

Accepted Manuscript

Exposure At Default Models with and without the Credit Conversion Factor

Edward N.C. Tong , Christophe Mues , Iain Brown ,
Lyn C. Thomas

PII: S0377-2217(16)00100-4
DOI: [10.1016/j.ejor.2016.01.054](https://doi.org/10.1016/j.ejor.2016.01.054)
Reference: EOR 13499



To appear in: *European Journal of Operational Research*

Received date: 16 February 2014
Revised date: 24 November 2015
Accepted date: 25 January 2016

Please cite this article as: Edward N.C. Tong , Christophe Mues , Iain Brown , Lyn C. Thomas , Exposure At Default Models with and without the Credit Conversion Factor, *European Journal of Operational Research* (2016), doi: [10.1016/j.ejor.2016.01.054](https://doi.org/10.1016/j.ejor.2016.01.054)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Highlights

- Direct EAD models ignore CCF formulation and select EAD as response variable
- Performance is compared to CCF and utilization change models
- Direct EAD model is more accurate in calibration than benchmark models
- Direct EAD and CCF based models can be combined to drive further performance uplift
- Direct EAD models without CCF formulation are an alternative for EAD modelling

ACCEPTED MANUSCRIPT

Exposure At Default Models with and without the Credit Conversion Factor

Edward N.C. Tong^{a, 1}*, Christophe Mues^b, Iain Brown^c,
Lyn C. Thomas^b

^a*Bank of America, New York, NY, USA*

^b*Southampton Business School, University of Southampton, Southampton, SO17
1BJ, United Kingdom*

^c*SAS UK, Wittington House, Henley Road, Marlow, SL7 2EB, United Kingdom*

* Corresponding author:

Edward Tong
Bank of America
One Bryant Park
New York, NY
10036
USA
Email: enctong@gmail.com

Email addresses:

enctong@gmail.com (E.N.C. Tong), c.mues@soton.ac.uk (C. Mues),

iain.brown@sas.com (I. Brown), l.thomas@soton.ac.uk (L.C. Thomas)

¹ The views expressed in the paper are those of the authors and do not represent the views of the Bank of America.

Abstract

The Basel II and III Accords allow banks to calculate regulatory capital using their own internally developed models under the advanced internal ratings-based approach (AIRB). The Exposure at Default (EAD) is a core parameter modelled for revolving credit facilities with variable exposure. The credit conversion factor (CCF), the proportion of the current undrawn amount that will be drawn down at time of default, is used to calculate the EAD and poses modelling challenges with its bimodal distribution bounded between zero and one. There has been debate on the suitability of the CCF for EAD modelling. We explore alternative EAD models which ignore the CCF formulation and target the EAD distribution directly. We propose a mixture model with the zero-adjusted gamma distribution and compare its performance to three variants of CCF models and a utilization change model which are used in industry and academia. Additionally, we assess credit usage - the percentage of the committed amount that has been currently drawn - as a segmentation criterion to combine direct EAD and CCF models. The models are applied to a dataset from a credit card portfolio of a UK bank. The performance of these models is compared using cross-validation on a series of measures. We find the zero-adjusted gamma model to be more accurate in calibration than the benchmark models and that segmented approaches offer further performance improvements. These results indicate direct EAD models without the CCF formulation can be an alternative to CCF based models or that both can be combined.

Keywords: exposure at default; credit cards; generalized additive models; regression; risk analysis

1. Introduction

The Basel II and III Accords define the standards for calculating regulatory capital requirements for banks across the world (Basel Committee on Banking Supervision, 2005, 2011). Under the Advanced Internal Ratings-Based approach (AIRB), banks are allowed to assess credit risk using their own internally developed models which target three key parameters for each credit facility: (i) Probability of Default, PD, (ii) Loss Given Default, LGD and (iii) Exposure at Default, EAD. These parameter estimates can be used to produce an estimate for the expected loss (EL) or to estimate the unexpected loss for which banks must hold capital. Beyond the purpose of calculating regulatory capital, these three

parameters have wide ranging uses for banks, serving as inputs into economic capital models, stress testing, impairment forecasting, pricing and informing portfolio management across retail, corporate and wholesale portfolios.

In retail credit risk, PD modelling has been the main focus of credit research for several decades and in recent years, LGD models with challenging bimodal distributions have also been the focus of research (Loterman et al., 2012). Although EAD distributions are comparatively as difficult to model, they have received much less attention in the literature.

For credit card and overdraft portfolios, EAD estimation has proven a hard problem to tackle in practice. For fixed exposures such as residential mortgages and personal loans, the estimate for EAD can simply be taken from the current on-balance amount and little if any modelling is required. For credit cards though, the revolving nature of the credit line poses challenges with regards to predicting the exposure at default time. As credit card customers may borrow more money in the months prior to default, simply taking the current balance for non-defaulted customers would not produce a conservative enough estimate for the amount drawn by the time of default. The EAD could partially be driven by current or recent customer behaviour (i.e. credit usage, drawn, undrawn amounts, changes to undrawn amounts over time). As an example of two distinct behaviour groups, some customers, classified as transactors, tend to pay off their entire balance at the end of each month while others, termed revolvers, tend to pay off only part of the monthly balance and hence incur interest charges (So et al., 2014).

To estimate the EAD for credit cards or other forms of revolving credit, the Basel II/III Accord has suggested the use of historic data to evaluate the Credit Conversion Factor (CCF), i.e. the proportion of the current undrawn amount that will likely be drawn down at time of default (Valvonis, 2008). The Accord did not explicitly require EAD models to use CCF calculations; however, CCFs are regularly referred to in the Accord. Once a CCF estimate is produced for a (segment of) variable exposure(s), the EAD is then given by: $EAD = \text{Current Drawn Amount} + (CCF \times \text{Current Undrawn Amount})$. With this (indirect) approach, the accuracy of EAD prediction is obviously linked to the quality of the CCF model and such modelling has posed substantial challenges because the distribution of CCF does not conform to standard statistical distributions. CCF distributions tend to be highly bimodal with a probability mass at zero (no change in balance), another at one (borrowing has gone up to the

credit limit), and a relatively flat distribution in between, not unlike some LGD distributions (Loterman et al., 2012). Furthermore, in many CCF datasets, one might see a substantial number of negative CCFs and CCFs greater than one (an example of the latter may be where the credit limit has increased between the point of observation and the time of default, allowing the customer to go over the original limit); since the final model estimates themselves would have to be constrained between zero and one, such individual observations are sometimes truncated to zero or one, respectively (Jacobs Jr, 2010).

Traditional regression modelling with ordinary least squares (OLS) may be less suitable for the CCF because predicted values may be less than zero or greater than one, leading to invalid CCF predictions. Additionally, the non-normality of the error term undermines many of the OLS tests. Standard logistic regression commonly used for PD models would also be inappropriate because the CCF response variable is proportional and not binary. Appropriate discretization of the CCF response would be necessary, which could result in some information loss, or alternatively, fractional response regression should be considered.

Taplin et al. (2007) have argued that the CCF formulation is problematic because the bounded CCF distribution forces EAD to be equal to the credit limit when CCF equals 1. In practice, it is common to find accounts with EAD greater than the credit limit from charges accrued due to additional purchases over the limit and interest charges, or credit limit changes. The authors instead suggested models that predict EAD directly and ignore the CCF formulation. However, Yang and Tkachenko (2012) have contended that CCF models are more suitable given that the EAD response variable may be too statistically difficult to model given the granular scale of currency amounts and that the CCF formula is less prone to such scaling issues with its range being limited to the unit interval.

The aims of this paper are to empirically assess alternative statistical methods for modelling the EAD by targeting the EAD distribution directly rather than focusing on the CCF; to evaluate this, we use a credit card portfolio from a large UK bank. We hypothesize that competitive EAD models can be developed by ignoring the CCF formulation and instead selecting EAD as the response variable in a statistical model. Two different direct EAD models are considered – an OLS model and a zero-adjusted gamma model (Rigby and Stasinopoulos, 2005, 2007).

The zero-adjusted gamma (ZAGA) model was explored to deal with the positively skewed nature of EAD and considering its prior use in predicting the LGD amount of residential mortgage loans (Tong et al., 2013). In this model, the EAD amount is modelled as a continuous response variable using a semi-parametric discrete-continuous mixture model approach with the zero-adjusted gamma distribution. Firstly, as the non-zero or positive EAD amount displays right-skewness, it is modelled with the gamma distribution. The mean and dispersion of the positive EAD amount are modelled explicitly as a function of explanatory variables. Secondly, the probability of the (non-)occurrence of a zero EAD amount is modelled with a logistic-additive model. All mixture components, i.e. the logistic-additive component for the probability of zero EAD and the log-additive components for the mean and dispersion of the EAD amount conditional on there being a non-zero EAD, can be estimated using account-level behavioural characteristics.

The performance of these direct EAD models are benchmarked against three CCF models (with CCF as the response variable) using OLS, Tobit and fractional response regression and the utilization change model. These approaches are established methods used in industry and/or academia for EAD and LGD modelling (Brown, 2011, Bellotti and Crook, 2012, Bijak and Thomas, 2015).

When borrowers are already close to maxing out the credit line and the undrawn amount is low, the CCF can become highly volatile and model performance may be compromised (Qi, 2009). Therefore, a combined approach is suggested that segments on credit usage (i.e. utilization rate, or the percentage of the committed amount that has been currently drawn) and then uses two separate models, with either the CCF or EAD as the response variable, depending on the utilization segment that the credit card falls into. We hypothesize that the combined use of CCF modelling for accounts with low utilization and direct EAD models for accounts with high utilization may improve the overall model performance.

