A Vertical Mixture Cure Model for Credit Risk Analysis

Ewa Wycinka and Tomasz Jurkiewicz

Abstract Credit risk assessment is one of the most important tasks of banks and other financial institutions. There are three main reasons of credit termination: maturity, early repayment and default. Credits that mature can be considered as not susceptible to early termination, whereas early repayments can be treated as competing risk to default. Most credits end on time (mature) or are repaid early, default happens only for a few percentage of credits. Modelling probability of default requires taking into account the probability of early repayment and maturity. We propose the use of a vertical mixture cure model with a cured fraction to analyse the probability of default. Empirical research was conducted on the sample of 5,000 consumer credit accounts of a Polish financial institution. Credits were observed 24 months since origination. The vertical mixture cure model was estimated with characteristics of borrowers as predictors. The discrimination ability of the model through 24 months of the credit life span was compared with a mixture model that has been earlier proposed in the literature.

Ewa Wycinka · Tomasz Jurkiewicz
University of Gdansk, Armii Krajowej street 101, 81-824 Sopot, Poland
✉ ewa.wycinka@ug.edu.pl
✉ tomasz.jurkiewicz@ug.edu.pl

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 4, No. 1, 2018

DOI: 10.5445/KSP/1000085951/21
ISSN 2363-9881

1 Introduction

As banks grant loans to consumers, they expect that the capital as well as the interest will be repaid at predefined times (usually in monthly instalments). For credits granted for a fixed term, paying back all of the instalments is called credit maturity (Dirick et al, 2015). Some credits are terminated earlier. Early repayment means that a credit is fully repaid before the predefined end date. The second reason for early termination is default, defined as payment of an instalment overdue by 90 days. The risk of default equates with the credit risk which is the chance that a borrower will be unable to make the required payments on his debt obligations.

Accords on capital adequacy, known as The Basel II and Basel III Accords, legally required more accurate credit risk calculations. This increased the banks' interest in statistical and operational research models to manage a borrower's account during its life, including any possible write-off. Additionally, the international financial reporting standard IFRS 9 *Financial instruments* that came into force in January 2018 extends the requirements in the area of credit risk analysis by introducing the obligation of probability of default estimation for more than one year ahead in order to evaluate lifetime expected credit losses (Vaněk and Hampel, 2017).

There are many statistical and operational research models to manage a borrower's account during its life, and possible write-off, but none of them has been proved to be much better than any other (Thomas et al, 2005). This legitimises searches for better methods. The purpose of this study is to propose the use of mixture cure models for competing risks with vertical approach for the probability of default in time under the control of the probability of early repayment and the probability of maturity. The rest of the paper is organised as follows. Section 2 is the review of the most important papers about mixture cure models, as well as presenting applications of survival methods to credit risk assessment. In section 3, basic definitions from the competing risks theory are given. The application of the vertical approach in the analysis of competing risks using mixture cure models is described in section 4. The next section comprises the description of the empirical study of the group of credits and the results of the application of a mixture cure model with competing risks to default risk assessment. The estimated model will be compared to the mixture cure model proposed by Tong et al (2012). The last part of the paper comprises a discussion of the results and suggestions for further research.

2 Literature Review

Mixture cure models belong to survival analysis methods. The concept of biostatistical models comprising a cured fraction of patients has a long history (Farewell, 1986). The first models were proposed by Boag (1949) and Berkson and Gage (1952). Larson and Dinse (1985) proposed regression models to assess the effect of covariates on the joint distribution of time and type of events based on the marginal distribution of the type of event (Equation 5). Nicolaie et al (2019) proposed a vertical modelling approach based on the marginal distribution of time to event (Equation 6) to model competing risks with a cured fraction. Nicolaie et al (2019) used the proposed method in medicine (the survival of patients with malignant melanoma).

Probably the first ones who adopted survival analysis methods for credit risk assessment were Green and Shoven (1986). They used a Proportional Hazards (PH) model to evaluate mortgage termination by refinance. However, after Narain (1992) applied accelerated failure time models to the risk of default, the interest in the use of methods of survival analysis to credit risk assessment increased considerably. Stepanova and Thomas (2002) gave attention to the presence of competing risks in credit risk assessment and evaluated separate Cox PH models for default and early repayment. Tong et al (2012) used mixture cure models to model not only the risk of default, but also the risk of maturity. Watkins et al (2014) proposed a parametric mixture cure model, based on the approach presented by Larson and Dinse (1985), to assess in one model the probability of default, early repayment and maturity. Dirick et al (2015) extended this approach by replacing the parametric hazards model by a semiparametric Cox PH model. In this paper, we propose to use vertical models with a cured fraction for credit risk analysis.

