The University of Sydney Business School
The University of Sydney

BUSINESS ANALYTICS WORKING PAPER SERIES

Multiple Event Incidence and Duration Analysis for Credit Data Incorporating Non-Stochastic Loan Maturity

John G. T. Watkins [a], Andrey L. Vasnev [b†] and Richard Gerlach [b]
[a] Risk Management RBS, Commonwealth Bank of Australia.
[b] Business School, University of Sydney, NSW 2006, Australia.

Summary

Applications of duration analysis in Economics and Finance exclusively employ methods for events of stochastic duration. In application to credit data, previous research incorrectly treats the time to pre-determined maturity events as censored stochastic event times. The medical literature has binary parametric 'cure rate' models that deal with populations that never experienced the modelled event. We propose and develop a Multinomial parametric incidence and duration model, incorporating such populations. In the class of cure rate models, this is the first fully parametric multinomial model and is the first framework to accommodate an event with pre-determined duration. The methodology is applied to unsecured personal loan credit data provided by one of Australia's largest financial services organizations. This framework is shown to be more flexible and predictive through a simulation and empirical study that reveals: simulation results of estimated parameters with a large reduction in bias; superior forecasting of duration; explanatory variables can act in different directions upon incidence and duration; and, variables exist that are statistically significant in explaining only incidence or duration.

December 2012

BA Working Paper No: 03/2013
http://sydney.edu.au/business/business_analytics/research/working_papers

Multiple Event Incidence and Duration Analysis for Credit Data Incorporating Non-Stochastic Loan Maturity

John G. T. Watkins [a], Andrey L. Vasnev [b†] and Richard Gerlach [b]

[a] *Risk Management RBS, Commonwealth Bank of Australia.*
[b] *Business School, University of Sydney, NSW 2006, Australia.*

December 2012

Summary Applications of duration analysis in Economics and Finance exclusively employ methods for events of stochastic duration. In application to credit data, previous research incorrectly treats the time to pre-determined maturity events as censored stochastic event times. The medical literature has binary parametric 'cure rate' models that deal with populations that never experienced the modelled event. We propose and develop a Multinomial parametric incidence and duration model, incorporating such populations. In the class of cure rate models, this is the first fully parametric multinomial model and is the first framework to accommodate an event with pre-determined duration. The methodology is applied to unsecured personal loan credit data provided by one of Australia's largest financial services organizations. This framework is shown to be more flexible and predictive through a simulation and empirical study that reveals: simulation results of estimated parameters with a large reduction in bias; superior forecasting of duration; explanatory variables can act in different directions upon incidence and duration; and, variables exist that are statistically significant in explaining only incidence or duration.

1. INTRODUCTION

Within industry, the practice of risk assessment for retail credit is dominated by logistic and probit regression techniques. These models are most commonly employed to establish the incidence of default over a twelve-month time horizon, see e.g. Altman and Saunders (1998) and Crook et al. (2007). More recently, researchers have investigated the use of survival analysis methods to assess credit risks. The papers of Banasik et al. (1999), Stepanova and Thomas (2002), Andreeva (2006) and Bellotti and Crook (2009) each examine time to prepayment and default events individually, treating all other failure times as censored observations. The risks of, and times to, these events are examined simultaneously in the papers of Deng et al. (2000), Pavlov (2001) and Ciochetti et al. (2002). Deng et al. (2000) emphasizes the importance of the jointness of the decision to default or prepay on mortgages in their option pricing framework.

The event of loan maturity is incorrectly treated as a censored observation in previous

[†]Correspondence to: Andrey L. Vasnev, the University of Sydney Business School, room 480, Merewether Building (H04), University of Sydney, NSW 2006, Australia. Email: andrey.vasnev@sydney.edu.au

credit risk research. The framework, common to the latter three papers above, used to simultaneously analyse the time to prepayment and default events is developed in Han and Hausman (1990), Sueyoshi (1992) and McCall (1996), which as a group has been coined HHSM. HHSM develop a proportional hazards survival analysis framework for the examination of labour market problems. The factors influencing the time to transition to non-terminal employment states are assessed using this framework.

In credit data, the events of prepayment, maturity and write off are terminal. Although the simultaneous estimation of the prepayment and write off risks is important, the treatment of the maturity events has not been adequate. A class of mixture models, known as cure rate models in the medical literature, provide motivation for the framework developed in this paper to address this issue. This class of survival analysis model mixes a binary distribution, usually logistic, with a typical distribution used for the analysis of failure time data, e.g. Weibull. This method is pioneered in Boag (1949) and Berkson and Gage (1952) for the analysis of the fraction of patients cured after experiencing cancer therapies, who were previously erroneously classified as censored observations. The method is further employed in Farewell (1982), Sy and Taylor (2000), Peng and Dear (2000) and Cancho et al. (2009). The use of such models to analyse failure times in medical research is motivated by a biological possibility of cure and often evidenced by heavy censoring and Kaplan-Meier (KM) non-parametric survival function estimates that plateau to values strictly greater than zero, see e.g. Sy and Taylor (2000). These latter papers extend the cure rate method in multiple ways, including to the non-parametric sphere of analysis.

Hoggart and Griffin (2001) uses a binary cure rate model to analyse customer attrition rates in the banking industry, adopting the Bayesian cure rate method developed in Chen et al. (1999). Cancho et al. (2009) uses the same framework in a clinical study on cancer patients. Tsodikov et al. (2003) extend the framework of Chen et al. (1999) to a multinomial non-parametric Bayesian cure rate method. In addition, Chen et al. (1999) argues that extending to a multinomial parametric cure rate model would be theoretically and computationally cumbersome. However, our paper reveals this is not the case, at least for credit data where all events can be observed and the events of interest, prepayment, maturity and write off, are all terminal, mutually exclusive and collectively exhaustive events.

The methods developed in this paper contribute to the current literature in three ways:

i.) the framework resolves the estimation bias in previous models caused by treating a pre-determined or non-stochastic terminal event as a censored observation;
ii.) the model is the first in its class to allow for the simultaneous incidence and duration modelling of a set of M mutually exclusive events, where up to $M-1$ of the events' duration times may be non-stochastic or pre-determined, and;
iii.) the application to unsecured credit data is the first empirical study to extend the seminal work of Deng et al. (2000) to the simultaneous and joint modelling of write off, prepayment and maturity events in credit data.

An empirical study utilises a unique data set of over one million unsecured personal loan observations provided by one of Australia's largest financial services organisations. The data contains limited application form fields, indicating the financial, demographic and risk characteristic of the loan applicant and also, if available, the final outcome of these loans, indicating if they were written off, closed good on maturity or prepaid before

maturity. This data is used for evaluation of the performance of the model in predicting loan lifetime outcomes, an essential part of any retail bank's originations framework. The application simultaneously estimates parameters for both incidence and duration of credit events. The results show:

i.) Superior forecasting of time to prepayment and write off over previous applications of duration analysis methods to credit data;
ii.) Regressors can act in opposite directions upon the incidence of the event and the conditional duration, and;
iii.) Regressors exist that are only significant in explaining incidence or conditional duration, but not both.

A simulation study compares additional models used in past research and finds the model developed in this paper more accurately estimates the true parameter values and does not suffer from biases caused by treating maturity observations as censored prepayment and write off events.

The following topics are left for further research. First, unobserved heterogeneity is not included in the current model. It could be incorporated in the spirit of Deng et al. (2000) with different parameters across the groups or with random heterogeneity factors as in Mealli and Pudney (1996). Second, dynamics are not incorporated due to the restricted time frame of our application. We refer the reader to McNeil and Wendin (2007), Duffie et al. (2009), Koopman et al. (2009), and Koopman et al. (2011) for a possible avenue to extend of our model with unobserved time specific frailty factors.

The rest of the paper is divided into the following sections: Section 2 examines survival analysis methods and cure rate models; Section 3 develops the proposed model; Section 4 presents the results of the simulation, empirical and forecasting studies; and, Section 5 concludes.

2. MOTIVATION AND PREVIOUS ADVANCES

The fundamental quantity under assessment is time to event data, from a risk assessment perspective, where the event of interest may be default or write off, and where the 'failure' time would be measured from loan origination to loan closure. The set of observable failure times are in the set of non-negative real numbers. In past research, each observed failure time, t_i, is treated as a random variable with a probability density function (pdf), $f(t)$. The cumulative density function (cdf), $F(t)$, also defines the survival function, via $S(t) = 1 - F(t)$. The focus of many applications is to specify and estimate the distribution for the failure time variable, though non-parametric estimation techniques are also frequently used.

In Banasik et al. (1999), Stepanova and Thomas (2002), Andreeva (2006) and Bellotti and Crook (2009), duration analyses with an independent competing risk assumption for the events of prepayment and write off are conducted, where the observed maturity events are treated as censored prepayment and default event times. Under this independent competing risks assumption the prepayment and default observations are analysed separately, treating all other observed failure times as censored default or prepayment times, respectively. The likelihood function ($L(\Theta)$) across observations $i = 1, ..., N$ is:

$$L(\Theta) = \prod_{i=1}^{N} f(t_i)^{1-\delta_i} S(t_i)^{\delta_i} \tag{2.1}$$

where δ_i takes the value of 1 for censored observations, and 0 otherwise, and Θ is a generic parameter vector.

Deng et al. (2000) simultaneously model the events of prepayment and default, using an extended framework originally developed in the series of seminal papers HHSM. Pavlov (2001) and Ciochetti et al. (2002) apply the same framework where the data is split into the mutually exclusive sets of prepayment, default, censoring and unknown event types. The set of censored events again contains all maturity observations. The log-likelihood function ($\mathfrak{L}(\Theta)$) is maximised and is written as:

$$\mathfrak{L}(\Theta) = \sum_{i=1}^{N} \{\delta_{Pi} \ln [F_P(t_i)] + \delta_{Di} \ln [F_D(t_i)] + \delta_{Ui} \ln [F_U(t_i)] + \delta_{Ci} \ln [F_C(t_i)]\} \quad (2.2)$$

where $F_j(t_i)$ for $j = P, D, U, C$ are the probabilities of mortgage termination due to (P)repayment, (D)efault, (U)nkown reason and (C)ensoring, respectively. The δ_{ji} for $j = P, D, U, C$ are indicator variables taking the value of 1 when the i^{th} individual experiences event j, and 0 otherwise.

