	Amount of cash withdrawal	9.6E-05	9.0589	0.0026	-	-	-	-	-	-
CRLM_lag6	Credit limit	-	-	-	0.1926	18.11	<.0001	0.8808	108.88	<.0001
JUMP_lag6	Rate of total jumps	0.5297	139.059	<.0001	169.4357	5.47	<.0001	-	-	-
PARR_lag6	Proportion of months in arrears	0.5224	98.2974	<.0001	-611.9711	-16.32	<.0001	-	-	-
PAYM_lag6	Repayment amount	-	-	-	-	-	-	0.0248	5.31	<.0001
SCBA_lag6	Outstanding balance	-	-	-	0.2748	24.65	<.0001	0.0258	3.78	<.0001
Macroeconomic variables, lagged 6 months										
AWEN_lag6	Average wage earnings	0.00028	0.2213	0.638	-	-	-	-	-	-
CIRN_lag6	Credit card interest rate	-	-	-	34.4334	6.62	<.0001	-93.6091	-45.33	<.0001

CONS_lag6	Consumer confidence	0.0107	93.7627	<.0001	14.1991	24	<.0001	-	-	-
HPIS_lag6	House Price Index	-0.0015	15.1903	<.0001	-4.5187	-20.25	<.0001	1.4577	24.58	<.0001
IOPN_lag6	Index of production	-0.004	85.2713	<.0001	-	-	-	-	-	-
IRMA_lag6	Base interest rate	-0.0212	13.4071	0.0003	-	-	-	-	-	-
LAMN_lag6	Amount outstanding, ln	0.884	29.4337	<.0001	-	-	-	-333.8840	-13.2	<.0001
LFTN_lag6	FTSE Index, ln	-0.2104	23.3489	<.0001	-221.5203	-10.15	<.0001	-	-	-
RPIN_lag6	Retail Price Index	0.00192	2.22	0.1362	18.6970	19.33	<.0001	-	-	-
UERS_lag6	Unemployment rate	-	-	-	74.0384	10.32	<.0001	3.1202	1.23	0.2170
Model specific required variables										
duration	Survival time (months) since last event	0.2009	1615.2273	<.0001	-	-	-	-	-	-
period	Number of times event has happened	-0.0376	4967.5318	<.0001	-	-	-	-	-	-
Time on books	Time on books (months)	-	-	-	5.4430	10.33	<.0001	3.8476	12.85	<.0001

4.3. Overall performance

After applying the mixture model onto the test sets, we compute overall r-square, Mean Absolute Error (MAE), Mean Error (ME) and the symmetric Mean Absolute Percentage Error (sMAPE) for the predicted versus the observed balance, given in Table 5. The sMAPE is able to circumvent the problem of having £0 balance that would mean dividing by 0 in the calculation of MAPE.

In the case where both observed and predicted balance are £0 (i.e. the prediction is accurate and there is 0 error), these observations are left out of the calculation of sMAPE as they do not contribute to the error. We see that the model is able to achieve a modest r-square of 0.57 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 0.61. This is a significant improvement from the r-square of 0.006 to 0.25 achieved by the CCF and LEQ models, which are commonly used in the industry. The rest of the performance measures also indicate that the CCF model has the worst performance with large MAE values of close to £1,000, compared to around £670 for the mixture model, and around the mid £700s in the LEQ model. The Mixture model also has the lowest sMAPE values.

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, on a sample of default time observations depending on whether the accounts were current or delinquent, and whether outliers were excluded from the model development. Jacobs Jr. (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.