3 Competing Risks

Let (T, C) be a bivariate random variable with T, a continuous variable representing the time to the first event, and $C = k$ $(k = 1, \dots, p)$, a discrete variable denoting the type of event. Due to the right censoring, the variable (T, C) is only partially observable. We observe a pair $(\min\{T, T_c\}, C)$. As the result, the joint distribution of (T, C) is difficult to identify.

However, the joint distribution is completely specified by the cumulative incidence function CIF (also called subdistribution) of the event k, which is the probability that an event of type k will occur until time t (Lindqvist, 2008)

$$F_k(t) = P(T \leq t, C = k). \tag{1}$$

The CIF is not a proper distribution function because

$$\lim_{t \to \infty} F_k(t) = P(C = k). \tag{2}$$

The sum of the cumulative incidence functions for all of the p types of events is equal to the overall distribution function (Pintilie, 2006)

$$F(t) = \sum_{k=1}^{p} F_k(t). \tag{3}$$

The CIF can be presented as

$$F_k(t) = \int_0^t h_k(u) S(u)\, du \tag{4}$$

where $h_k(t)$ is the cause-specific hazard for event k at time t and $S(t)$ is the overall survival function (the probability of being free of any event prior to time t), that is $S(t) = 1 - F(t)$ (Pintilie, 2006).

For the bivariate random variable (T, C), marginal and conditional distributions can be delineated. Conditional distributions of the bivariate random variable can be expressed as

$$P\left(T = t | C = k\right) = \frac{P(T = t, C = k)}{P(C = k)} \tag{5}$$

and

$$P\left(C = k | T = t\right) = \frac{P(T = t, C = k)}{P(T = t)} \tag{6}$$

where $P(C = k)$ is the marginal distribution of event types and $P(T = t)$ is the marginal distribution of the time of the first event.

4 Vertical Approach to Mixture Cure Models for Competing Risks

Mixture cure models (cure rate models) assume that an analysed population is not homogeneous and consists of two subpopulations. The first subpopulation comprises units that are not susceptible to the event (units that will never experience the event). These are long event-free survivors. The second subpopulation includes units that are susceptible to the event. These units experience the event during follow-up or they will experience the event in time.

Let us consider all the events in spite of the type of an event. Let Y be the indicator of susceptibility with $Y = 1$ if the unit is susceptible to the event, with the probability $P(Y = 1) = p$ and $Y = 0$ otherwise. The Y variable is only partially observable. If the event occurs during the follow-up then $Y = 1$, in the opposite case the unit could be not susceptible or the event would occur out of the follow-up. The unconditional distribution function in a mixture cure model approach is

$$F(t) = pF(t|Y = 1) + (1 - p)F(t|Y = 0), \tag{7}$$

where $F(t|Y = 0) \equiv 0$ is the degenerate distribution function (Peng and Taylor, 2014). Therefore the second element of the sum can be omitted. If the interest is in the effect of covariates on T, then the unconditional distribution function can be expressed as

$$F(t|x, z) = p(z)F(t|Y = 1, x) \tag{8}$$

where x and z may or may not be the same vectors of covariates measured at time zero, related respectively to the probability that the event occurs and the probability of time to the event. The $p(z)$ is the probability of being susceptible and can be evaluated by a logit model

$$\ln\left(\frac{p(z)}{1 - p(z)}\right) = z^T\beta. \tag{9}$$

This part of the mixture cure model is called an incidence model. The conditional (on $Y = 1$) distribution function for the susceptible units can be modelled by the Cox PH model, through the relation $F\,(t, Y = 1) = 1 - S(t, Y = 1)$, as

$$S\,(t\,|Y=1,x)=S_0(t|Y=1)^{\exp(x^Tb)} \tag{10}$$

where $S_0(t)$ is a baseline hazard function. This part of the mixture cure model is called a latency model. Maximum likelihood estimators of the parameters (β, b) and the estimator of the function $S_0(t)$ are evaluated in an iterative maximisation algorithm (EM algorithm; Peng and Dear, 2000).