Our dataset included time to default as a variable. In practical model development, this variable would be considered unknown a priori for each customer and would not typically be used as a candidate covariate in predictive model fitting. Nonetheless, it has been used in previous empirical studies to study explanatory drivers of CCF (Moral, 2011, Brown, 2014, Jacobs Jr, 2010). Therefore, discarding it would make our results less comparable to those reported by others. Furthermore, it would be interesting to explore this time effect on the

various components of the ZAGA model, particularly the dispersion component as one would intuitively expect the error variance to increase the more time elapses between the point of observation and default.

To allow a model with time to default as one of its explanatory variables to be applied to a prediction task, we propose an additional survival analysis model component. Survival analysis has previously been employed to model time to default in retail loan portfolios, providing insight into factors that predict *when* consumers are more likely to default (Stepanova and Thomas, 2002, Malik and Thomas, 2010, Tong et al., 2012). Similarly, we develop a PD model using the Cox proportional hazards model (Cox, 1972, Hosmer et al., 2008) with time to default as the event of interest but with the length of the cohort period as time horizon. We then show how the resulting monthly PD estimates can be combined with an EAD model that has time to default included as a covariate. This method for modelling EAD using a consistent probabilistic definition and a direct EAD estimation approach was proposed by Witzany (2011). Their research termed this method the ‘weighted PD approach’ and suggested the use of default intensities to estimate EAD by considering the time to default. Our paper extends their work by using a real banking dataset and explicit use of the Cox proportional hazards model. Leow and Crook (2015) have also combined survival and panel modelling methods comprising credit limit and drawn balance models to predict EAD for credit cards. We suggest this method could further incorporate the time to default as a predictive covariate in an EAD model to improve model performance.

The novel aspects of our study thus are that we (1) evaluate whether competitive EAD models can be developed by targeting the EAD distribution directly without using a CCF component, (2) assess credit usage as a segmentation criterion allowing us to combine two types of EAD models to further improve performance, (3) compare the performance of these new approaches to CCF and utilization change models commonly used in industry and/or academia and (4) propose an additional survival analysis component to allow the use of time to default as a predictive covariate in EAD modelling. All models will be assessed out-of-sample using cross validation on a series of discrimination and calibration measures.

The remainder of the paper is organized as follows. In section 2, an overview of the dataset along with the application and behavioural characteristics used for the EAD models will be presented. The statistical and validation methods used in our experiments are discussed in

section 3. Next, the results of the models are discussed in section 4. Section 5 will conclude the paper and suggest some further avenues for research.

2. Data

The dataset consisted of 10,271 observations of accounts from a major UK bank. The dataset derived from a credit cards portfolio observed over a three year period from January 2001 to December 2004. In the absence of additional data about other potential default triggers, for the purpose of this study, a default occurred when a charge off or closure was incurred on the credit card account. A charge off in this case was defined as the declaration by the creditor that an amount of debt is unlikely to be collected, declared at the point of 180 days or 6 months without payment. To compute the observed CCF value, the original data set was divided into two twelve-month cohorts. The first cohort ran from November 2002 to October 2003 and the second cohort from November 2003 to October 2004. In the cohort approach for CCF, discrete calendar periods are used to group defaulted facilities into 12-month periods, according to the date of default. Data was then collected on candidate EAD risk factors and drawn/undrawn amounts at the beginning of the calendar period and drawn amount at the default date.

Figure 1 shows the empirical CCF distribution after truncation; the mean CCF value here was 0.515 (sd = 0.464). The value is similar to that of S&P and Moody's defaulted borrowers' revolving lines of credit from 1985 to 2007, as reported by Jacobs Jr (2010); there, the truncated mean was 0.422 (sd = 0.409). Note that the bimodal nature of Figure 1 shows similarities to reported LGD distributions (Loterman et al., 2012, Bellotti and Crook, 2012). Figure 2 displays the distribution we observed for the EAD, clearly showing the positively skewed nature of this variable. Please note that some of the scales on the figures in this study have been removed for data confidentiality reasons.

Figure 1. Distribution of the Credit Conversion Factor (after truncation).

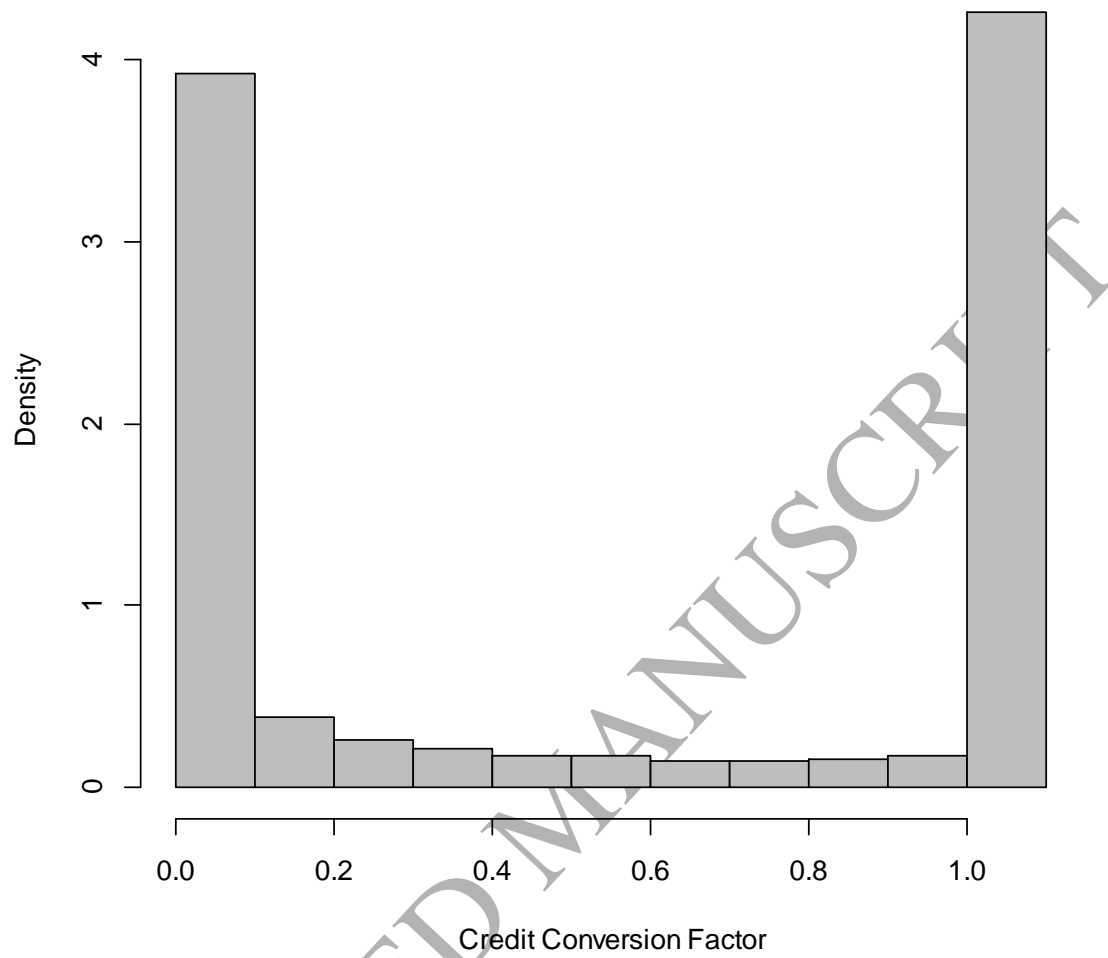
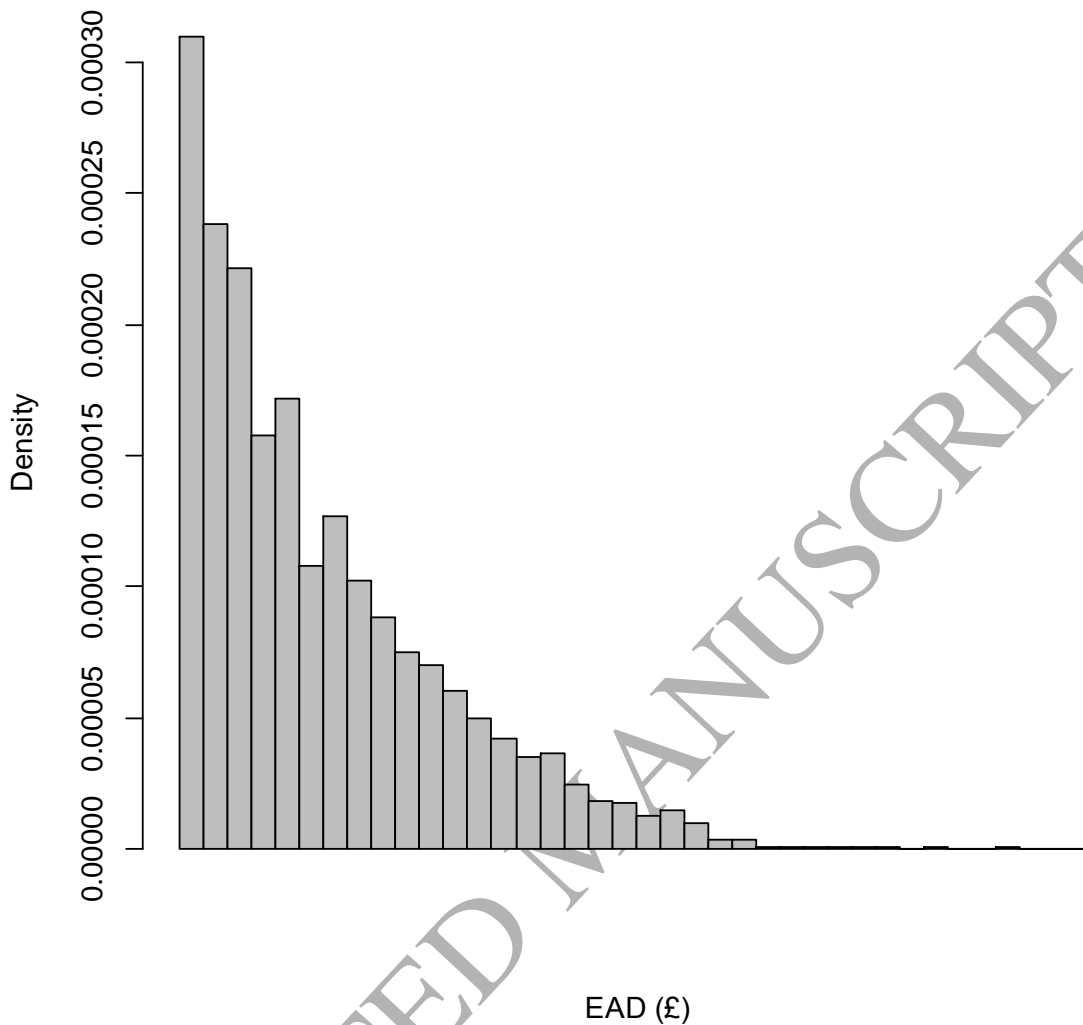


