International Journal of Forecasting 29 (2013) 548–562

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A zero-adjusted gamma model for mortgage loan loss given default

Edward N.C. Tong *, Christophe Mues, Lyn Thomas

University of Southampton - Southampton Management School, Southampton, United Kingdom

ARTICLE INFO

Keywords:
Regression
Finance
Credit risk modelling
Mixture models
LGD
Basel II

ABSTRACT

The Internal Ratings Based (IRB) approach introduced in the Basel II Accord requires financial institutions to estimate not just the probability of default, but also the Loss Given Default (LGD), i.e., the proportion of the outstanding loan that will be lost in the event of a default. However, modelling LGD poses substantial challenges. One of the key problems in building regression models for estimating the loan-level LGD in retail portfolios such as mortgage loans relates to the difficulty of modelling their distributions, as they typically contain extensive numbers of zeroes. In this paper, an alternative approach is proposed where a mixed discrete-continuous model for the total loss amount incurred on a defaulted loan is developed. The model accommodates the probability of a zero loss and the loss amount given that a loss occurs simultaneously. The approach is applied to a large dataset of defaulted home mortgages from a UK bank and compared to two well-known industry approaches. Our zero-adjusted gamma model is shown to present an alternative and competitive approach to LGD modelling.

1. Introduction

The advanced Internal Ratings Based (IRB) approach outlined in the Basel II and Basel III Accords allows banks to calculate their own regulatory capital requirements based on internal credit risk model estimates (Basel Committee on Banking Supervision, 2005).

It requires banks to develop suitable methods for estimating three key parameters for each segment of their loan portfolios: PD (probability of default in the next 12 months), LGD (loss given default, i.e., the proportion of the outstanding loan that will be lost in the event of a default) and EAD (exposure at default).

For consumer credit, probability of default modelling has been a main objective of credit scoring for several decades. However, the additional IRB requirement of having to model LGD has posed substantial challenges, partly because of the properties of its distribution. Datasets of defaulted loan observations for residential mortgage portfolios or other retail portfolios usually exhibit a large probability mass at zero where no losses have been incurred, either because the account has cured and returned to performing status, or, in the case of mortgage loans, because the property has subsequently been repossessed and the sale price covered the loan balance at default adequately (Leow & Mues, 2012; Loterman, Brown, Martens, Mues, & Baesens, 2012; Thomas, Matuszyk, & Moore, 2012). Also, whereas the actual LGD observations of some individual loan defaults may fall outside the (0, 1) range, as LGD is supposed to include all economic costs (e.g., additional collection costs) and recoveries (e.g., penalties paid), the model estimates themselves are expected to be constrained to this interval.

Although the LGD research literature has traditionally focused more on corporate loan portfolios, LGD modelling for residential mortgages is a growing research area, given the impact of the new Accords on consumer lending, and the importance of mortgage loss estimation in the current financial context. One published approach for mortgages involved modelling LGD directly using ordinary

* Corresponding author.
E-mail address: E.Tong@soton.ac.uk (E.N.C. Tong).

0169-2070
http://dx.doi.org/10.1016/j.ijforecast.2013.03.003

least squares regressions (Qi & Yang, 2009). Their approach was developed using data from private mortgage insurance companies for a set of high loan-to-value loans, and although the ordinary least squares model was used, its use could be criticized because of the non-normal distribution of LGD.

Alternatively, a two-stage approach has been introduced in industry by Lucas (2006) and further investigated in the academic literature by Leow and Mues (2012). The method incorporates a probability of repossession (foreclosure) model developed with logistic regression and a haircut model using an OLS regression. The haircut represents the discount factor to be applied to the estimated sale price of the property, given that repossession occurs. The two models are then combined to produce an expected loss percentage given default. They showed the two-stage approach to perform better than the single-stage approach with a standard OLS regression.

There has also been interest recently in using quantile regressions or quantiles of model estimates to obtain LGD predictions (Somers & Whittaker, 2007; Zhang & Thomas, 2012). Somers and Whittaker (2007) have argued that using the low tail of the property value is more predictive of probable losses than the average value estimates in those settings where no loss is incurred in most accounts.

Where censored approaches to LGD modelling are concerned, Tobit regressions have been suggested as one of the methods to be used for modelling the restricted range of the LGD distribution (Bellotti & Crook, 2012). In a Tobit regression, the observed range of the response variable has a Gaussian distribution; however, Sigrist and Stahel (2012) introduced a censored model which allows the response to be Gamma distributed. For LGD data fitted with the Tobit model, the Gaussian assumption may not be suitable, and therefore the censored Gamma regression was developed to overcome the skewed nature of this interval. They also proposed a zero-inflated Gamma model for excess zeroes which were dealt with in probit regression.

Most of the existing literature on LGD modelling in the consumer and corporate credit risk domains has focused on modelling the LGD distribution directly. However, this distribution is known to be challenging to model accurately, due partly to its strongly unimodal or sometimes bimodal nature and lack of predictive characteristics. Therefore, in this paper, rather than modelling LGD (i.e. the loss as a *proportion* of the exposure) directly, we propose to model the incurred financial loss *amount*. Once an estimate of the amount has been obtained, one can then simply infer the LGD parameter by dividing the predicted loss by the loan balance or exposure.

In our proposed approach, the loss amount is modelled as a continuous response variable using a semi-parametric discrete-continuous mixture model approach with the zero-adjusted gamma distribution. Firstly, since the non-zero or positive loss amount exhibits heavy right-skewness, it is modelled using the gamma distribution. Both the mean and the dispersion of the positive loss amount are modelled explicitly as functions of explanatory variables. Secondly, the probability of the (non-)occurrence of a zero loss amount is modelled using a logistic-additive model. All of the mixture model components, i.e., the logistic-additive component for the probability of a zero loss and the log-additive components for the mean and dispersion of the loss amount conditional on there being a positive loss, are estimated using loan-level application and behavioural characteristics and house price index (HPI) covariates. The LGD parameter is then estimated by dividing the predicted loss amount by the loan balance.

When modelling the relationship between the response variable and continuous covariates, past credit risk research has focused on categorizing such covariates using binning methods. Such techniques can be arbitrary and result in a loss of information and precision for the estimated coefficients (Harrell, 2001; Royston, Altman, & Sauerbrei, 2006). Categorization also assumes that the relationship between the response and the covariate is flat within intervals, which may be unreasonable. Another common method would be to assume that continuous covariates are related to the response variable linearly, which would be incorrect for non-linear relationships. For example, such a method would not allow either the magnitude or the sign of coefficients to vary according to the range of covariate values. Our approach adopts a semi-parametric route by allowing non-linear relationships with the loss amount response variable through the use of regression splines (Eilers & Marx, 1996). Exploiting such non-linear relationships will reduce the bias in the estimates, improve the predictive performance of the model, and offer additional insights into the effects of covariates, while still retaining a fair level of model interpretability (Harrell, 2001; Hastie, Tibshirani, & Friedman, 2009).

Although the proposed approach has not yet been attempted in the context of consumer lending (to the best of our knowledge), the concept of estimating the expected loss amount for the exposures in a portfolio has been proposed previously in insurance modelling for policy claim amounts. Heller, Stasinopoulos, Rigby, and De Jong (2007) developed a discrete-continuous mixture model for estimating the total claim amount at a policy level from a portfolio of motor insurance policies. They used two components—the negative binomial distribution for modelling the number of claims for individual policies, and the inverse Gaussian for the claim amount given that a claim occurred. With the risk factors for prospective policy holders, the expected total claim size is then obtained from the product of the expected number of claims and the expected claim size for an individual claim.

Other discrete-continuous mixture models which use a mixture of Bernoulli and beta random variables have also been developed for the recovery rate modelling of corporate loans by Calabrese (2010) and Calabrese and Zenga (2010). They propose two logistic regression models for the recovery rates at the 0 and 1 end-points. For the (0, 1) interval, a joint beta regression model is developed to accommodate skewness and heteroscedastic errors by modelling the mean and dispersion of the response variable jointly. However, note that these methods can only be used to model LGD directly, not to model the loss amount itself.

To validate our approach empirically, the zero-adjusted gamma model is applied to a large dataset of defaulted

home mortgages from a UK bank. The results are compared to the ordinary least squares (OLS) with beta transformation method, which is the parametric regression approach adopted by LossCalc (Gupton & Stein, 2005), a well-known industry model developed by Moody's KMV. The LossCalc approach has also been used or referred to in other comparative studies reported in the literature (Bellotti & Crook, 2012; Hlawatsch & Reichling, 2010; Loterman et al., 2012; Qi & Zhao, 2011; Thomas et al., 2012). In addition, the results are also compared to the Tobit regression, a model that treats any LGD values which are below zero or above one as censored (Greene, 1997; Tobin, 1958). Tobit regressions or variants thereof have been used both in LGD benchmarking studies and in industry (Bellotti & Crook, 2012; Sigrist & Stahel, 2012). The predictive performances of the three approaches are assessed using walk-forward validation with out-of-sample and out-of-time testing. A series of discrimination and calibration measures are presented for comparison. The zero-adjusted gamma model is shown in order to present an alternative and competitive approach to LGD modelling.