Nicolaie et al (2019) proposed an extension of the above mixture model to the competing risks by putting the contribution of the cause-specific hazard of the cause k in an overall hazard in the model. The conditional relative cause-specific hazard of cause k at time t is defined as

$$\pi_k(t|Y=1)=P(C=k|T=t,Y=1) \tag{11}$$

and is the probability that the event of type k occurs given any event occurs and given time t. This probability can be estimated by a multinomial logit model

$$\ln\left(\frac{\pi_k(t|u)}{1-\pi_k(t|u)}\right)=\gamma^TB(t)+\upsilon^Tu \tag{12}$$

where $B(t)$ is a vector of predefined time functions, e.g. B-spline functions, and u is a vector of covariates (Nicolaie et al, 2019). To evaluate the conditional on $Y=1$ cumulative incidence function (CIF) on cause k, Nicolaie et al (2019) proposed to compute the cause-specific hazard function from Equation 4 as the product of the conditional relative cause-specific hazard of cause k ($\pi_k(t)$) and the cause specific-hazard for all causes (overall hazard), denoted by $h.(t)$:

$$F_k(t|Y=1,x,u)=\int_0^t \pi_k(v|Y=1,u)\,h.(v|Y=1,x)\,S(v|Y=1)\,dv \tag{13}$$

where $h.(t|Y=1,x)=h_0.(t,Y=1)\exp(\sum_{k=1}^{m}\beta_kx_k)$ and $S(t,Y=1)$ are evaluated by the Cox PH model. Nicolaie et al (2019) proved that the maximum likelihood estimators of the parameters (γ,υ) in Equation 12 can be estimated separately from estimators of the parameters (β, b) and an estimator of the function $S_0(t)$.

Finally, the unconditional cumulative incidence function (for the whole population) can be expressed as

$$F_k(t|z,x,u)=p(z)\,F_k(t|Y=1,x,u). \tag{14}$$

5 Data Analysis

Empirical research was conducted on a sample of 5,000 consumer credit accounts from a portfolio of 60-month personal loans of a Polish financial institution. All of the credits were granted in five subsequent months. Each credit was observed for 24 months or until early repayment or default if that occurred earlier. There were 2,188 creditors (43.8 %) who repaid all 24 instalments (or had a delay in payment that was shorter than 90 days), 297 creditors (5.9 %) who defaulted during the first 24 months, and 2,515 creditors (50.3 %) who repaid the credit (early repayments). Default is the event of interest, whereas earlier repayment is considered to be a competing risk. Borrowers who repaid all 24 instalments were considered to be long-survivors.

The dataset contains typical application characteristics used in credit scoring such as: amount of credit, amount of the instalments, the purpose of the loan, age of the applicant, property and educational level. For the requirements of the financial institution sharing the data, the names of the variables were anonymised. Variables are denoted by letter X and numbers. Numbers preceded by an underscore denote the number of the variable's attributes. Variables without underline are binary. All the variables were categorised in order to maximise Kaplan-Meier survival curves between the distinct attributes (Wycinka, 2015). Because application characteristics were highly correlated between each other, the association structure of the data was revealed with the use of Markov network structures (Edera et al, 2014). Variables associated with the highest number of other variables were included in the model as predictors whereas variables correlated to them were not selected. Subsequently, a stepwise Akaike Information Criterion (AIC) procedure was applied to check if any of the explanatory variable should be removed from the model. Ultimately, 5 predictors were left in the model. The choice of the variables proposed in this paper is partially subjective but the method is easy to apply and allows to identify correlated variables.

Different methods of variable selection were used for mixture cure models by other authors: Tong et al (2012) used backward variable selection to leave only significant covariates in the model. This method, however, is not proper in the case of correlated variables (Harrell, 2015). Dirick et al (2015) described the use of a genetic algorithm as well as some modification of AIC for variable selection in mixture cure models. The drawback of this method is that it is time-consuming.