Figure 2. Distribution of observed Exposure at Default.



As shown in Table 1, a total of 11 candidate variables were considered for the models. The first six candidate variables in Table 1 were suggested by Moral (2011). They were generated from the monthly data in each of the cohorts, where t_d is the default date and t_r is the reference date (i.e. the start of the cohort). The latter five variables were previously suggested in Brown (2011), with the aim of improving the predictive performance of the models.

The credit conversion factor for account i , CCF_i , was calculated as the ratio of the observed EAD minus the drawn amount at the start of the cohort over the credit limit at the start of the cohort minus the drawn amount at the start of the cohort, i.e.:

$$CCF_i = \frac{E(t_d)_i - E(t_r)_i}{L(t_r)_i - E(t_r)_i} \quad (1)$$

Table 1. Candidate variables considered for EAD models.

| Variable(s) | Notation | Description |
|-------------------------|----------------------------------|--|
| Committed amount | $L(t_r)$ | Advised credit limit at start of cohort |
| Drawn amount | $E(t_r)$ | Exposure at start of cohort |
| Undrawn amount | $L(t_r) - E(t_r)$ | Limit minus exposure at start of cohort |
| Drawn percentage | $\frac{E(t_r)}{L(t_r)}$ | Exposure at start of the cohort divided by credit limit at start of the cohort (also commonly referred to as utilization rate or credit usage) |
| Time to default | $t_d - t_r$ | Default date minus reference date (months) |
| Rating class | $R(t_r)$ | Behavioural score at start of cohort grouped into 4 bins: (1) AAA-A, (2) BBB-B, (3) C, (4) Unrated |
| Average days delinquent | | Average number of days delinquent in previous 3, 6, 9, or 12 months |
| Undrawn percentage | $\frac{L(t_r) - E(t_r)}{L(t_r)}$ | Undrawn amount at start of cohort divided by credit limit at start of cohort |
| Limit increase | | Binary variable indicating increase in committed amount since 12 months prior to start of cohort |
| Absolute change drawn | | Absolute change in drawn amount: variable amount at t_r minus variable amount 3, 6 or 12 months prior to t_r |
| Relative change drawn | | Relative change in drawn amount: variable amount at t_r minus variable amount 3, 6 or 12 months prior to t_r , divided by variable amount 3, 6 or 12 months prior to t_r , respectively. |

3. Statistical models

The following sections outline the different statistical modelling approaches used to regress the EAD, CCF or utilization change against the candidate drivers listed in Table 1. The direct EAD models (i.e. those with EAD as the response variable) are described in Section 3.1. The three types of CCF models used are outlined in Section 3.2. The utilization change model is described in Section 3.3. The segmented models are introduced in Section 3.4 and the survival model add-on is outlined in Section 3.5. Finally, the process of model validation and testing is described in Section 3.6.

3.1. Direct EAD models

3.1.1. Zero-adjusted gamma model

The credit cards portfolio is stratified into two groups, the first group having zero EAD (in the absence of further data, we have to assume these may potentially include a number of special or technical default cases, charge-offs related to other accounts, truncated/rounded observations, transfers of the outstanding amount to other repayment arrangements, or they could be the result of late payments subsequent to the default trigger entering the EAD calculation) and a second group having non-zero EADs. The latter appears to have a continuous positively skewed distribution (see Figure 3) and accounts for the large majority of cases.

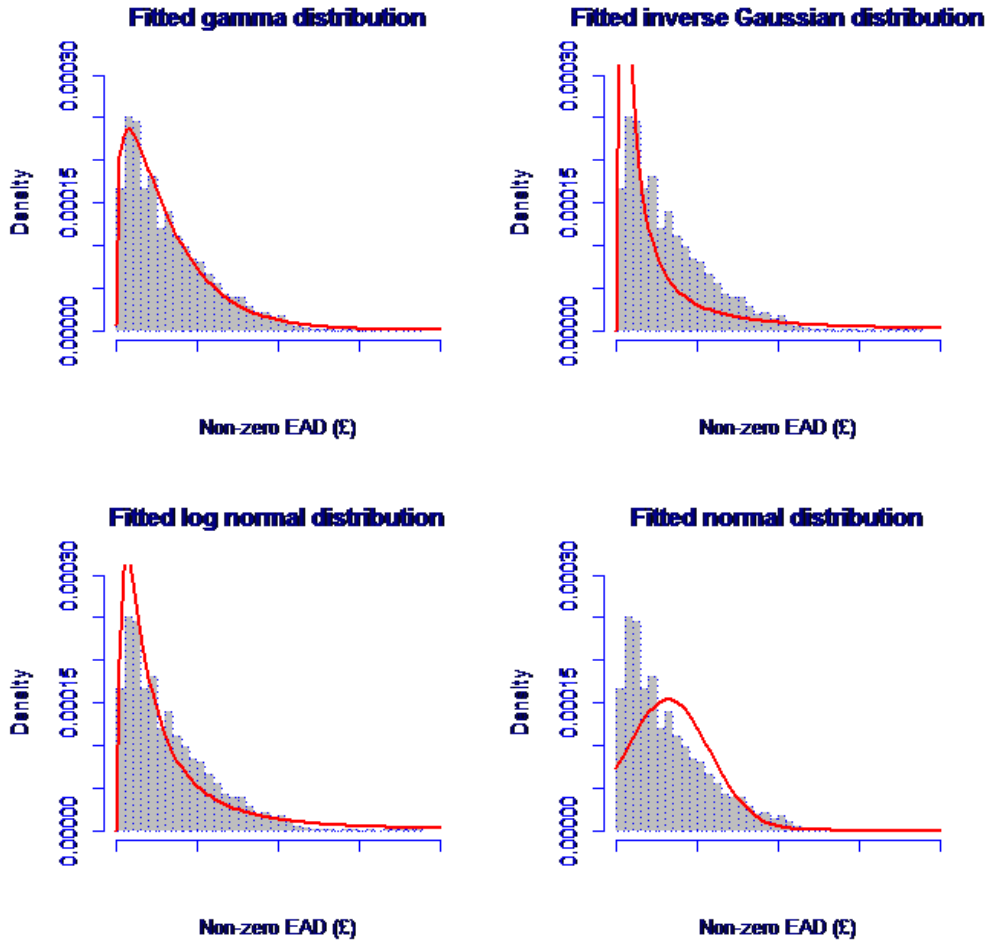
Let y_i denote the EAD observed for the i th account, $i = 1, \dots, n$ (for simplicity, the index i will be omitted from here on); \mathbf{x} will be used to denote the vector of covariates observed for the account. A mixed discrete-continuous probability function for y can then be specified as:

$$f(y) = \begin{cases} \pi & \text{if } y = 0 \\ (1-\pi)g(y) & \text{if } y > 0 \end{cases} \quad (2)$$

where $g(y)$ is the density of a continuous distribution and π is the probability of zero EAD.

Figure 3 shows different candidate distributions for $g(y)$ fitted to the non-zero EADs. Three positively skewed distributions were explored: the gamma, inverse Gaussian and log normal distributions; the normal distribution is shown as a reference comparison. The candidate distributions were fitted onto a training set of a random representative sample. Figure 3 indicates that the gamma distribution produced the most suitable fit for the histogram of positive EADs. There was further support for the fitted gamma distribution as it produced the lowest Akaike Information Criteria (AIC) when compared to the inverse Gaussian and log normal distributions. The zero-adjusted gamma distribution was hence selected to model $f(y)$. The resulting model will be referred to in this paper as ZAGA-EAD.

Figure 3. Candidate continuous distributions for non-zero EAD on training set.