The novel aspects of our study are that we: (1) consider modelling of the loss amount directly for estimating the LGD parameter; (2) consider the use of regression splines for modelling non-linear effects between the response variable and covariates for LGD modelling; (3) benchmark the results of our model with two well-known approaches which are used in industry; and (4) develop our models on a very large sample of defaulted residential mortgages and evaluate their performances using an out-of-sample and out-of-time validation process. The remainder of the paper is organized as follows. In Section 2, an overview of the data is presented, along with the application and behavioural characteristics used. The statistical and validation methods used in our experiments are discussed in Section 3. Next, the results of the statistical models are discussed in Section 4. Section 5 will conclude the paper and suggest some areas for further research.

2. Data

Our LGD dataset, provided by a UK bank, contains account-level observations of defaulted loans from a residential mortgage portfolio between 1988 and 2000 (see also Leow & Mues, 2012). Observations were collected from all parts of the United Kingdom. The total sample contained over 113,000 accounts, with 21 application, behavioural and house price index (HPI) predictor variables in the dataset. The dataset consisted of observations through to the year 2002, but we allowed a two-year exposure window to give time for possible repossessions, and analyzed default events up to 2000.

Fig. 1 shows the distribution of LGD for the entire sample, with a large proportion of zeroes. Please note that some of the scales on the figures in the present study have been removed for data confidentiality reasons.

After a mortgage loan defaults, one potential outcome is that the property undergoes repossession by the bank, and legal, administrative and holding costs are incurred. The process of repossession and sale of the property may take a few years to complete. The present analysis, however, considers only the nominal LGD, which does not include discounting. Any extra costs and interest which are incurred are also excluded from the analysis, because the dataset did not contain any information on such costs at an account level.

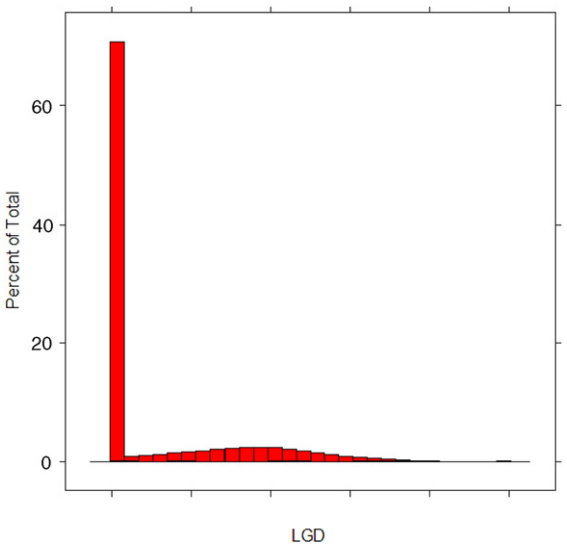

Fig. 1. Distribution of observed LGD on the entire sample of residential mortgages.

Table 1 lists the 21 candidate predictors which are considered for use in the analysis for modelling the loss amount and LGD. The time on books variable was computed as the time between the start date of the loan and the approximate date of default.[1]

The Basel II Accord requires financial institutions to estimate the risk of default and the corresponding losses over a 12-month horizon from a given time point (termed the observation time). Thus, LGD models should not contain information that is only available at the time of default. However, the dataset had limitations, in that information on the state of the account in the months prior to default (e.g., the loan balance at the observation time) was unavailable, as the data were cross-sectional. Hence, we observed behavioural data at the (approximate) default time instead of at the observation time, provided that a reasonable forward-looking adjustment could be used to convert the current value of a variable, such as the outstanding balance, to an estimate at the time of default.

Table 1 contains several variables that include the indexed valuation at default in their definition. As reassessing the value of each property through various time points would be a costly process, an approximate valuation at default was derived by updating the initial security value (which was available in our dataset) using the publicly available Halifax House Price index[2] (all houses, all buyers, non-seasonally adjusted, quarterly, regional). The indexed valuation of the property at default was computed

[1] The date of default was estimated by the bank using the arrears status and the amount of cumulated arrears at the end of each year for each account. The exact default date was not provided in the dataset.

[2] Available from: http://www.lloydsbankinggroup.com/media1/economic_insight/halifax_house_price_index_page.asp.

Table 1
List of application, behavioural and house price index predictor variables for modelling the loss amount and LGD.

Variable	Type
Balance at default (exposure at default)	Continuous
Original loan amount	Continuous
Original property valuation	Continuous
Indexed valuation of property at default quarter	Continuous
HPI growth rate at start quarter (%)	Continuous
HPI growth rate at default quarter (%)	Continuous
HPI index at start quarter (non-seasonally adjusted, quarter, region)	Continuous
HPI index at default quarter (non-seasonally adjusted, quarter, region)	Continuous
Additional mortgage security cover value (AMS)	Continuous
AMS as a percentage of indexed valuation at default (%)	Continuous
Time on books (years)	Continuous
Initial loan to value (LTV)	Continuous
Debt to value (DTV)	Continuous
Indexed valuation at default quarter to average valuation for region	Continuous
Previous default indicator	Binary
Second applicant indicator	Binary
Insurance indicator	Binary
Loan term (years)	Discrete
Property age	Categorical
Security type (flat, detached, semi-detached, terraced, other)	Categorical
Geographical region (13 levels)	Categorical

as follows:

$$\text{Valuation of security}_{\text{default}} = \frac{\text{HPI}_{\text{def yr,def qtr,region}}}{\text{HPI}_{\text{start yr,start qtr,region}}} \times \text{Valuation of security}_{\text{start}}. \quad (1)$$

The valuation at default was then used to calculate some of the variables in Table 1, such as the debt to value ratio (DTV) at default and the ratio of indexed valuation at the default quarter to the average valuation for the region.

3. Statistical models

The three types of models which are fitted to the data are outlined in Sections 3.1–3.3. In order to set up the model for the zero-adjusted gamma approach, an investigation into the loss amount distribution was considered. The loss amount distribution has excess zeroes and a positively skewed distribution. One way of dealing with excess zeroes and positive skewness is to apply a mixed discrete-continuous model for the total loss amount.

Such an approach would involve the assumption that the portfolio is stratified into two groups: the first group has zero loss amounts, and the second group has non-zero losses which are assumed to have a continuous distribution that accommodates heavy right skewness.

Let y_i be the loss amount on the ith account, $i = 1, \ldots, n$. The mixed discrete-continuous probability function of y can then be written 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, right skewed distribution, and π is the probability of zero loss.

Candidate distributions for the non-zero loss amounts, $g(y)$, are given in Fig. 2. Three right-skewed distributions were considered: the gamma, inverse Gaussian and log normal distributions. The inverse Gaussian has been shown to be a suitable fit for total claim sizes in motor insurance policies (Heller et al., 2007). The normal distribution was also presented as a baseline comparison for the other, right skewed distributions. All of the candidate distributions were subsequently fitted on a training set of a random two-thirds sub-sample. Fig. 2 suggests that the gamma distribution had the best fit for the histogram of non-zero loss amounts. There was also support for the fitted gamma distribution, as it produced a lower Akaike Information Criterion (AIC) than either the inverse Gaussian or log normal distribution (Akaike, 1974). The zero-adjusted gamma distribution was therefore considered for modelling $f(y)$; i.e., the gamma distribution was selected for modelling $g(y)$ and a binomial distribution was used to model π.

3.1. Zero adjusted gamma model

The probability function of the zero-adjusted gamma distribution, denoted by ZAGA (μ, σ, π), is defined by Rigby and Stasinopoulos (2010):

$$f(y|\mu,\sigma,\pi) = \begin{cases} \pi & \text{if } y = 0 \\ (1-\pi)\left[\dfrac{1}{\left(\sigma^2\mu\right)^{1/\sigma^2}} \dfrac{y^{\frac{1}{\sigma^2}-1} e^{-y/(\sigma^2\mu)}}{\Gamma\left(1/\sigma^2\right)}\right] & \text{if } y > 0 \end{cases}$$
$$\text{for } 0 \le y < \infty, \text{ where } 0 < \pi < 1, \quad \text{mean } \mu > 0, \text{ dispersion } \sigma > 0, \quad (3)$$

with:

$$E(Y) = (1-\pi)\,\mu \quad \text{and} \quad \text{Var}(Y) = (1-\pi)\,\mu^2\left(\pi + \sigma^2\right). \quad (4)$$

The ZAGA model is implemented using the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework (Rigby & Stasinopoulos, 2005). This method allows for a wide range of skewed and kurtotic distributions by explicitly modelling various distributional parameters, which may include the location/mean, scale/dispersion, skewness and kurtosis as functions of predictor variables. Such an approach allows for the fitting of distributions that do not belong to the exponential family, as featured in the Generalized Linear Model (GLM) (Nelder & Wedderburn, 1972) and Generalized Additive Model (GAM) (Hastie & Tibshirani, 1990; Wood, 2006) frameworks.