As the next step in model building, assumption of the proportionality of hazards in the Cox model was verified with the test proposed by Lin et al (1993) and Li et al (2015). All the calculations have been made in R (packages: Survival (Therneau, 2015), goftte (Sfumato and Boher, 2017), SMcure (Chao Cai and Zhang, 2015), Hmeasure (Anagnostopoulos and Hand, 2019), bnlearn (Scutari, 2010)). Estimates of the parameters of the model are presented in Table 1.

Table 1: Estimates of the mixture cure model for default and early repayment (vertical approach).

Covariates	Latency Part				Incidence Part				Relative Hazard			
	(Logit Model)				(Cox Model)				(Logit Model)			
	OR	95 % CI		p-value	HR	95 % CI		p-value	OR	95 % CI		p-value
		Lower	Upper			Lower	Upper			Lower	Upper	
Intercept	3.93	2.44	6.35	0.1397	.	.	.	.	0.12	0.06	0.23	0.0000
X 1_1	1.79	1.09	2.92	0.0287	1.24	1.02	1.51	0.0287	2.06	1.45	2.92	0.0000
X 1_2	0.92	0.72	1.18	0.2153	0.88	0.78	1.00	0.0465	0.45	0.34	0.60	0.0000
X 2_1	0.81	0.58	1.13	0.1500	0.90	0.77	1.06	0.2000	0.87	0.61	1.25	0.4529
X 2_2	0.79	0.55	1.14	0.5130	0.85	0.71	1.01	0.0706	0.65	0.44	0.96	0.0309
X3	1.26	0.96	1.66	0.4578	1.03	0.90	1.20	0.6417	2.17	1.66	2.84	0.0000
X4	1.14	0.70	1.86	0.0739	0.99	0.76	1.28	0.9267	2.66	1.76	4.02	0.0000
X5	0.67	0.48	0.94	0.0363	1.09	0.94	1.26	0.2532	2.36	1.62	3.42	0.0000
bs(Time)1	.	.	.	.	.	.	.	.	0.12	0.03	0.56	0.0064
bs(Time)2	.	.	.	.	.	.	.	.	1.34	0.58	3.08	0.4874
bs(Time)3	.	.	.	.	.	.	.	.	0.99	0.47	2.09	0.9780

(.) = variables not included in the model, bs = B-spline basis function

In the next step, empirical relative cause-specific hazards were calculated as

$$\widehat{\pi}_k(t|Y=1) = \frac{d_{tk}}{d_t} \tag{15}$$

where d_{tk} is the number of defaults in time t (events of type k) and d_t is the number of all types of events (both defaults and early repayments) in time t. The results are presented in Figure 1 (dotted line). Due to the definition of the default, it could appear for the first time in the third month. In the analysed sample, the first early repayment was recorded in the fourth month. As a result, relative cause-specific hazard for the default in the third month is equal to one.

In subsequent months, values of the relative hazard vary around 0.107 and slightly grow each month. The effect of time on relative hazard was smoothed by B-spline functions (Figure 1, solid line).

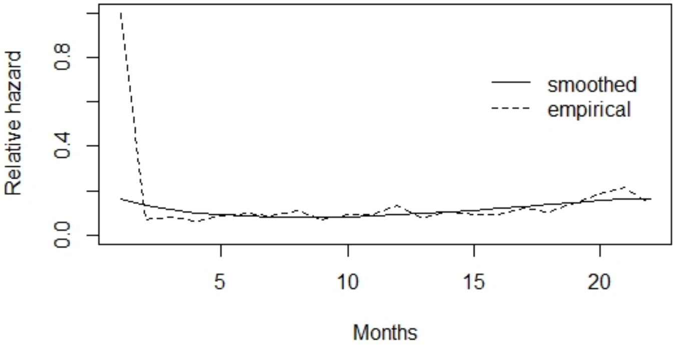

Figure 1: Relative cause specific hazard of default.

Splines, as well as other covariates, were included in the logit model for relative hazard (Equation 12). Because there were only two competing risks analysed (default and early repayment), the multinomial logit model was reduced to a binary logit model.The results are given in Table 1.

Each part of the model uses the same covariates. However, their role is different. Let us analyze the variable $X2\_2$ and its parameters. In the latency part, OR = 0.79, which means that the borrower with $X2\_2 = 1$ has a 21 % lower risk of being susceptible than the reference group. In the incidence part, HR = 0.85, which means that for a susceptible borrower with $X2\_2 = 1$ the risk of early termination is 25 % lower than for the reference group. Additionally, OR=0.65 means that the susceptible borrower with $X2\_2 = 1$ has a 45 % lower risk of termination due to default than the reference group. The discrimination measures for the model given in Table 1 are: McFadden's pseudo R-squared 0.085 and $AUC = 0.715$.