The treatment of maturity as a censored observation, whose true observed 'failure' time is then treated as an under-estimate of the actual time to maturity (i.e. of itself), can potentially lead to bias in the parameter estimates and is clearly not appropriate or optimal. The class of cure rate models motivate a solution to this issue, as developed here, being a class of mixture model, where a binary distribution is mixed with a typical failure time distribution, with positive support, and applied to time-to-event data where there are individuals who never experienced, and will never experience, one or more of the events under study. In addition, it is not known ab initio to which group an individual belongs, and in fact the incidence of all the events are treated as stochastic. Tsodikov et al. (2003) define the surviving proportion (i.e. those 'cured') as the non-zero asymptotic value, p, of the survival function, $S(t)$, as t tends to infinity. In the medical literature the term survival analysis is used, while in economic applications the term duration analysis is more frequent. In this paper we do not distinguish between the two.

This then leads to the two-component (binary) mixture model that Tsodikov et al. (2003) show can be characterized as:

$$S(t) = E\left\{[S(t|\,\zeta = 1)]^{\zeta}\right\} = (1-p) + pS(t|\,\zeta = 1) \quad (2.3)$$

where ζ is a binary variable taking the value 1 with probability p and 0 otherwise. The surviving fraction is $(1-p)$ and the incidence of susceptible individuals is p, with duration described by the conditional survival function, $S(t|\,\zeta = 1)$.

Hoggart and Griffin (2001) apply the cure rate methodology to customer attrition from banks, assuming there are N *iid* Poisson risks with mean θ, so that the probability an individual does not attrite becomes $\exp\{-\theta\}$. This method is applied to clinical data on patients suffering from cancer in Cancho et al. (2009). Farewell (1982) parameterises the incidence proportion using logistic regression and the duration distribution using the Weibull density function. Sy and Taylor (2000) and Peng and Dear (2000) develop semi-parametric techniques for the binary cure rate model. Tsodikov et al. (2003) develop non-parametric and semi-parametric Bayesian multinomial methods for cure rate models. In all these situations, and for credit data, hurdle rate or zero-inflated models would be inadequate to resolve the issue of treating pre-determined terminal events as censored observations. This is due to the density mass of each pre-determined variable being observable; all these density masses do not occur at the zero value; and, these models

would not appropriately incorporate censoring. In the following section a fully-parametric model incorporating cure rate techniques is developed.

3. MODEL FOR SIMULTANEOUS DURATION ESTIMATION OF M MUTUALLY EXCLUSIVE EVENTS WITH A SUBSET OF NON-STOCHASTIC EVENTS

Assume there are M mutually exclusive terminal events in the set $\mathbf{M}$, where the first J events have non-stochastic or pre-determined event times and where $J \leq M-1$. The actual failure times are denoted $\widetilde{T}_{ij}$, where $j \in \mathbf{M}$ indicates the j-th event and $i = 1, ..., N$ indicates the i-th individual. The time to the pre-determined events is $\overline{a}_j$ and with this we can define

$$\widetilde{T}_{ij} = \overline{a}_j \text{ for } j = 1, ..., J \tag{3.4}$$

$$\widetilde{T}_{ij} \in [0, \infty) \text{ for } j = J+1, ..., M \tag{3.5}$$

Next, define $\widetilde{\mathbf{q}}$ as a vector of labels where the i-th element, $\widetilde{q}_i$, indicates which event in $\mathbf{M}$ was observed for the i-th individual. Then, M binary indicators are defined as:

$$y_{ij} = \left\{ \begin{array}{ll} 1 & \text{if } \widetilde{q}_i = j \\ 0 & \text{otherwise} \end{array} \right\}, \quad \text{for} \quad j \in \mathbf{M} \quad \text{and} \quad i = 1, ..., N. \tag{3.6}$$

The density of $\widetilde{\mathbf{q}}$, being the observed incidences of each event, follows a multinomial distribution, characterized with likelihood: $\prod_{i=1}^{N} \prod_{j=1}^{M} p_{ij}^{y_{ij}}$.

Here, the probability of incidence for each event is: $\Pr(y_{ij} = 1) = p_{ij} = F_j(\mathbf{x}_i, \beta_{Ij})$; where $\mathbf{x}_i$ and β_{Ij} are $(k \times 1)$ column vectors of individual specific regressors and corresponding coefficients, respectively. The subscript "I" indicates these effect parameters solely pertain to the incidence of events. The function, F_j, must satisfy the following conditions: $p_{ij} \in [0, 1]$ and $\sum_{j=1}^{M} p_{ij} = 1$, so that the events form a set of mutually exclusive and exhaustive events. Note that events $j = 1, ..., J$ occur stochastically, even though they have non-stochastic durations. Thus, we do not know which loans or observations will have non-stochastic durations ab initio, but conditional on observing an event in $j = 1, ..., J$, we know the duration exactly, as detailed below.

The actual failure times or durations, $\widetilde{T}_{ij}$, are assumed to be conditionally *iid* across i, with pdf $f_j(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1)$ for $j = 1, ..., M$, where $\phi_j = \left(\beta'_{Lj}, \gamma_j\right)'$ is the set of parameters for distribution f_j. Here, the subscript "L" indicates parameters pertaining to duration (latency) only; $\mathbf{x}_i$ and y_{ij} are as above. For the events with fixed duration, i.e. $j = 1, ..., J$, it follows that $\Pr\left[\widetilde{T}_{ij} = \overline{a}_j\right] = 1$. In the case that $j = J+1, ...M$, a density with positive support is used for the duration of these events.

Our method also deals with censored observations, as follows. Let C_i be the time to censoring for the i^{th} individual. Each individual will have a censoring time, however, only a subset of individuals will have censoring times without also having an observed failure time, i.e. will actually be a censored observation. The observed time to an event or censoring is defined as:

$$T_i = \widetilde{T}_{ij} \wedge C_i \tag{3.7}$$

Let $\mathbf{T} = (T_1 \cdots T_N)'$ be the vector of observed failure or censoring times for all individuals. A binary indicator variable is then defined, to signal if an event has an observed

duration or is still active, i.e it is then censored, as

$$\delta_i = \begin{cases} 1 & \text{if } \widetilde{T}_{ij} \leq C_i, \\ 0 & \text{if } \widetilde{T}_{ij} > C_i, \end{cases} \tag{3.8}$$

which is defined conditionally on either observing event j (so that $\delta_i = 1$), or on loan i being still active but censored ($\delta_i = 0$). In the latter case, the information on which event will occur is not available in the sample. This is taken into account in further derivations with the help of the binary indicators, y_{ij}, for the observed component in the likelihood function. In the censored component of the likelihood function, all event survival densities are used with the corresponding probabilities.

Under this "right" censoring mechanism, the random variables C_i are iid across i, with pdf v_i and cdf V_i. Conditional on the observed regressors for individual i, the data pairs (T_i, δ_i) are assumed independent over i. The censoring mechanism here is consistent with the non-informative censoring mechanisms detailed in Kalbfleisch and Prentice (2002).

The uncensored events, including those non-stochastic durations, have incidence probabilities:

$$\begin{aligned} &\Pr\left[T_i \in [t, t+dt), \delta_i = 1 \mid \mathbf{x}_i, \phi_j\right] \\ &\quad = \Pr\left[C_i \geq t + dt\right] \Pr\left[y_{ij} = 1\right] \Pr\left[\widetilde{T}_{ij} \in [t, t+dt) \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right] \\ &\quad \simeq \left[1 - V_i(t)\right] p_{ij} f_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right) dt \qquad \text{for } j = J+1, ..., M, \end{aligned} \tag{3.9}$$

while for $j = 1, ..., J$, we obtain

$$\Pr\left[y_{ij} = 1\right] \Pr\left[T_i \in [t, t+dt), \delta_i = 1 \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right] \simeq \left[1 - V_i(t)\right] p_{ij} \tag{3.10}$$

For the censored observations we do not know which event will occur, thus their probability is:

$$\begin{aligned} &\Pr\left[T_i \in [t, t+dt), \delta_i = 0 \mid \mathbf{x}_i, \phi_j\right] \\ &\quad = \Pr\left[C_i \in [t, t+dt)\right] \sum\nolimits_{j=1}^{M} \Pr\left[y_{ij} = 1\right] \Pr\left[\widetilde{T}_{ij} \geq t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right] \\ &\quad \simeq v_i(t) \left\{ \sum_{j=1}^{J} p_{ij} + \sum_{j=J+1}^{M} p_{ij} S_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right) \right\} dt \qquad \text{for } j = 1, ..., M \end{aligned} \tag{3.11}$$

Note that for $j = 1, ..., J$:

$$\begin{aligned} S_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right) &= \Pr\left[\widetilde{T}_{ij} \geq t = \overline{a}_j \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right] \\ &= 1 - \Pr\left[\widetilde{T}_{ij} < t = \overline{a}_j \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right] = 1 \end{aligned} \tag{3.12}$$

since conditional on $y_{ij} = 1$, t is the maturity date.

Given that the censoring mechanism is noninformative, the terms relating to the pdf and cdf of the censoring variables can be ignored, as constants of proportionality. The resulting likelihood for the set of parameters $\mathbf{\Theta} = (\beta_{I1}', ..., \beta_{IM}', \phi_1', ..., \phi_M')'$, with independent and noninformative right censoring times is:

$$L\left(\mathbf{\Theta} \mid \mathbf{X}, \widetilde{\mathbf{q}}, \mathbf{T}, \delta\right) \propto \prod_{i=1}^{N} \left\{ \prod_{j=1}^{J} p_{ij}^{y_{ij}} \prod_{j=J+1}^{M} \left[p_{ij} f_j\left(t\right)\right]^{y_{ij}} \right\}^{\delta_i} \left\{ \sum_{j=1}^{J} p_{ij} + \sum_{j=J+1}^{M} p_{ij} S_j\left(t\right) \right\}^{1-\delta_i} \tag{3.13}$$

where $\mathbf{X} = \left(\mathbf{x}_1 \cdots \mathbf{x}_N\right)'$; δ has typical element δ_i as defined in equation 3.8. For $j = J + 1, ..., M$, $f_j\left(t\right)$ is $f_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right)$ and $S_j\left(t\right)$ is the survival function $S_j\left(t \mid \mathbf{x}_i, \phi_j, y_{ij} = 1\right)$ where the conditional statements have been dropped for notational ease. In addition, for $j = 1, ...J$, $f_j\left(t\right)$ and $S_j\left(t\right)$ take values of unity as in equation 3.12. This model does not decompose when censored observations are present and must be estimated jointly. When all events under observation have occurred, i.e. there is no censoring, the likelihood naturally decomposes into the individual event components.