The GAMLSS approach is also a semi-parametric method that allows the relationship between the predictor variables and the response variable to be modelled either parametrically or non-parametrically using spline smoothers, the latter of which are a key feature of the GAM approach.

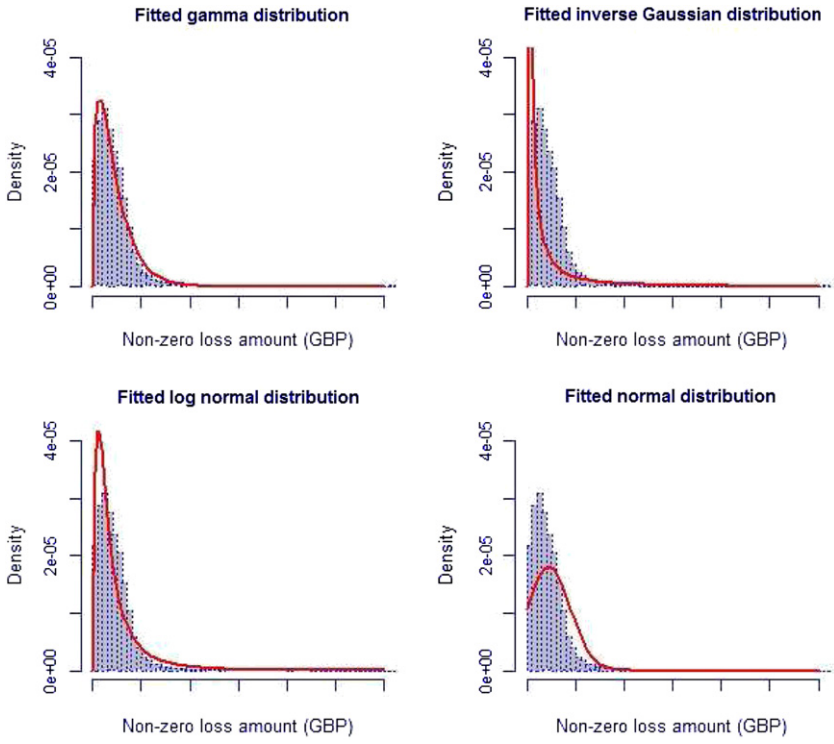

Fig. 2. Candidate continuous distributions for non-zero loss amounts on the training set.

3.1.1. Modelling the probability of zero loss and expected non-zero loss in terms of predictor variables

The ZAGA model includes three components. The mean, μ, and dispersion, σ, of a non-zero loss amount, and the probability of a zero loss, π, are modelled in terms of predictor variables using suitable link functions:

$$\begin{aligned}\log(\mu) &= \eta_1 = X_1^T\beta_1 + \sum_{j=1}^{J_1} h_{j1}\left(x_{j1}\right)\\ \log(\sigma) &= \eta_2 = X_2^T\beta_2 + \sum_{j=1}^{J_2} h_{j2}\left(x_{j2}\right)\\ \operatorname{logit}(\pi) &= \eta_3 = X_3^T\beta_3 + \sum_{j=1}^{J_3} h_{j3}\left(x_{j3}\right),\end{aligned} \tag{5}$$

where $X_k^T\beta_k$ denote parametric terms, $h_{jk}(x_{jk})$ are non-parametric terms such as smoothing splines, and the distribution parameters are $k = 1, 2, 3$. The dispersion of the non-zero loss amount is the squared coefficient of variation, δ^2/μ^2, from the exponential family for the gamma density function (McCullagh & Nelder, 1989), where δ^2 denotes the variance of the non-zero loss amount distribution.

The predictor variables X_k and x_{jk} may differ based on the parameter being modelled. This allows for other predictors to have an impact on whether or not there is a loss (cf. the model component for π), the size of that loss (μ), and the precision of the corresponding estimate (cf. the model part for dispersion σ). Also, the $h_{jk}(x_{jk})$ term allows the predictors to be modelled non-parametrically. The log link functions imply the existence of multiplicative effects on the response variable of the non-zero loss amount, and also ensure that the predictions will be non-negative.

The $h_{jk}(x_{jk})$ functions in this study are modelled using penalized B-splines (Eilers & Marx, 1996). The inclusion of such non-parametric smoothing terms has several advantages, including the ability to identify 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 being the log (penalized) likelihood, k the penalty parameter of 2, and N the number of parameters in the fitted model (Akaike, 1974).

3.1.2. Maximum likelihood estimation

According to the model, each account is associated with a probability of zero loss, π, and a loss amount, y, given that a loss occurs, which produces a pair $(1-\pi, y)$. These pairs are then used to form the following log-likelihood function term:

$$\ln f(y) = \ln f(\pi) + \ln f\left(y|\left(1-\pi\right)\right). \tag{6}$$

The log-likelihood is then the sum of Eq. (6) over all accounts. The maximization of the likelihood proceeds in two separate maximizations, one for the component based on $f(\pi)$ and one for the component based on $f\left(y|\left(1-\pi\right)\right)$. We used an algorithm which was described by Rigby and Stasinopoulos (2005) and is based on penalized likelihood estimation. The estimates of the probability of

zero loss, and of the mean and dispersion of $g(y)$, are used to compute an estimate of $f(y)$ which combines the probability of loss with the loss amount given that there is a loss. The independent estimation for $f(\pi)$ and $f(y|(1-\pi))$ avoids the difficulty of having to estimate $f(y)$ directly.

3.1.3. LGD prediction from loss amount

Finally, the predicted value of LGD for each observation is defined by:

$$\text{LGD} = \frac{E(Y)}{\text{EAD}}, \tag{7}$$

where $E(Y)$ is the account-level fitted value of the loss amount from the ZAGA model, and *EAD* is the exposure at default, or final loan balance.

3.1.4. Variable selection and goodness of fit

A parsimonious ZAGA model was sought where variable selection was performed through stepwise selection with backward elimination. The minimization of the AIC statistic (Akaike, 1974) was used during this backward elimination.

To assess the goodness-of-fit of the model, the independence of the normalized quantile residuals and their normality were assessed, in order to verify whether the model described the systematic part, with the remaining information being independent and identically distributed random noise. The residuals were checked by observing the mean, variance, skewness and kurtosis, and by inspecting the residual versus fitted value plots, residual density plots and *qq*-plots, as described by Rigby and Stasinopoulos (2007). In addition, a series of discrimination and calibration measures were computed, with the results shown in Section 4.

The model was developed and implemented using gamlss package (Rigby & Stasinopoulos, 2007) in the R 2.13.1 software (R Development Core Team, Vienna, Austria).

3.2. Ordinary least squares with beta transformation model

As another reference model against which to compare the ZAGA model, an ordinary least squares with beta transformation model (OLS-beta) was also fitted (Gupton & Stein, 2005).[3] This is a well-known technique that is used in industry for LGD modelling, and which has also been used for comparison purposes in the academic literature (Bellotti & Crook, 2012; Loterman et al., 2012; Qi & Zhao, 2011; Thomas et al., 2012).

This approach assumes that LGD is beta distributed. The α and β parameters of the beta distribution are derived from the empirical LGD response variable distribution, and are then used to compute cumulative probabilities. Subsequently, the inverse standard normal transformation is used to convert these cumulative probabilities from a (0, 1)-scale to $(-\infty, \infty)$ in order to meet the OLS normality assumption better. Next, OLS regression is performed on this transformed dependent variable. The values fitted by the OLS model can be transformed back from $(-\infty, \infty)$ to (0, 1) using the normal distribution; these probabilities are finally transformed back to the starting distribution using the inverse beta distribution.

The beta transformation of LGD is given by:

$$Z = \Phi^{-1}\left[\text{Beta}\left(\text{LGD}, \alpha, \beta, \varepsilon\right)\right], \tag{8}$$

where Φ^{-1} is the inverse standard normal distribution, α and β are positive shape parameters, and ε is a small adjustment for zero LGD values.

For the OLS-beta approach, an adjustment was necessary, where a small value ε was added to observed zero values of LGD before the first transformation step. This was essential because the inverse normal and beta transformations are undefined at zero. However, we found that the fit of the OLS regression was quite sensitive to the choice of ε. Hence, following an approach similar to that of Qi and Zhao (2011), a sensitivity analysis was performed to select an optimal value of ε; further details are provided in the results reported in Section 4.