Finally, conditional and unconditional cumulative incidence functions were calculated for all of the units. In order to explore how the model fits at each time t, a set of binary data was created for each moment t, in which one denoted the unit for which default occurred up to time t and zero denoted the unit for which default did not occur till time t. The unconditional cumulative incidence function for default at time t, calculated for each unit, was used as a score function. Four different discrimination measures were used: Area under the Receiver Operating Characteristic Curve (AUC), Kolmogorov-Smirnov statistic, Gini coefficient and H-measure. Confidence intervals for the measures of discrimination were calculated as 2.5 and 97.5 percentiles from 10,000 bootstrapped samples. The results are given in Table 2.

Table 2: Discrimination measures of mixture cure model for default and early repayment (vertical approach) at different months.

Month	H (95 % CI)	Gini (95 % CI)	AUC (95 % CI)	KS (95 % CI)
3	0.410 (0.210-0.640)	0.650 (0.440-0.840)	0.820 (0.720-0.920)	0.580 (0.400-0.780)
4	0.410 (0.260-0.580)	0.660 (0.500-0.820)	0.830 (0.750-0.910)	0.610 (0.450-0.760)
5	0.407 (0.213-0.636)	0.647 (0.436-0.839)	0.823 (0.718-0.920)	0.584 (0.395-0.780)
6	0.414 (0.261-0.585)	0.663 (0.499-0.820)	0.832 (0.749-0.910)	0.605 (0.446-0.762)
7	0.348 (0.226-0.493)	0.617 (0.472-0.757)	0.808 (0.736-0.878)	0.555 (0.425-0.691)
8	0.310 (0.214-0.435)	0.608 (0.494-0.726)	0.804 (0.747-0.863)	0.521 (0.412-0.648)
9	0.247 (0.166-0.354)	0.541 (0.432-0.654)	0.770 (0.716-0.827)	0.447 (0.351-0.562)
10	0.211 (0.137-0.303)	0.488 (0.384-0.597)	0.744 (0.692-0.799)	0.407 (0.311-0.512)
11	0.193 (0.125-0.277)	0.460 (0.357-0.566)	0.730 (0.678-0.783)	0.381 (0.290-0.476)
12	0.207 (0.142-0.283)	0.485 (0.388-0.581)	0.742 (0.694-0.791)	0.405 (0.321-0.494)
13	0.197 (0.137-0.270)	0.469 (0.381-0.563)	0.734 (0.691-0.782)	0.393 (0.313-0.476)
14	0.187 (0.133-0.254)	0.462 (0.376-0.549)	0.731 (0.688-0.774)	0.384 (0.31-0.463)
15	0.178 (0.126-0.242)	0.449 (0.369-0.536)	0.725 (0.684-0.768)	0.366 (0.298-0.441)
16	0.158 (0.108-0.216)	0.418 (0.339-0.503)	0.709 (0.669-0.751)	0.335 (0.271-0.403)
17	0.156 (0.109-0.211)	0.416 (0.338-0.497)	0.708 (0.669-0.748)	0.331 (0.268-0.396)
18	0.159 (0.113-0.212)	0.424 (0.349-0.501)	0.712 (0.674-0.751)	0.337 (0.276-0.400)
19	0.167 (0.120-0.219)	0.437 (0.366-0.511)	0.718 (0.683-0.755)	0.345 (0.287-0.405)
20	0.168 (0.124-0.219)	0.439 (0.370-0.510)	0.720 (0.685-0.755)	0.345 (0.289-0.404)
21	0.173 (0.127-0.223)	0.444 (0.377-0.514)	0.722 (0.688-0.757)	0.348 (0.293-0.408)
22	0.171 (0.126-0.219)	0.442 (0.376-0.512)	0.721 (0.688-0.756)	0.347 (0.290-0.403)
23	0.167 (0.125-0.212)	0.441 (0.377-0.508)	0.721 (0.689-0.754)	0.345 (0.292-0.398)
24	0.169 (0.128-0.212)	0.444 (0.381-0.506)	0.722 (0.691-0.753)	0.345 (0.291-0.399)