4. SIMULATION STUDY AND EMPIRICAL APPLICATION

4.1. Model Application to Credit Data

There are three terminal events and one event, maturity, has predetermined duration $(J = 1)$ in credit data applications. Let $\mathbf{M} = \{1, 2, 3\}$ for maturity, write off and prepayment events, respectively. The failure time variables have the following restrictions:

$$\widetilde{T}_{i1} = \overline{a}_1, \;\; \widetilde{T}_{i2} \in [0, \overline{a}_1 + \varepsilon), \;\; \widetilde{T}_{i3} \in [0, \overline{a}_1), \tag{4.14}$$

where ε is a small and positive, to account for the write off process in banks' collections departments. The chosen functional form for the incidence probabilities, F_j, will be the alternative-invariant form of the Multinomial Logit (MNL), characterised as:

$$F_j\left(\mathbf{x}_i, \beta\right) = \frac{\exp\left(\mathbf{x}_i^T \beta_{Ij}\right)}{\sum_{l=1}^{3} \exp\left(\mathbf{x}_i^T \beta_{Il}\right)}. \tag{4.15}$$

For identification, the regression parameters for the incidence of maturity are set to zero, i.e. $\beta_{I1} = \mathbf{0}$, making $(j = 1)$ the base category for comparison.

The distributional assumptions applied to the events with stochastic duration, $\widetilde{T}_{ij}$ where $j = 2, 3$, are

- $\mathbf{G}amma\left(\gamma_{Lj}, \theta_{Lj}\right)$

$$\frac{1}{\Gamma(\gamma_{Lj})} \exp\left(-\mathbf{x}_i^T \beta_{Lj} \gamma_{Lj}\right) t_i^{\gamma_{Lj}-1} \exp\left\{-\exp\left[\ln(t_i) - \mathbf{x}_i^T \beta_{Lj}\right]\right\}$$

  with $\theta_{Lj} = \exp\left(\mathbf{x}_i^T \beta_{Lj}\right)$,
- $\mathbf{W}eibull\left(\gamma_{Lj}, \theta_{Lj}\right)$

$$\gamma_{Lj} \exp\left(-\mathbf{x}_i^T \beta_{Lj} \gamma_{Lj}\right) t_i^{\gamma_{Lj}-1} \exp\left\{-\exp\left[\gamma_{Lj}\left(\ln(t_i) - \mathbf{x}_i^T \beta_{Lj}\right)\right]\right\}$$

  with $\theta_{Lj} = \exp\left(\mathbf{x}_i^T \beta_{Lj}\right)$, and
- *Log-*$\mathbf{N}ormal\left(\mu_{Lj}, \sigma_{Lj}^2\right)$

$$\frac{1}{\sqrt{2\pi}\gamma_{Lj} t_i} \exp\left(-\frac{\left[\ln(t_i) - \mathbf{x}_i^T \beta_{Lj}\right]^2}{2\gamma_{Lj}^2}\right)$$

  with $\theta_{Lj} = \exp\left(\mu_{Lj}\right) = \exp\left(\mathbf{x}_i^T \beta_{Lj}\right)$ and $\sigma_{Lj} = \gamma_{Lj}$.

These distributions are respectively represented by the labels: G_j, W_j, and N_j, for $j = 2, 3$. For example, the application of the Gamma and Weibull distributions to write off and prepayment durations, respectively, will subsequently be denoted by "G_2W_3". Alternatively, density functions that specifically account for the upper bound in (4.14) could be employed. However, each density had negligible weight at this upper bound in each case for our data. This is because all loans that were prepaid were done so well before the maturity date and the same is observed for the vast majority of loans that were written off.

The empirical application of this model will focus on the origination decision of retail banking firms. The core problem facing the origination decision of a financial institution is whether to extend credit to an applicant based on information available at that point in time. Lifetime forecasts of applicant behavior are made to determine whether the business is profitable.

The aims of the simulation and empirical application are:

1 Perform simulations to explore the biases of each model applied in the literature.
2 Model lifetime account behavior using application data exclusively. This is to replicate the modelling used to inform the credit origination decision.
3 Explore the risk factors specific to this application data
4 Compare the forecasts of the most common competing risk survival analysis empirical studies against those of the model developed in this paper.

4.2. Simulation Study

The extent of the biases in the competing risks (CR) model and other models applied in the literature are explored through a simulation study, where random event times are generated from Log-Normal and Weibull distributions for write off and prepayment events, respectively (N_2W_3). Four studies are detailed below, differing in chosen parameter values and chosen incidence proportions. The fixed proportions are created using a uniform (0,1) random variable labelled U_I. Five models (labelled I to V) are used to estimate the parameters on sets of 1,000 observations, replicated 20,000 times. Once an incidence of maturity, write off, or prepayment has been randomly observed for each of the 1,000 data points, a duration vector, $\mathbf{T}$, can be constructed. Each element of $\mathbf{T}$ will correspond to the elements of the randomly and independently generated (N_2W_3) event times. The vector of event times and the corresponding indicator vectors are then used as observations in the likelihood equation, e.g. as in equation (3.13), which is maximised via the simplex method in Matlab software. The following models from the literature are estimated for each simulated data set:

- Model I: developed in this paper;
- Model II: same as model I except treats maturity events as censored;
- Model III: simultaneous estimation of prepayment and default without separation of incidence and duration;
- Model IV: examines prepayment individually, treating all other events as censored observations (Competing Risks (CR) for prepayment); and,
- Model V: examines write off individually, treating all other events as censored observations (CR for write off).

Scenario	p_1	p_2	p_3	$\ln(\sigma_2)$	β_2	γ_3	β_3
Sim01 $\text{Ł}_2\text{W}_3$	0.05	0.20	0.75	ln(0.6)	4.50	7.00	5.50
Sim02 $\text{Ł}_2\text{W}_3$	0.05	0.75	0.20	ln(0.6)	4.50	7.00	5.50
Sim03 $\text{Ł}_2\text{W}_3$	0.40	0.20	0.40	ln(0.6)	4.50	7.00	5.50
Sim04 $\text{Ł}_2\text{W}_3$	0.60	0.10	0.30	ln(0.6)	4.50	7.00	5.50

Table 1. Parameter values used to generate simulation scenarios 01 to 04

		$\widehat{p}_2$	$\widehat{p}_3$	$\ln(\widehat{\sigma}_2)$	$\widehat{\beta}_2$	$\widehat{\gamma}_3$	$\widehat{\beta}_3$
Sim01	True Values	0.20	0.75	-0.5108	4.50	7.00	5.50
Model I	Mean	0.2000	0.7500	-0.5150	4.5002	7.0129	5.4999
	Std	0.0126	0.0137	0.0507	0.0421	0.2004	0.0054
	Prct Err.	-0.02%	0.01%	0.82%	0.00%	0.18%	0.00%
Model II	Mean	0.2500		-0.0630	4.8769	7.0129	5.4999
	Std	0.0137		0.0525	0.0650	0.2004	0.0054
	Prct Err.	24.98%		-87.67%	8.38%	0.18%	0.00%
Model III	Mean			-0.0630	4.8769	7.0129	5.4999
	Std			0.0525	0.0650	0.2004	0.0054
	Prct Err.			-87.67%	8.38%	0.18%	0.00%
Model IV	Mean					4.3454	5.5711
	Std					0.2784	0.0102
	Prct Err.					-37.92%	1.29%
Model V	Mean			0.4791	6.7995		
	Std			0.0403	0.1105		
	Prct Err.			-193.78%	51.10%		

Table 2. Summary of parameter estimates from Sim01; where 'Prct Err.' is the percent difference of the mean simulation value from the true parameter value.

Table 1 shows the actual parameter and distribution settings for the simulation study. The results for the simulation scenario Sim01 are displayed in Table 2. Even at only 5% maturity events being incorrectly classed as censored observations, these results clearly show significant bias in the parameter estimates for models II through to V, while indicating that there is negligible bias in the estimates from the model developed in this paper, being less than 1% in each case, and usually far less than that, when properly accounting for maturity events. The bias in the estimates from the other models is demonstrably larger, between 10 to 200 times larger in direct comparisons between the same parameters, thus indicating the framework developed in this paper assists in controlling the clear and significant types of bias caused by not accounting for the introduction of a non-random terminal event time that is part of a set of mutually exclusive terminal events. The other three simulation scenarios have quite similar results and are available from the authors upon request.

Note that the 0.82% bias observed in the estimate of the Log-Normal distribution shape parameter is also viewed as negligible for the following reasons: the absolute bias of this shape parameter (0.0042) is 3 times smaller than the Weibull distribution shape parameter; the absolute value of this shape parameter is 14 times smaller than that of the Weibull distribution whilst the relative bias is only 4 times larger.

4.3. Application Data Example

The data set of over one million observations contains information on unsecured personal loans that originated between 1 March 2001 and 31 March 2008, and were provided by one of Australian's largest financial services institutions. These loans could be contracted for terms of whole years ranging from 12 months to 84 months, and we emphasise that these are not mortgages but unsecured fixed term credit facilities. In addition to application data at the account level, sufficient information on opening and closing dates of accounts and the reason for their terminations was also provided. The list of personal loan application data/variables provided for this research is outlined in Table 3.

Number of Applicants per Loan	Time with Current Employer
Total Assets	Time with Previous Employer
Total Liabilities	Current State
Other Bank Home Loan	Time at Current Address
Other Bank Liabilities	Time at Previous Address
House Value	Guarantor
Other Value	Number of Installments
Accommodation Status	Total Loan Amount
Gender	Interest Rate at Application
Age at application	Repayment Amount

Table 3. List of Application Data.