Note that we chose to model the transformed response variable using a polynomial OLS regression, which allowed quadratic and cubic effects for continuous variables. Such an approach allows non-linear effects to be estimated and tested. Variable selection was performed through stepwise selection, with backward elimination based on minimizing the AIC. The OLS-beta model was also developed in R software.

3.3. Tobit regression model

The LGD distribution is bounded by zero and one, and a large proportion of accounts for residential mortgages have zero LGDs. It has been suggested that the Tobit model may be more appropriate for such data, because any values below zero or above one are treated as censored. The LGD response is only observed in the interval [0, 1]. The standard 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 then given by:

$$y* = \beta\mathbf{x} + \varepsilon, \quad \text{where } y* | \mathbf{x} \sim N\left(\mu, \sigma^2\right), \tag{9}$$

and

$$\begin{aligned} y &= 0, && \text{if } y* \le 0, \\ &= y*, && \text{if } 0 < y* < 1, \\ &= 1, && \text{if } y* \ge 1. \end{aligned} \tag{10}$$

Maximum likelihood estimates can be obtained for the β coefficients; for further details, refer to Greene (1997). Similar to the previous models, variable selection was performed through stepwise selection. The model was implemented in Stata 10.1 software (StataCorp, College Station, TX, USA).

[3] A comparison with a standard OLS regression was also considered. However, the OLS model produced an excess of negative predictions for LGD (28%–39% of negative predictions, depending on the validation year used). The OLS method may therefore be more useful with LGD distributions that are strongly bimodal, such as those observed in credit card portfolios (Bellotti & Crook, 2012).

3.4. Model validation and testing

The method of walk-forward validation (Gupton & Stein, 2005) was used to evaluate model performances in terms of discrimination and calibration. This procedure involves repeatedly fitting a model on one time period and testing its performance on a subsequent time period. This can be considered a special case of cross validation which features both out-of-sample and out-of-time validation. Out-of-time validation allows one to assess whether the modelling approach (as opposed to an individual model) is robust throughout time and credit cycles. The method also helps to prevent overfitting, and could be used to check how reliable the models were.

For the three approaches (ZAGA, OLS-beta and Tobit), a model was initially fitted to data from the years 1988 to 1993 and validated on the year 1994, which provided the first validation fold. Next, the training sample was extended to the period 1988–1994 and the resulting models were validated on the year 1995. This process was repeated, moving forward one year at a time, until the last validation year of 2000 was reached, thus providing seven years' worth of validation folds. Using this procedure, the data used to train the model were never used to validate the model, which guaranteed proper out-of-sample and out-of-time testing. Discrimination and calibration measures were then computed for these seven years of validation folds in order to test for model performance differences between the ZAGA, OLS-beta and Tobit approaches.

A series of discrimination measures for assessing the risk rankings of accounts were used. In addition to the Area Under the Receiver Operating Characteristic Curve (AUC), an industry standard measure, we also report Spearman's ρ, a ranked correlation coefficient, Pearson's r correlation coefficient, and the H measure, proposed more recently by Hand (2009). Hand (2009) argues that the H measure is superior to the AUC, because it is a coherent estimator of the discrimination performance. Unlike the AUC, it is not sensitive to the empirical score distributions of the default and non-default groups in the sample. The AUC has a deficiency, in that it uses different misclassification cost distributions for different classifiers, which implies that the AUC uses different metrics to evaluate different classification algorithms. However, the H measure maintains coherence, because the misclassification cost distribution functions are given by a pre-specified beta distribution, for which the symmetric beta$(x; 2, 2)$ has been proposed as a default. If there are any disagreements between these discrimination measures, Hand (2009) argued that the H measure should be considered the measure of choice for comparing the performances of the various methods. Note that a dichotomous response or gold standard variable was necessary for computing the AUC and the H-measure; this was created using the average of the validation year as the cutoff. In other words, both metrics indicate the extent to which the models are able to distinguish between higher and lower than average losses in the validation fold.

As a calibration measure, the concordance correlation (Lin, 1989, 2000) was used to assess the agreement between the predicted LGD from a given model and the observed LGD. The concordance correlation, being a measure of agreement, is different to the Pearson r correlation, which is a measure of linear association. If the observed LGD were plotted against the model-based LGD estimates, a well-calibrated model would produce estimates that fall on a 45° line through the origin. The Pearson correlation would fail to detect departures from the 45° line, but agreement measures such as the concordance correlation will correct for this. Finally, the root mean square error (RMSE), a measure commonly used in benchmarking studies, was also provided to assess the calibration performance further (Bastos, 2010).

4. Results

The following subsections describe the results of our experiments. Section 4.1 outlines and discusses a model fitted to a two-thirds training sample. The subsequent model validation and testing were done using walk-forward validation over the entire dataset, producing seven years' worth of validation folds, as was described in Section 3.3.

4.1. Zero-adjusted gamma model

Table 2 lists the ZAGA model parameter results obtained from the training set. Backward elimination resulted in a total of 14 predictors being selected across the three components of the ZAGA model. In Figs. 3–5, partial effects plots are shown for a selection of predictors fitted with smoothing splines (denoted by $s()$ in Table 2). The solid lines denote the penalized B-spline smoothing estimates and the dashed lines represent the point-by-point standard errors. The smoothing splines estimated non-linear relationships between the response of the respective model component (μ, σ, π) and predictor, and as such, there was no single regression coefficient or slope associated with the splines. The splines themselves represent the 'slope'. The fitted values for each predictor capture the average changes in the response variable of the model component as a result of small changes in a predictor. In Table 2, the p-values associated with the smoothing splines denoted by $s()$ were a test of their non-linearity (Hastie et al., 2009; Rigby & Stasinopoulos, 2005).

As the partial effects plots shown in Fig. 3 are on a logit scale, and considering that the logit link function was chosen for the occurrence of a zero loss (π) model component, these plots can also be used to identify potential non-linear relationships between the predictor and response. Hence, Fig. 3 clearly indicates that several of the predictors are non-linearly related to the response.

The debt-to-value (DTV) ratio was one of the stronger predictors for the occurrence of a zero loss, exhibiting a non-linear negative relationship with the response; i.e., higher values of DTV result in lower odds of zero losses. This is an intuitive result, since the risk of having to repossess and subsequently sell the property at a value that is lower than the remaining loan amount is expected to be greater when the loan size is close to or greater than the estimated market value of the property. The HPI growth rate at default also had a non-linear but positive

Table 2
The zero adjusted gamma model for the occurrence of loss and the loss amount based on a two-thirds training sample.

Model component	Estimate	SE	*p*-value
logit(π) *for occurrence of zero loss*			
Intercept	3.618	0.100	<0.001
s(HPI growth at default)	0.019	0.001	<0.001
s(AMS to valuation at default)	−0.723	0.148	<0.001
s(HPI at start quarter)	0.005	<0.001	<0.001
s(Time on books)	0.098	0.004	<0.001
s(Debt-to-value)	−3.783	0.064	<0.001
s(Valuation at default to average valuation for region)	−0.277	0.022	<0.001
Previous default indicator (yes vs. no)	0.183	0.041	<0.001
Security type (detached vs. flat)	0.737	0.042	<0.001
Security type (semi-detached vs. flat)	0.648	0.029	<0.001
Security type (terraced vs. flat)	0.434	0.025	<0.001
Security type (other vs. flat)	0.808	0.155	<0.001
Loan term (16–25 years vs. 0–15 years)	−0.775	0.060	<0.001
Loan term (26–40 years vs. 0–15 years)	−0.217	0.072	0.003
Property age (1919–45 vs. <1919)	0.126	0.029	<0.001
Property age (1945+ vs. <1919)	0.082	0.022	<0.001
Property age (missing vs. <1919)	0.185	0.297	0.534
Region (Northern Ireland vs. England and Wales)	0.801	0.102	<0.001
Region (Scotland vs. England and Wales)	0.298	0.051	<0.001
log(μ) *of loss amount given loss occurred*			
Intercept	−0.367	0.130	0.005
s(log[Balance at default])	0.851	0.014	<0.001
s(HPI growth at default)	−0.005	0.001	<0.001
s(AMS to valuation at default)	0.146	0.055	<0.001
s(HPI at start quarter)	0.001	<0.001	<0.001
s(Time on books)	−0.015	0.002	<0.001
s(Debt-to-value)	0.785	0.031	<0.001
s(Valuation at default to average valuation for region)	0.119	0.013	<0.001
Previous default indicator (yes vs. no)	−0.058	0.021	0.006
Security type (detached vs. flat)	−0.193	0.015	<0.001
Security type (semi-detached vs. flat)	−0.230	0.010	<0.001
Security type (terraced vs. flat)	−0.176	0.008	<0.001
Security type (other vs. flat)	0.015	0.062	0.806
Second applicant indicator (yes vs. no)	−0.020	0.007	0.005
Property age (1919–45 vs. <1919)	−0.100	0.010	<0.001
Property age (1945+ vs. <1919)	−0.171	0.008	<0.001
Property age (missing vs. <1919)	−0.050	0.132	0.703
Region (Northern Ireland vs. England and Wales)	0.168	0.074	0.024
Region (Scotland vs. England and Wales)	−0.165	0.028	<0.001
log(σ) *of loss amount given loss occurred*			
Intercept	−0.697	0.008	<0.001
s(HPI growth at default)	0.019	0.001	<0.001
s(Time on books)	0.035	0.002	<0.001
Insurance policy at default (yes vs. no)	0.079	0.016	<0.001
Region (Northern Ireland vs. England and Wales)	0.393	0.050	<0.001
Region (Scotland vs. England and Wales)	0.155	0.025	<0.001