The probability function of the ZAGA (μ, σ, π) model, a mixed discrete-continuous distribution, is defined by Rigby and Stasinopoulos (2010):

$$f(y | \mu, \sigma, \pi) = \begin{cases} \pi & \text{if } y = 0 \\ (1 - \pi) \text{Gamma}(\mu, \sigma) & \text{if } y > 0 \end{cases}$$

for $0 \leq y < \infty$,

where $0 < \pi < 1$, mean $\mu > 0$, dispersion $\sigma > 0$,

$$\text{Gamma}(y, \mu, \sigma) = \frac{1}{(\sigma^2 \mu)^{1/\sigma^2}} \frac{y^{\frac{1}{\sigma^2}-1} e^{-y/(\sigma^2 \mu)}}{\Gamma(1/\sigma^2)}$$

with:

$$E(y) = (1 - \pi)\mu \quad \text{and} \quad \text{Var}(y) = (1 - \pi)\mu^2(\pi + \sigma^2) \quad (4)$$

The ZAGA-EAD model is implemented using the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework developed by Rigby and Stasinopoulos (2005). Their approach allows a range of skewed and kurtotic distributions to explicitly model distributional parameters that may include the location/mean, scale/dispersion, skewness and kurtosis as functions of explanatory variables. GAMLSS also allows fitting of distributions that do not belong to the exponential family as provided in the Generalized Linear Model (GLM) (Nelder and Wedderburn, 1972) and Generalized Additive Model (GAM) frameworks (Hastie et al., 2009).

The GAMLSS approach is a semi-parametric method that allows the relationship between the explanatory variables and response variable to be modelled either parametrically (e.g. where linearity is met), or non-parametrically, using spline smoothers, the latter of which is a key feature of the GAM approach.

There are three components to the ZAGA-EAD model. The mean, μ , and dispersion, σ , of a non-zero EAD and the probability of zero EAD, π , are modelled as a function of the explanatory variables using appropriate respective link functions:

$$\begin{aligned} \log(\mu) &= \eta_1 = \mathbf{x}'_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \\ \log(\sigma) &= \eta_2 = \mathbf{x}'_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}(x_{j2}) \\ \text{logit}(\pi) &= \eta_3 = \mathbf{x}'_3\beta_3 + \sum_{j=1}^{J_3} h_{j3}(x_{j3}) \end{aligned} \quad (5)$$

where $\mathbf{x}'_k\beta_k$ denote parametric terms, $h_{jk}(x_{jk})$ are non-parametric terms such as smoothing splines and with $k = 1, 2, 3$ for the distribution parameters (hence, each model component can have its own selection of covariates). The dispersion of non-zero EAD is the squared coefficient of variation, δ^2/μ^2 , from the exponential family for the gamma density function (McCullagh and Nelder, 1989) where δ^2 denotes the variance of the non-zero EAD

distribution. The $h_{jk}(x_{jk})$ functions are modelled with penalized B -splines (Eilers and Marx, 1996). Such non-parametric smoothing terms have the ability to find non-linear relationships between the response and predictor variables (Hastie et al., 2009). Penalized B -splines were chosen because they are able to select the degree of smoothing automatically using penalized maximum likelihood estimation. This selection was done by minimizing the Akaike Information Criterion, i.e. $AIC = -2L + kN$, with L the log (penalized) likelihood, k the penalty parameter (set to 2), and N the number of parameters in the fitted model (Akaike, 1974). Automatic selection of smoothing may suggest non-linear or linear relationships to the response variable as discovered in the data.

Each account, i , in this model is associated with a probability of zero EAD, π_i , and a non-zero EAD amount, y_i . These pairs are then used to form the following likelihood function:

$$\begin{aligned} L &= \prod_{i=1}^n f(y_i) \\ &= \prod_{y_i=0} \pi_i \prod_{y_i>0} (1-\pi_i) \text{Gamma}(\mu_i, \sigma_i) \end{aligned} \quad (6)$$

An algorithm developed by Rigby and Stasinopoulos (2005) was used, which is based on penalized (maximum) likelihood estimation. The estimates of the probability of zero EAD, mean and dispersion of $g(y)$ are used to compute an estimate for $f(y)$ which combines the probability of EAD and the EAD amount given that there is a non-zero EAD.

The model was developed and implemented using the *gamlss* package by Rigby and Stasinopoulos (2007) in R 3.0.1 software (R Development Core Team, Vienna, Austria).

3.1.2. Ordinary least squares

The second direct EAD model was based on a standard OLS regression of the EAD response (untransformed) against the explanatory variables. We denote this model as OLS-EAD. A parsimonious model was selected through stepwise selection and backward elimination based on a 5 percent α -level.

3.2. CCF models

Three models comprising OLS, Tobit and fractional response regression were developed to predict the CCF (rather than the EAD directly). An account-level estimate for EAD is then derived from the predicted CCF as follows:

$$EAD = \text{Current Drawn Amount} + (CCF \times \text{Current Undrawn Amount}) \quad (7)$$

Firstly, a standard OLS regression model, denoted OLS-CCF, was fitted for the CCF target. Secondly, a Tobit regression model (Tobin, 1958, Greene, 1997), denoted Tobit-CCF, was developed, which treats observations with CCF below zero and above one as censored with the response only observed in the interval $[0, 1]$. The Tobit model assumes a latent variable y^* , for which the residuals conditional on covariates \mathbf{x} are normally distributed. The two-sided Tobit model is given by:

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \quad (8)$$

$$\text{where } y^* | \mathbf{x}' \sim N(\mu, \sigma^2)$$

and

$$\begin{aligned} y &= 0, \text{ if } y^* \leq 0, \\ &= y^*, \text{ if } 0 < y^* < 1, \\ &= 1, \text{ if } y^* \geq 1 \end{aligned} \quad (9)$$

Maximum likelihood estimates are obtained for the β coefficients; for further details we refer to Greene (1997).

Thirdly, a fractional response regression (denoted FRR-CCF) was run. This model has been used for modelling bimodal LGD distributions of credit cards and corporate loan portfolios (Bellotti and Crook, 2012, Qi and Zhao, 2011). FRR is a quasi-likelihood method proposed by Papke and Wooldridge (1996) to model a fractional continuous response variable bounded between zero and one, with valid asymptotic inference and is given by:

$$E(CCF | \mathbf{x}') = F(\mathbf{x}'\boldsymbol{\beta}) \quad (10)$$

where \mathbf{x} is a vector of explanatory variables, $\boldsymbol{\beta}$ is a vector of coefficients and $F()$ represents the logistic functional form which ensures that predicted values are constrained between zero and one.

$$F(\mathbf{x}'\boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{x}'\boldsymbol{\beta})} \quad (11)$$

To estimate the $\boldsymbol{\beta}$ coefficients, the log-likelihood function is maximized, i.e. the sum over all accounts of:

$$l(\boldsymbol{\beta}) = CCF \times \log[F(\mathbf{x}'\boldsymbol{\beta})] + (1 - CCF) \times \log[1 - F(\mathbf{x}'\boldsymbol{\beta})] \quad (12)$$

Similarly to the direct EAD models, variable selection for all CCF models was performed through stepwise selection and backward elimination. The OLS-EAD and all three CCF models were developed with SAS 9.3 software (SAS Institute Inc., Cary, NC, USA).

3.3. Utilization change model

An alternative benchmark model, which has been popular in industry, was developed based on the facility utilization change (Yang & Tkachenko, 2012). The utilization change models the outstanding dollar amount change as a fraction of the current commitment amount and is defined for account i as

$$util = \frac{E(t_d)_i - E(t_r)_i}{L(t_r)_i} \quad (13)$$

A Tobit model, denoted Tobit-UTIL, was fitted as in Equation (8) which treats observations with $util$ below zero and above one as censored hence the response is only observed in the interval $[0, 1]$.

3.4. Credit usage segmentation model

Segmented models were developed using the credit usage variable to partition accounts into low and high utilization accounts. A CCF model was then fitted to the low usage subset of the data, an EAD model to the latter. Sensitivity analysis was used to identify an optimal credit usage cut-off for the partitioning. Model calibration performance was evaluated by varying the credit usage segmentation cut-point from 10% to 95%. The cut-off that produced the highest calibration performance (i.e. lowest MAE, RMSE; cf. section 3.6) was selected. When a cut-off was identified, low usage accounts were modelled with an FRR-CCF model since this is the model that achieved the highest calibration performance among the CCF models considered earlier. High usage accounts were tackled with OLS-EAD and ZAGA-EAD models. We denote the two resulting segmentation models by OLS-USE (the one comprising FRR-CCF and OLS-EAD) and ZAGA-USE (i.e. FRR-CCF combined with ZAGA-EAD), respectively.

3.5. Survival EAD model

To allow the time to default variable to be used as an explanatory variable in practical model development, we propose that a Survival PD model be developed and applied in conjunction with the EAD model, we term this combination the Survival EAD model. The time to default variable is unknown a priori and cannot be used for predictive modelling with conventional EAD model frameworks. To avoid having to discard the variable, a Survival PD model component was developed with the Cox proportional hazards (PH) approach. Several aforementioned models, Tobit-CCF, FRR-CCF, Tobit-UTIL, ZAGA-EAD and ZAGA-USE, with time to default as an explanatory variable, were considered for the EAD component.

The semi-parametric approach in hazard form for the Cox PH model is given by:

$$h(t | \mathbf{x}') = h_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}) \quad (14)$$

where $h(t | \mathbf{x}')$ is the hazard or default intensity at time t conditional on a vector of explanatory variables \mathbf{x}' , and in which $h_0(t)$ is the baseline hazard, i.e., the propensity of a default occurring around t (given that it has not occurred yet) when all explanatory variables are zero. The baseline hazard is left unspecified for the Cox PH model.

