Zero-inflated cure rate regression models for time-to-default with applications

Francisco Louzada[1]

Institute of Mathematical Science and Computing at the University of São Paulo (USP), Brazil

Mauro R. de Oliveira Jr.

Caixa Econômica Federal and Federal University of São Carlos, Brazil

Fernando F. Moreira

University of Edinburgh Business School, Scotland, UK

Abstract

In this paper, we introduce a methodology based on zero-inflated long-term survival data in order to deal with fraud rate estimation in bank loan portfolios. Our approach enables us to accommodate three different types of loan borrowers, i.e., fraudsters, those who are susceptible to default and finally, those who are not susceptible to default. Regarding to the survival analysis framework, an advantage of our approach is to accommodate zero-inflated times, which is not possible in the standard cure rate model introduced by Berkson & Gage [7]. To illustrate the proposed method, a real dataset of loan survival times is fitted by the zero-inflated Weibull cure rate model. The parameter estimation is reached by maximum likelihood estimation procedure and Monte Carlo simulations are carried out to check its finite sample performance.

Keywords: `bank loans, fraud rate, portfolios, survival, zero-inflated, Weibull`

1. Introduction

In some cases, banks and financial institutions completely lose contact with clients as soon as their loans are granted and, therefore, all amount lent is lost. These kind of borrowers are considered fraudsters with, by definition, loan survival time equal to zero.

On the other hand, there are customers who no longer honour their loan instalments, but unlike fraudsters, they honour their instalments for a while. At some future moment, by private financial reasons, they no more meet their debts with the bank and, so, become a defaulter. However now, the client has a positive loan survival time, represented by the elapsed time until the occurrence of the default.

The term "default" used throughout this paper means the financial distress event when clients lose the creditworthiness to meet their commitments in respect a loans granted by financial institutions. Such criteria may vary from bank to bank by conservative reasons. For instance, a bank may declare a default condition

[1] Corresponding author: louzada@icmc.usp.br

of a customer if he or she has not been paying any instalments of his or her mortgage loan for more than two or three consecutive months.

To complete the picture, and ensure the survival of bankers, there are good customers, those who keep up to date with their obligations and, therefore, there will not be records of events of default.

Thus, for the survival of bankers and mainly the maximization of profits, they must seek to maintain high rate of non-defaulting loans, while the rates of fraudsters and defaulters must be very low.

These types of banking customers described above motivate the models we propose in this paper. The dataset analysed comprises customers who, in one way or another, have not honoured their contractual obligations with the bank, either by fraud in the application process, or the loss of creditworthiness over time, along with good clients who honour their obligations and have never experienced the event of default.

Such data analysis must be addressed to make a holistic risk management of the banking loan portfolio, that is, dealing with fraud prevention, control and mitigation of default and, finally, ensure the customer loyalty growth within the group of clients non-susceptible to default.

As we see in Figure 1, these considerations delimit the data we cover in this work: a set of zeros, positives and unrecorded banking loan survival times.

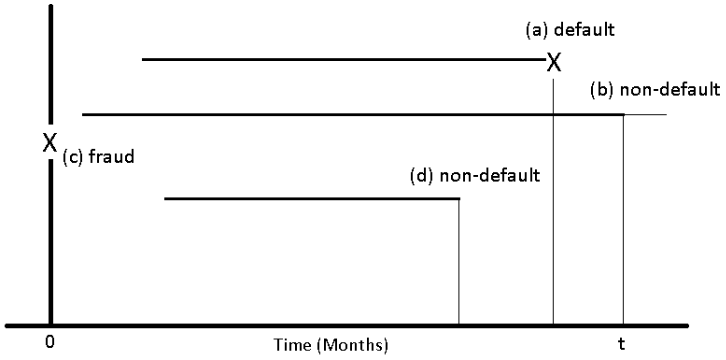

Figure 1: Loan survival time data.

According to the survival analysis terminology applied to credit risk analysis, the dataset here studied comprises three different types of non-negative survival times: survival times equal to zero from fraudsters, positive default times from defaulters, and the absence of registration, or censored times, from non-defaulted clients.

Thus, the aim of this paper is to propose a model that jointly accommodate three types of survival times broadly present in banking loan portfolios. The chosen framework to be enhanced is survival analysis, since it is already an established methodology for dealing with non-negative data with censored records.

2. Real data sets

This section presents a statistical summary of the data sets made available by a major Brazilian bank and exploited throughout this paper.

It is important to note that the presented data sets, amounts, rates and levels of the available covariates, do not necessarily represent the actual condition of the financial institution's customer base.

That means, despite being a real database, the bank may have sampled the data in order to change the current status of its loan portfolios.

2.1. Loan survival time data

Tables 1 and 2 present, respectively, the frequency and quantitative summary of a personal loan lifetime data, with a subdivision considering one available covariate.

For reasons of confidentiality, we will refer to the characteristic of the client by a covariate with x label. In this particular case, we can see below that x has three levels, referring to a specific characteristic of a bank customer profile.

We observe in the first table, the total number of customers in each level of x, the frequencies of loan lifetimes equal to zero by fraudsters, the frequency of observed defaults and the amount of censored data.

The second table shows a simple statistical description of the observed time to default, without considering the life time of the fraudsters, and censored as well.

Customers	Number of customers	Number of fraudsters	Number of defaulters	Number of censored
$x = 1$	1,626	137 (8.43%)	305 (18.76%)	1,184 (72.82%)
$x = 2$	1,574	127 (8.07%)	242 (15.37%)	1,205 (76.56%)
$x = 3$	938	30 (3.19%)	93 (9.92%)	815 (86.88%)
Total	4,138	294 (7.11%)	640 (15.46%)	3,204 (77.43%)

Table 1: Frequency of the bank loan lifetime data.

Customers	Mean	Median	Standard deviation	Interquartile range
$x = 1$	15.016	11.000	11.874	17
$x = 2$	15.640	12.500	12.134	16.75
$x = 3$	18.182	15.000	12.868	18
Total	15.712	12.000	12.148	17

Table 2: Summary of time-to-default event.

Figure 2 presents the Kaplan-Meier (K-M) survival curves, according to the segmentation given by the covariate x.

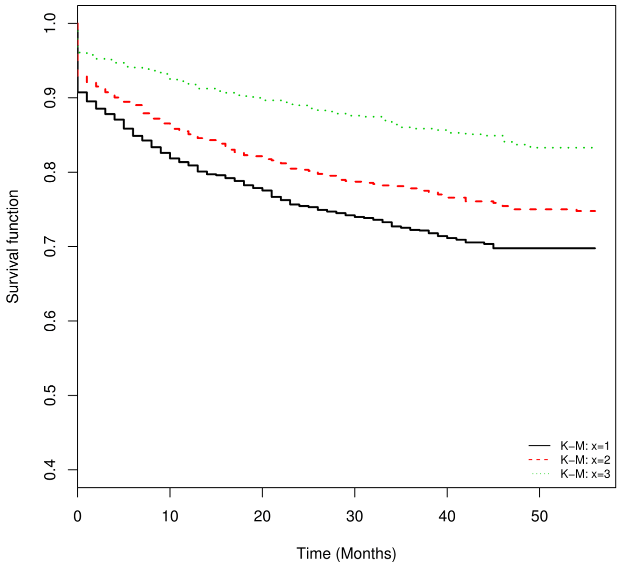

Figure 2: Kaplan-Meier survival according to customer profile: $x = 1$, $x = 2$ and $x = 3$.

The plateaus confirms an evidence of considerable presence of clients non-susceptible to the event of default. We also observe an increased cure rate as x increasingly ranges from $x = 1$ to $x = 3$, thereby, demonstrating that the $x = 3$ customer group is much less likely to default than the remaining group of customers.

We can see that the survival curve falls instantly from the zero time. This challenge we propose to deal with in this paper. Such behaviour, representing a concentration of zero inflated events, to the best of our knowledge, has not yet been incorporated into the long-term survival model framework.

3. Literature review

For using survival analysis techniques in credit risk settings, we must consider the modelling outcome of interest be the survival time after the loan concession, also mentioned as customer or loan survival time, which is represented by the time span to occurrence of an event of default.

In order to perform such an approach, survival data are generally modelled by a continuous probability distribution supported on the real non-negative interval $[0, +\infty)$.

This has been done in different papers, such as Banasik *et al.* [3], Stepanova & Thomas [32], Abad *et al.* [1], Bellotti & Crook [6], Tong *et al.* [33], Louzada *et al.* [20] and Barriga *et al.* [4].

The reason for the widespread use of survival analysis in credit risk rather than other modelling techniques, besides allowing monitor over time the loan portfolio credit risk, is that it can accommodate censored data, which is not supported, for example, in credit scoring techniques purely based on good and bad client classification, see for instance Hand & Henley [14], Abreu [2], Louzada-Neto [21] and Lessmann *et al.* [17].

Notwithstanding survival analysis deals with non-negative and censored data, however, generally without excess, or even, the presence of zeros.

Unlike survival data analysis, in other areas we can observe most commonly the existence of non-negative data with presence of zeros, sometimes with excess. Usually, it occurs frequently in count data analysis, see for instance Lambert [16], Barry & Welsh [5], Lord *et al.* [19], Conceição *et al.* [12].

In Vieira *et al.* [34] and Ospina & Ferrari [25], the authors deal with zero-inflated proportion data models. Therefore, it is already a commonplace the expression "zero-inflated data".

In Liu *et al.* [18], the occurrence of zeros excess is exploited within two longitudinal medical follow-ups. In the first one, a SIDA study, the zero data comes from records of non-recurrence of opportunistic diseases, while in the second study, zero data are recorded as the number of non-recurrent tumours in a soft tissue sarcoma study.