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

relationship with the response, indicating that zero losses are more common in time periods where the housing market is doing well. The time on books predictor had a positive relationship with the response (which again is intuitive, as a larger part of the loan will have been paid off by the time of default); the effect was most pronounced for the shorter periods, with a being plateau observed after a certain time point. The indexed property valuation at default over the average valuation for the region had a negative relationship. This suggests that relatively higher priced properties had greater chances of incurring losses after default. The bumpiness observed in some of the plots was due to the small numbers of observations in certain ranges of the covariates.

Next, Fig. 4 shows a selection of partial effects plots on a log scale for the mean component, μ, of the gamma distribution, i.e. the loss amount given that a loss occurs. The solid line again represents the penalized *B*-spline smoothing estimate, while the dashed lines represent the point-by-point standard errors. As the mean component was developed using the log link function, these plots can be used to check for possible non-linear relationships between the predictor and the response. Please note that other partial effects plots of predictors from Table 2 have been omitted for the sake of brevity.

The plots in Fig. 4 suggest that the four predictors again had non-linear relationships with the response. The indexed debt-to-value ratio (DTV) exhibited a positive relationship with the response, with larger DTVs contributing greater factor changes to the mean. As expected, the final loan balance was one of the most important predictors for the mean component, and had a positive relationship with the response (i.e., larger loan balances implied greater losses), with the effect being more pronounced for the lower range of loan balances. The plot for the time on books predictor suggests a bathtub relationship, with the lowest and highest values yielding somewhat larger losses, whilst the plot is relatively flat for the middle part of the range. Overall, the predictor for the valuation at default relative to the average valuation for the region also had a positive relationship with the response. However, there appeared to be two distinct modes at the range limits. Such an effect may be observed because the lowest and highest priced properties were more difficult to sell, thus resulting in larger loss amounts.

Fig. 5 displays another selection of partial effects plots on a log scale, this time for the dispersion component (σ) of the gamma distribution for loss amount given that a loss occurred. For example, similarly to Leow and Mues (2012), we observe that the dispersion for time on books is shown to mostly have a positive effect on the dispersion (i.e., the exact loss is harder to estimate the longer an account has been on the books prior to default—this is intuitive, as a valuation based on market information may be less accurate for properties sold a fairly long time ago), until the dispersion seemed to reduce again after a certain point.

4.2. Ordinary least squares with beta transformation model

In order to develop the OLS-beta model, an adjustment ε for zero LGDs was necessary. The sensitivity of ε was investigated by fitting the model on a wide range of ε values from 1e−11 to 0.06 for each available year. Results are shown in Table 3 for three representative years. The bootstrapped standard errors were computed on 100 bootstrapped samples, and suggested that the in-sample

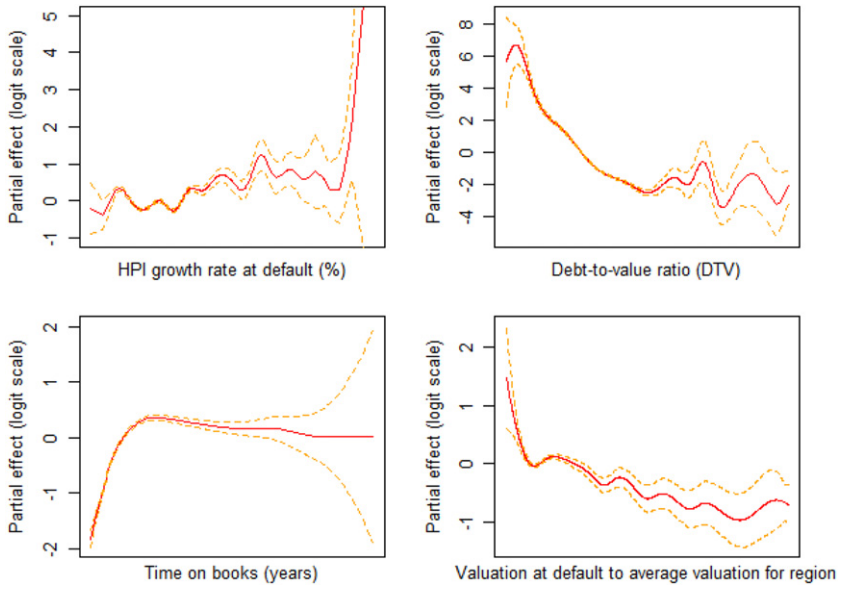

Fig. 3. Occurrences of zero loss amounts for the zero adjusted gamma model.

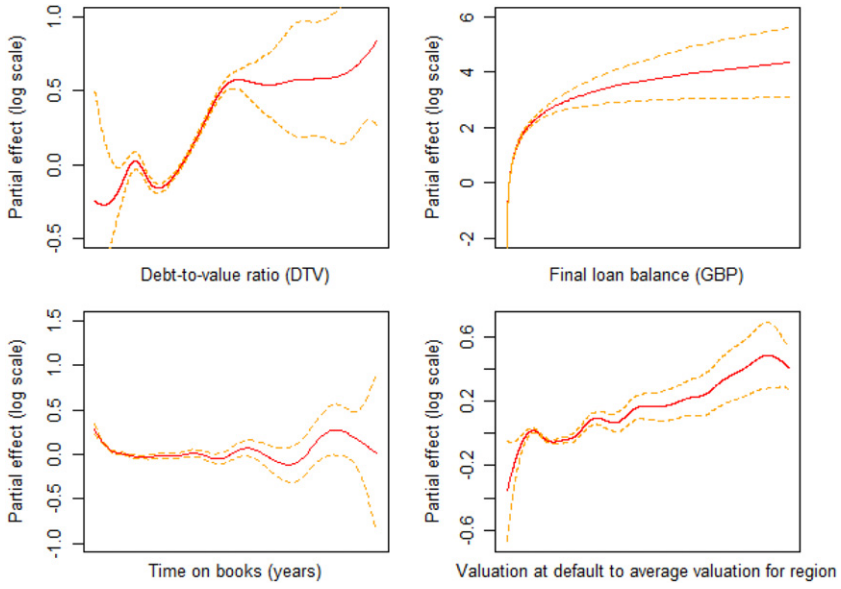

Fig. 4. Mean of the loss amount given that a loss occurred for the zero adjusted gamma model.

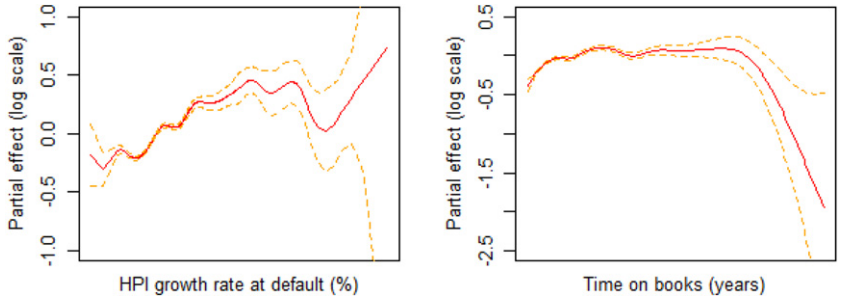

Fig. 5. Dispersion of the loss amount given that a loss occurred for the zero adjusted gamma model.

Table 3
Model diagnostics of the ε values for representative years.