Table 4 details the proportion of maturity, write off, prepayment and censored observations across each contracted loan term. Both censoring and maturity form a large proportion of the events in most loan term stratum. The treatment of maturity events as censored observations would likely have significant impacts on parameter estimation.

Models were estimated for each loan term strata excluding the 72 and 84 month loan term data sets. The omission is due to the length of the observation window relative to the time to the predetermined maturity events. The period under consideration covers 85 months, leaving only 13 and 1 month, respectively, where observations with 72 and 84 month maturity could be observed. Thus, small sample sizes, in conjunction with the heavy censoring, led us to omit these loan term strata. Since loan term strata are analyzed separately, this causes no selection bias in our results.

Kaplan-Meier (KM) survival curves were constructed on the data set, see Figure 1. Each event type was examined treating the others as censored in this figure. Panels (a) to (c) show, respectively, the effects of: non-stochastic duration events with probability mass at that date; heavy censoring of observations impacting survival curve estimates; and, KM survival curves that plateau to values strictly greater than zero, a trait that indicates the use of cure rate mixture models is recommended, as described in Sy and Taylor (2000). Each of these issues are managed under the model framework developed in this paper and applied in the following section.

4.4. Results of Model Fitting

Maximum Likelihood Estimation (MLE) is performed utilising the Nelder-Mead method of simplexes, which is the preferred method due to its generally more optimal conver-

TERM	Full Term	Write Off	Prepayment	Censored	TOTAL
12	43.69%	1.63%	40.87%	13.79%	1.68%
24	20.04%	2.44%	63.37%	14.16%	8.20%
36	8.56%	3.10%	69.08%	19.27%	12.40%
48	3.41%	4.23%	70.55%	21.79%	10.10%
60	1.38%	5.28%	62.98%	30.37%	24.10%
72	0.54%	5.96%	63.99%	29.49%	4.03%
84	0.01%	6.70%	52.89%	40.38%	39.50%
TOTAL	4.14%	5.18%	60.20%	30.50%	100.00%

Table 4. Percent of Accounts Experiencing Defined Permanent Events and Censoring with Contracted Term as Stratum

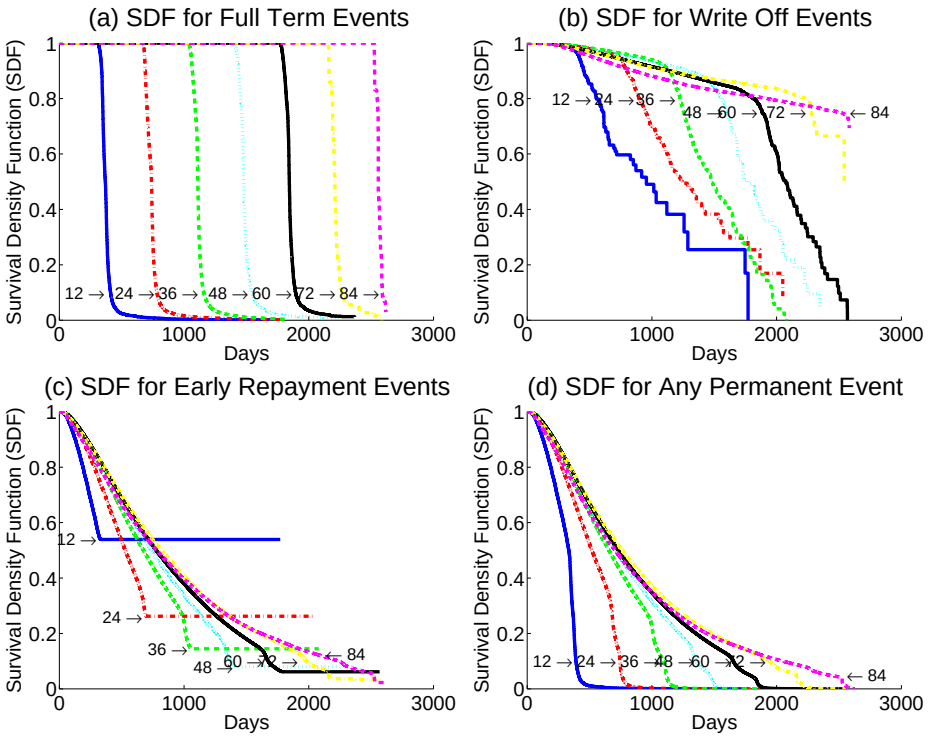

Figure 1. Kaplan Meier survival function estimates on the full data set with 12 to 84 month loan term stratum labels.

gence properties, compared to a Newton-Raphson type search, for these models and data sets. The best fitting model, with a separate set of regressors for incidence and duration, and distributions for duration for write-offs and pre-payment, was chosen based on its Bayesian Information Criteria (BIC). The parameter estimates for this model are subsequently presented and briefly discussed. Each regression effect estimate is interpreted with particular attention to the new results showing risk factors that can act in different directions upon incidence and duration whilst other risk factors are significant for either incidence or duration, but not both. All results are, in our opinion, logically consistent

with expectation and intuition and are explored by examining conditional log odds and impacts of increases in regressors on the mean duration.

LAR	ln (Loan Amount/ Total Assets)	HV	ln (House Value in 1000's)
TL	ln (Total Liabilities in 1000's)	Lamt	Total Loan Amount
TA	ln (Total Assets in 1000's)	TCA	Time at Current Address
TCE	Time with Current Employer	TPA	Time at Previous Address
TPE	Time with Previous Employer	GEN	1 if Female, 0 otherwise
PCR	1 if Period 36 to 56 of low credit quality, else 0	AddYrs	TCA + TPA
Guar	1 if guarantor on loan, 0 otherwise	EmpYrs	TCE + TPE
Int	Interest Rate at Application	Age	Age in years

Table 5. Variables used in empirical application

The regressors used in this empirical application capture financial, demographic, social stability and collective responsibility aspects of the applicants and are described in Table 5. The set of parameters for the k regressors can be charactised as $\boldsymbol{\Theta} = (\beta'_{I2}, \beta'_{I3}, \phi'_2, \phi'_3)'$, where $\phi_j = \left(\beta'_{Lj}, \gamma_j\right)'$ and each β vector is $k \times 1$ and the γ's are scalar, bringing the total number of parameters in the model to $(k \times 4) + 2$.

	12	24	36	48	60
G_2G_3	130,550.09	999,829.49	1,691,494.65	1,457,727.29	3,233,059.25
G_2N_3	132,062.25	1,011,948.31	1,709,943.73	1,470,431.32	3,252,580.28
G_2W_3	129,316.02	⧾992,198.73	⧾1,683,291.80	⧾1,454,077.57	⧾3,230,943.09
N_2G_3	130,518.20	999,958.15	1,691,834.76	1,458,044.18	3,233,504.08
N_2N_3	132,029.61	1,012,086.80	⧇1,710,278.91	⧇1,470,751.17	⧇3,253,084.87
N_2W_3	⧾129,285.01	992,336.37	1,683,622.32	1,454,381.49	3,231,329.38
W_2G_3	130,645.22	999,976.88	1,691,509.26	1,457,749.99	3,233,488.50
W_2N_3	⧇132,155.98	⧇1,012,089.73	1,709,957.94	1,470,447.48	3,252,984.85
W_2W_3	129,411.40	992,348.72	1,683,307.10	1,454,103.83	3,231,386.54

Table 6. Bayesian Information Criteria for distribution pairs across 12 to 60 month term unsecured personal loans (⧾: minimum; and; ⧇:maximum for each term data set)

The BIC for each of the nine distribution combinations across the five data sets is displayed in Table 6. These indicate that the combination of the Gamma for for write-off and the Weibull distributional for pre-prepayment, resulted in the best model in the largest four of the five data sets. The Weibull was always the optimal distribution for pre-payment, whilst the log-normal was the least optimal distribution for pre-payment. The parameter estimates of the models with the optimal BIC values are presented in Tables 8 to 11.

Estimated conditional log odds ratios are used to interpret the results of the incidence models. Further, impacts of the parameter estimates on the mean duration are used to interpret conditional duration components of the model. The conditional log odds ratio of the k^{th} regressor between any two events (j and l) is $\frac{\partial \log(p_j/p_l)}{\partial x_k} = \beta_{jk} - \beta_{lk}$. Direct interpretations of the parameters yield comparison of relative risk to the base category

event of maturity (since $\beta_{I1} = \emptyset$). The means of each duration distribution are presented below in Table 7:

Distribution	Gamma	Weibull	Log-Normal
Mean	$\gamma \exp(x_i^T \beta_L)$	$\exp(x_i^T \beta_L)\Gamma(1 + 1/\gamma)$	$\exp(x_i^T \beta_L + \sigma^2/2)$

Table 7. Mean durations for distributions used in this study

The estimates of β_L all act directly on the mean duration of an event allowing for interpretation that a positive point estimate's effect is to increase the time to the event occurring, holding all other variables constant.

The financial variables in this model are the Loan to Asset Ratio (LAR) and Total Liabilities (TL). The LAR is a measure of how leveraged the applicant is at loan origination, whilst the TL variable is more closely correlated to wealth and borrowing power. The conditional log odds ratios are consistent with expectation, where the more geared an applicant is the more likely they are to write off than mature or prepay (Table 8, $\beta_{I2(LAR)} > 0$ and $\beta_{I2(LAR)} > \beta_{I3(LAR)}$ from Table 9 for all terms). Whereas the relative risk of write off to prepayment or maturity is decreased as TL increases (Table 8, $\beta_{I2(TL)} < 0$ and $\beta_{I2(TL)} < \beta_{I3(TL)}$ from Table 9 for all terms). Despite the relative risk of write off decreasing as TL increases, conditional on experiencing write off the higher the applicant's TL the sooner they will write off in the shorter loan term data sets (Table 10, $\beta_{L2(TL)} = -0.0011$ for 24 month term loans). This represents a risk to the revenue line on the profit and loss statement of the lending institution. The impact of LAR to the revenue line can be interpreted similarly. The higher the gearing of an applicant, the sooner they will write off if they were to actually experience write off (Table 10, $\beta_{L2(LAR)} = -0.0115$ for 60 month term loans).