CI = Confidence Interval

We also estimated a mixture cure model for default as the only type of event, as proposed by Tong et al (2012). In this approach early repayments are considered as censoring. The estimates of the model are given in Table 3. The latency part of the model evaluates the probability of early termination due to default whereas the incident part of the model evaluates the distribution of time to default. In this approach the relative hazard is not estimated. Comparing these results with Table 1.1, we can observe changes in the values of parameters of variables. Let us focus again on the variable $X2\_2$. In the model given in Table 3 in the latency part OR=0.37. This means that borrower with $X2\_2 = 1$ has a 63 % lower risk of default than the reference group. In the latency part HR=1.32 which means that the hazard for a susceptible borrower with $X2\_2 = 1$ is 32 % higher than for the reference group. The difference from the model given in Table 1 is caused by the fact that in the model given in Table 3, borrowers who have made early repayments are treated as censored observations (still susceptible to default), not as competing risk. Discrimination measures for this model are shown in Table 4.

Table 3: Estimates of the mixture cure model for default only.

Covariates	Latency Part (Logit Model)				Incidence Part (Cox Model)			
	OR	95 % CI		p-value	HR	95 % CI		p-value
		Lower	Upper			Lower	Upper	
Intercept	0.33	0.17	0.62	0.0007	.	.	.	.
X1_1	2.33	1.42	3.84	0.0008	1.32	0.77	2.24	0.3140
X1_2	0.37	0.25	0.55	0.0000	0.93	0.57	1.52	0.7634
X2_1	0.43	0.25	0.76	0.0033	1.27	0.66	2.41	0.4734
X2_2	0.37	0.22	0.65	0.0005	1.32	0.71	2.47	0.3857
X3	1.81	1.28	2.56	0.0008	1.30	0.86	1.96	0.2134
X4	1.22	0.64	2.33	0.5442	1.45	0.58	3.62	0.4259
X5	1.24	0.78	1.97	0.3719	1.52	1.00	2.31	0.0478

(.) variables not included in the model

Table 4: Discrimination measures of mixture cure model for default and early repayment (vertical approach) at different months (1/2).

Month	H (95 % CI)	Gini (95 % CI)	AUC (95 % CI)	KS (95 % CI)
3	0.406 (0.216-0.634)	0.649 (0.41-0.837)	0.825 (0.705-0.918)	0.587 (0.402-0.783)
4	0.412 (0.255-0.575)	0.664 (0.489-0.812)	0.832 (0.745-0.906)	0.606 (0.447-0.759)
5	0.343 (0.223-0.482)	0.614 (0.455-0.754)	0.807 (0.727-0.877)	0.548 (0.412-0.678)
6	0.320 (0.222-0.436)	0.615 (0.494-0.727)	0.808 (0.747-0.864)	0.533 (0.423-0.648)
7	0.252 (0.170-0.351)	0.546 (0.434-0.654)	0.773 (0.717-0.827)	0.454 (0.355-0.555)
8	0.219 (0.143-0.303)	0.499 (0.385-0.601)	0.749 (0.692-0.800)	0.416 (0.321-0.509)
9	0.200 (0.131-0.282)	0.469 (0.358-0.572)	0.735 (0.679-0.786)	0.388 (0.297-0.476)
10	0.212 (0.145-0.286)	0.489 (0.388-0.584)	0.745 (0.694-0.792)	0.408 (0.322-0.494)
11	0.202 (0.140-0.274)	0.474 (0.377-0.568)	0.737 (0.689-0.784)	0.397 (0.314-0.482)
12	0.192 (0.134-0.258)	0.467 (0.378-0.554)	0.734 (0.689-0.777)	0.389 (0.313-0.466)
13	0.183 (0.129-0.244)	0.454 (0.369-0.537)	0.727 (0.685-0.769)	0.373 (0.300-0.442)

Table 4: Discrimination measures of mixture cure model for default and early repayment (vertical approach) at different months (2/2).