Table 5: Performance measures for mixture model, LEQ model and CCF model, for test sets[4]

Model	Test set Index	Test set	Number of obs	R-square	Mean Absolute Error (MAE)	Mean Error (ME): Obs - Pred	sMAPE
Mixture model developed on default accounts	Test set I	default accounts, all observations	66,460	0.5722	672.82	-118.78	0.5738
	Test set II	accounts at time of default	11,734	0.6131	684.10	97.79	0.4260
LEQ model developed on default accounts at time of default, undrawn limit > £5	Test set I	default accounts, all observations	66,460	0.2583	715.02	212.70	0.5814
	Test set II	accounts at time of default	11,734	0.2538	789.29	489.66	0.4707
CCF model (ln CCF) developed on default accounts at time of default, CCF > 0 and truncated at 95 percentile	Test set I	default accounts, all observations	66,460	0.1035	891.12	348.52	0.9593
	Test set II	accounts at time of default	11,734	0.0061	1028.72	464.04	0.8441

We next look at the distributions of predicted balances, $\tilde{B}_{i\tau}$ from the CCF, LEQ and Mixture models and compare them against the distributions of observed balance over time, $B_{i\tau}$. Figure 4 compares the distributions of observed and predicted balances of Test set I, i.e. all observations of default accounts, where for the purpose of a clearer illustration, we look at values of balance between £0 and £20,000. Due to confidentiality agreements, all balance

[4] In order to have comparable MAE, ME and sMAPE values, the predicted CCF and predicted LEQ values are converted into predicted balances, according to equations 2 and 3 respectively.

values had to be indexed (on observed balance). From Figure 4, we see that although the distribution of predicted balance from the LEQ model most closely matches that of observed balance, the LEQ model actually underestimates balance, with a mean indexed balance of 0.0635 compared to an observed mean of 0.0741. The CCF model further underestimates balance with a mean indexed balance of 0.0556, and the Mixture model slightly overestimates balance, with a slightly higher mean indexed balance of 0.0808.