Zero-inflated data also appears in the context of left censored data. In Blackwood [8], for example, left censored data are generate in experiments related to the presence of toxic products in the environment. Due to inaccuracy of the tools used for measurement, it is not always possible to fully observe some results and only an upper limit is recorded.

Also dealing with the presence of left censored data, Braekers & Grouwels [9] reviewed a laboratory experiment with mice conducted by Markel *et al.* [22], where the outcome of interest is the induced sleep time measured after ingestion of a dose of ethanol.

As some mice present immunity for the administered dose of ethanol, the analysed data set contains a proportion of sleep time zero.

In the statistical approach proposed to re-analyse the data obtained in the earlier conducted experiment, i.e., in order to reinvestigate the influence of covariates on the outcome of interest, Braekers & Grouwels [9] proposed a logistic regression model for the probability of a zero outcome value and the Cox regression model for the non-zero outcomes.

Perhaps it is unhelpful, or cruelly insensitive, if we consider human survival times equal to zero in clinical trials or medical studies. Hence, it might be why, to the best of our knowledge, we have not found study that is willing to account for zero-inflated data in the medical specialized literature and that aims to analyse human patient survival time.

However, the same sense of respect expended with clinical trials, in some way, does not seem to be the same required when we deal with credit risk events. On the contrary, information about zero-inflated time should be taken into account in credit risk analysis, and must be useful for identifying clients who apply for loans only for the purpose of defrauding the financial institution by, since the beginning, not honouring its obligations under the credit granted.

3.1. Preliminary

In survival analysis, the random variable T of interest is the time span until the occurrence of an expected event. Depending on the context in which it appears, T might be called by lifetime or failure time.

In industry it is customarily associated with the time up to failure of a machine. In the medical area, for example, can be associated with the span time to recurrence of a disease under treatment, or even the death of a patient.

The focus of interest in credit risk setting is the failure time related to the span time up to the occurrence of a loan default. Obviously, in all cases T is non-negative and, generally, is treated as a continuous random variable.

According to Colosimo & Giolo [11] and Rinne [29], there are several functions which completely specify the distribution of a random variable in survival analysis, since they are mathematically equivalent functions.

They are the probability density function (PDF), cumulative distribution function (CDF), the complementary cumulative distribution function (CCDF), the hazard rate, the cumulative hazard rate and, finally, the mean residual life function.

Within a survival analysis context, the complementary cumulative distribution function (CCDF) is known as survival function and, commonly, is denoted by $S(\cdot)$.

The downside of considering the standard survival analysis in credit risk is the mathematical fact that the survival function is a proper survival function, i.e., goes to zero as time progresses indefinitely. That means, the survival function, $S(t) = P(T > t)$, satisfies:

$$\lim_{t \to \infty} S(t) = 0 \tag{1}$$

Unlike what happens in many real situations, in this standard framework is not contemplated the presence of immunity to the effects that lead to the occurrence of the concerned event.

Indeed, returning to examples in medical field, there are patients suffering from disease who, once submitted to treatment, they recover completely. They are known as cured or long-term survivors.

Similarly, in credit risk studies on loan portfolios of financial institutions, most customers never experience the condition of being delinquent. In this financial context, they are also known as non-defaulting clients or long-term clients.

Therefore, when it is needed consider the presence of cure or long-term data, the traditional survival analysis is not at all suitable for modelling failure time. In those cases, where there are immunity to the occurrence of failures, new statistical tools have been proposed.

To handle this aforementioned challenge, Berkson & Gage [7] proposed a simple way that added the fraction of cured ($p > 0$) into the survival function.

The authors have introduced the following survival expression based on two sub-populations of individuals susceptible and non-susceptible to the occurrence of the event of interest:

$$S(t) = p + (1 - p)S^*(t), \qquad t \geq 0, \tag{2}$$

where S^* is the survival function of the individuals susceptible to failure and $p > 0$ is the proportion of the individuals immune to failure (cured).

This model is called cure rate model or long-term survival model.

Unlike S^*, S is an improper survival function, since it satisfies:

$$\lim_{t\to\infty} S(t) = p > 0. \tag{3}$$

The advantage of the cure rate model, according to Othus *et al.* [26], among others authors, is that it allows associate covariates in both parts of the model.

Indeed, it allows covariates to have different influence on cured patients, linking covariates with p, and on patients who are not cured, i.e., susceptible to the event, linking covariates with the parameters of the proper survival function S^*.

3.2. Proposal

We are at the point of presenting the gap and the contribution of our proposed model to the literature on survival analysis techniques.

To the best of our knowledge, there is no literature considering a cure rate model that accounts for the excess of individuals who have already experimented the event of interest at the beginning of the considered study, i.e., with survival time equal to zero.

In this sense, and focusing on the portfolio credit risk context, we define the following proportions to be accommodated in the new proposed model:

- γ_0: the proportion of zero-inflated times, i.e., related to fraudsters;
- γ_1: the proportion of immune to failure, i.e., related to non-defaulters (cured).

Thus, we have the following expression for the improper survival function of a dataset comprised by all possible loan survival times:

$$S(t) = \gamma_1 + (1 - \gamma_0 - \gamma_1)S^*(t), \qquad t \geq 0, \tag{4}$$

where S^* is the survival function related to the $(1 - \gamma_0 - \gamma_1)$ proportion of subjects susceptible to failure, γ_1 is the proportion of subjects immune to failure (cured or non-defaulted) and finally, γ_0 is the proportion of individuals fraudsters (with survival time equal to zero). This model in 4 is called zero-inflated cure rate model.

The important fact that differentiates the inflated cure rate version from the standard cure rate approach in 2, once they share the fact that both are based on improper survival functions, is expressed in the second of the following satisfied properties:

$$\lim_{t\to\infty} S(t) = \gamma_1 > 0, \tag{5}$$

$$S(0) = 1 - \gamma_0 < 1. \tag{6}$$

The above properties can be viewed in the particular trait of the proposed survival function expression:

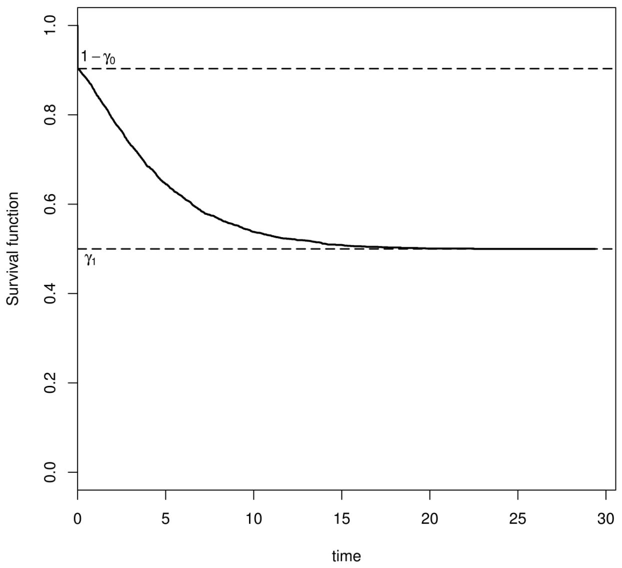

Figure 3: The (improper) survival function of the zero-inflated log-term survival data.

Note that, if $\gamma_0 = 0$, i.e., without the excess of zeros, we have the cure rate model of Berkson & Gage [7].

3.3. Justification

In this paper, we have justified the need for the zero-inflated cure rate model based on a credit risk setting. The purposes is to deal with the problem of assessing the propensity to defraud in loan applications in terms of estimating fraud rate in banking loan portfolios, according to the characteristics of all borrower clients.

To reach this goal, we proposed a jointly modelling of zero-inflated time in loan survival data with cure rate, where we link together covariate in all parts of the proposed model.

Therefore, this is why we propose a methodology based on an enhanced cure rate approach, i.e., to account for zero excess in a long-term survival environment.

To exemplify the application of the proposed approach, we analyse a portfolio of loans made available by a large Brazilian commercial bank. In order to check the propensity of a customer be a fraudster based on personal characteristics, we propose a methodology based on a regression model framework.

3.4. Organization

The remainder of this paper is organized as follows. In Section 4, we formulate the model and present the approach for parameter estimation. A study based on Monte Carlo simulations with a variety of parameters is presented in Section 5. An application to a real data set of a Brazilian bank loan portfolio is presented in Section 6. Some general remarks are presented in Section 7.

4. Model specification

In what follows, we consider the zero-inflated cure rate model as defined in equation 4. The associated (improper) cumulative distribution function (CDF) and failure density function (PDF) are given by:

$$F(t) = \gamma_0 + (1 - \gamma_0 - \gamma_1)F^*(t), \qquad t \geq 0, \tag{7}$$

$$f(t) = \begin{cases} \gamma_0, & \text{if} \quad t = 0, \\ (1 - \gamma_0 - \gamma_1)f^*(t), & \text{if} \quad 0 < t, \end{cases} \tag{8}$$

where the parameters γ_0 and γ_1 are as defined in Section 3.2.

F^* and f^* are, respectively, the cumulative distribution function and probability density function underpinning the $(1 - \gamma_0 - \gamma_1)$ proportion of subject susceptible to failure.

