SURVIVAL ANALYSIS OF BANK LOANS IN THE PRESENCE OF LONG-TERM SURVIVORS

Mercy Marimo [1]
School of Statistics and Actuarial Science, University of the Witwatersrand
e-mail: *Mercy.Marimo@yahoo.com*

Charles Chimedza
School of Statistics and Actuarial Science, University of the Witwatersrand
e-mail: *charles.chimedza@wits.ac.za*

Key words: Cause specific, Competing risk, Consumer credit, Proportional hazards.

Abstract: In this paper we model competing risks, default and early settlement events, in the presence of long term survivors and compare survival and logistic methodologies. Cause specific Cox regression models were fitted and adjustments were made to accommodate a proportion of long term survivors. Methodologies were compared using ROC curves and area under the curves. The results show that survival methods perform better than logistic regression methods when modelling lifetime data in the presence of competing risks and in the presence of long term survivors.

1. Introduction

Conventional models of credit are often built on static variables obtained from application data. Logistic regression (LR) has been the cornerstone of credit models. It plays a very important role in building scorecards, which determine whether an applicant should be granted a loan or not. Even though LR methods have been in use for model building, Bellotti and Crook (2007) showed that survival analysis methods are more competitive and often superior to the LR approach as they use more information, including details of censoring as well as survival time. Consumer credit data is analogous to lifetime data as it concerns the credit status of a cohort of customers with different loan repayment behaviours over a given observation period. A single money lending product offering instalment loans is considered in this case, whereby a customer repays vehicle finance loan in instalments, over a predetermined repayment period. Survival methodology is applied to the prediction of two mutually exclusive events, default and Early Settlement (ES). The occurrence of these two events over the observation period impacts negatively on profitability (Stepanova and Thomas, 2002).

[1] Corresponding author.

AMS: 62F99

2. Methods

In this paper survival analysis is used to model loan survival in the presence of competing risks. The performance of the survival model is then compared to the more commonly used logistic regression.

The multinomial approach in the context of competing risks can be applied if the time points are continuous (Jenkins, 2005). In this case (bank loan data), the time points are discrete. Thus, applying multinomial logistic regression requires making assumptions about the discrete time (interval) hazard to relate the process to continuous time.

It was proven in Xue et al. (2013) that the proportional hazards model outperforms polytomous logistic regression. This study also suggests that binary logistic regression too is weaker than the proportional hazards model.

There is also one equation for predicting each outcome in multinomial regression. When considering competing risks, there is a different equation for each outcome in each time point, this means the multinomial probabilities are the time-specific hazards of each outcome. Therefore, none of the time point specific equations will provide a single predictive summary.

Treatment of long term survivors in the context of competing risks may also need a separate study to establish an appropriate link function. Because of the issues raised above the multinomial approach will not be considered at this stage.

2.1. Logistic Regression

Logistic regression is a type of generalised linear model used to predict an event based on a set of predictors. It uses a logit transformation on the dependent variable expressed as follows (for a simple logistic model):

$$\text{logit}(Y) = \log_e(odds) = \log_e\left(\frac{p}{1-p}\right) = \alpha + \beta X, \tag{1}$$

where the response variable Y is coded 1 for event and 0 for non-event, and X is the independent variable. The odds of an event is defined as $\frac{p(\text{events})}{p(\text{non-events})}$, p represents the probability of event and is defined as $\frac{\text{number of events}}{total(\text{events,non-events})}$, α is the intercept, β is the regression coefficient of X, and $e = 2.71828$ is the base of the system of natural logarithms. The odds of a non-event is then $\frac{p(\text{non-events})}{p(\text{events})}$. The probability of a non-event, $p(\text{non-event})$, is $1 - p = \frac{\text{number of non-events}}{total(\text{events,non-events})}$, hence $p(\text{event}) + p(\text{non-event}) = 1$. The odds of an event, odds(event), is the reciprocal of the odds of a non-event, and thus odds(event) multiplied by odds(non-event) is equal to 1. An odds ratio, which is a measure of effect in LR, is a quotient of two odds and is used to compare the two odds. An odds ratio greater than 1 indicates an increased likelihood of an event, while an odds ratio of less than 1 indicates a decreased likelihood of an event (Lottes, DeMaris and Adler, 1996). Taking the antilog on both sides of (1), the logistic regression equation to predict the probability of the outcome of interest given x, a specific value of X, becomes a nonlinear relationship between the probability of Y and X, i.e.,

$$p = P(Y = \text{event of interest}|X = x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$