Month	H (95 % CI)	Gini (95 % CI)	AUC (95 % CI)	KS (95 % CI)
14	0.162 (0.112-0.220)	0.423 (0.343-0.503)	0.711 (0.671-0.752)	0.340 (0.275-0.406)
15	0.158 (0.110-0.214)	0.419 (0.342-0.500)	0.710 (0.671-0.750)	0.334 (0.271-0.400)
16	0.161 (0.115-0.217)	0.427 (0.355-0.506)	0.714 (0.677-0.753)	0.340 (0.279-0.404)
17	0.168 (0.123-0.225)	0.439 (0.369-0.516)	0.719 (0.684-0.758)	0.347 (0.290-0.412)
18	0.168 (0.125-0.222)	0.440 (0.372-0.516)	0.720 (0.686-0.758)	0.347 (0.292-0.408)
19	0.173 (0.129-0.226)	0.446 (0.377-0.519)	0.723 (0.688-0.759)	0.350 (0.294-0.411)
20	0.170 (0.127-0.221)	0.442 (0.373-0.513)	0.721 (0.686-0.756)	0.346 (0.291-0.406)
21	0.165 (0.123-0.214)	0.441 (0.374-0.507)	0.721 (0.687-0.754)	0.344 (0.289-0.401)
22	0.167 (0.125-0.215)	0.443 (0.378-0.507)	0.721 (0.689-0.754)	0.345 (0.292-0.400)
23	0.157 (0.117-0.201)	0.430 (0.365-0.493)	0.715 (0.682-0.746)	0.333 (0.280-0.384)
24	0.158 (0.119-0.201)	0.431 (0.367-0.492)	0.715 (0.684-0.746)	0.330 (0.278-0.382)

6 Conclusions and Further Research

An application of vertical modelling with a cured fraction was used to evaluate the lifetime probability of default under the control of the probabilities of early repayment and maturity. The discrimination power of the above method seems to be quite satisfactory at all analysed time points and is comparable to the methods proposed earlier in the literature. Better discrimination ability observed in the first months compared to later periods is combined with wider confidence intervals. This is due to a low number of cumulative defaults in the first months.

Covariates used in all of the parts of the mixture cure model were categorised in order to maximise the difference in survival between the units belonging to distinct attributes of categorised variables. This strategy seems to be appropriate in the latency part of the model. However, in the incidence part, categorisation should be made in favour of a maximisation of the odds ratio of early terminated credits (both defaults and early repayments) to long survivals. Finally, in the logit model for relative cause-specific hazards, categorisation of the models could be prepared only on the set of early terminated credits and should aim to maximise the odds ratio of defaults to early repaid credits. Since the method of estimation of the parameters of a mixture cure model with competing risks allows for different sets of covariates, the above proposition of implementing different categorisation methods should be considered. Dirick et al (2015) applied multiple event mixture cure models based on Equation (5) to credit risk assessment. For further research, it would be interesting to compare that approach, using the same data set, with the one presented in this paper.

References

Anagnostopoulos C, Hand DJ (2019) Package "hmeasure". URL: `http://www.hmeasure.net`. R package version 1.0-2.

Berkson J, Gage RP (1952) Survival Curve for Cancer Patients Following Treatment. Journal of the American Statistical Association 47(259):501–515. DOI: 10.1080/01621459.1952.10501187.

Boag JW (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. URL: `http://www.jstor.org/stable/2983694`.

Chao Cai YP Yubo Zou, Zhang J (2015) Package "smcure". URL: `https://CRAN.R-project.org/package=smcure`. R package version 2.0.

Dirick L, Claeskens G, Baesens B (2015) An Akaike Information Criterion for Multiple Event Mixture Cure Models. European Journal of Operational Research 241(2):449–457. DOI: 10.1016/j.ejor.2014.08.038.

Edera A, Strappa Y, Bromberg F (2014) The Grow-Shrink Strategy for Learning Markov Network Structures Constrained by Context-Specific Independences. In: Advances in Artificial Intelligence (IBERAMIA 2014), Bazzan AL, Pichara K (eds), Springer International Publishing, Cham (Switzerland), Lecture Notes in Computer Science, Vol. 8864, pp. 283–294. DOI: 10.1007/978-3-319-12027-0_23.

Farewell VT (1986) Mixture models in survival analysis: Are they worth the risk? Canadian Journal of Statistics 14(3):257–262. DOI: 10.2307/3314804.