Year	ε	R^2	Bootstrap SE	RMSE	Bootstrap SE
1997	1.00E−11	0.301	0.002	1.391	0.003
	0.0001	0.310	0.002	0.815	0.002
	0.0005	0.312	0.002	0.731	0.002
	0.001	0.312	0.002	0.693	0.001
	0.005	0.314	0.002	0.610	0.001
	0.01	**0.313**	**0.002**	**0.583**	**0.001**
	0.05	0.305	0.002	0.593	0.002
	0.06	0.301	0.002	0.606	0.002
1998	1.00E−11	0.295	0.002	1.385	0.003
	0.0001	0.304	0.002	0.811	0.002
	0.0005	0.305	0.002	0.728	0.002
	0.001	0.306	0.002	0.691	0.002
	0.005	0.307	0.002	0.608	0.002
	0.01	**0.307**	**0.003**	**0.581**	**0.002**
	0.05	0.298	0.003	0.593	0.002
	0.06	0.295	0.003	0.606	0.002
1999	1.00E−11	0.293	0.002	1.367	0.003
	0.0001	0.301	0.002	0.801	0.002
	0.0005	0.303	0.002	0.719	0.002
	0.001	0.303	0.002	0.682	0.002
	0.005	0.304	0.002	0.601	0.002
	0.01	**0.304**	**0.003**	**0.575**	**0.002**
	0.05	0.294	0.003	0.591	0.002
	0.06	0.291	0.003	0.604	0.002

estimates of R^2 and RMSE were reasonably precise. The findings indicated that the optimal value of ε was 0.01, as this was consistently the value at which the R^2 was arguably maximized and the RMSE was minimized.

Given this optimal value of ε, the estimated parameters of an OLS-beta polynomial regression model based on a random two-thirds training sample are displayed in Table A.1 of the Appendix. The fitted OLS-beta model achieved an adjusted R^2 value of 0.298. Walk-forward validation results were also produced using the same ε value of 0.01 and are discussed further in the following subsection.

4.3. Tobit regression model

The fitted parameters of a Tobit regression model based on a random two-thirds training sample are listed in Table A.2 of the Appendix. The resulting model achieved a McFadden's Pseudo R^2 of 0.338, and its walk-forward validation results are discussed in the next section.

4.4. Discrimination and calibration measures based on walk-forward validation

Walk-forward validation was performed and discrimination and calibration measures were computed for each validation year from 1994 to 2000 (see Fig. 6). The performance of the ZAGA approach was measured as a whole for predicting LGD (not the loss amounts). The discrimination results (i.e., the Pearson r and Spearman's ρ, the AUC and the H-measure) indicated that the ZAGA models fairly consistently showed better discrimination performances than the OLS-beta models. Only in 1996 did the OLS-beta approach discriminate marginally better according to Spearman's ρ and the AUC. Tobit performed similarly to ZAGA according to Spearman's ρ measure, and according to the Pearson r for some years. The AUC and the H-measure indicated that ZAGA discriminated better than Tobit for most years.

Calibration results (which included the concordance correlation r_c and RMSE) suggested that the ZAGA model also showed a superior calibration performance for most of the validation years relative to OLS-beta. The year 2000 was the only year in which OLS-beta had a similar calibration performance to ZAGA in terms of RMSE. The Tobit model displayed a poorer concordance correlation for most years, but was competitive with the ZAGA model according to the RMSE measure.

In addition to the statistical measures of discrimination and calibration, Fig. 7 also shows the mean LGD of the observed and expected values, grouped into 10 risk bands of equal frequency for four representative years. If the expected values were perfect, the plotted points would lie on the 45° line of perfect prediction. For the years shown, the ZAGA model had average expected LGD values which were closer to the line of perfect prediction, suggesting a higher calibration ability than the OLS-beta model. The competitive calibration of the Tobit model is apparent in Fig. 7, where ZAGA provides a modest improvement in calibration performance. In 1994, all three models appeared to underestimate LGD, with the underestimation being most noticeable for the OLS-beta model, where there were more predictions near zero (ε). Generally, the ZAGA model discriminated well between the lower and higher risk bands.

Finally, Fig. 8 plots the mean observed minus the forecast (expected) LGD for the validation years 1994–2000. If the models were perfect, the plotted values would lie on the horizontal line where $y = 0$. Again, the ZAGA model was calibrated better than the OLS-beta for all seven years, and was improved upon modestly for a majority of years by the Tobit model. All three of the models showed larger mean errors in 1997. The Tobit model was calibrated better than the ZAGA model in 1996 and 1997, whereas ZAGA had higher calibrations for 1998, 1999 and 2000.

5. Conclusions and future research

This paper develops and empirically validates a zero-adjusted gamma (ZAGA) model with a semi-parametric formulation for estimating loss given default amounts in a residential mortgage loan portfolio. The model includes log-additive components for the mean and dispersion of loss amounts given that a loss occurs, as well as a logistic-additive component for the probability of a zero loss. These model components are estimated independently, and can be fitted with either the same set of covariates or different selections. The relationship between the response variable and the covariates can be modelled either parametrically or non-parametrically. In order to estimate LGD, we then take the predicted loss amount values from the model and divide them by the exposure or loan balance at the observation time. In that sense, we estimate LGD through a direct estimate of the loss amount.

One of the benefits of the suggested approach is that the three components of the mixture model provide the

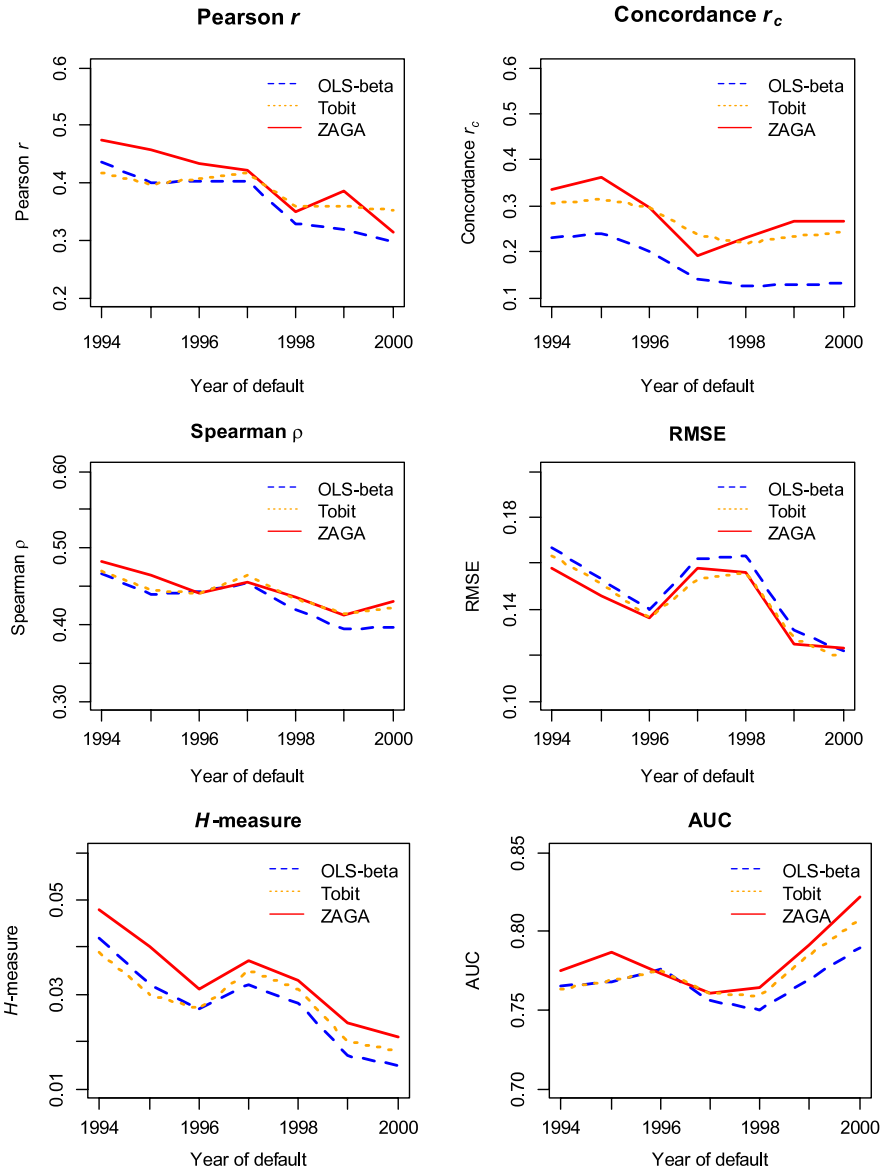

Fig. 6. Discrimination and calibration measures with walk-forward validation for LGD model performance.

analyst with a three-way interpretation, by estimating (i) the factors that predict the occurrence of a loss, (ii) the factors that influence the size of the resulting loss amount, and (iii) the factors that influence the dispersion of the loss amount. The dispersion estimates could be used to provide more conservative estimates where the parameters are less certain, which would be useful for managing model risk in a banking environment. The ZAGA model allows further insights into the factors that predict LGD than the interpretations offered by the OLS-beta and Tobit models, which may not be as intuitive.