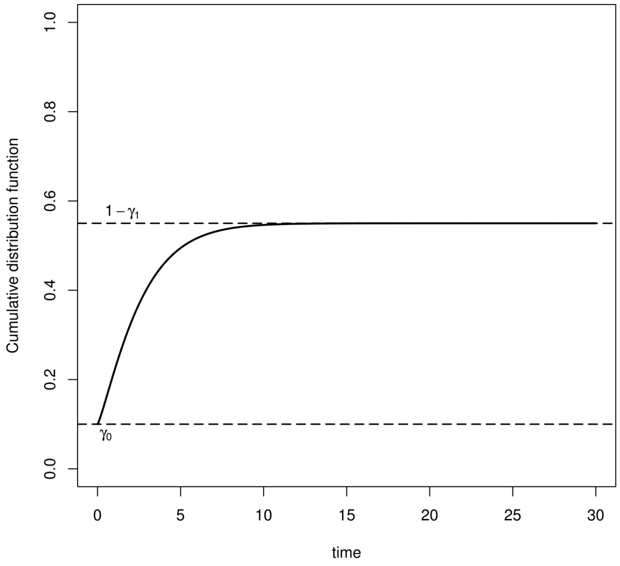

Figure 4: The (improper) cumulative distribution function (CDF) of the zero-inflated log-term survival data.

Note that, the improper CDF of the zero-inflated cure rate model, $F(t)$, has the advantageous property of accommodating the excess of zeros, γ_0, since it satisfies:

$$F(0) = \gamma_0$$

Moreover, it does not undermine the fraction of cured, γ_1, since it also satisfies:

$$\lim_{t\to\infty} F(t) = 1 - \gamma_1.$$

4.1. The zero-inflated Weibull cure rate model

In this section, we associate the Weibull distribution as the probability density function for the subjects susceptible to failure. We choose the Weibull function since it has been widely used to model survival data, and also has served as motivation for the proposal of various types of generalizations, see for example, Cooner *et al.* [13], Rinne [29], Rodrigues *et al.* [31], Ortega *et al.* [24] and Cancho *et al.* [10].

Then, let the Weibull distribution represents the survival behaviour of the non-negative random variable T^*, which denotes the time-to-default for the susceptible subjects. The CDF of the Weibull distribution is given by

$$F_w^*(t) = 1 - e^{-\left(\frac{t}{\lambda}\right)^\alpha}, \qquad t \geq 0, \tag{9}$$

where $\alpha > 0$ and $\lambda > 0$ are, respectively, shape and scale parameters. The PDF of the Weibull distribution is obtained from the equation 9 as

$$f_w^*(t) = \frac{d}{dt}F_w^*(t) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} e^{\left(-\frac{t}{\lambda}\right)^\alpha}, \qquad t \geq 0. \tag{10}$$

The zero-inflated cure rate model proposes to distinguish between three sub-populations of banking borrowers:

1. a segment of those who will not honour any instalment of the loan, i.e., fraudsters with failure time zero;
2. a segment of those are susceptible to default;
3. a segment of those who are not susceptible to default, i.e., cured (non-defaulters or long-term survivors).

Consequently, as in the standard cure rate modelling, there are two possibilities for the customer who is not a fraudster:

- information about the default time is right observed, that is, the client has defaulted during the monitoring of the loan; ;
- information about the default time is right censored, that is, the client will probably become a defaulter if given enough time, or the client is really a good payer and will never default, regardless of the monitoring period term.

4.2. Likelihood function

For the likelihood contribution of a client i, we should pay attention to fact that there are different sub-group of clients. Therefore, the likelihood contribution of each client i for the zero-inflated cure rate model, obtained from 4 and all considerations we have done above, must assume three different values:

$$\begin{cases} \delta_0, & \text{if} \quad i \text{ is a fraudster,} \\ (1-\gamma_0-\gamma_1)f^*(t_i), & \text{if} \quad i \text{ is not censored,} \\ \gamma_1+(1-\gamma_0-\gamma_1)S^*(t_i), & \text{if} \quad i \text{ is censored,} \end{cases} \tag{11}$$

Let the data take the form $\mathcal{D} = \{t_i, \delta_i\}$, where $\delta_i = 1$ if t_i is a observable time to default, and $\delta_i = 0$ if it is right censored, for $i = 1, 2, \cdots n$. Let (α, λ) denote the parameter vector of the Weibull distribution and, finally, (γ_0, γ_1) be the parameters associated, respectively, with the proportion of fraudsters (inflation of zeros) and the proportion of long-term survivors (cure rate).

The likelihood function of the zero-inflated Weibull cure rate model, with a four-parameter vector $\vartheta = (\alpha, \lambda, \gamma_0, \gamma_1)$, is based on a sample of n observations, $\mathcal{D} = \{t_i, \delta_i\}$. Then, following Klein & Moeschberger [15], we can write the likelihood function under non-informative censoring as:

$$L(\vartheta; \mathcal{D}) \propto \prod_{t_i=0} \{\gamma_0\} \prod_{t_i>0} \left\{ [(1-\gamma_0-\gamma_1)f_w^*(t_i)]^{\delta_i} [\gamma_1+(1-\gamma_0-\gamma_1)S_w^*(t_i)]^{1-\delta_i} \right\}. \tag{12}$$

The log-likelihood function for $\vartheta = (\alpha, \lambda, \gamma_0, \gamma_1)$, corresponding to the observed data is given by

$$\begin{aligned} \log\{L(\vartheta;\mathcal{D})\} \quad = \quad & const + \sum_{t_i=0} \log\{\gamma_0\} + \sum_{t_i>0} \log\left\{ [(1-\gamma_0-\gamma_1)f_w^*(t_i)]^{\delta_i} \right\} \\ & + \sum_{t_i>0} \log\left\{ [\gamma_1+(1-\gamma_0-\gamma_1)S^*(t_i)]^{1-\delta_i} \right\} \\ = \quad & const + \sum_{t_i=0} \log\{\gamma_0\} + \delta_i \sum_{t_i>0} \log\{1-\gamma_0-\gamma_1\} \\ & + \delta_i \sum_{t_i>0} \log\{f_w^*(t_i)\} + (1-\delta_i) \sum_{t_i>0} \log\{\gamma_1+(1-\gamma_0-\gamma_1)S^*(t_i)\} \end{aligned}$$

4.3. The model regression version

Here, we introduce a way to link covariates with the parameters set in the zero-inflated Weibull cure rate model. This modelling allows us to determine the effect of covariates all at once on the zero-inflated times, on the cure rate and on the failure times.

Therefore, we propose to relate the set of three parameters $\{\gamma_0, \gamma_1, \lambda\}$, respectively, proportion of zeros, proportion of cured and scale parameter of the Weibull distribution, with a set of three-covariate vectors $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$.

These covariate vectors, as occurs in practice, may be the same, i.e., $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3$.

Following the setting made in Pereira *et al.* [27], p. 128, the regression version of the zero-inflated Weibull cure rate model is defined by 4 up to 11, and by the following link functions:

$$\begin{cases} \gamma_{0i} &= \frac{e^{\mathbf{x}_{1i}^{\top}\beta_1}}{1+e^{\mathbf{x}_{1i}^{\top}\beta_1}+e^{\mathbf{x}_{2i}^{\top}\beta_2}}, \\ \gamma_{1i} &= \frac{e^{\mathbf{x}_{2i}^{\top}\beta_2}}{1+e^{\mathbf{x}_{1i}^{\top}\beta_1}+e^{\mathbf{x}_{2i}^{\top}\beta_2}}, \\ \lambda_i &= e^{\mathbf{x}_{3i}^{\top}\beta_3}, \end{cases} \tag{13}$$

where β_j's are three vectors of regression coefficients to be estimated.

Note that, as required, the component link functions ensure that $0 < \gamma_{0i}, \gamma_{1i} < 1$, $0 < 1 - \gamma_{0i} - \gamma_{1i} < 1$ and $\lambda_i > 0$ are always satisfied.

4.4. Parameter estimation

Parameter estimation is performed by straightforward use of maximum likelihood estimation (MLE), where, as we will see, its simple application is supported by our simulation studies. Hence, the maximum likelihood estimator $\hat{\vartheta}$, regarding the parameter vector ϑ, both considering the model with or without covariates, are obtained through maximization of $L(\vartheta; \mathcal{D})$ or $\ell(\vartheta; \mathcal{D}) = \log\{L(\vartheta; \mathcal{D})\}$.

According to Migon *et al.* [23], under suitable regularity conditions, the asymptotic distribution of the maximum likelihood estimator, $\hat{\vartheta}$, is a multivariate normal with mean vector ϑ and covariance matrix $\mathbf{I}^{-1}(\hat{\vartheta})$, where $\mathbf{I}(\vartheta)$ is the Fisher information. Since it is not possible to compute the Fisher information matrix $\mathbf{I}(\vartheta)$, due to the censored observations, instead it is possible to use the matrix of second derivatives of the log likelihood, what are computed numerically. Then, the Fisher information matrix can be estimated by the observed information matrix $\mathbf{I}(\vartheta) = \{-\partial^2\ell(\vartheta)/\partial\vartheta\partial\vartheta^T\}^{-1}$, evaluated at $\vartheta = \hat{\vartheta}$.

Thereafter, let I^{ii} be the *ith* diagonal element of the inverse of $\mathbf{I}$, evaluated in $\hat{\vartheta}$. An approximate $100(1-\alpha)\%$ confidence interval for $\hat{\vartheta}_i$, based on assumed regularity conditions, is given by $\left(\hat{\vartheta}_i - z_{\alpha/2}\sqrt{I^{ii}},\ \ \hat{\vartheta}_i + z_{\alpha/2}\sqrt{I^{ii}}\right)$, where z_α denotes the $100(1-\alpha)$ percentile of the standard normal random variable. In the application section we set $\alpha = 0.05$, where we get a 95% confidence interval for each ML estimation.