Combining estimates from the Cox PH and EAD models, we calculate the expected EAD for account i , as follows:

$$EAD = \sum_{t=1}^{12} \left(\frac{[S(t-1) - S(t)]}{1 - S(12)} \times EAD(t) \right) \quad (15)$$

where $S(t)$ is the survival function at time t , $[S(t-1) - S(t)]$ thus gives the probability of default occurring in the t -th month according to the Cox PH model, and $EAD(t)$ is the EAD model estimate (according to Tobit-CCF, FRR-CCF, Tobit-UTIL, ZAGA-EAD or ZAGA-USE) conditional on the time to default being t . Hence (15) allows us to produce estimates of EAD without any prior knowledge of the time to default variable.

Note that, to produce valid EAD estimates, the horizon length for the Cox model must be the length of each cohort period (12 months) and the origin of time is taken to be the start of the cohort period in which default occurs; this means no event time censoring is observed in the data and each of the produced default probabilities are indeed conditional on the account defaulting over the cohort period (i.e. $S(12)=0$). One could argue that, in the absence of censoring, other (non-survival) regression methods could also be considered, but its flexible baseline hazard still makes the Cox PH model an attractive candidate for modelling time to default. The Cox PH model was developed with SAS 9.4 software (SAS Institute Inc., Cary, NC, USA). Table 8 displays the results of the models fitted using (15).

3.6. Model validation and testing

To assess the out-of-sample performance of the models thoroughly, 10-fold cross validation was conducted on the entire sample of accounts on a series of discrimination and calibration measures. All measures were derived from account-level EAD predictions (either direct ones

or produced indirectly through a predicted CCF) to have a common base of comparison. To evaluate discriminatory power (i.e. the models' ability to discriminate between different levels of EAD risk), the Pearson r and Spearman's ρ correlation were computed. The Pearson r measures linear association and the Spearman's ρ correlation measures the correlation between the rank orderings of observed and expected EADs. Calibration performance (here seen as the model's ability to come up with accurate account-level estimates of EAD) was assessed with the mean absolute error (MAE) and the root mean square error (RMSE). A normalized version of these measures was also produced, where MAE and RMSE were calculated for EAD / Commitment Amount, which facilitated a percentage interpretation. These measures were termed MAE_{norm} and RMSE_{norm} respectively.

ACCEPTED MANUSCRIPT

4. Results

Next, we present the results obtained for a direct EAD model, two competing CCF models, the sensitivity analysis of the segmented model and the cross-validated performance measures to compare all models (reported values here are averages over each 10 runs). Finally, we show the findings for the added survival component.

4.1. ZAGA-EAD model parameters

The parameters of a representative ZAGA-EAD model are shown in Table 2, with the three sub-components for the occurrence of zero EAD, mean of non-zero EAD and dispersion of non-zero EAD: π , μ and σ respectively. The parameters fitted with splines are denoted by $s(\cdot)$ in the table. The other estimates without spline functions are either fitted as categorical variables or linearly as continuous variables.

Figure 4 shows the partial effect plots on log odds scale for the occurrence of zero EAD. These plots can be useful for interpreting coefficient estimates. For example, larger undrawn amounts are associated with a higher propensity (and probability) of zero EAD (see top-left plot) and larger exposure commitment corresponds to a lower propensity of zero EAD (see top-right plot). Precision of the estimates can be gauged with 95 percent confidence intervals represented as dashed lines.

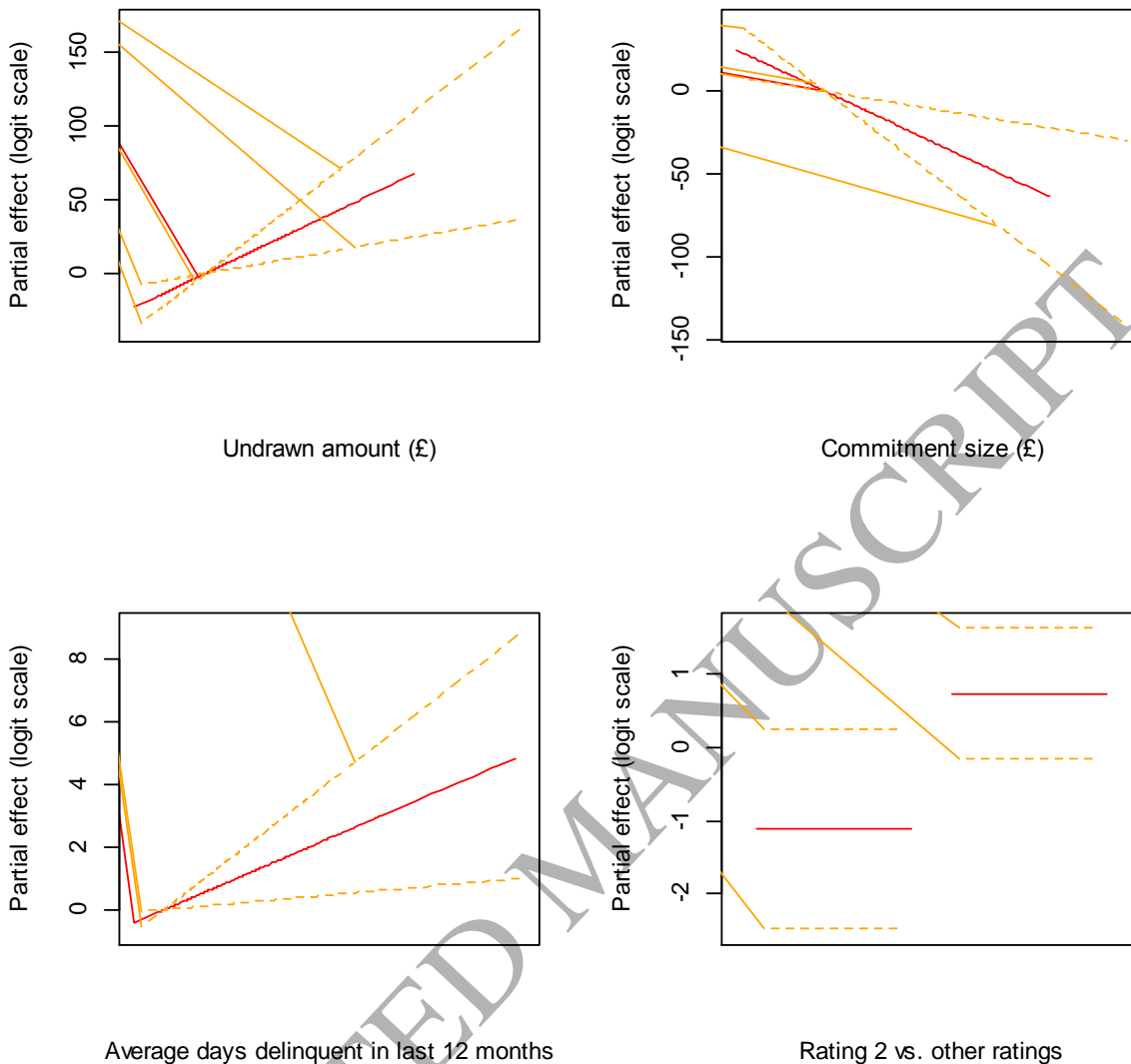
For example, in Figure 4, the partial effect of Rating 2 versus Other Ratings is shown as approximately -1 on the logit or log odds scale, which represents the propensity of zero EAD after adjustment for the effect of other covariates in the model. Hence the odds for the occurrence of zero EAD would reduce by 63% ($1 - e^{-1}$) for Rating 2 versus other Ratings.

Table 2. Zero-adjusted gamma model (ZAGA-EAD) based on a representative training sample.

| Model component | Estimate | SE | <i>p</i> -value |
|---|----------|--------|-----------------|
| log(μ) for non-zero EAD | | | |
| Intercept | 6.949 | 0.007 | <.001 |
| <i>s</i> (Commitment amount) | 0.0003 | 8.0e-7 | <.001 |
| <i>s</i> (Undrawn percentage) | -1.561 | 0.015 | <.001 |
| Time to default | 0.003 | 0.001 | <.001 |
| Average days delinquent (last 6 months) | -0.0004 | 2.1e-4 | 0.055 |
| Rating class 1 versus others | 0.038 | 0.020 | 0.064 |
| log(σ) for non-zero EAD | | | |
| Intercept | -3.630 | 0.055 | <.001 |
| Undrawn percentage | 3.497 | 0.048 | <.001 |
| Time to default | 0.033 | 0.007 | <.001 |
| logit(π) for occurrence of zero EAD | | | |
| Intercept | -6.000 | 1.259 | <.001 |
| Undrawn amount | 0.008 | 0.002 | 0.002 |
| Commitment amount | -0.007 | 0.002 | 0.002 |
| Average days delinquent (last 12 months) | 0.128 | 0.051 | 0.012 |
| Rating class 2 versus others | 1.848 | 1.120 | 0.099 |

s() is a penalized *B*-spline smoothing function

Figure 4. Propensity of zero EAD for zero-adjusted gamma model.



Importantly, Figure 5 shows the partial effects for the mean and Figure 6 the dispersion of non-zero EAD. For example, the commitment size / amount plot in Figure 5 suggests higher committed exposure is linked to larger EAD, but the relationship is non-linear (which could in part be explained by the log link function used). A longer time to default is also associated with higher EAD. All of the effects encountered appear to be intuitive.

In Figure 6, the undrawn percentage plot shows a strong positive linear relationship whereby higher undrawn proportions are associated with higher dispersion in the non-zero mean of EAD. This implies the ZAGA-EAD model has greater uncertainty in EAD prediction for accounts with low credit usage, which provides some justification for including our segmented models (i.e. OLS-USE, ZAGA-USE) into the study. Also, time to default has the

expected positive relationship with both conditional mean (as the drawn down amount can accumulate over time) and dispersion (the farther from default, the harder to predict the final balance) – hence, there is potential value in the survival component proposed earlier.