Another advantage of the semi-parametric nature of the model is that the method does not imply a 'black box' approach for interpreting the effects of individual covariates. The relationships between the response variable and the covariates are modelled using flexible non-parametric splines, and the interpretation of the effect size remains explicit. This transparent feature of the method may be useful when explaining or defending an implemented model to regulators.

When tested empirically on a large dataset of UK mortgage defaults using a walk-forward validation procedure, the proposed method was shown to perform favourably, in terms of both discrimination and calibration, relative to two well-known industry methods, namely the OLS-beta approach adopted by LossCalc (Gupton & Stein, 2005) and the Tobit regression model (Greene, 1997; Tobin, 1958). We therefore conclude that the zero adjusted gamma model presents a powerful alternative to existing LGD modelling approaches, by allowing one to produce LGD estimates, not by modelling the often inconvenient distribution of the LGD rate itself, but instead by modelling the loss *amount* using a semi-parametric mixture model.

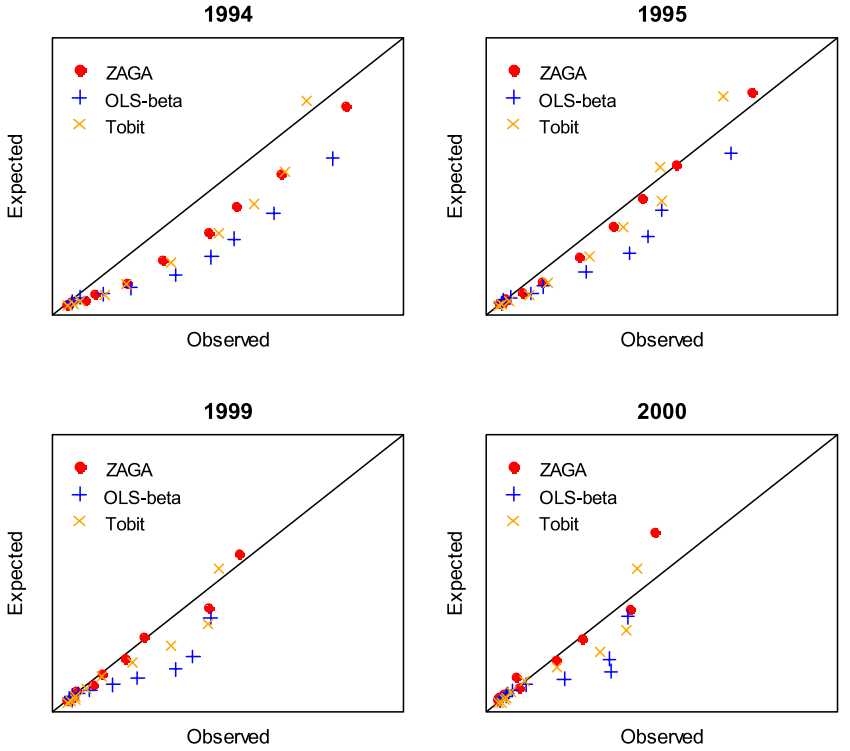

Fig. 7. Mean LGD by decile risk bands with walk-forward validation for four representative years.

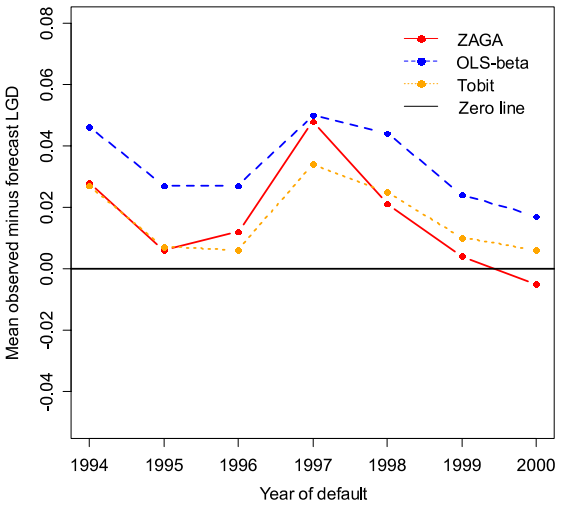

Fig. 8. Difference between the mean observed and forecast LGD with walk-forward validation.

One limitation of our study is that it is possible that the model's estimate of the loss amount will exceed the actual loan balance, and hence predict LGD values beyond the value of 1 if left untruncated. If such a situation would be expected to occur regularly, unlike in our experiments, and the model developer considers it to be an issue, the choice of the distribution for the loss amount should be re-examined, because the gamma distribution may not be suitable for such datasets. Instead, other types of skewed distributions could be considered, for example the inverse Gaussian or log normal.

We suggest several opportunities for future research. In the Basel II Accord, a downturn and stressed LGD estimate is required for regulatory capital requirements. One approach to the computation of downturn estimates for our model would be to provide stressed values of key covariates that enter the model. For example, loan-to-value, Halifax HPI growth and HPI index values may be scaled to appropriate downturn economic conditions prior to model estimation. As the gamma component for positive loss amounts estimates the conditional mean and dispersion, the variance and quantiles of the total loss amounts can be computed from the fitted mixture density. Hence, quantiles or percentiles of predicted LGD can be produced, which may also be useful for conservative, downturn or stressed estimates of LGD for Basel regulations. In addition, the conditional quantiles of the loss amount may be better estimates of LGD than the conditional mean which we have focussed on. Indeed, there has been research in credit risk modelling that has suggested that quantiles may be more useful for prediction (Somers & Whittaker, 2007; Zhang & Thomas, 2012).

Another possible extension would be to explore whether the performance of the zero-adjusted gamma model could be improved further by including a more comprehensive set of macroeconomic variables, as our current study has only included house price index covariates. Due to limitations in the data, we did not have a panel dataset structure. Hence, lagged covariates were not used in the models, although they would be expected to improve the

predictive performance further. Also, a fourth avenue for research would be to estimate the total loss arising from accounts at the portfolio level, at various time points. Such an approach was considered by Heller et al. (2007), who estimated the total claim size for a portfolio of Australian motor insurance policies using their discrete-continuous mixture model.

Finally, our study has used a logistic-additive model to develop the model component for the occurrence of zero loss. There are further opportunities to develop potentially superior models by considering other types of binary classification methods, perhaps based on data mining or ensemble methods such as regression trees (Bastos, 2010), support vector machines, neural networks (Baesens et al., 2003), or random forests (Breiman, 2001). However, it should be noted that several of these methods are 'black box' approaches, where the effects of individual explanatory variables cannot be interpreted conveniently.

Acknowledgment

We would like to thank the bank which provided the dataset that enabled this research to be conducted.

Appendix

See Tables A.1–A.5.

Table A.1
Ordinary least squares regressions with a beta transformation on LGD, based on a two-thirds training sample.

Predictor	Estimate	SE	*p*-value
Intercept	1.698	0.135	<0.001
Indexed valuation at default	−1.92E−06	1.35E−07	<0.001
Indexed valuation at default^2	6.79E−13	6.90E−14	<0.001
Indexed valuation at default^3	−6.33E−20	9.00E−21	<0.001
Balance at default	2.57E−06	1.91E−07	<0.001
Balance at default^2	−1.40E−12	1.45E−13	<0.001
Balance at default^3	1.87E−19	2.47E−20	<0.001
HPI growth at default	−0.003	3.70E−04	<0.001
HPI growth at default^2	−3.27E−04	3.17E−05	<0.001
HPI growth at default^3	6.59E−06	6.83E−07	<0.001
AMS to valuation at default	0.768	0.053	<0.001
AMS to valuation at default^2	−3.053	0.610	<0.001
HPI at start quarter	−0.015	0.002	<0.001
HPI at start quarter^2	7.23E−05	1.19E−05	<0.001
HPI at start quarter^3	−1.12E−07	2.05E−08	<0.001
Time on books (years)	−0.240	0.005	<0.001
Time on books (years)^2	0.033	0.001	<0.001
Time on books (years)^3	−0.001	4.45E−05	<0.001
Debt-to-value (DTV)	−1.245	0.057	<0.001
Debt-to-value (DTV)^2	1.807	0.053	<0.001
Debt-to-value (DTV)^3	−0.505	0.015	<0.001
Valuation at default to average valuation for region	−0.202	0.028	<0.001
Valuation at default to average valuation for region^2	0.117	0.014	<0.001
Valuation at default to average valuation for region^3	−0.015	0.002	<0.001
Previous default (yes vs. no)	−0.028	0.008	<0.001
Security type (detached vs. flat)	−0.215	0.009	<0.001
Security type (semi-detached vs. flat)	−0.234	0.007	<0.001
Security type (terraced vs. flat)	−0.173	0.006	<0.001
Security type (other vs. flat)	−0.209	0.028	<0.001
Loan term (16–25 years vs. 0–15 years)	0.117	0.009	<0.001
Loan term (26–40 years vs. 0–15 years)	0.018	0.012	0.151
Second applicant (yes vs. no)	−0.018	0.005	<0.001
Geographical region (Northern Ireland vs. England and Wales)	−0.067	0.018	<0.001
Geographical region (Scotland vs. England and Wales)	−0.090	0.011	<0.001

$R^2 = 0.298$; Adj. $R^2 = 0.298$; AIC = 129,912; BIC = 130,236.