There are various software and routines available to approximate the parameter estimate and confidence interval described above. We choose the method "BFGS", see details in R Core Team [28], which comes within the **R** routine `optim`.

5. Simulation studies

We proceed a parameter estimation based on a maximum likelihood principle and use the method of maximization "BFGS" of the R routine optim() for that.

In order to check the behaviour of the asymptotic theory for increasing sample size, we performed a simulation study for examining the coverage probabilities of the 95% confidence intervals for the MLEs. The simulation study also provides the results for root mean square errors, bias and standard deviations for the estimated parameters, to ensure that they decrease as expected with increasing sample size.

The simulation study is based on 100 sample replications, where the sample size increases according to the nature of the real data sets in which the model has been applied in this paper. So, we perform Monte Carlo simulations where the sample size varies as $n = 250$, 500, 1000 and 2000.

Three simulation studies are performed for the proposed zero-inflated Weibull cure rate regression model, which is introduced by the improper survival function in 4 and the following link functions in 14 below.

For this purpose of simulation, we let x be a random variable that represents a consumer characteristic. Hence, the configuration of parameters on a single covariate x is replaced by the following expression:

$$\begin{cases} \gamma_{0i} &= \frac{e^{\beta_{10}+x_i\beta_{11}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}, \\ \gamma_{1i} &= \frac{e^{\beta_{20}+x_i\beta_{21}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}, \\ \alpha &= e^{\beta_3}, \\ \lambda_i &= e^{\beta_{40}+x_i\beta_{41}}, \end{cases} \tag{14}$$

Note that we do not link covariates to the shape Weibull parameter (α). However, note that we made the following re-parametrization $\alpha = e^{\beta_3}$, in order to facilitate the process of estimation of the parameter within the routine optim().

The description of sample generation, i.e., all details of the simulated survival time distribution, and results obtained regarding to the proposed estimation method for the model parameters are described in the next sections.

5.1. Simulation algorithm

Suppose that the time of occurrence of an event of interest has the improper cumulative distribution function $F(t)$ given by 7, i.e.:

$$F(t) = \gamma_0 + (1 - \gamma_0 - \gamma_1)F^*(t), \qquad t \geq 0.$$

We aim to simulate random samples of size n posing as loan survival times, where each sample comprises a proportion γ_0 of zero-inflated times, a cure fraction of γ_1 and with a proportion $(1 - \gamma_0 - \gamma_1)$ of failures times drawn from a Weibull distribution with α and λ parameters.

The following step-by-step algorithm is proposed for this purpose, which is based on the link functions 14, with a x covariate drawn from a Bernoulli distribution with parameter 0.5, representing a consumer feature.

1. Set β_{10} and β_{11} related to the value of the desired proportion of zero-inflated times, γ_0, along with β_{20} and β_{21} related to the value of the desired cure fraction, γ_1; finally, set the Weibull parameters β_3 related to α, as well as, β_{40} and β_{41} related to λ;
2. Drawn x_i from $x \sim$ Bernoulli(0.5) and calculate γ_{0i}, γ_{1i}, α and λ_i as in 14;

3. Generate u_i from a uniform distribution U(0,1);
4. If $u_i \leq \gamma_{0i}$, set $s_i = 0$;
5. If $u_i > 1 - \gamma_{1i}$, set $s_i = \infty$;
6. If $\gamma_{0i} < u_i \leq 1 - \gamma_{1i}$, generate v_i from a uniform distribution $\text{U}(\gamma_{0i}, 1 - \gamma_{1i})$ and take s_i as the root of $F(t) - v_i = 0$, where $F(t)$ is given as in 7;
7. Generate w_i from a uniform $\text{U}(0, max(s_i))$, considering only finites s_i;
8. Calculate $t_i = min(s_i, w_i)$, if $t_i < w_i$, set $\delta_i = 1$, otherwise, set $\delta_i = 0$.
9. Repeat as necessary from step 2 until you get the desired amount of sample (t_i, δ_i).

Note that the censoring distribution chosen is a uniform distribution with limited range in order to keep the censoring rates reasonable, see Rocha *et al.* [30], p.12.

5.2. Parameter scenarios

Considering the parameters established in the regression model defined by 14, we set three different scenarios of parameters for the simulation studies here performed. Playing the role of covariate, we assume x as a binary covariate with values drawn from a Bernoulli distribution with parameter 0.5.

For scenario 1, β_{10} assumes -3 and β_{11} assumes 1. β_{20} assumes -2.5 and β_{21} assumes 0.3. Given the average value of x is 0.5, we have that γ_0 assumes on average a value of 0.0697, and γ_1 assumes on average a value of 0.0809. Compared to the other scenarios 2 and 3, scenario 1 has the characteristic of having a **low rate of fraudsters and cured**, respectively, 6,97% and 8,09%. Regarding to the Weibull parameters, β_3 assumes 0.5, β_{40} assumes 1.5 and β_{41} assumes 2. This implies that the Weibull parameters α and λ in average are, respectively, equal to 1.64 and 12.18. In this scenario 1, the mean and standard deviation of the defaulted time are respectively equal to 10.89 and 6.81.

For scenario 2, β_{10} assumes -2 and β_{11} assumes 2. β_{20} assumes -1.5 and β_{21} assumes 1.5. Given the average value of x is 0.5, we have that γ_0 assumes on average a value of 0.1999, and γ_1 assumes on average a value of 0.256. Compared to the other scenarios 1 and 3, scenario 2 has the characteristic of having a **moderate rate of fraudsters and cured**, respectively, 19.99% and 25.6%. Regarding to the Weibull parameters, β_3 assumes 1.5, β_{40} assumes -0.5 and β_{41} assumes 3. This implies the Weibull parameters α and λ in average are, respectively, equal to 4.48 and 2.71. In this scenario 2, the mean and standard deviation of the defaulted time are respectively equal to 2.47 and 0.62.

Finally, for scenario 3, β_{10} assumes -0.5 and β_{11} assumes 0.75. β_{20} assumes -0.35 and β_{21} assumes 1.75. Given the average value of x is 0.5, we have that γ_0 assumes on average a value of 0.2469, and γ_1 assumes on average a value of 0.4731. Compared to the other scenarios 1 and 2, scenario 3 has the characteristic of having a **high rate of fraudsters and cured**, respectively, 24.69% and 47.31%. Regarding to the Weibull parameters, β_3 assumes 2.5, β_{40} assumes 1.25 and β_{41} assumes 3.5. This implies that the Weibull parameters α and λ in average are, respectively, equal to 12.12 and 20.08. In this scenario 3, the mean and standard deviation of the defaulted time are respectively equal to 19.24 and 1.93.

The following Kaplan-Meier plots feature the survival distinction between the three scenarios set for the regression parameters, trying to simulate, not at all, a range of scenarios consistent with a probable current condition of real loan portfolios.

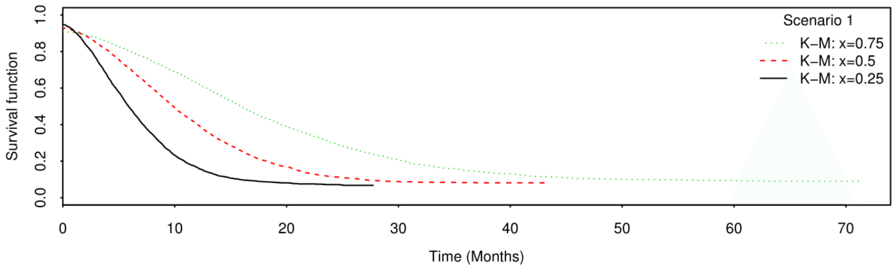

Figure 5: Kaplan-Meier (K-M) survival curves according parameter scenario 1.

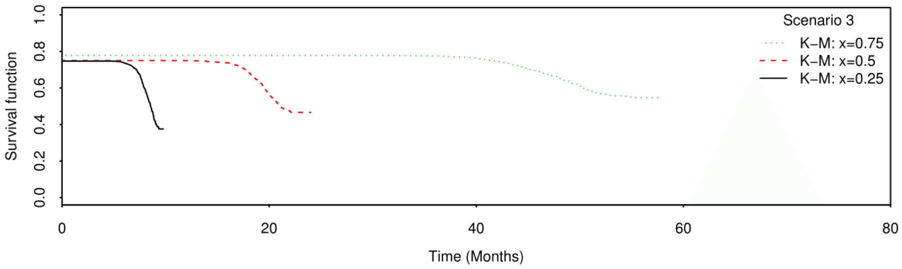

Figure 6: Kaplan-Meier (K-M) survival curves according parameter scenario 2.

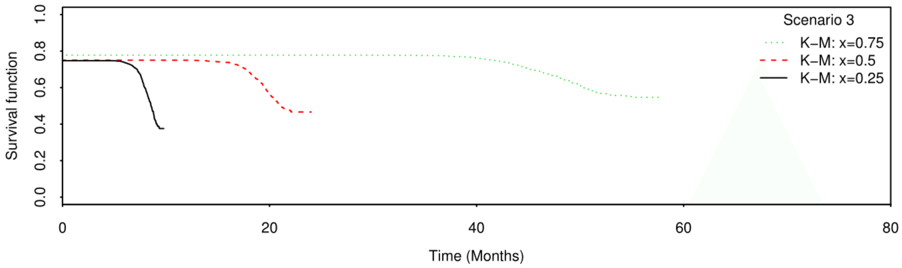

Figure 7: Kaplan-Meier (K-M) survival curves according parameter scenario 3.