Figure 5. Mean of non-zero EAD for zero-adjusted gamma model.

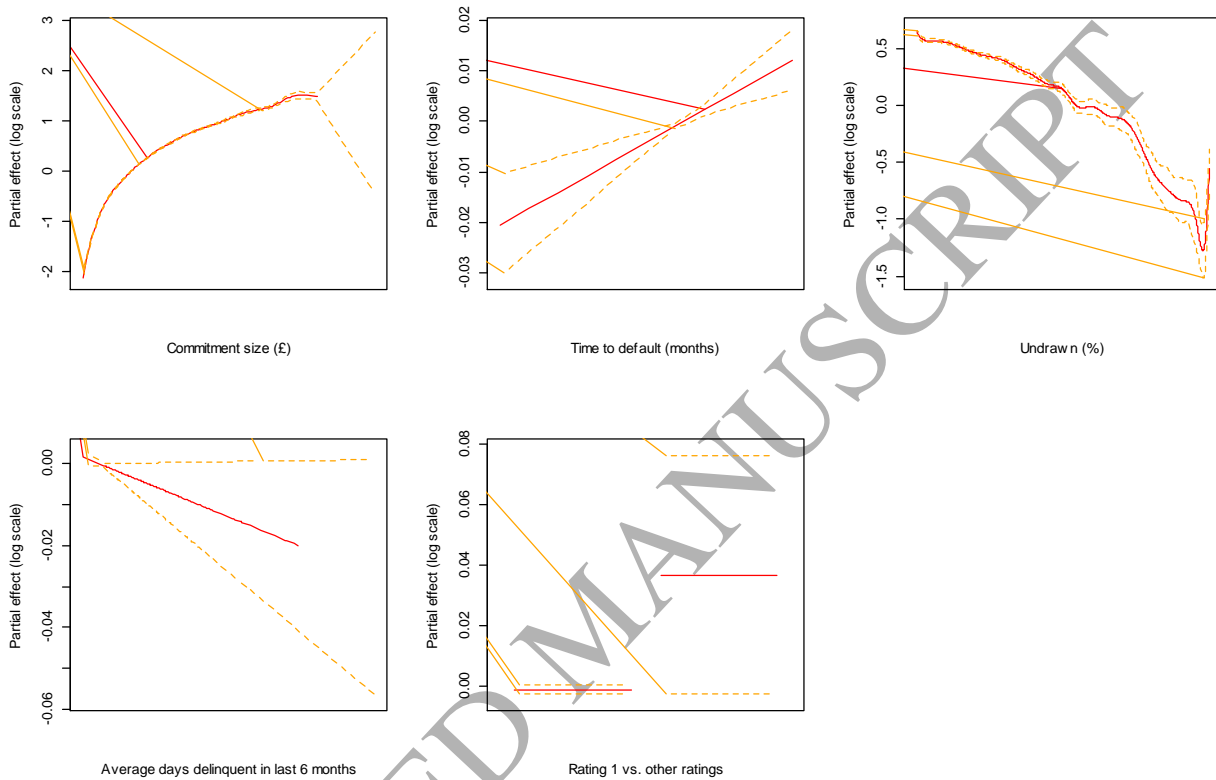
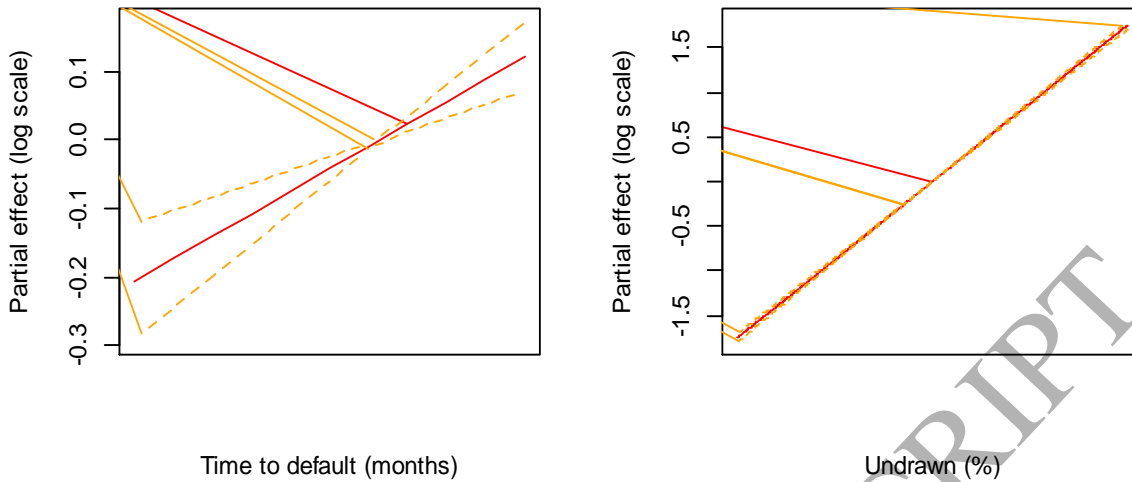


Figure 6. Dispersion of non-zero EAD for zero-adjusted gamma model.



4.2. OLS-CCF and FRR-CCF model parameters

The OLS and FRR models with CCF as the response variable were two of the benchmark models. The parameter estimates for a representative training sample are shown in Table 3 for OLS-CCF and Table 4 for FRR-CCF. For brevity reasons, coefficient estimates for the Tobit-CCF, Tobit-UTIL and OLS-EAD models are not shown but can be made available on request.

Stepwise variable selection for both models resulted in similar choices of covariates. The direction of the coefficient estimates from both models is consistent and confirms previous findings by Jacobs Jr (2010) where the effect of credit usage was negative while commitment amount and time to default were positive in sign.

Table 3. CCF model with ordinary least squares regression (OLS-CCF) based on a representative training sample.

| Parameter | Estimate | SE | <i>p</i> -value |
|---|----------|--------|-----------------|
| Intercept | 0.152 | 0.030 | <.001 |
| Commitment amount | -5.8e-5 | 5.5e-6 | <.001 |
| Drawn amount | 7.9e-5 | 6.8e-6 | <.001 |
| Credit usage (%) | -0.128 | 0.026 | <.001 |
| Time to default | 0.036 | 0.002 | <.001 |
| Rating class 1 vs 4 | 0.241 | 0.037 | <.001 |
| Rating class 2 vs 4 | 0.244 | 0.018 | <.001 |
| Rating class 3 vs 4 | 0.091 | 0.018 | <.001 |
| Average days delinquent (last 6 months) | 0.003 | 0.001 | 0.0019 |

Table 4. CCF model with fractional response regression (FRR-CCF) based on a representative training sample.

| Parameter | Estimate | SE | <i>p</i> -value |
|---|----------|--------|-----------------|
| Intercept | -1.497 | 0.146 | <.001 |
| Commitment amount | -2.7e-4 | 2.8e-5 | <.001 |
| Drawn amount | 3.6e4 | 3.5e-5 | <.001 |
| Credit usage (%) | -0.591 | 0.125 | <.001 |
| Time to default | 0.158 | 0.011 | <.001 |
| Rating class 1 vs 4 | 1.058 | 0.177 | <.001 |
| Rating class 2 vs 4 | 1.055 | 0.089 | <.001 |
| Rating class 3 vs 4 | 0.407 | 0.089 | <.001 |
| Average days delinquent (last 6 months) | 0.012 | 0.004 | 0.004 |

4.3. Sensitivity analysis of credit usage based segmentation models

For the OLS-USE and ZAGA-USE segmented models, a line search was required to determine an appropriate cut point for segmenting the accounts into low- and high-usage segments. Table 5 shows this sensitivity analysis for the OLS-USE model. The optimal cut point (i.e. the one yielding the model combination with the lowest MAE) was found to be at 90% credit usage, which also happened to be the median of the variable.

Table 5. Performance measures from 10-fold cross validation by varying credit usage segmentation cut-off used by the OLS-USE model.

| Measure | Credit usage percentage cut-off | | | | | | | |
|-----------------|---------------------------------|--------|--------|--------|--------|--------|--------|--------|
| | 10% | 20% | 30% | 50% | 70% | 80% | 90% | 95% |
| Pearson r | 0.790 | 0.792 | 0.794 | 0.801 | 0.808 | 0.808 | 0.804 | 0.796 |
| Spearman ρ | 0.733 | 0.739 | 0.743 | 0.747 | 0.753 | 0.752 | 0.750 | 0.743 |
| MAE | 920.1 | 911.3 | 902.0 | 873.6 | 847.8 | 837.9 | 829.9 | 839.2 |
| RMSE | 1623.9 | 1620.3 | 1615.3 | 1597.2 | 1582.9 | 1575.9 | 1565.7 | 1571.7 |

4.4. Discrimination and calibration performance

The discrimination and calibration performance of the models, all in terms of the EAD predictions produced by them, were assessed with 10-fold cross validation and are shown in Table 6. There was broad similarity of discriminatory performance across models based on the Pearson r and Spearman ρ . There did not appear to be a model that was superior based on these measures.