The number of months with current and previous employer (EmpYrs) and the months at current and previous address (AddYrs) are proxies for the employment and residential stability of an applicant. Neither of these characteristics is significant in explaining the incidence of a credit event (see Tables 8 and 9 for incidence parameters), holding the other factors constant. Whilst residential stability is also not significant in explaining the conditional duration of any event (see Tables 10 and 11 for duration parameters), EmpYrs does significantly impact the conditional write off duration (see Table 10): the more one has been employed with their last two employers, the sooner they may write off in the 24 and greater month loan term data sets (Table 10, $\beta_{L2(EmpYrs)} = 0.0005$ for 60 month term loans), conditional on write off occurring. Though this seems somewhat counter intuitive, it could be related to long term employment, perhaps in the same role with a set of (now) redundant skills. However, the very small magnitude of the parameter estimates (times the actual values for Empyrs) for this variable ensure it does not have an economically significant effect on the portfolio.

Age and gender are the two personal demographic variables in this best fitting model. Age is insignificant in explaining both incidence and conditional duration (see Tables 8 to 11 for the age variable significance) as would be expected in a model that is already conditioned on financial variables where higher LAR is concentrated in the youth and higher TL is strongly correlated with age in this data. This is not the case for gender which is significant in all parts of the models, except two (Table 8, $\beta_{I2(GEN)}$ for 48 month term and from Table 11, $\beta_{L3(GEN)}$ for 12 month term). Following, if the applicant

is female, the relative risk of write off to prepayment or maturity is significantly reduced (Table 8, $\beta_{I2(GEN)} < 0$ for 24, 36 and 60 month terms, and $\beta_{I2(GEN)} < \beta_{I3(GEN)}$ from Table 9 for 24 and greater month terms) and conditional on experiencing write off the mean time to write off is increased if the applicant is female (Table 10, $\beta_{I2(GEN)} > 0$ for all terms). The effects of an applicant being female are to increase the relative incidence of prepayment (Table 9, $\beta_{I3(GEN)} = 0.1615$ for 48 month terms and $\beta_{I3(GEN)} > 0$ for 24 and greater terms) whilst increasing the conditional mean duration to prepayment (Table 11, $\beta_{L3(GEN)} > 0$ for all statistically significant parameter estimates). In industry application, gender would likely not be used, despite being a powerful discriminator of financial risks, given that it is not legal to discriminate based on gender alone. It is possible there are omitted variables not available in this study, that could be substituted for gender, such as cash savings balances or risk based pricing to control for the impact of adverse selection.

The last variable in the models is the guarantor (Guar) indicator and gives insights into moral obligation and collective responsibility. At the origination of a loan, for there to be a guarantor it is most likely a parent or guardian of a young adult guaranteeing to support the servicing of the loan. This creates a powerful relationship between the applicant and the guarantor in their focus toward the responsibilities of servicing the loan. With the repercussions of being liable for the loan should the main party fail, an individual would normally not guarantee a loan they believed they would end up paying for. This is also evidenced in the relative risk of write off to prepayment and maturity being significantly reduced with the presence of a guarantor (Table 8, $\beta_{I2(Guar)} = -1.4184$ for 60 month term, $\beta_{I2(Guar)} < 0$ and $\beta_{I2(Guar)} < \beta_{I3(Guar)}$ from Table 9 for all terms). Furthermore, in the presence of a guarantor the write off conditional mean duration is also increased (Table 10, $\beta_{L2(Guar)} > 0$ for all terms). The presence of a guarantor is also significant in explaining the incidence of prepayment but not the conditional mean duration (see Tables 9 and 11 for significance of $\beta_{I3(Guar)}$ and $\beta_{L3(Guar)}$, respectively). These results indicate that the effect of the guarantor will significantly reduce the losses and the relative risk of prepayment to write off is increased (Tables 8 and 9, $\beta_{I3(Guar)} > \beta_{I2(Guar)}$ for all terms) whilst the relative risk of prepayment to maturity is decreased (Table 9, $\beta_{I3(Guar)} < 0$ for all terms), making them likely to be more profitable applicants for a financial institution.

Overall these results are logically consistent with expectations. The flexibility of this framework has enabled this through allowing parameters to act freely in any direction on incidence and latency whilst simultaneously and jointly estimating them under one likelihood function. The next section of the paper explores how well the model discriminates between the incidence of the events and how well it forecasts the conditional durations of the events.

4.5. Forecasting Results

This section of the paper examines forecast comparisons between a typical competing risks (CR) survival analysis framework (see equation 2.1) and the model developed in this paper (see equation 3.13). A Multinomial Logit (MNL) (see equation 4.15) is also employed so as to compare with the MNL component of the model developed in this paper. The data was split into sample and forecast periods: the first 80% of the data, is the sample period for fitting, being all loans resolved or completed in the period 1 March 2001 to 31 March 2008. The last 20% of the data points are reserved as an

	12 N_2W_3	24 G_2W_3	36 G_2W_3	48 G_2W_3	60 G_2W_3
Constant	-2.7773^a	-1.3675^a	-0.3377^a	0.6039^a	1.2310^a
	(0.0180)	(0.0124)	(0.0083)	(0.0080)	(0.0052)
LAR	0.2108^a	0.1739^a	0.1291^a	0.1542^a	0.1643^a
	(0.0023)	(0.0013)	(0.0008)	(0.0007)	(0.0004)
TL	-0.0671^a	-0.0418^a	-0.0332^a	-0.0175^a	-0.0232^a
	(0.0009)	(0.0005)	(0.0004)	(0.0004)	(0.0002)
EmpYrs	-0.0048	-0.0043	-0.0029	-0.0041	-0.0031
	(0.0079)	(0.0048)	(0.0034)	(0.0033)	(0.0021)
AddYrs	-0.0027	-0.0032	-0.0032	-0.0032	-0.0027
	(1.9394)	(0.2976)	(0.1916)	(0.1847)	(0.1435)
Age	-0.0002	-0.0028	-0.0098	-0.0058	-0.0052
	(0.1564)	(0.1056)	(0.0704)	(0.0675)	(0.0443)
Guar	-1.4361^a	-1.1311^a	-1.1199^a	-1.1441^a	-1.4184^a
	(0.0465)	(0.0233)	(0.0263)	(0.0458)	(0.0452)
GEN	0.0466^a	-0.1988^a	-0.1012^a	0.0008	-0.0926^a
	(0.0111)	(0.0060)	(0.0068)	(0.0119)	(0.0117)

Table 8. Write off incidence parameter estimates $\{\beta_{I2}\}$; a, b, & c indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively. Standard errors are displayed for each parameter estimate in brackets.

	12 N_2W_3	24 G_2W_3	36 G_2W_3	48 G_2W_3	60 G_2W_3
Constant	0.3850^a	1.6696^a	2.3822^a	3.2075^a	3.6645^a
	(0.0038)	(0.0021)	(0.0023)	(0.0041)	(0.0038)
LAR	-0.0544^a	-0.0733^a	-0.0621^a	-0.0469^a	0.0239^a
	(0.0003)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
TL	0.0017^a	0.0124^a	0.0110^a	0.0046^a	0.0310^a
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
EmpYrs	0.0001	-0.0001	0.0000	-0.0002	-0.0003
	(0.0017)	(0.0008)	(0.0009)	(0.0015)	(0.0015)
AddYrs	-0.0004	-0.0003	-0.0002	-0.0003	-0.0000
	(0.1166)	(0.0374)	(0.0376)	(0.0583)	(0.0596)
Age	-0.0180	-0.0214	-0.0199	-0.0224	-0.0244
	(0.0314)	(0.0167)	(0.0188)	(0.0319)	(0.0306)
Guar	-0.2471	-0.1028^a	-0.0887^a	-0.1864^a	-0.0525^a
	(0.2552)	(0.0043)	(0.0049)	(0.0070)	(0.0055)
GEN	-0.0675^a	0.0564^a	0.1094^a	0.1615^a	0.1078^a
	(0.0021)	(0.0007)	(0.0010)	(0.0014)	(0.0014)

Table 9. Prepayment incidence parameter estimates $\{\beta_{I3}\}$; a, b, & c indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively. Standard errors are displayed for each parameter estimate in brackets.

out of sample forecast period, being all censored observations and all loans completed or continuing in the period 1 April 2008 to 30 June 2010. The forecasts for incidence,

	12 N_2W_3	24 G_2W_3	36 G_2W_3	48 G_2W_3	60 G_2W_3
Constant	5.7465^a	4.4881^a	5.0458^a	5.3545^a	5.5759^a
	(0.0013)	(0.0003)	(0.0004)	(0.0005)	(0.0005)
LAR	0.0159^a	-0.0007^a	-0.0099^a	-0.0151^a	-0.0115^a
	(0.0002)	(0.0000)	(0.0000)	(0.0001)	(0.0001)
TL	-0.0001^c	-0.0011^a	0.0029^a	0.0035^a	0.0081^a
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
EmpYrs	0.0004	0.0006^a	0.0003^c	0.0005^b	0.0005^a
	(0.0006)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
AddYrs	−0.0002	0.0001	0.0002	0.0003	0.0001
	(0.2644)	(0.0081)	(0.0091)	(0.0137)	(0.0153)
Age	0.0036	−0.0008	−0.0015	−0.0004	0.0000
	(0.0114)	(0.0029)	(0.0033)	(0.0047)	(0.0037)
Guar	0.1307^a	0.0996^a	0.1779^a	0.2550^a	0.3292^a
	(0.0145)	(0.0060)	(0.0054)	(0.0068)	(0.0053)
GEN	0.1034^a	0.0489^a	0.0117^a	0.0290^a	0.0642^a
	(0.0036)	(0.0016)	(0.0014)	(0.0018)	(0.0014)
$\ln(\gamma_{L2})$	-0.9542^a	1.7555^a	1.4760^a	1.2425^a	1.0702^a
	(0.3478)	(0.1549)	(0.1040)	(0.0990)	(0.0652)

Table 10. Write off duration parameter estimates $\{\beta_{L2}, \gamma_{L2}\}$; *a*, *b*, & *c* indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively. Standard errors are displayed for each parameter estimate in brackets.

conditional duration and unconditional duration are compared between models on the basis of forecast accuracy, discriminatory power of the model and bias removal.