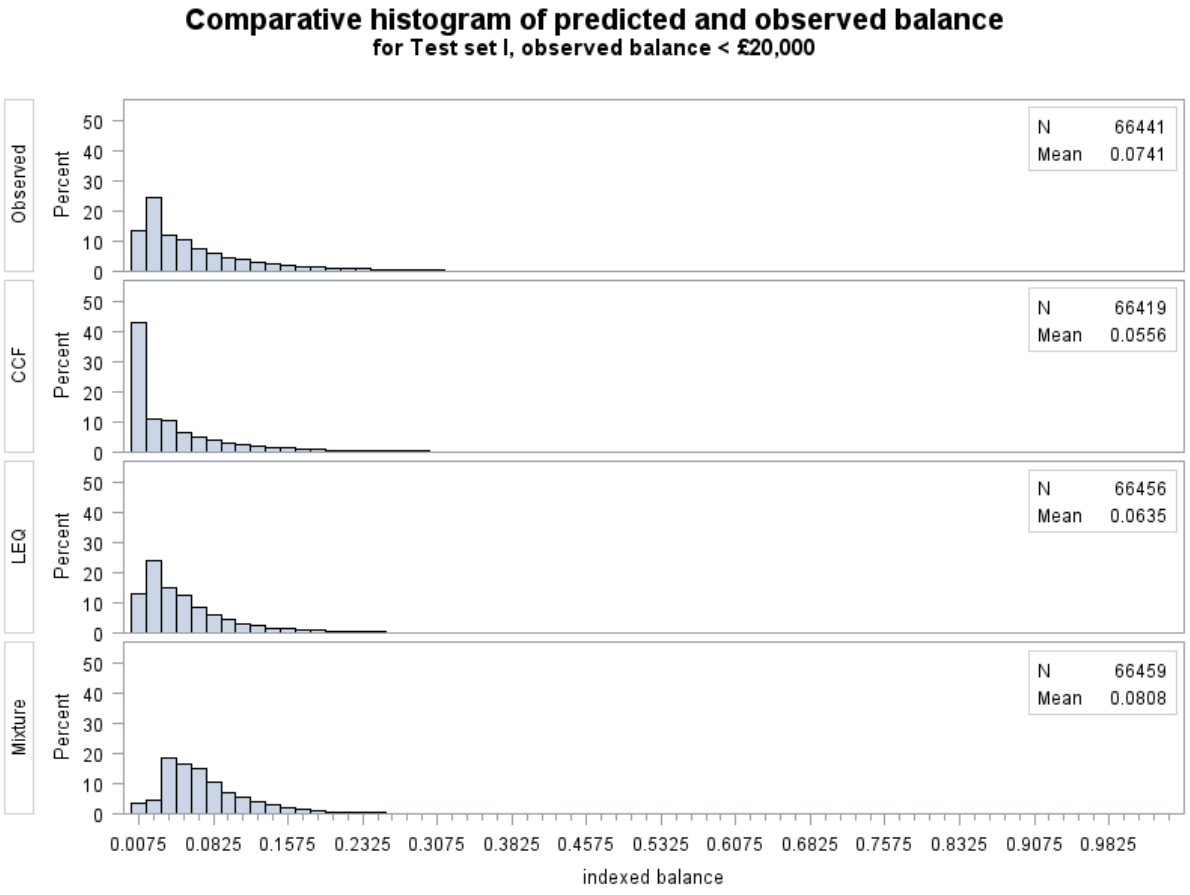

Figure 4: Comparative histogram of predicted and observed balances, indexed on observed balance, for Test set I, all observations for default accounts (where observed balance lies between £0 and £20,000). The top panel gives the distribution of observed balance, followed by the distribution of predicted balance from the CCF model, the LEQ model and the Mixture model in the second, third and fourth panel respectively. The number of observations and the mean indexed balance of each panel are also given.

Figure 5 compares the distributions of predicted and observed balances for Test set II, i.e. only default time observations for all default accounts. Again, the values of balance are

limited to between £0 and £20,000 for clearer representation of the distributions and all values of balances are indexed on observed balance. All three models underestimate balance here, where the LEQ and CCF models underestimate it by a larger extent than the mixture model, having a mean indexed balance of 0.0706, 0.0709 and 0.0915 respectively, compared to an observed indexed mean of 0.0957.