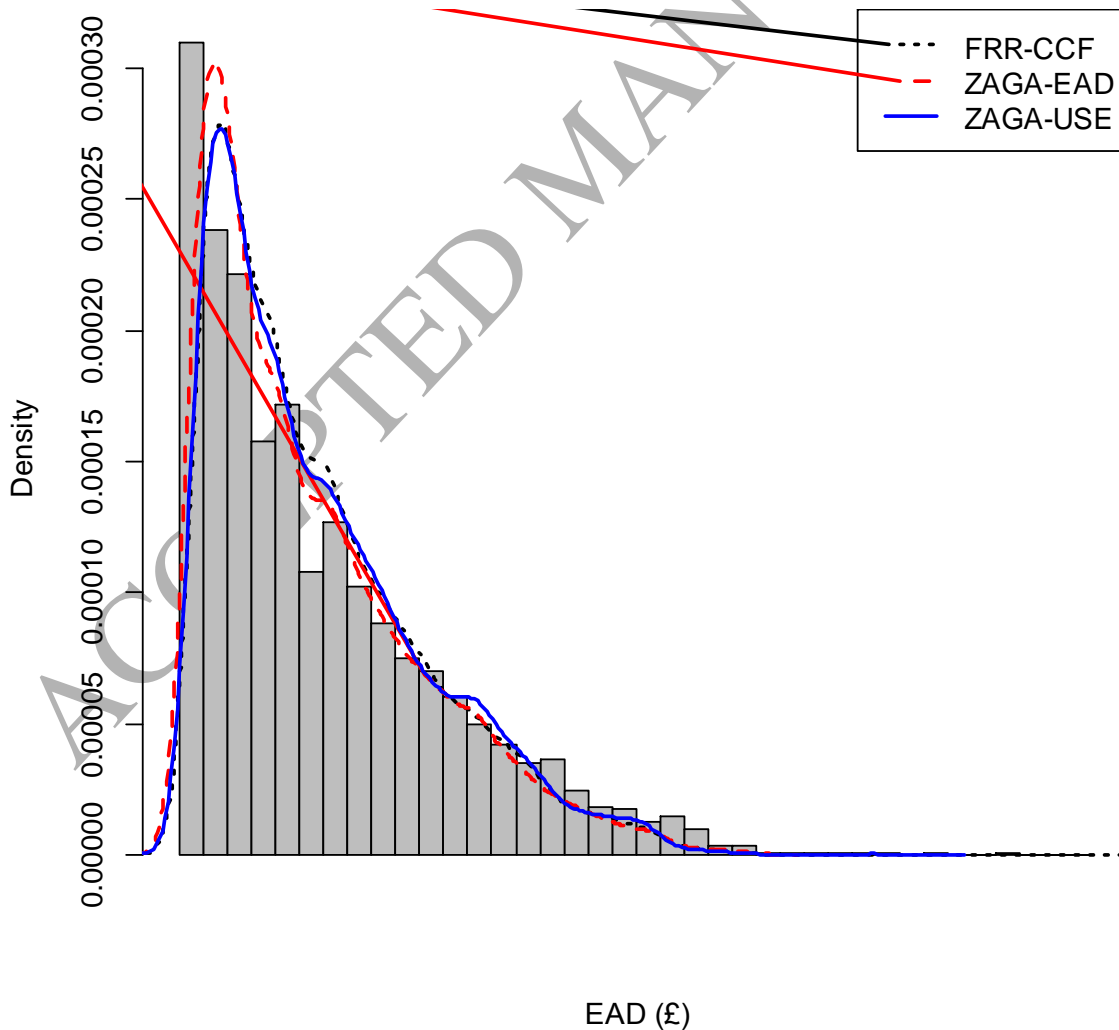
The results did reveal performance differences based on the MAE, MAE_{norm}, RMSE and RMSE_{norm} calibration measures. Among the CCF models, the FRR-CCF had the lowest MAE and RMSE (best performance). Among all models, the OLS-EAD had the highest MAE (worst performance). Although the RMSE was higher than for two of the CCF models, ZAGA-EAD had the lowest MAE and MAE_{norm} at 833.5 and 0.268 respectively, among all non-segmented models. Segmentation by credit usage, i.e. using the OLS-USE and ZAGA-USE approach, reduced the MAE further. The ZAGA-USE had the lowest MAE of all model approaches with 819.2.

Figure 7 shows the observed EAD histogram along with fitted EAD densities for FRR-CCF, ZAGA-EAD and ZAGA-USE. The fitted values were computed through 10-fold cross validation. Importantly, the ZAGA-EAD model is able to reproduce the large peak at the lower bound of EAD more closely than the other models. This also provides a plausible explanation as to why ZAGA-EAD was characterised by a highly competitive MAE but a somewhat disappointing RMSE, as producing a wider distribution may result in some larger residuals that are heavily penalised by the latter criterion.

Table 6. Performance measures from 10-fold cross validation for CCF, direct EAD and segmented credit usage models using observed time to default.

| Measure | OLS- CCF | Tobit- CCF | FRR- CCF | Tobit- UTIL | OLS- EAD | ZAGA- EAD | OLS- USE | ZAGA- USE |
|----------------------|-------------|---------------|-------------|----------------|-------------|--------------|-------------|--------------|
| Pearson r | 0.792 | 0.799 | 0.801 | 0.808 | 0.809 | 0.798 | 0.804 | 0.803 |
| Spearman ρ | 0.741 | 0.737 | 0.743 | 0.746 | 0.744 | 0.742 | 0.750 | 0.749 |
| MAE | 859.0 | 870.6 | 856.1 | 925.2 | 883.3 | 833.5 | 829.9 | 819.2 |
| RMSE | 1614.8 | 1586.3 | 1577.7 | 1654.3 | 1546.1 | 1602.5 | 1565.7 | 1571.0 |
| MAE _{norm} | 0.273 | 0.276 | 0.273 | 0.294 | 0.301 | 0.268 | 0.269 | 0.260 |
| RMSE _{norm} | 0.432 | 0.430 | 0.430 | 0.442 | 0.448 | 0.454 | 0.430 | 0.429 |

Figure 7. Histogram of observed EAD and predicted EAD densities from 10-fold cross validation for FRR-CCF, ZAGA-EAD and ZAGA-USE models.



4.5. Survival model component

A survival model was trained to show how an EAD model having time to default as a covariate could still be applied in a prediction setting. The results of fitting the Cox proportional hazards model onto a representative training sample are shown in Table 7. Positive coefficient estimates imply that a unit increase in the variable is associated with an increased hazard (and thus shorter time to default) and, conversely, negative values indicate reduced hazards of defaulting (default tends to occur later).

The estimated survival probabilities produced by this Cox model were then combined with: the ZAGA-EAD model described in Section 3.1; ZAGA-USE, i.e. the best performing segmentation model (cf. section 3.4); several of the competing CCF models and the UTIL model, against which both ZAGA models were benchmarked in the previous section. For each resulting model configuration, predicted EAD values were computed according to Equation (15), i.e. by weighting the EAD estimates produced for different default time intervals by the monthly PD estimates from the survival component. As all accounts were guaranteed to default within a 12 month time horizon, the estimated survival function was set to zero at $t = 12$, i.e. no accounts survived beyond 12 months in the sample. Each such model combination (referred to as Cox-ZAGA, Cox-ZAGA-USE, Cox-Tobit-CCF, Cox-FRR-CCF, and Cox-Tobit-UTIL, respectively) is a particular instance of the Survival EAD model described in Section 3.5. This approach eliminates the need for any prior knowledge of time to default, and thus allows us to verify whether the performance improvements obtained with the ZAGA approaches are maintained in a practical deployment setting where forward-looking predictions are required. Table 8 provides the model performance comparison for all such selected Survival EAD model configurations.

The Cox-ZAGA model demonstrated good discrimination ability with a Pearson r of 0.798 and Spearman ρ of 0.741. The MAE and MAE_{norm} were 830.3 and 0.266. The RMSE and $RMSE_{norm}$ were 1603.2 and 0.455. These results showed very competitive explanatory power compared to a setting where time to default would be allowed to enter the EAD calculation directly. In fact, this combined model, which does not rely on time to default, performed better in terms of MAE than most of the previous model components that used observed time to default, except for the segmented credit usage models (see Table 6). Furthermore, Table 8

corroborates our earlier findings by showing both ZAGA model combinations (cf. the two right-most columns) still outperformed the competing models in terms of MAE.

Table 7. Cox proportional hazards PD model component of Survival EAD model on a representative training sample.

| Parameter | Estimate | SE | <i>p</i> -value |
|--|----------|--------|-----------------|
| Credit usage (%) | 0.239 | 0.038 | <.001 |
| Rating class 1 vs 4 | -0.527 | 0.081 | <.001 |
| Rating class 2 vs 4 | -0.635 | 0.039 | <.001 |
| Rating class 3 vs 4 | -0.315 | 0.041 | <.001 |
| Average days delinquent (last 3 months) | 0.007 | 0.002 | <.001 |
| Average days delinquent (last 12 months) | -0.020 | 0.003 | <.001 |
| Relative change drawn (last 3 months) | -2.6e-6 | 1.1e-6 | 0.021 |
| Absolute change drawn (last 3 months) | 4.4e-5 | 1.3e-5 | 0.001 |

Table 8. Performance measures from 10-fold cross validation for Survival EAD models using PD weighting method of Equation (15).

| Measure | Cox-Tobit-CCF | Cox-FRR-CCF | Cox-Tobit-UTIL | Cox-ZAGA | Cox-ZAGA-USE |
|----------------------|---------------|-------------|----------------|----------|--------------|
| Pearson <i>r</i> | 0.792 | 0.792 | 0.801 | 0.798 | 0.798 |
| Spearman ρ | 0.721 | 0.723 | 0.733 | 0.741 | 0.736 |
| MAE | 908.8 | 903.5 | 1176.2 | 830.3 | 860.1 |
| RMSE | 1617.5 | 1615.2 | 1865.3 | 1603.2 | 1593.4 |
| MAE _{norm} | 0.287 | 0.286 | 0.375 | 0.266 | 0.273 |
| RMSE _{norm} | 0.437 | 0.437 | 0.497 | 0.455 | 0.435 |

5. Conclusions and future research

Our study considered the development of EAD models which target the EAD distribution directly in lieu of the CCF. Two such direct models were developed using OLS and the zero-adjusted gamma approach. These were compared to more commonly known CCF variants using OLS, Tobit and fractional response regression and the utilization change model. Segmentation by credit usage was also attempted, which involves combining CCF and direct EAD models based on a suitable cut-off level.