Table A.2
Tobit regression on LGD based on a two-thirds training sample.

Predictor	Estimate	SE	*p*-value
Intercept	0.905	0.150	<0.001
Indexed valuation at default	−3.20E−06	1.77E−07	<0.001
Indexed valuation at default^2	1.96E−12	1.41E−13	<0.001
Indexed valuation at default^3	−2.86E−19	2.60E−20	<0.001
Balance at default	3.04E−06	2.29E−07	<0.001
Balance at default^2	−2.40E−12	2.35E−13	<0.001
HPI growth at default	−0.001	0.000	0.031
HPI growth at default^2	−0.0004	0.000	<0.001
HPI growth at default^3	7.05E−06	6.82E−07	<0.001
AMS to valuation at default	1.421	0.117	<0.001
AMS to valuation at default^2	−5.320	0.796	<0.001
AMS to valuation at default^3	6.762	1.589	<0.001
HPI at start quarter	−0.017	0.002	<0.001
HPI at start quarter^2	0.0001	0.0000	<0.001
HPI at start quarter^3	−1.20E−07	2.08E−08	<0.001
Time on books (years)	−0.181	0.005	<0.001
Time on books (years)^2	0.026	0.001	<0.001
Time on books (years)^3	−0.001	0.000	<0.001
Debt-to-value (DTV)	1.948	0.095	<0.001
Debt-to-value (DTV)^2	−0.696	0.074	<0.001
Debt-to-value (DTV)^3	0.061	0.018	0.001
Valuation at default to average valuation for region	−0.204	0.026	<0.001
Valuation at default to average valuation for region^2	0.128	0.013	<0.001
Valuation at default to average valuation for region^3	−0.015	0.002	<0.001
Previous default (yes vs. no)	−0.032	0.008	<0.001
Security type (detached vs. flat)	−0.162	0.008	<0.001
Security type (semi-detached vs. flat)	−0.171	0.006	<0.001
Security type (terraced vs. flat)	−0.110	0.005	<0.001
Security type (other vs. flat)	−0.112	0.025	<0.001
Loan term (16–25 years vs. 0–15 years)	0.144	0.011	<0.001
Loan term (26–40 years vs. 0–15 years)	0.043	0.013	0.001
Second applicant (yes vs. no)	−0.012	0.004	0.001
Geographical region (Northern Ireland vs. England and Wales)	−0.168	0.019	<0.001
Geographical region (Scotland vs. England and Wales)	−0.092	0.010	<0.001

Pseudo $R^2 = 0.338$; AIC = 56,079; BIC = 56,393; σ (SE) = 0.356 (0.002).

Table A.3
R code for fitting a zero adjusted gamma model with penalized *B*-splines.

```
library(gamlss)
mzaga < - gamlss(Loss2 ∼ pb(x1, method =“GAIC″) +
pb(x2, method =“GAIC″) + · · · + factor(x5),
                 sigma.fo = ∼
pb(x1, method =“GAIC″) + · · · + factor(x5),
                 nu.fo = ∼ pb(x2, method =“GAIC″) + · · ·,
                 data = train, family = ZAGA)
summary(mzaga)
```

Table A.4
R code for fitting an ordinary least squares with beta transformation model.

```
# Add epsilon to LGD values of zero or less
LGD_adj[LGD<=0] <- 0.01

# Compute alpha and beta for beta distribution
Mu <- mean(LGD_adj)
Var <- var(LGD_adj)
Alpha < - (Mu^2*(1-Mu)/Var)-Mu
Beta <- Alpha*(1/Mu-1)

# Transform LGD to beta distributed variable and
then transform to a normal distributed variable
LGD_cumBeta = pbeta(LGD_adj, shape1=Alpha,
shape2=Beta)
LGD_invNorm = qnorm(LGD_cumBeta, mean=Mu,
sd=sqrt(Var))
LGD_invstdNorm = qnorm(LGD_cumBeta)

mols_beta < - lm(LGD_invstdNorm ∼
x1 + x2 + ... + factor(x5), data = train)
summary(mols_beta)
```

Table A.5
Stata code for fitting a two-sided Tobit regression model.

```
xi: tobit lgd x1 x2 x3... i.x5 i.x6, ll(0) ul(1)
```

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*, 627–635.

Basel Committee on Banking Supervision (2005). *International convergence of capital measurement and capital standards: a revised framework*. Basel, Switzerland: Bank for International Settlements.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking and Finance*, *34*, 2510–2517.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *28*, 171–182.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Calabrese, R. (2010). *Predicting bank loan recovery rates in a mixed continuous-discrete model*. University of Milano-Bicocca.

Calabrese, R., & Zenga, M. (2010). Bank loan recovery rates: measuring and nonparametric density estimation. *Journal of Banking and Finance*, *34*, 903–911.

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.

Gupton, G. M., & Stein, R. M. (2005). *LossCalc v2: Dynamic Prediction of LGD*. Moody's KMV.

Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Harrell, F. (2001). *Regression modeling strategies*. New York, NY: Springer.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Taylor and Francis.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer.

Heller, G. Z., Stasinopoulos, D. M., Rigby, R. A., & De Jong, P. (2007). Mean and dispersion modelling for policy claims costs. *Scandinavian Actuarial Journal*, *2007*, 281–292.

Hlawatsch, S., & Reichling, P. (2010). A framework for loss given default validation of retail portfolios. *The Journal of Risk Model Validation*, *4*, 23–48.

Leow, M., & Mues, C. (2012). Predicting loss given default (LGD) for residential mortgage loans: a two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting*, *28*, 183–195.

Lin, L. I. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255–268.

Lin, L. I. K. (2000). A note on the concordance correlation coefficient. *Biometrics*, *56*, 324–325.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, *28*, 161–170.

Lucas, A. (2006). Basel II problem solving. In *Conference on Basel II and credit risk modelling in consumer lending*. Southampton, UK.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. New York: Chapman and Hall.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, *135*, 370–384.

Qi, M., & Yang, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance*, *33*, 788–799.

Qi, M., & Zhao, X. (2011). Comparison of modeling methods for loss given default. *Journal of Banking and 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*.

Rigby, R. A., & Stasinopoulos, D. M. (2010). *A flexible regression approach using GAMLSS in R*. London Metropolitan University.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, *25*, 127–141.

Sigrist, F., & Stahel, W.A. (2012). *Using the censored gamma distribution for modeling fractional response variables with an application to loss given default*. ETH Zurich.

Somers, M., & Whittaker, J. (2007). Quantile regression for modelling distributions of profit and loss. *European Journal of Operational Research*, *183*, 1477–1487.

Thomas, L. C., Matuszyk, A., & Moore, A. (2012). Comparing debt characteristics and LGD models for different collections policies. *International Journal of Forecasting*, *28*, 196–203.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall/CRC.

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*, 204–215.

Edward N.C. Tong is portfolio analytics manager in corporate banking risk at the Royal Bank of Scotland in London. He is a part-time doctoral candidate in credit risk modelling at the University of Southampton. He previously worked in consumer credit risk at Suncorp Bank in Brisbane, Australia. He has also been employed as a statistician in the pharmaceutical and healthcare industry. His research has led to publications in journals including the *European Journal of Operational Research*, *BMC Infectious Diseases* and *BMJ Quality and Safety*. Edward has a Masters degree in statistics from the University of Queensland, Australia.

Christophe Mues is a senior lecturer at the Southampton Management School at the University of Southampton (UK). Prior to his appointment at the University of Southampton, he was employed as a researcher at K.U. Leuven (Belgium), where he obtained the degree of Doctor of Applied Economics in November 2002. His main research interests are in the domains of business intelligence and data mining. In recent years, he has developed a particular interest in applying data mining techniques to credit risk modelling in the context of Basel II and credit scoring. His findings have been published in various international journals and conference proceedings.

Lyn Thomas is Professor of Management Science at the University of Southampton. His interests are in applying operational research and statistical ideas in the area of finance, particularly in credit scoring and risk modelling in consumer lending. He is a founding member of the Credit Research Centre at the University of Edinburgh and one of the principal investigators for the Quantitative Financial Risk Management Centre based at Southampton. He has authored or co-authored four books in the area, including *Consumer Credit Models; Pricing, Profit and Portfolios* and *Credit Scoring and its Applications*. He is a Fellow of the Royal Society of Edinburgh and a past President of the Operational Research Society, and was awarded the Beale Medal of that Society in 2008.