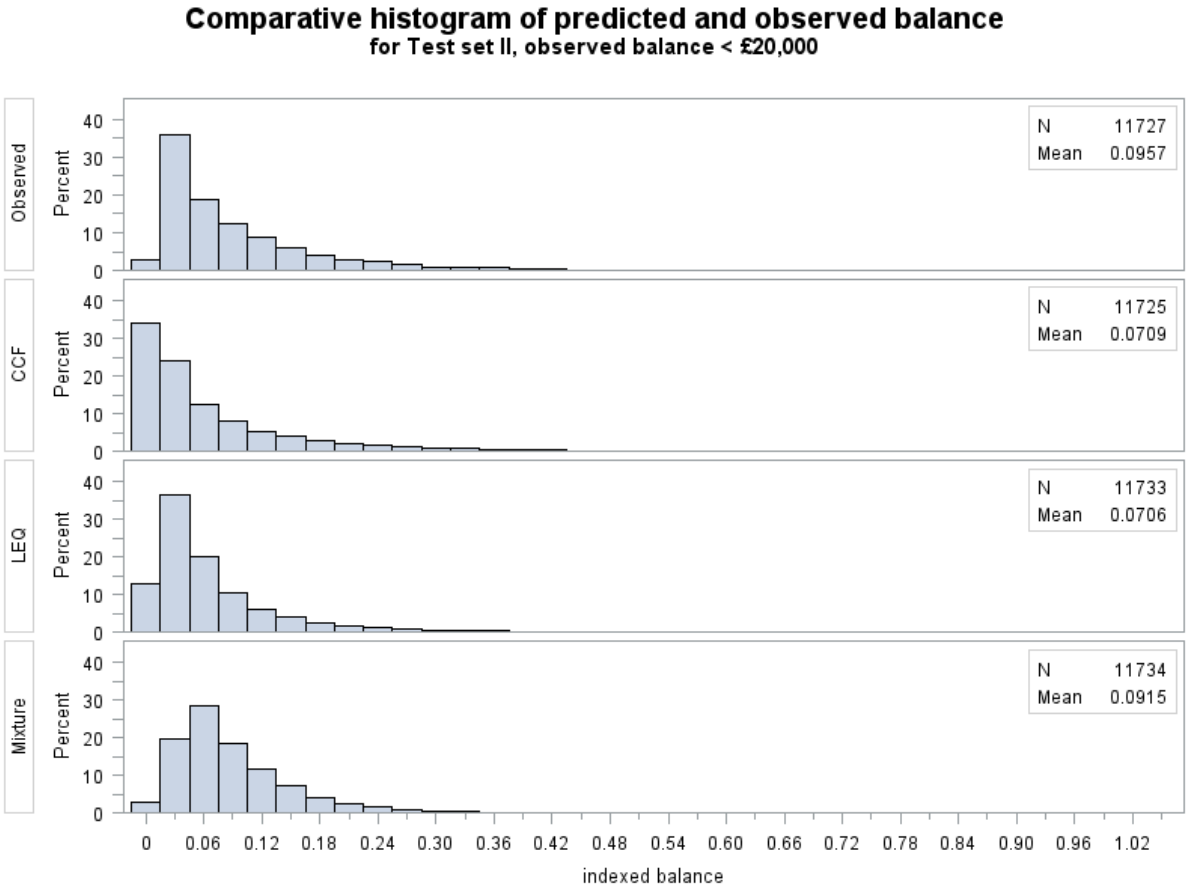

Figure 5: Comparative histogram of predicted and observed balances, indexed on observed balance, for Test set II, only observations at time of default (where observed balance lies between £0 and £20,000). The top panel gives the distribution of observed balance, followed by the distribution of predicted balance from the CCF model, the LEQ model and the Mixture model in the second, third and fourth panel respectively. The number of observations and the mean indexed balance of each panel are also given.

In the context of the calculation of economic capital, an overestimation of balance (and EAD) is preferable, although any overestimation also signifies opportunity cost. We also note that we are trying to predict 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 τ, given an account defaulted. 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 τ. 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 τ. The final prediction for balance at duration time τ is then said to be either the estimated limit or balance, depending on how likely the borrower is to be overstretched at that time τ. This is the sum of two products: the probability of being overstretched multiplied by the estimated limit; and the probability of not being overstretched multiplied by the estimated balance in both cases given default (c.f. Equation 10).

Using this mixture model, we find that we are able to get good predictions for outstanding balance for accounts that at some time default, not only at the time of default, but at any time over the entire loan period. This would allow us to make predictions for outstanding balance and hence EAD before default occurs, for delinquent accounts. Overall r-square values achieved are 0.57 when looking over the entire loan period for all delinquent accounts, and 0.61 when only looking at accounts at default time only. We also estimated regression models for CCF and LEQ, and found that not only could we only use observations at default time (whereas in the mixture model, we used all observations of default accounts), we were also forced to exclude a number of genuine observations in order to get sensible parameter estimates. This issue was not unique to our data and was also documented in Qi (2009), Jacobs Jr. (2008) and Taplin et al. (2007). The CCF and LEQ

models were also not able to predict for balance given default as accurately as our mixture model, achieving r-square values ranging from 0.006 to only 0.25. We also looked at the distributions of observed and predicted balances from all models, and find that the LEQ model consistently underestimates balance given default.

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 and consistently 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 by the mixture model are commendable. Overall, we believe the mixture model proposed here is a much better and more practical alternative to the LEQ and CCF models currently being used.

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 yet retain the knock-on effects in an adverse economic situation, if any. 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 $PD_{i\tau}$, $LGD_{i\tau}$ and $EAD_{i\tau}$ computed would enable us to get a distribution for $loss_{i\tau}$, 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*.

BASEL COMMITTEE ON BANKING SUPERVISION 2004. International Convergence of Capital Measurement and Capital Standards: A Revised Framework.

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 JR., M. 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. 2014. Intensity Models and Transition Probabilities for Credit Card Loan Delinquencies. *European Journal of Operational Research,* In Press.

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*.

RISK MANAGEMENT ASSOCIATION 2004. Industry Practices in Estimating EAD and LGD for Revolving Consumer Credits - Cards and Home Equity Lines of Credit.

TAPLIN, R., TO, H. M. & HEE, J. 2007. Modelling Exposure at Default, Credit Conversion Factors and the Basel II Accord. *Journal of Credit Risk,* 3**,** 75-84.

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.