The following histograms, as well, feature the data distribution distinction between the three scenarios set for the regression parameters.

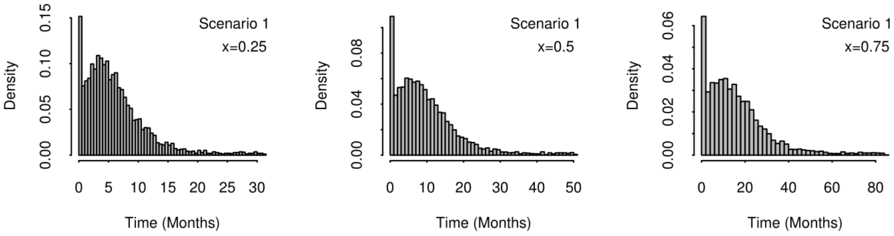

Figure 8: Histogram of loan survival data according parameter scenario 1.

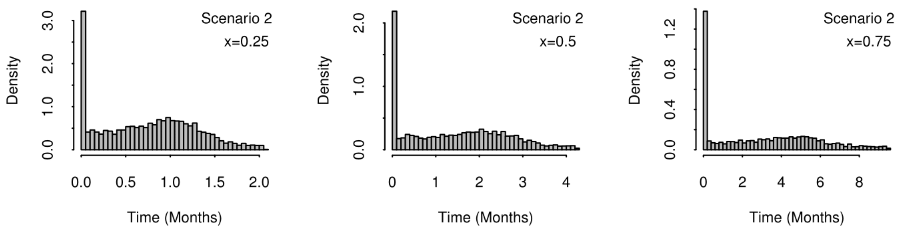

Figure 9: Histogram of loan survival data according parameter scenario 2.

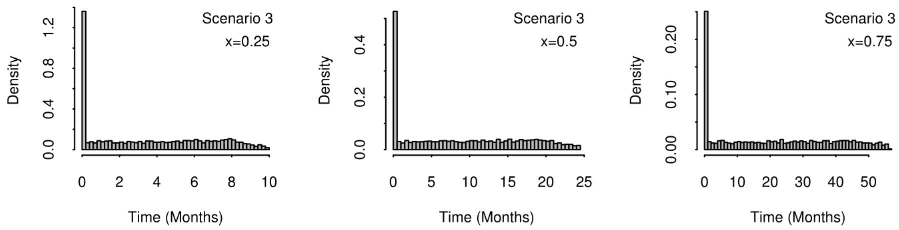

Figure 10: Histogram of loan survival data according parameter scenario 3.

5.3. Results of Monte Carlo simulations

The followings tables and figures describe the simulation results for the three simulated scenarios of parameters. As described in the sections above, the model parameters are linked on a single covariate x, according to the following expression:

$$\begin{cases} \gamma_{0i} &= \frac{e^{\beta_{10}+x_i\beta_{11}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}, \\ \gamma_{1i} &= \frac{e^{\beta_{20}+x_i\beta_{21}}}{1+e^{\beta_{10}+x_i\beta_{11}}+e^{\beta_{20}+x_i\beta_{21}}}, \\ \alpha &= e^{\beta_3}, \\ \lambda_i &= e^{\beta_{40}+x_i\beta_{41}}. \end{cases} \tag{15}$$

The parameter values are selected in order to assess the estimation method performance under different scale parameters (related to the Weibull time-to-default distribution), and also under a composition of different proportions of fraudsters and non-defaulters (cured).

It can be seen from the Tables that:

1. in general, the maximum likelihood estimation in average, MLEA, is close to the parameters set in the simulated parameter scenarios;
2. in general, the root mean square error, biases and standard deviations decrease as sample size increases;
3. in the scenarios with the greatest presence of fraudsters and cured, i.e., scenario 2 (Moderate) and 3 (High), the MLEA, and the measures of RMSE, Bias and SD of the regression parameters related to γ_0 and γ_1, performs better compared to scenario 1 (Low), due, of course, to greater presence of fraudsters and censored data.

n (CD)	Parameter	MLEA	SD	RMSE	Bias	CP
250 (24.07%)	β_{10}	-3.121	0.571	0.581	-0.121	0.96
500 (22.11%)	[-3.00]	-3.027	0.381	0.380	-0.027	0.92
1000 (21.19%)		-3.051	0.255	0.259	-0.051	0.94
2000 (20.07%)		-2.980	0.149	0.150	0.020	0.95
250	β_{11}	1.143	0.625	0.638	0.143	0.96
500	[1.00]	0.999	0.453	0.451	-0.061	0.92
1000		1.055	0.286	0.290	0.055	0.94
2000		0.964	0.178	0.181	-0.036	0.95
250	β_{20}	-2.539	0.408	0.408	-0.039	0.99
500	[-2.5]	-2.515	0.278	0.277	-0.015	0.95
1000		-2.493	0.209	0.208	0.007	0.93
2000		-2.506	0.126	0.126	-0.006	0.93
250	β_{21}	-0.148	1.712	1.761	-0.448	0.98
500	[0.3]	0.127	0.827	0.841	-0.173	0.99
1000		0.299	0.417	0.415	-0.001	0.92
2000		0.264	0.242	0.243	-0.036	0.97
250	β_3	0.500	0.053	0.053	0.001	0.96
500	[0.5]	0.496	0.044	0.044	-0.004	0.96
1000		0.501	0.032	0.032	0.001	0.94
2000		0.500	0.022	0.022	0.000	0.93
250	β_{40}	1.499	0.066	0.066	-0.001	0.94
500	[1.5]	1.496	0.042	0.042	-0.003	0.95
1000		1.503	0.034	0.034	0.003	0.90
2000		1.501	0.021	0.021	0.001	0.95
250	β_{41}	1.997	0.141	0.140	-0.003	0.89
500	[2.0]	2.001	0.084	0.083	0.001	0.95
1000		1.987	0.060	0.061	-0.013	0.91
2000		2.004	0.034	0.034	0.003	0.98

Table 3: Maximum likelihood estimation in average, standard deviation, square root of mean squared error, Bias and coverage probability of the parameters of zero-inflated Weibull cure rate regression model for simulated data under the first scenarios of parameters, obtained from Monte Carlo simulations with 100 replications, increasing sample size (n) and 21.86% in average of censored data (CD).

n (CD)	Parameter	MLEA	SD	RMSE	Bias	CP
250 (38.40%)	β_{10}	-2.053	0.373	0.375	-0.053	0.93
500 (36.45%)	[-2.0]	-2.032	0.217	0.218	-0.031	0.96
1000(35.89%)		-2.010	0.169	0.168	-0.010	0.95
2000(36.05%)		-2.004	0.109	0.109	-0.003	0.94
250	β_{11}	2.005	0.557	0.555	0.004	0.91
500	[2.0]	2.024	0.304	0.304	0.023	0.96
1000		2.017	0.234	0.233	0.017	0.96
2000		2.015	0.147	0.147	0.015	0.96
250	β_{20}	-1.524	0.256	0.256	-0.024	0.97
500	[-1.5]	-1.476	0.158	0.159	0.023	0.93
1000		-1.528	0.101	0.104	-0.027	0.98
2000		-1.505	0.088	0.088	-0.004	0.94
250	β_{21}	1.261	1.272	1.288	-0.239	0.96
500	[1.5]	1.362	0.611	0.624	-0.137	0.96
1000		1.547	0.267	0.270	0.046	0.93
2000		1.506	0.181	0.180	0.005	0.95
250	β_3	1.518	0.074	0.075	0.017	0.97
500	[1.5]	1.503	0.061	0.061	0.002	0.89
1000		1.512	0.037	0.039	0.001	0.94
2000		1.504	0.029	0.029	0.003	0.92
250	β_{40}	-0.502	0.028	0.028	-0.002	0.92
500	[-0.5]	-0.502	0.018	0.018	-0.001	0.94
1000		-0.500	0.013	0.013	-0.0001	0.92
2000		-0.500	0.008	0.008	-0.0001	0.97
250	β_{41}	3.004	0.115	0.115	0.004	0.90
500	[3.0]	3.002	0.080	0.080	0.001	0.95
1000		3.000	0.047	0.047	-0.0003	0.96
2000		2.998	0.032	0.032	-0.001	0.93

Table 4: Maximum likelihood estimation in average, standard deviation, square root of mean squared error, Bias and coverage probability of the parameters of zero-inflated Weibull cure rate regression model for simulated data under the second scenarios of parameters, obtained from Monte Carlo simulations with 100 replications, increasing sample size (n) and 36.69% in average of censored data (CD).

n (CD)	Parameter	MLEA	SD	RMSE	Bias	CP
250 (51.54%)	β_{10}	-0.538	0.212	0.215	-0.038	0.94
500 (53.18%)	[-0.5]	-0.486	0.166	0.165	0.014	0.92
1000(55.18%)		-0.493	0.118	0.118	0.006	0.96
2000(52.90%)		-0.508	0.081	0.081	-0.007	0.92
250	β_{11}	0.733	0.970	0.965	-0.016	0.87
500	[0.75]	0.654	0.768	0.770	-0.096	0.92
1000		0.737	0.413	0.411	-0.013	0.95
2000		0.784	0.252	0.253	0.034	0.96
250	β_{20}	-0.372	0.208	0.209	-0.022	0.91
500	[-0.35]	-0.341	0.150	0.149	0.008	0.92
1000		-0.323	0.093	0.096	0.026	0.98
2000		-0.357	0.076	0.076	-0.006	0.94
250	β_{21}	1.223	2.036	2.094	-0.526	0.96
500	[1.75]	1.409	1.476	1.507	-0.341	0.97
1000		1.677	0.763	0.763	-0.073	0.97
2000		1.768	0.309	0.308	0.018	0.97
250	β_3	2.537	0.109	0.114	0.037	0.92
500	[2.5]	2.514	0.078	0.079	0.0143	0.93
1000		2.512	0.060	0.061	0.0124	0.89
2000		2.502	0.031	0.031	0.0019	0.98
250	β_{40}	1.249	0.010	0.010	-0.0004	0.97
500	[1.25]	1.248	0.008	0.008	-0.0016	0.94
1000		1.250	0.006	0.006	0.0004	0.97
2000		1.250	0.004	0.004	0.0001	0.93
250	β_{41}	3.576	0.305	0.313	0.076	0.78
500	[3.5]	3.507	0.081	0.081	0.007	0.88
1000		3.501	0.050	0.049	0.0005	0.91
2000		3.503	0.024	0.024	0.0025	0.98

Table 5: Maximum likelihood estimation in average, standard deviation, square root of mean squared error, Bias and coverage probability of the parameters of zero-inflated Weibull cure rate regression model for simulated data under the third scenarios of parameters, obtained from Monte Carlo simulations with 100 replications, increasing sample size (n) and 53.56% in average of censored data (CD).