Lorenz curves (Figure 2) and Gini coefficients (Table 12) are used to compare MNL estimates and the incidence component of the model developed in this paper (Watkins-Vasnev-Gerlach: WVG). The events of write off and prepayment are banded into groups of equal size and ordered by the score $(x_i^T\beta)$. The x and y axes plot the cumulative percent of the event of interest against the cumulative percent of all other events, respectively. The further these curves depart from a unit (1) gradient, the better is the model's discriminatory ability. The power of discrimination is directly measured by the Gini coefficient, which is the ratio of the area between the curve with unit slope and the Lorenz curve, to the area under the curve with unit slope.

The WVG and MNL Gini coefficients are all relatively similar. Both models are better at discriminating write off events than maturity or prepayment. The incidence of maturity is better discriminated as the loan term increases for both models. Neither model, using this set of application data, offers a level of discrimination above a 0.5 Gini coefficient which would be a benchmark for industry standard models. Results for the Gini coefficients are displayed in Table 12.

The parameter estimates for the competing risks (CR) model (see equation 2.1 for the likelihood function specification) are displayed for 36 month term loans in Table 13. The MNL (columns 1 and 2 of Table 13) and WVG incidence estimates (Tables 8 and 9) are similar in sign and magnitude, particularly for Guarantor and LAR characteristics. The MNL incidence parameter estimates are all highly significant for the CR model.

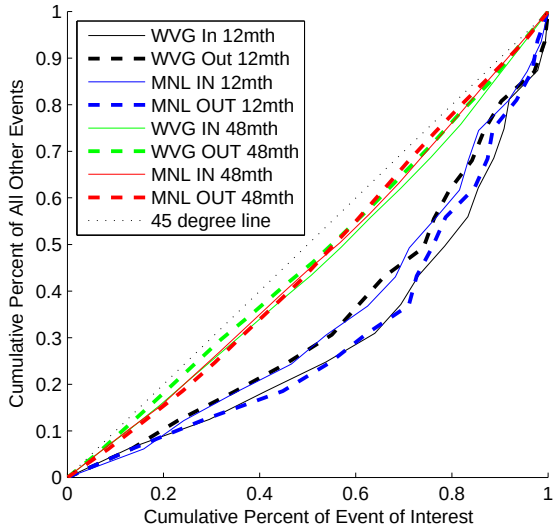

Figure 2. Lorenz curves for 12 months loan write off and 48 months loans prepayment; where WVG = Watkins-Vasnev-Gerlach, MNL = Multinomial Logit, In = Fitting Sample, Out = Out Of Sample

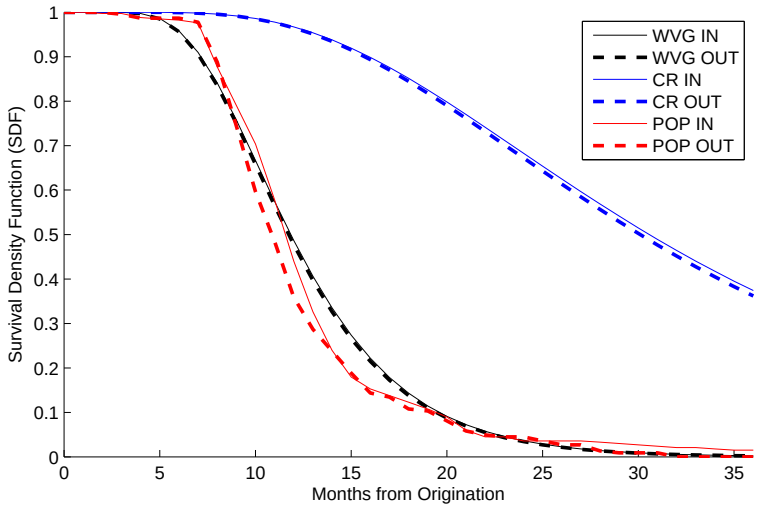

Figure 3. Write Off SDF for 12 months loans; where WVG = Watkins-Vasnev-Gerlach, CR = Competing Risks, POP = Population, IN = Fitting Sample, OUT = Out Of Sample

	12 N_2W_3	24 G_2W_3	36 G_2W_3	48 G_2W_3	60 G_2W_3
Constant	5.4251^a	6.0404^a	6.3416^a	6.4742^a	6.5554^a
	(0.0012)	(0.0005)	(0.0005)	(0.0006)	(0.0005)
LAR	0.0102^a	0.0159^a	0.0258^a	0.0341^a	0.0577^a
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
TL	0.0005^a	-0.0015^a	-0.0041^a	-0.0063^a	-0.0069^a
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
EmpYrs	-0.0001	0.0001	0.0002	0.0002	0.0002
	(0.0005)	(0.0002)	(0.0002)	(0.0003)	(0.0002)
AddYrs	0.0000	0.0000	0.0001	0.0001	0.0001
	(0.0501)	(0.0102)	(0.0085)	(0.0100)	(0.0082)
Age	0.0001	0.0032	0.0050	0.0066	0.0081^b
	(0.0099)	(0.0042)	(0.0037)	(0.0045)	(0.0035)
Guar	0.0560	0.0240^b	0.0293	0.0333	0.0650
	(0.0983)	(0.0111)	(0.0183)	(0.0211)	(0.0119)
GEN	-0.0034	0.0242^a	0.0357^a	0.0361^a	0.0493^a
	(0.0271)	(0.0087)	(0.0060)	(0.0058)	(0.0035)
γ_{L3}	2.7903^a	2.4056^a	2.0529^a	1.8044^a	1.6060^a
	(0.0383)	(0.0272)	(0.0218)	(0.0218)	(0.0168)

Table 11. Prepayment duration parameter estimates $\{\beta_{L3}, \gamma_{L3}\}$; a, b, & c indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively. Standard errors are displayed for each parameter estimate in brackets.

EVENT		Write Off		Prepayment		Maturity	
SAMPLE		In	Out	In	Out	In	Out
WVG	12	0.3031	0.3912	0.0901	0.0648	0.1003	0.0758
WVG	24	0.3661	0.2565	0.1205	0.0908	0.1263	0.0920
WVG	36	0.2326	0.2297	0.1152	0.0742	0.1495	0.0877
WVG	48	0.2062	0.1970	0.0975	0.0600	0.1997	0.0708
MNL	12	0.3099	0.3807	0.0970	0.0743	0.1050	0.0783
MNL	24	0.2635	0.2664	0.1153	0.0932	0.1248	0.0961
MNL	36	0.2372	0.2409	0.1112	0.0745	0.1522	0.0917
MNL	48	0.2060	0.1964	0.0784	0.0534	0.1936	0.0762

Table 12. Gini coefficients for the MNL and WVG models for 12 to 48 month terms

There are three notable high-level comparisons between the parameter estimates for the CR and WVG models. First, the CR and WVG models have estimates that are similar in sign for all location parameter estimates. Second, there are significant differences in magnitude for each location parameter for the write off duration estimates. Third, there are significant differences in the shape parameter estimates for both write off and prepayment. The WVG model's shape parameters are all above unity whereas the opposite is true of the CR shape parameters. All CR duration parameter estimates are significant for the model specified in equation 2.1. These differences have led to similar incidence discriminatory power, similar duration discriminatory power and a significantly lower

ability to accurately forecast duration for the CR model.

	Incidence		Duration	
	Write Off	Prepayment	Write Off	Prepayment
Constant	0.1079^{b}	2.6554^{a}	7.6190^{a}	6.5162^{a}
	(0.0505)	(0.0269)	(0.0216)	(0.0052)
LAR	0.1523^{a}	-0.0410^{a}	-0.0595^{a}	0.0304^{a}
	(0.0130)	(0.0076)	(0.0047)	(0.0015)
TL	-0.0856^{a}	0.0353^{a}	0.0341^{a}	-0.0184^{a}
	(0.0126)	(0.0064)	(0.0047)	(0.0012)
EmpYrs	-0.0034^{a}	0.0003^{b}	0.0014^{a}	0.0001^{a}
	(0.0004)	(0.0002)	(0.0002)	(0.0000)
AddYrs	-0.0032^{a}	-0.0001^{c}	0.0012^{a}	0.0000^{c}
	(0.0002)	(0.0001)	(0.0001)	(0.0000)
Age	-0.0101^{a}	-0.0209^{a}	0.0014^{a}	0.0073^{a}
	(0.0017)	(0.0009)	(0.0006)	(0.0002)
Guar	-1.2581^{a}	-0.3034^{a}	0.4296^{a}	0.0347^{a}
	(0.0881)	(0.0380)	(0.0338)	(0.0074)
GEN	-0.1037^{a}	0.1205^{a}	0.1002^{a}	0.0049
	(0.0359)	(0.0194)	(0.0128)	(0.0035)
γ_{L_j}			0.4169^{a}	0.5558^{a}
			(0.0049)	(0.0014)

Table 13. CR model parameter estimates for 36 month term data set with Gamma and Weibull distribution for Write Off and Prepayment, respective; a, b, & c indicate the parameter estimates are significant at the 1%, 5% and 10% levels, respectively. Standard errors are displayed for each parameter estimate in brackets.

The CR model forecasts for the write off conditional survival density function (SDF) show severe over estimation of the time to write off. The forecasts from the WVG model are significantly closer estimates of the observed write off conditional SDF; see Figure 3 and Figure 4 for a comparison between the population, WVG and CR models for both in and out of sample. In longer term structure loan strata the CR model shows under estimation and then over estimation of the time to write off; Figure 4. The WVG forecasts are close to the development data sample (IN), however, the speed to write off greatly increased over the out of sample data set which corresponded to the global financial crisis (GFC). Neither model responded to the change as the same types of people were now proceeding to write off faster.

The forecasts for prepayment conditional SDFs in Figure 5 show similar results as for the write off event. The CR model again shows under estimation then significant over estimation. By removing the bias created by censoring all other terminal events, the WVG model predicts the time to prepayment with greater accuracy; see Figure 5 for a comparison of the population, WVG and CR models for both in and out of sample.