Green J, Shoven JB (1986) The Effects of Interest Rates on Mortgage Prepayments. Journal of Money, Credit and Banking 18(1):41–59. DOI: 10.2307/1992319.

Harrell F (2015) Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd edn. Springer, Cham (Switzerland). ISBN: 978-3-319194-25-7.

Larson MG, Dinse GE (1985) A Mixture Model for the Regression Analysis of Competing Risks Data. Journal of the Royal Statistical Society. Series C (Applied Statistics) 34(3):201–211. DOI: 10.2307/2347464.

Li J, Scheike TH, Zhang MJ (2015) Checking Fine and Gray Subdistribution Hazards Model With Cumulative Sums of Residuals. Lifetime Data Analysis 21(2):197–217. DOI: 10.1007/s10985-014-9313-9.

Lin DY, Wei LJ, Ying Z (1993) Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals. Biometrika 80(3):557–572. DOI: 10.2307/2337177.

Lindqvist BH (2008) Competing Risks. In: Encyclopedia of Statistics in Quality and Reliability. John Wiley & Sons, Ltd, Hoboken (USA). DOI: 10.1002/9780470061572.eqr067.

Narain B (1992) Survival Analysis and the Credit-granting Decision. In: Credit Scoring and Credit Control, Thomas L, Edelman D, Crook J (eds). Oxford University Press, pp. 109–122.

Nicolaie M, Taylor J, Legrand C (2019) Vertical Modeling: Analysis of Competing Risks Data With A Cure Proportion. Lifetime Data Analysis 25(1):1–25. DOI: 10.1007/s10985-018-9417-8.

Peng Y, Dear KB (2000) A Nonparametric Mixture Model for Cure Rate Estimation. Biometrics 56(1):237–243. DOI: 10.1111/j.0006-341X.2000.00237.x

Peng Y, Taylor JM (2014) Cure Models. In: Handbook of Survival Analysis, Klein J, von Houwelingen H, Ibrahim JG, Scheike TH (eds). Chapman & Hall / CRC, pp. 113–134. ISBN: 978-0-367330-96-5.

Pintilie M (2006) Competing Risks: A Practical Perspective, Statistics in Practice, Vol. 58. John Wiley & Sons, Hoboken (USA). ISBN: 978-0-470870-68-6.

Scutari M (2010) Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software 35(3):1–22. DOI: 10.18637/jss.v035.i03.

Sfumato P, Boher JM (2017) goftte: Goodness-of-Fit for Time-to-Event Data. URL: `https://CRAN.R-project.org/package=goftte`. R package version 1.0.5.

Stepanova M, Thomas L (2002) Survival Analysis Methods for Personal Loan Data. Operations Research 50(2):277–289. DOI: 10.1287/opre.50.2.277.426.

Therneau TM (2015) A Package for Survival Analysis in S. URL: `https://CRAN.R-project.org/package=survival`. Version 2.38.

Thomas LC, Oliver RW, Hand DJ (2005) A Survey of the Issues in Consumer Credit Modelling Research. Journal of the Operational Research Society 56(9):1006–1015. DOI: 10.1057/palgrave.jors.2602018.

Tong EN, Mues C, Thomas LC (2012) Mixture Cure Models in Credit Scoring: If and When Borrowers Default. European Journal of Operational Research 218(1):132–139. DOI: 10.1016/j.ejor.2011.10.007.

Vaněk T, Hampel D (2017) The Probability of Default Under IFRS 9: Multi-Period Estimation and Macroeconomic Forecast. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis 65(2):759–776. DOI: 10.11118/actaun201765020759.

Watkins JGT, Vasnev AL, Gerlach R (2014) Multiple Event Incidence and Duration Analysis For Credit Data Incorporating Non-stochastic Loan Maturity. Journal of Applied Econometrics 29(4):627–648. DOI: 10.1002/jae.2329.

Wycinka E (2015) Time to Default analysis in Personal Credit Scoring. In: Financial Investments and Insurance – Global Trends and the Polish Market, Jajuga K, Ronka-Chmielowiec W (eds), Publishing House of Wrocaw University of Economics, Wrocaw (Poland), Vol. 381, pp. 527–536. ISBN: 978-8-376954-63-9, URL: http://www.dbc.wroc.pl/dlibra/doccontent?id=29335.