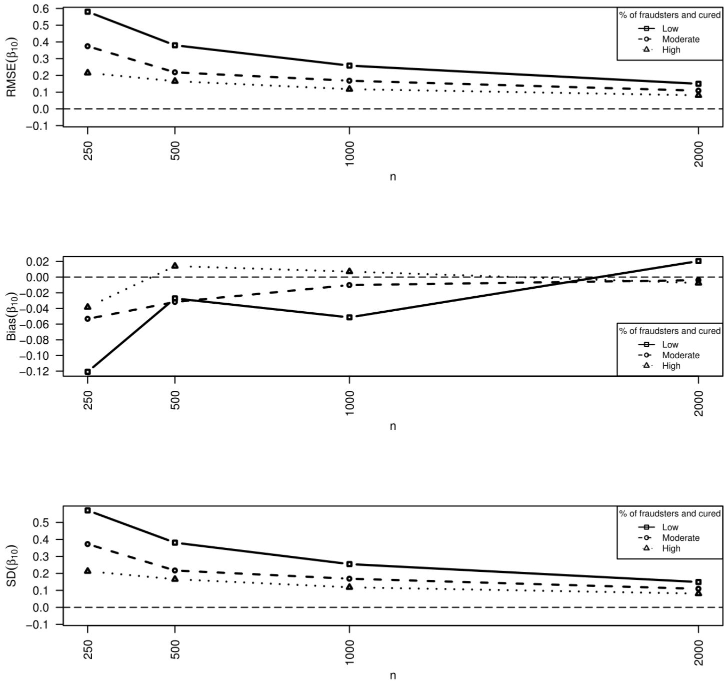

Figure 11: Root mean square error, bias and standard deviation for β_{10}.

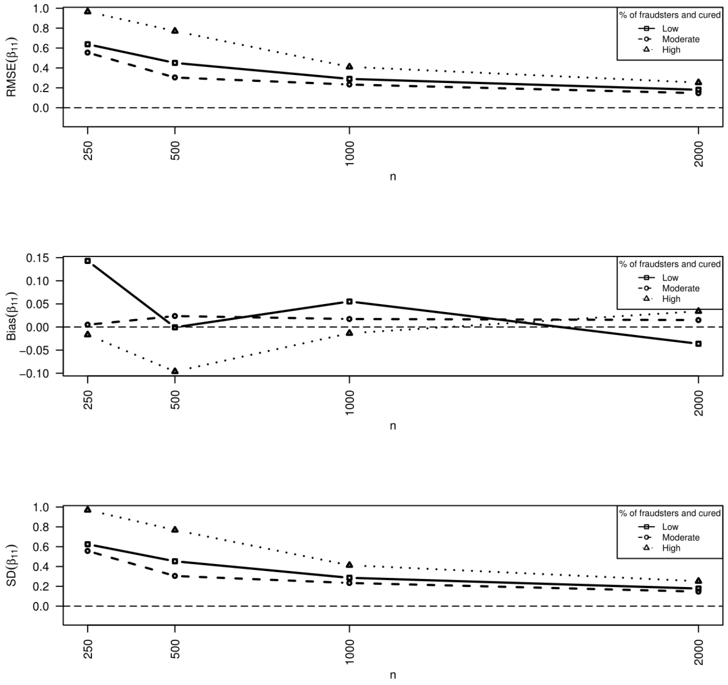

Figure 12: Root mean square error, bias and standard deviation for β_{11}.

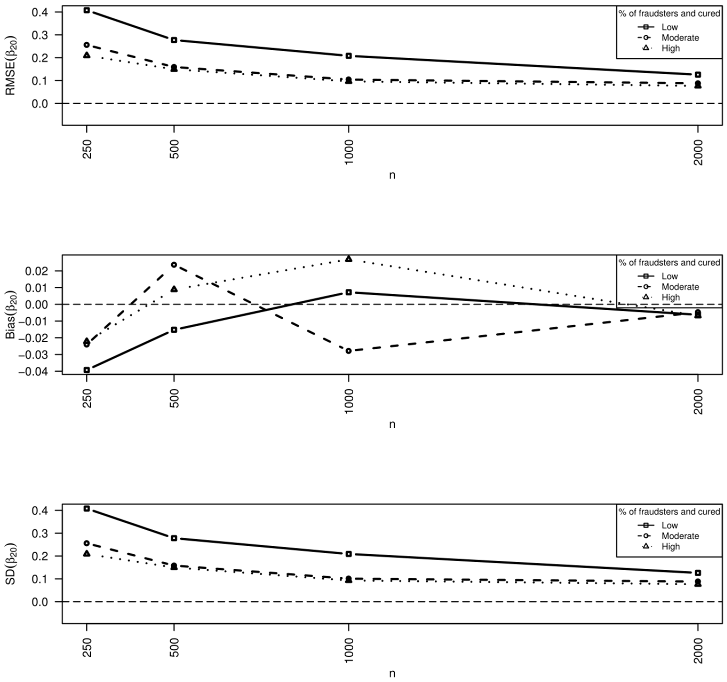

Figure 13: Root mean square error, bias and standard deviation for β_{20}.

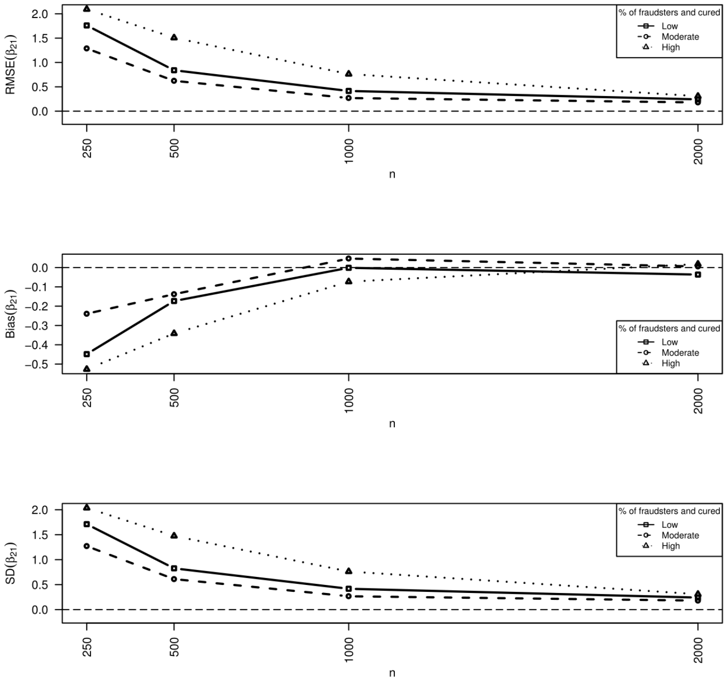

Figure 14: Root mean square error, bias and standard deviation for β_{21}.

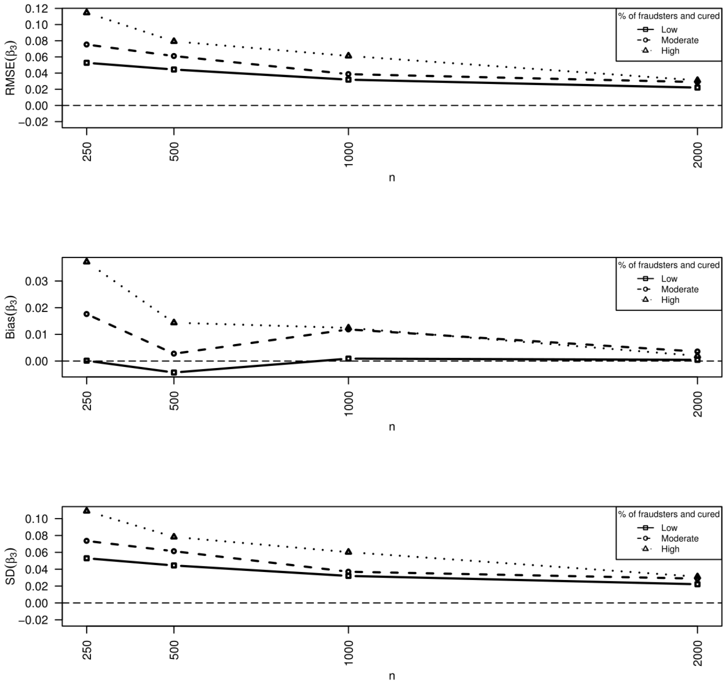

Figure 15: Root mean square error, bias and standard deviation for β_3.