Despite the WVG model providing demonstrably better forecasts upon visual inspec-

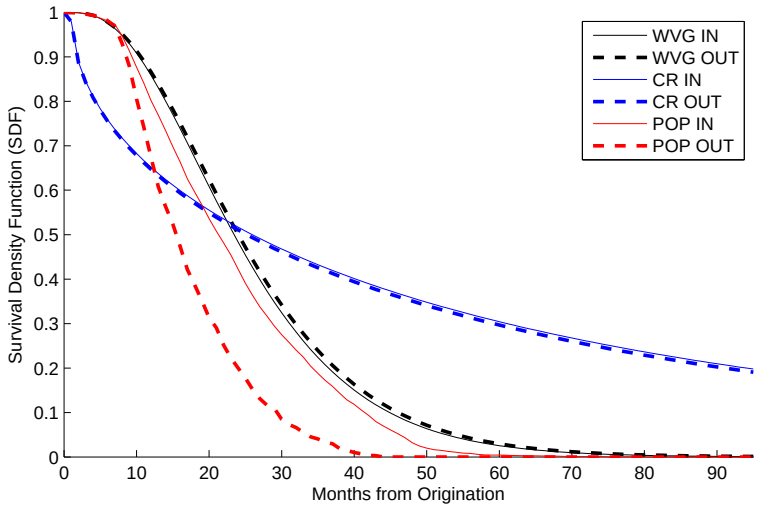

Figure 4. Write Off SDF for 48 months loans; where WVG = Watkins-Vasnev-Gerlach, CR = Competing Risks, POP = Population, IN = Fitting Sample, OUT = Out Of Sample

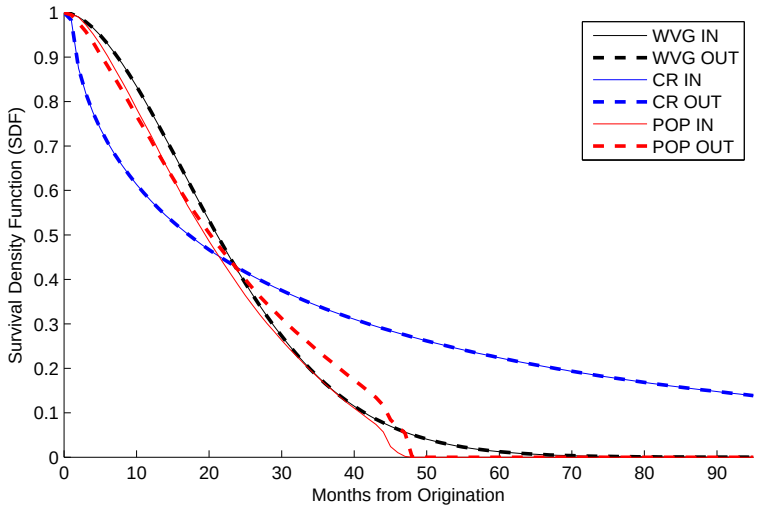

Figure 5. Prepayment SDF for 48 months loans; where WVG = Watkins-Vasnev-Gerlach, CR = Competing Risks, POP = Population, IN = Fitting Sample, OUT = Out Of Sample

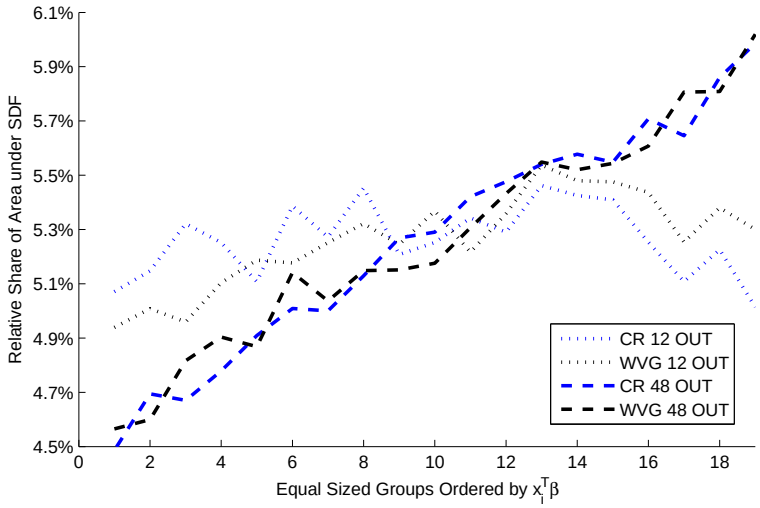

Figure 6. Rank Ordered Relative Area under SDF for prepayment events 12 and 48 months.

tion, there is little difference between the CR model and the WVG model's ability to order the fastest and slowest observations together; see Figure 6. In Figure 6 the observations were grouped uniformly by the score ($x_i^T \beta$) from the SDF estimates and empirical SDFs were plot for each group. The area under each empirical SDF was summed to find the total area. The relative share each group had of the total area was plot by the ordered groups. The lowest scores had the least area and the highest score the most area under the prepayment empirical SDFs, with the discriminatory power becoming stronger as term increased, but differing little between models. Figure 6 plots these discriminatory power results for the 12 and 48 month prepayment loans.

Combining all components of incidence and conditional duration, a forecast of the unconditional SDF can be created, see Figures 7 and 8. The MNL results were combined with the CR model results and compared against the unconditional SDF WVG model forecast in Figure 7. The results again visually indicate a significantly closer estimate of the empirical unconditional SDF. Forecast errors in the CR model are compounded when brought together with the MNL estimates. The estimates from the WVG model tend to match the in sample data set trends, with in and out of sample forecasts almost identical in this model built on limited data from the application form.

The forecasting results indicate there is little difference in discriminatory power between the models. However, the accuracy of the SDF forecasts is visually superior to those of the typical CR model. These forecasting biases in the application of the CR models reflects the extent of bias shown in the simulation study. The removal of this bias is seen in Figures 3 to 8 in the forecasts performed using the WVG model.

4.6. Diagnostics

This section examines the distributional assumptions and standard errors of estimates to assess if the model specifications, distributional assumptions and optimisation were

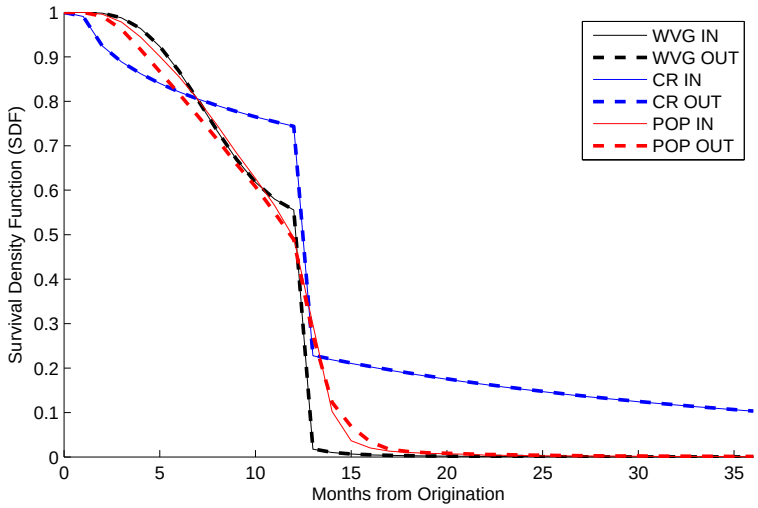

Figure 7. Unconditional SDF for 12 months loans; where WVG = Watkins-Vasnev-Gerlach, CR = Competing Risks, POP = Population, IN = Fitting Sample, OUT = Out Of Sample

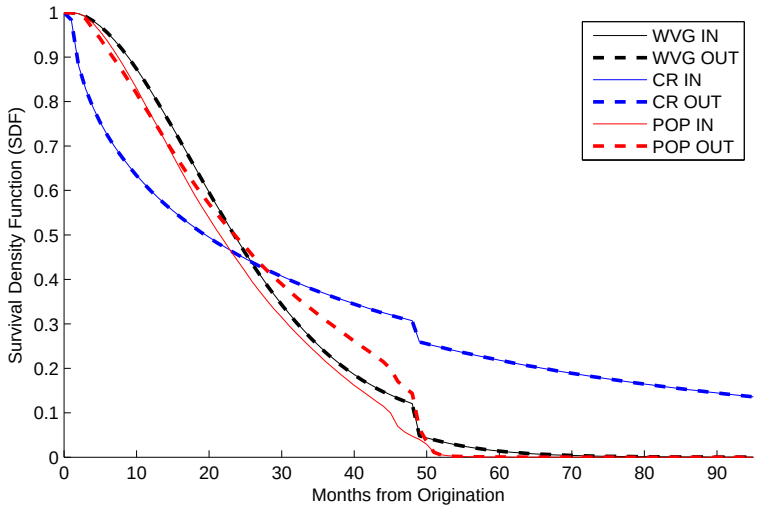

Figure 8. Unconditional SDF for 48 month loans; where WVG = Watkins-Vasnev-Gerlach, CR = Competing Risks, POP = Population, IN = Fitting Sample, OUT = Out Of Sample

adequate. Anderson Darling (AD) and Kolmogorov-Smirnov (KS) statistics (available upon request from the authors) found sufficient evidence to reject the null hypothesis that the distribution assumptions were correct. However, these classical statistics are sensitive to large data sets such as the one used in this empirical application.

Under correct Weibull distributional assumption the negative log of the KM survival curve function against log time should appear linear and should have an intercept term of $-\widehat{\beta}_{Lj}\widehat{\gamma}_{Lj}$ and a slope coefficient of $\widehat{\gamma}_{Lj}$. A regression was performed to test the linearity KM survival curve estimates, such that:

$$y_i = \widehat{\eta}_a + \widehat{\eta}_b \ln(t_i) + u_i \tag{4.16}$$

where $y_i = \ln\left\{-\ln\left[\widehat{S}_{KM}(t_i)\right]\right\}$ and was regressed on the natural logarithm of the observed failure times (t_i) separately for each event. Should the time to the particular event be distributed Weibull then the estimates should hold the following relationship with the optimised parameters of the parametric survival estimation:

$$\widehat{\eta}_a = -\widehat{\gamma}_{Lj}\widehat{\beta}_{Lj} \quad \text{and} \quad \widehat{\eta}_b = \widehat{\gamma}_{Lj} \tag{4.17}$$

A Wald test for the linearity of the log negative log KM survival curve plots was performed. The results (available upon request from the authors) indicate that only two instances do not reject the null hypothesis that the linear regression coefficients are equal to the appropriate combination of MLE parameter estimates. This was exclusive to the Weibull-Weibull and Log-Normal-Weibull distribution pairs for the prepayment event with 48 month term.