The cross validated discrimination measures reported in Table 6 broadly showed that direct EAD models and CCF models risk ranked similarly. In terms of calibration measures (the MAE, MAE_{norm} , RMSE and $RMSE_{norm}$), the FRR-CCF model had the highest performance among CCF variants. The OLS-EAD model had the lowest RMSE; however, the model produced 15 negative fitted values of EAD as its output is not constrained to be a positive value. Although these values could in theory be truncated, this may be considered a drawback of using the OLS model for targeting EAD directly. The OLS-CCF model had the second lowest MAE among CCF models but it also produced 10 negative fitted values of CCF predictions below zero. The utilization change model, Tobit-UTIL, did not perform as well as the other models.

When comparing the non-segmented models, the ZAGA-EAD showed the highest performance among the CCF and direct EAD models, having the lowest MAE and MAE_{norm} from the cross validated findings (see Table 6). Additionally, the ZAGA-EAD model does not produce negative EAD values as the zero-adjusted gamma distribution only predicts values of zero and above.

The notion that CCF models perform better for low credit usage accounts and that direct EAD models perform better for high credit usage accounts appeared to be supported by various findings. The positive relationship of the undrawn percentage with the dispersion parameter in the ZAGA-EAD model indicated that direct EAD models provide less precise estimates for low credit usage. Also, the segmented credit usage models developed by combining CCF and direct EAD models provided further performance improvements (see Table 6). Although the discrimination results remained broadly similar relative to non-segmented models, the calibration performance was improved with lower MAE and RMSE

values observed for both types of segmented models. The OLS-USE model produced the second lowest MAE among all model types; the ZAGA-USE model had the lowest MAE which represented the most accurate model for this study.

Our models from Table 6 included the observed time to default as a predictive covariate. We showed that the time to default variable, which is unknown a priori for a credit line, can nonetheless be applied in a prediction context by using a survival model component alongside a direct EAD model approach. This provides EAD estimates for each month weighted by the respective PD. According to Table 8, the Survival EAD models were competitive and had similar performance compared to the use of a model with observed values of time to default. When combined with the weighted PD approach, the Cox-ZAGA-EAD model had the lowest MAE while the Cox-ZAGA-USE model had the second lowest MAE and the lowest RMSE. In other words, the ZAGA approach proved highly competitive, not just with observed time of default but equally when combined with a survival model component that does not require prior knowledge of this variable.

The direct EAD models had some limitations with respect to drawn balances. Basel compliance requires estimated EAD to be at least equal to or above the drawn balance of the credit line. Some accounts from direct EAD models could have predicted a value of EAD that is less than the drawn balance. Thus during model implementation, appropriate overrides could be used to floor such account-level EAD predictions at the observed drawn balance. This effect would not occur for truncated CCF models where the CCF cannot take values below zero. We note however that if the direct EAD model is used to pool accounts into different EAD risk grades, account-level estimates of EAD that fall outside of the expected range would present less of a problem and the ZAGA-EAD model's better calibration performance would likely imply better grade-level estimates of EAD. However, the direct EAD models are more complex with more parameters to estimate; hence, for use in industry, model developers should consider potential implications of this level of complexity for model implementation and auditing.

Future avenues for research could explore further improving the segmented credit usage models by considering alternative CCF model components, for example, using a beta inflated mixture model (Rigby and Stasinopoulos, 2010) to accommodate the highly bimodal nature

of the CCF distribution. Other EAD distributions with long tails may also be tried as part of the direct EAD models, e.g. using two component gamma distributions for two underlying subpopulations of low and high EAD amounts. The survival model component may be further developed using parametric survival models with truncated survival distributions which allows for fixed maximum time horizons given defaulted accounts have done so within a 12 month horizon.

In summary, our results suggest direct EAD models using the gamma model without the CCF formulation offer a competitive alternative to CCF or utilization change based models. The findings also indicate model segmentation by credit usage may improve calibration performance further, which implies direct EAD models are a complement to CCF based models. This is a positive finding for EAD models as the focus of prediction is not only on risk ranking ability but on the concordance of the observed and predicted values. We suggest EAD model developers consider exploring the use of direct EAD models and credit usage as a segmentation criterion for uplift in calibration performance and to improve risk sensitivity of credit risk models.

Acknowledgments

We thank several anonymous reviewers, as well as Professor Joe Whittaker (University of Lancaster) and Dr Katarzyna Bijak (University of Southampton) for their recommendations which improved the manuscript.

References

- AKAIKE, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19, 716-723.
- BASEL COMMITTEE ON BANKING SUPERVISION 2005. International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Bank for International Settlements. In: SETTLEMENTS, B. F. I. (ed.). Basel, Switzerland: Bank for International Settlements.
- BASEL COMMITTEE ON BANKING SUPERVISION 2011. Basel III: A global regulatory framework for more resilient banks and banking systems. In: SETTLEMENTS, B. F. I. (ed.). Basel, Switzerland: Bank for International Settlements.
- BELLOTTI, T. & CROOK, J. 2012. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28, 171-182.
- BIJAK, K. & THOMAS, L. C. 2015. Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of the Operational Research Society*, 66, 342-352.
- BROWN, I. 2011. Regression Model Development for Credit Card Exposure At Default (EAD) using SAS/STAT® and SAS® Enterprise Miner™ 5.3. *SAS Global Forum*. Las Vegas, NV.
- BROWN, I. L. J. 2014. *Developing Credit Risk Models Using SAS Enterprise Miner and SAS/STAT: Theory and Applications*, Carey, NC, SAS Institute.
- COX, D. R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187-220.
- EILERS, P. H. C. & MARX, B. D. 1996. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11, 89-102.
- GREENE, W. H. 1997. *Econometric analysis*, London, Prentice Hall International.
- HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY, Springer.
- HOSMER, D., LEMESHOW, S. & MAY, S. 2008. *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley-Interscience.
- JACOBS JR, M. 2010. An Empirical Study of Exposure at Default. *Journal of Advanced Studies in Finance*, 1, 31-59.
- LEOW, M. & CROOK, J. 2015. A New Mixture Model for the Estimation of Credit Card Exposure at Default. *European Journal of Operational Research*, In press.
- LOTERMAN, G., BROWN, I., MARTENS, D., MUES, C. & BAESESENS, B. 2012. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, 28, 161-170.
- MALIK, M. & THOMAS, L. 2010. Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, 61, 411-420.
- MCCULLAGH, P. & NELDER, J. A. 1989. *Generalized linear models*, New York, Chapman and Hall.
- MORAL, G. 2011. EAD Estimates for Facilities with Explicit Limits. In: ENGELMANN BERND, R. R. (ed.) *The Basel II Risk Parameters: Estimation, Validation, Stress Testing - with Applications to Loan Risk Management*. 2nd ed. Berlin: Springer.
- NELDER, J. A. & WEDDERBURN, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135, 370-384.
- PAPKE, L. E. & WOOLDRIDGE, J. M. 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11, 619-632.
- QI, M. 2009. Exposure at Default of Unsecured Credit Cards. *Economics Working Paper 2009-2*. Office of the Comptroller of the Currency.
- QI, M. & ZHAO, X. 2011. Comparison of modeling methods for Loss Given Default. *Journal of Banking & Finance*, 35, 2842-2855.
- RIGBY, R. A. & STASINOPOULOS, D. M. 2005. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C*, 54, 507-554.

- RIGBY, R. A. & STASINOPOULOS, D. M. 2007. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23, 1-46.
- RIGBY, R. A. & STASINOPOULOS, D. M. 2010. A flexible regression approach using GAMLSS in R.
- SO, M. C., THOMAS, L. C., SEOW, H.-V. & MUES, C. 2014. Using a transactor/revolver scorecard to make credit and pricing decisions. *Decision Support Systems*, 59, 143-151.
- STEPANOVA, M. & THOMAS, L. 2002. Survival Analysis Methods for Personal Loan Data. *Operations Research*, 50, 277-289.
- TAPLIN, R., MINH TO, H. & HEE, J. 2007. Modeling exposure at default, credit conversion factors and the Basel II Accord. *The Journal of Credit Risk*, 3, 75-84.
- TOBIN, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36.
- TONG, E. N. C., MUES, C. & THOMAS, L. 2013. A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, 29, 548-562.
- TONG, E. N. C., MUES, C. & THOMAS, L. C. 2012. Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, 218, 132-139.
- VALVONIS, V. 2008. Estimating EAD for retail exposures for Basel II purposes. *The Journal of Credit Risk*, 4, 79-109.
- WITZANY, J. 2011. Exposure at Default Modeling with Default Intensities. *European Financial and Accounting Journal*, 6, 20-48.
- YANG, B. H. & TKACHENKO, M. 2012. Modeling exposure at default and loss given default: empirical approaches and technical implementation. *The Journal of Credit Risk*, 8, 81-102.