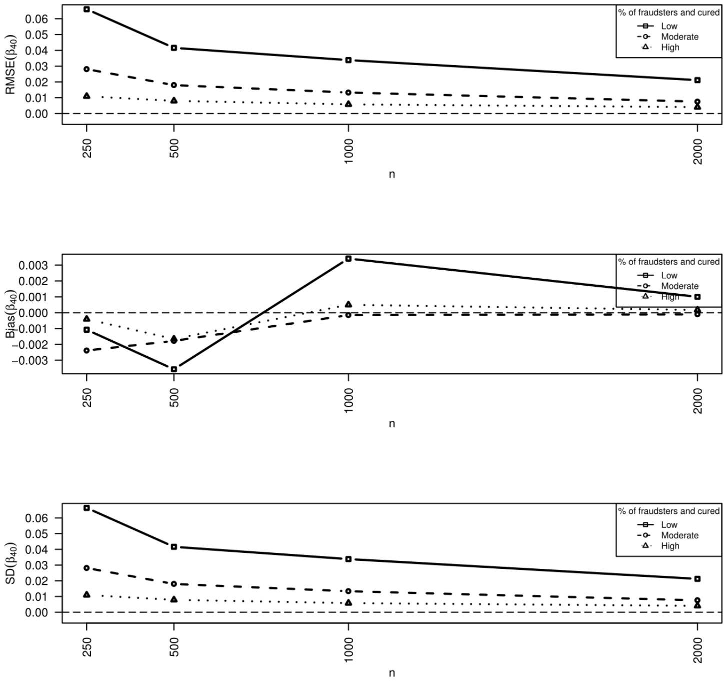

Figure 16: Root mean square error, bias and standard deviation for β_{40}.

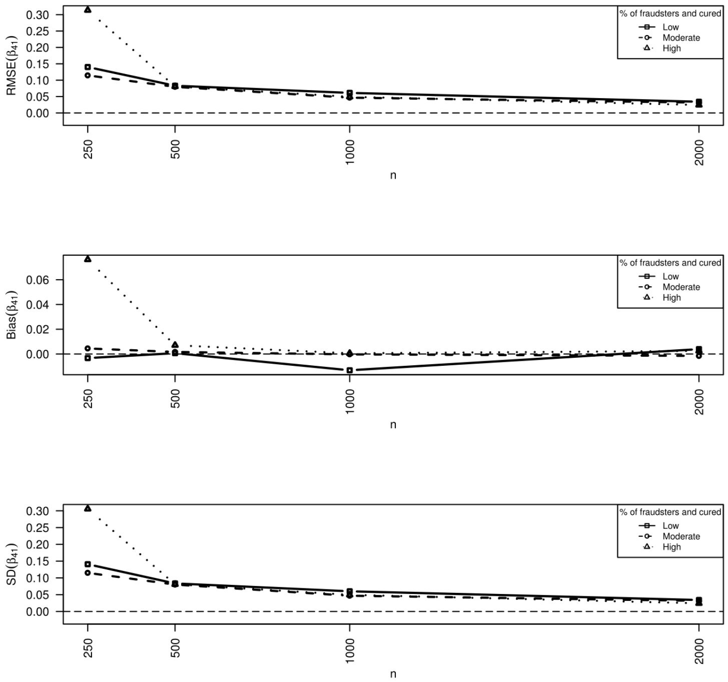

Figure 17: Root mean square error, bias and standard deviation for β_{41}.

6. Brazilian bank loan portfolio

In this section we present an application of the proposed model in a database made available by one of the largest Brazilian banks. Our objective is to check if customer characteristics are associated with consumer propensity of being fraudsters, defaulter or long-term customers, i.e., with no chance of becoming defaulter ahead.

It is important to note that the given database, quantities, rates and levels of the available covariate, do not necessarily represent the actual condition of the financial institution's customer base. Despite being a real database, the bank may have sampled the data in order to change the current status of its loan portfolio.

Tables 6 and 7 present a summary of the portfolio together with a portfolio segmentation considering only one available covariate. For reasons of confidentiality, we will refer to it by covariate x.

In this case, we can see below that x has three levels concerning to a particular characteristics of a banking customer profile.

Customers	Number of customers	Number of defaulters	Average of time to default	Standard deviation of time to default
$x = 1$	1,646	305 (18.530%)	15.016	11.874
$x = 2$	1,561	242 (15.503%)	15.641	12.134
$x = 3$	942	93 (9.873%)	18.183	12.868
Total	4,149	640 (15.425%)	15.712	12.148

Table 6: Summary of the bank loan lifetime data.

Customers	Number of fraudsters	Number of censored
$x = 1$	157 (9.538%)	1184 (71.940%)
$x = 2$	114 (7.303%)	1205 (77.194%)
$x = 3$	34 (3.609%)	815 (86.518%)
Total	305 (7.351%)	3204 (77.223)

Table 7: Summary of fraud and censored lifetime data.

6.1. Application model 1

To proceed with the application, we will deal with dummy covariates. As x has three levels, then we have two dummy covariates, $dx1$ and $dx2$. Let $dx1 = 1$ if $x = 1$, and $dx1 = 0$, otherwise. Similarly, $dx2 = 1$ if $x = 2$, and $dx2 = 0$, otherwise. The customer group such that $x = 3$ is characterized by setting $dx1 = dx2 = 0$.

According to the parameter configuration given by 16, next, we do not link covariates to the Weibull parameters, which we leave to the next subsection. However, note that we have made the following re-parametrization, $\alpha = e^{\alpha^{'}}$ and $\lambda = e^{\lambda^{'}}$ to facilitate the use of routine optim of the software R.

Therefore, we have the following set of parameters $\left\{\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}, \alpha^{'}, \lambda^{'}\right\}$ to be estimated by MLE approach:

$$\begin{cases} \gamma_{0i} & = & \left(\frac{e^{\beta_{10}+dx1_i\beta_{11}+dx2_i\beta_{12}}}{1+e^{\beta_{10}+dx1_i\beta_{11}+dx2_i\beta_{12}}+e^{\beta_{20}+dx1_i\beta_{21}+dx2_i\beta_{22}}}\right), \\ \gamma_{1i} & = & \left(\frac{e^{\beta_{20}+dx1_i\beta_{21}+dx2_i\beta_{22}}}{1+e^{\beta_{10}+dx1_i\beta_{11}+dx2_i\beta_{12}}+e^{\beta_{20}+dx1_i\beta_{21}+dx2_i\beta_{22}}}\right), \\ \alpha & = & e^{\alpha^{'}}, \\ \lambda & = & e^{\lambda^{'}}. \end{cases} \tag{16}$$

From the results presented in Table 8, through the analysis of the intercept parameter β_{10}, we see that being part of group $x = 3$ is significant for differentiation on the propensity to commit a fraud. Customers with this characteristic, i.e., $x = 3$, are less likely to be fraudster among customers of the other groups, since its estimated fraud rate is the lowest.

Parameter	Estimate (est)	Standard error (se)	$\lvert$est$\rvert$/ se
β_{10}	-1.2626	0.2031	6.215
β_{11}	0.3760	0.2229	1.686
β_{12}	0.2787	0.2301	1.211
β_{20}	1.8794	0.1196	15.70
β_{21}	-0.8117	0.1320	6.147
β_{22}	-0.5587	0.1350	4.137
$\alpha^{'}$	0.1124	0.0432	2.600
$\lambda^{'}$	3.1683	0.0764	41.43

Table 8: Maximum likelihood estimation for the zero-inflated Weibull cure rate regression.

The parameters β_{20}, β_{21} and β_{22}, also confirm that the segmentation given by the covariate x is significant to establish a descending order of long-term survival rates, from the group $x = 3$ with higher cure rate to the lowest cure rate group, $x = 1$.

Wherefore, the jointly analysis of zero inflation data with the fraction of cured, provided us with the valuable information that customers less likely to commit fraud, they are also more likely to be long-term survivors.

The analysis of the following figures confirms the aforementioned statement, as we can see in the Kaplan-Meier (K-M) survival curves, according to customer profiles $x = 1$, $x = 2$ and $x = 3$, see the Figure 18. It also presents the fitted survival functions as defined by the zero-inflated Weibull cure rate regression (ZIWCRR) model in 4.

We observed that the improper survival function falls instantly at $t = 0$ to the values $1 - \hat{\gamma}_0{}^1$, $1 - \hat{\gamma}_0{}^2$, $1 - \hat{\gamma}_0{}^3$, respectively, when $x = 1$, $x = 2$ or $x = 3$. So, the model accommodated well the zero-inflated survival times and, as expected, the plateaus at points $\hat{\gamma}_1^1$, $\hat{\gamma}_1^2$ and $\hat{\gamma}_1^3$ highlighted the presence of cure rate.