The residuals for time to each event were plotted to compare to the Extreme Value Minimum (EVM) distribution to examine the Weibull distribution model assumption and the domain of the errors is examined to determine relative accuracy. The histogram plots in Figure 9 indicate that the Weibull Weibull distribution assumption appears to match the EVM pdf most closely of the nine distribution pairs. The Gamma Weibull and Log-Normal Weibull (distribution pair assumptions for write off and prepayment, respectively) depart most from the pdf plot of the EVM distribution, characterised by an approximate tenfold decrease in the domain of the residuals. However, these plots correspond to the estimates with the lowest BIC. In addition, there may be correlation between the write off and prepayment events that needs to be addressed.

These diagnostics give evidence to support the Weibull distribution assumptions for prepayment despite the classical statistics such as the Wald test, AD and KS statistics rejecting all distribution assumptions. The Gamma and Log-Normal assumptions for the write off event deliver the lowest residual domains whilst still having consistent shape to the EVM distribution. These results provide some support to the appropriateness of applying fully parametric models to these credit data sets. The smooth, uniform profile likelihoods with clear maxima indicate the model specification results in unique global solutions in this application to credit data.

5. SUMMARY AND CONCLUSION

Research into the application of duration analysis to credit data has become increasingly abundant in recent years. Typical applications examine the credit events of default and prepayment individually. There have been applications treating the aforementioned events as dependent competing risks and have simultaneously estimated their parameters, arguing the option theoretic is the motivation for loan termination. However, all

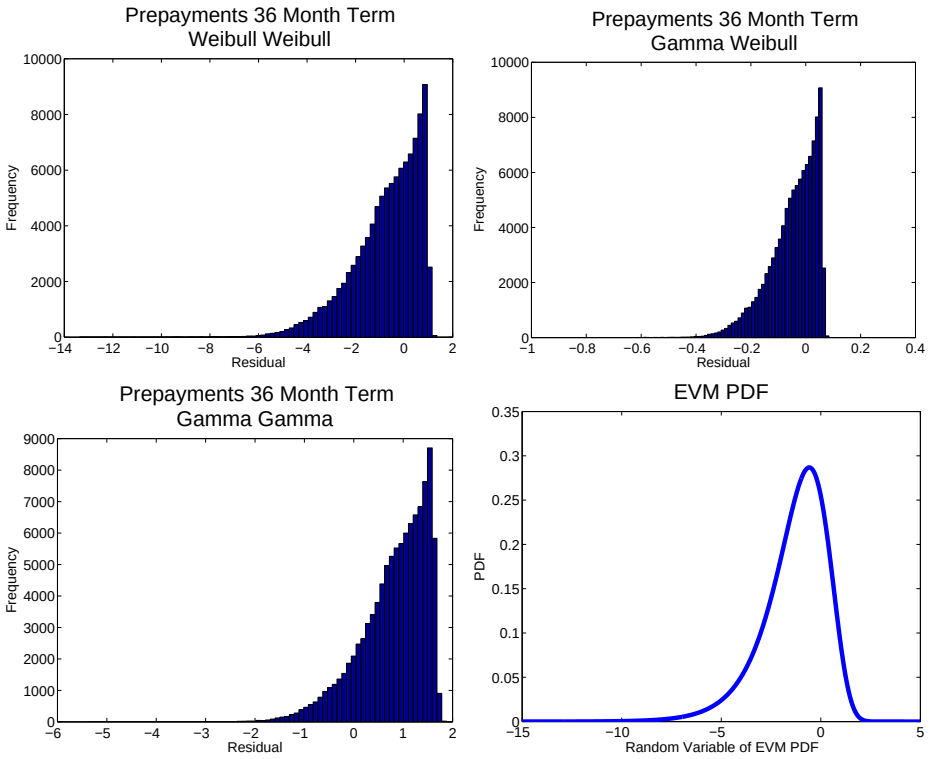

Figure 9. Time to prepayment residual histograms and EVM pdf plot

applications have failed to adequately treat credit maturity events which will lead to biases in parameter estimation.

This paper has developed the first integrated methodology for the analysis of a set of mutually exclusive events, where the duration time to a subset of events may be non-stochastic or pre-determined. It has been motivated by the cure rate method in the medical literature, augmenting these binary models to a fully parametric multinomial mixture model framework, best applied to credit data. Incidence and duration of each event in the system are estimated simultaneously.

The empirical application of the model used Australian unsecured retail personal loan credit data. The key aim is to build models using data available at application to examine: relative risk of events; impacts of risks to mean duration; compare forecasts and simulation parameter estimates for the most commonly applied models. The results found that the model developed in this paper lead to:

- Significant reduction of bias in parameter estimates
- Improved forecast accuracy over the typical competing risk applications
- Risk factors that can work in opposite directions upon incidence and duration
- Risk factors that are significant in explaining incidence or duration but not both.

The study utilised information on personal financials, employment stability, residential stability, demographics and moral obligations. All variables except age and residential

stability were found to be significant in explaining some part of the credit terminations. The financial variables relating to gearing and total liabilities increased and decreased the relative risk of write off as they increased in value. However, conditional on experiencing write off higher values in either variable lead to write off occurring sooner. This poses a risk of loss and a risk to the revenue line as the less time an account is on the books the less interest that is paid back to the lending institution.

The results within this paper were unattainable using existing methodologies. This aspect of the model allows for a deeper and more rigorous examination of credit data. There are far reaching applications of this model ranging from profit scoring to portfolio funding optimisation. Future research can extend this framework to explicitly to examine the dependence structure between prepayment and write off through either a copula or bivariate framework, in addition to an exclusively empirical study employing techniques to deal with unobserved heterogeneity and dynamics.

REFERENCES

Abramowitz, M. and I. Stegun (1965). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. New York: Dover.

Allen, L., G. DeLong, and A. Saunders (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking and Finance 28*(4), 727–752.

Altman, E. and A. Saunders (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking and Finance 21*(12), 1721–1742.

Anderson, T. and D. Darling (1954). A test of goodness of fit. *Journal of the American Statistical Association 49*(268), 765–769.

Andreeva, G. (2006). European generic scoring models using survival analysis. *The Journal of the Operational Research Society 57*(10), 1180–1187.

Banasik, J., J. Crook, and L. Thomas (1999). Not if but when will borrowers default. *The Journal of the Operational Research Society 50*(12), 1185–1190.

Bellotti, T. and J. Crook (2009). Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society 60*(12), 1699–1707.

Berkson, J. and R. Gage (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association 47*(259), 501–515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B (Methodological) 11*(1), 15–53.

Cameron, A. and P. Trivedi (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cancho, V. G., E. M. M. Ortega, and H. Bolfarine (2009). The exponentiated-Weibull regression models with a cure rate. *Journal of Applied Probability and Statistics 4*(2), 125–156.

Chen, M. H., J. G. Ibrahim, and D. Sinha (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association 94*, 909–919.

Ciochetti, D., Y. Deng, B. Gao, and R. Yao (2002). The termination of commercial mortgage contracts through prepayment and default: A proportional hazards approach with competing risks. *Real Estate Economics 30*(4), 595–633.

Crook, J., D. Edelman, and L. Thomas (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research 183*(3), 1447–1465.

Deng, Y., J. Quigley, and R. Van Order (2000). Mortgage terminations, heterogeneity, and the exercise of mortgage options. *Econometrica 68*(2), 275–307.

Duffie, D., A. Eckner, G. Horel, and L. Saita (2009). Frailty correlated default. *the Journal of Finance LXIV*(5), 2086–2123.

Farewell, V. T. (1982). The use of micture models for the analysis of survival data with long-term survivors. *Biometrics 38*(4), 1041–1046.

Han, A. and J. Hausman (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics 5*(1), 1–28.

Hoggart, C. and J. E. Griffin (2001). A bayesian partition model for customer attrition. In E. I. George (Ed.), *Bayesian Method with Applications to Science, Policy, and Official Statistics, Selected Papers from the ISBA 2000*, pp. 223–232.

Ibrahim, J., M.-H. Chen, and D. Sinha (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag.

Kalbfleisch, J. and R. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2 ed.). New Jersey: John Wiley & Sons.

Klein, J. and M. Meoschberger (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

Koopman, S. J., R. Kräussl, A. Lucas, and A. B. Monteiro (2009). Credit cycles and macro fundamentals. *Journal of Empirical Finance 16*, 42–54.

Koopman, S. J., A. Lucas, and B. Schwaab (2011). Modeling frailty-correlated defaults using many macroeconomic covariates. *Journal of Econometrics 162*, 312–325.

McCall, B. (1996). Unemployment insurance rules, joblessness, and part-time work. *Econometrica 67*(3), 647–682.

McNeil, A. J. and J. P. Wendin (2007). Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance 14*, 131–149.

Mealli, F. and S. Pudney (1996). Occupational pensions and job mobility in Britain: Estimation of a random-effects competing risks model. *Journal of Applied Econometrics 11*(3), 293–320.

Pavlov, A. (2001). Competing risks of mortgage termination: Who refinances, who moves and who defaults. *Journal of Real Estate Economics and Finance 23*(2), 185–211.

Peng, Y. and K. Dear (2000). A nonparametric micture model for cure rate estimation. *Biometrics 56*(1), 227–236.

Stepanova, M. and L. Thomas (2001). PHAB scores - proportional hazards analysis behavioural scores. *The Journal of the Operational Research Society 41*(9), 1007–1016.

Stepanova, M. and L. Thomas (2002). Survival analysis methods for personal loan data. *Operations Research Quarterly 50*(2), 277–289.

Sueyoshi, G. (1992). Semiparametric proportional hazards estimation of competing risks models with time-varying covariates. *Journal of Econometrics 51*(1-2), 25–58.

Sy, J. and J. Taylor (2000). Estimation in a Cox proportional hazards cure model. *Biometrics 56*(1), 227–236.

Thomas, L. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting 16*(2), 149–172.

Tsodikov, A. D., J. G. Ibrahim, and A. Y. Yakovlev (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association 98*(464), 1063–1078.