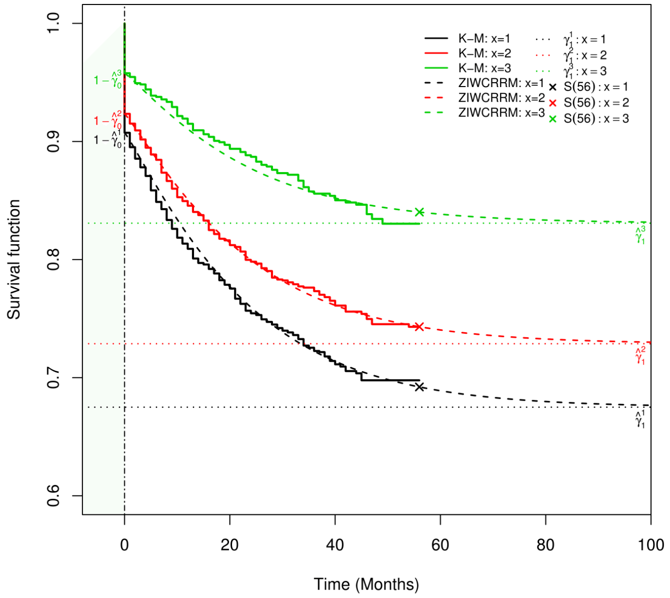

Figure 18: The Kaplan-Meier estimator according to customer $x = 1$, $x = 2$ and $x = 3$.

Table 9 shows the model results, considering the configuration parameters as given in 16. We can see that the outcomes are very satisfactory when compared with the actual data found out in the Table 7.

Estimated parameter	Consumer with $x = 1$	Consumer with $x = 2$	Consumer with $x = 3$
$\hat{\gamma}_0$	0.0953670	0.0730215	0.0361181
$\hat{\gamma}_1$	0.6731908	0.7316650	0.8362142
$\hat{S}(56)$	0.6902227	0.7460382	0.8456093

Table 9: Modelling outcomes of the model application 1.

Once the maximum observed time up to default for a bank customer is 56, the values for $\hat{S}(56)$, represent the estimated survival rate of the each sub-portfolio at this point. These values are compared with the cure rates shown in the third column of table 7.

6.2. Application model 2

Finally, we present the extended set of regression parameters that can be linked to the covariate x. In other words, we also connect the scale parameter of the Weibull distribution.

Thus, we have the set of parameters $\left\{\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}, \alpha^{'}, \beta_{30}, \beta_{31}, \beta_{32}, \right\}$ to be estimated by

MLE approach, given the following link functions:

$$\begin{cases} \gamma_{0i} & = & \frac{e^{\beta_{10}+dx1_i\beta_{11}+dx2_i\beta_{12}}}{1+e^{\beta_{10}+dx1_i\beta_{11}+dx2_i\beta_{12}}+e^{\beta_{20}+dx1_i\beta_{21}+dx2_i\beta_{22}}}, \\ \gamma_{1i} & = & \frac{e^{\beta_{20}+dx1_i\beta_{21}+dx2_i\beta_{22}}}{1+e^{\beta_{10}+dx1_i\beta_{11}+dx2_i\beta_{12}}+e^{\beta_{20}+dx1_i\beta_{21}+dx2_i\beta_{22}}}, \\ \alpha & = & e^{\alpha^{'}}, \\ \lambda_i & = & e^{\beta_{30}+dx1_i\beta_{31}+dx2_i\beta_{32}}, \end{cases} \tag{17}$$

In the model 2 outcomes, most of the estimated parameters are statistically significant, furthermore, they are consistent with those obtained in the model 1.

Parameter	Estimate (est)	Standard error (se)	|est|/ se
β_{10}	-1,3299	0,2501	5,315
β_{11}	0,4486	0,2676	1,676
β_{12}	0,4031	0,2731	1,475
β_{20}	1,6102	0,2292	7,025
β_{21}	-0,4925	0,2365	2,081
β_{22}	-0,2823	0,2399	1,176
$\alpha^{'}$	0,1149	0,0432	2,660
β_{30}	3,6228	0,2725	13,29
β_{31}	-0,5611	0,2773	2,023
β_{32}	-0,4689	0,2816	1,664

Table 10: Maximum likelihood estimation results for the zero-inflated Weibull cure rate regression model.

As in the previous application, the regression parameter β_{20}, reinforce the fact that the group $x = 3$ is an important feature for discriminating customer profiles regarding to the propensity to be a long-term survival costumer.

The extra information obtained through the parameters linked with the Weibull scale parameter, confirms the statement that the group $x = 3$, besides presents lower default rate and larger fraction of cured, it will also presents longer survival times.

7. Conclusion

We have presented a methodology in which we modify the simple cure rate model introduced by Berkson & Gage [7] to a credit risk setting, allowing us to estimate the proportions of the following loan applicants in a given portfolio: fraudsters, defaulters, and non-defaulters, i.e., long-term survivors.

At the heart of our methodology, the improper survival function is adapted to account for the excess of zeros, which represents the rate of loan fraudsters.

An advantage of our approach is to accommodate zero-inflated times, which is not possible in the standard cure rate model. In this scenario, information from fraudsters is exploited through the joint modelling of the survival time of fraudsters, who are equal to zero, along with the survival times of the remaining portfolio.

In this way, the model fits a survival curve for the defaulter sub-population and simultaneously estimates the proportion of zero-inflated times, which are related to fraudsters, and the proportion of non-defaulters.

To illustrate the proposed method, a loan survival times of a Brazilian bank loan portfolio is modelled by the proposed zero-inflated Weibull cure rate model, and the results showed satisfactory.

The importance of the jointly analysis of zero inflation data with the fraction of cured, is that it could provide us with the information that applicants less likely to commit fraud are also more likely to be long-term survivors.

Acknowledgement

The research was sponsored by CAPES - Process number: BEX 10583/14-9, Brazil.

References

[1] Abad, R. C., Fernández, J. M. V. & Rivera, A. D. (2009). Modelling consumer credit risk via survival analysis. *SORT: Statistics and Operations Research Transactions*, **33**(1), 3–30.

[2] Abreu, H. (2004). *Aplicação da Análise de Sobrevivência em um problema de Credit Scoring e comparação com a Regressão Logística*. Ph.D. thesis, Dissertação de Mestrado, UFSCar.

[3] Banasik, J., Crook, J. N. & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, pages 1185–1190.

[4] Barriga, G. D., Cancho, V. G. & Louzada, F. (2015). A non-default rate regression model for credit scoring. *Applied Stochastic Models in Business and Industry*.

[5] Barry, S. C. & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, **157**(2), 179–188.

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

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

[8] Blackwood, L. G. (1991). Analyzing censored environmental data using survival analysis: single sample techniques. *Environmental monitoring and assessment*, **18**(1), 25–40.

[9] Braekers, R. & Grouwels, Y. (2015). A semi-parametric cox's regression model for zero-inflated left-censored time to event data. *Communications in Statistics - Theory and Methods*, (just-accepted).

[10] Cancho, V. G., de Castro, M., Dey, D. K. *et al.* (2013). Long-term survival models with latent activation under a flexible family of distributions. *Brazilian Journal of Probability and Statistics*, **27**(4), 585–600.

[11] Colosimo, E. A. & Giolo, S. R. (2006). Análise de sobrevivência aplicada. In *ABE-Projeto Fisher*. Edgard Blücher.

[12] Conceição, K. S., Andrade, M. G. & Louzada, F. (2013). Zero-modified poisson model: Bayesian approach, influence diagnostics, and an application to a brazilian leptospirosis notification data. *Biometrical Journal*, **55**(5), 661–678.

[13] Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**(478).

[14] Hand, D. J. & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 523–541.

[15] Klein, J. & Moeschberger, M. (2003). Survival analysis: statistical methods for censored and truncated data. *Springer, New York*.

[16] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

[17] Lessmann, S., Baesens, B., Seow, H. & Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research*, **247**, 124–136.

[18] Liu, L., Huang, X., Yaroshinsky, A. & Cormier, J. N. (2015). Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics*, (just-accepted).

[19] Lord, D., Washington, S. P. & Ivan, J. N. (2005). Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, **37**(1), 35–46.

[20] Louzada, F., Cancho, V. G., Oliveira, M. R. & Yiqi, B. (2014). Modeling time to default on a personal loan portfolio in presence of disproportionate hazard rates. *Journal of Statistics Applications & Probability*, **3**(3), 295–305.

[21] Louzada-Neto, F. (2006). Lifetime modeling for credit scoring: A new alternative to traditional modeling via survival analysis. *Tecnologia de Crédito (Serasa)*, **56**, 8–22.

[22] Markel, P. D., DeFries, J. C. & Johnson, T. E. (1995). Ethanol-induced anesthesia in inbred strains of long-sleep and short-sleep mice: a genetic analysis of repeated measures using censored data. *Behavior genetics*, **25**(1), 67–73.

[23] Migon, H. S., Gamerman, D. & Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press.

[24] Ortega, E. M., Cordeiro, G. M. & Kattan, M. W. (2012). The negative binomial–beta weibull regression model to predict the cure of prostate cancer. *Journal of Applied Statistics*, **39**(6), 1191–1210.

[25] Ospina, R. & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, **56**(6), 1609–1623.

[26] Othus, M., Barlogie, B., LeBlanc, M. L. & Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, **18**(14), 3731–3736.

[27] Pereira, G. H., Botter, D. A. & Sandoval, M. C. (2013). A regression model for special proportions. *Statistical Modelling*, **13**(2), 125–151.

[28] R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[29] Rinne, H. (2008). *The Weibull distribution: a handbook*. CRC Press.

[30] Rocha, R., Nadarajah, S., Tomazella, V., Louzada, F. & Eudes, A. (2015). New defective models based on the kumaraswamy family of distributions with application to cancer data sets. *Statistical Methods in Medical Research*, pages 1–23.

[31] Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada-Neto, F. (2009). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**, 753–759.

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

[33] Tong, E. N., Mues, C. & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, **218**(1), 132–139.

[34] Vieira, A., Hinde, J. P. & Demétrio, C. G. (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, **27**(3), 373–389.