Extending the above logic to multiple predictors, the expression for logistic regression given a vector of predictors X_1 to X_k is thus,

$$p = P(Y = \text{event of interest}|\mathbf{X}) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_k x_k}}. \tag{2}$$

The LR model (2) constrains the predicted probabilities to lie within the range [0,1] and it allows the predictors to have a diminishing effect at extreme values of the dependent variable (Lottes et al., 1996). Considering the right side of model (2), the exponential function is always non-negative and always falls between 0 and 1. However, while LR tells us if the customer will default or settle early, survival methods suggest not only if, but when customers will experience an event (Stepanova and Thomas, 2002).

2.2. Survival Analysis

Survival analysis comprise a pool of specialised methods used to analyse lifetime data. The response variable is time until an event occurs and/or time to censorship. Censorship is the unique feature of survival analysis where survival experience is partly known. Survival analysis dates back to life and mortality tables mainly used in actuarial science and demography from around the 17th century. It led to the true meaning of "survival" through mortality rates. According to Odd et al. (2009), the original life tables method was based on wide time intervals and large data sets. Around the 1950s Kaplan and Meier proposed an estimator of survival curves (Odd et al., 2009). They developed a method for short time intervals and smaller sample sizes as opposed to those used in the actuarial and demographic studies. The 20th century saw further developments in handling survival data. The survivor function, also referred to as the *reliability function*, denoted by $S(t)$, is the probability that a respondent survives beyond a specified time t. Survival probabilities at different time lags help in summarizing survival data (Kleinbaum and Klein, 2005). The expression for the survivor function is given by:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx.$$

Theoretically, $S(t)$ is a monotone decreasing probability function, that is: at $t = 0$, $S(t) = S(0) = 1$ and at $t = \infty$, $S(t) = S(\infty) = 0$. Thus, $S(t) \in [0,1]$. Therefore $S(t)$ is essentially a probability of surviving beyond time t. The hazard function, also called the *mortality rate* or *conditional failure rate*, is the measure of potential failure at time t given that the respondent has survived up to some time t. The hazard function $h(t)$ is a rate expressed as the ratio of $f(t)$ to $S(t)$ and it is not a probability. Hence $h(t)$ takes non-negative infinite values $[0, \infty)$. The hazard function is mathematically expressed as follows:

$$h(t) = lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)},$$

where Δt represents a small time interval. The hazard rate gives the key/primary information in survival analysis as it determines the occurrence and timing of events. The empirical Kaplan Meier (KM) survival estimator considers n_i, the number of observations in the risk set, and d_i, the number of subjects failing at failure time t_j. The KM survival estimator is expressed as follows:

$$\widehat{S}(t_j) \quad = \prod_{i=1}^{j} \left(\frac{n_i - d_i}{n_i} \right).$$

2.2.1. The Cox Proportional Hazards Regression Model

Cox (1972) developed a regression model which produces *adjusted survival curves* by including covariates in the computation of survival estimates. The Cox proportional hazards (PH) regression methodology has gained popularity because of its flexibility and use of a small number of assumptions to obtain the basic information required from survival analysis. The Cox model hazard function calculates the hazard at time t of a subject, adjusted for possible explanatory variables. The formula is expressed as the product of the baseline hazard function of time and an exponential function of covariates. The baseline hazard is an unspecified form of the Cox model and the distribution of the outcome (survival time) is unknown. This makes the Cox PH regression a **semiparametric** model. The semiparametric property of the Cox PH model makes it a robust model which can closely approximate parametric models (Devarajan and Ebrahimi, 2011). The Cox PH model hazard function is:

$$h(t, \mathbf{X}) = h_0(t) \times \exp\left[\sum_{i=1}^{p} \beta_i X_i\right],$$

where $\mathbf{X}$ is a vector of predictor variables $X_1, X_2, \ldots, X_p$, $h_0(t)$ is the **baseline hazard** which involves t only and no covariates, and $\exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$ is an exponential component of the model that involves time independent covariates $\mathbf{X}$. Time independent variables do not change over time, for example population group and nationality. In the absence of explanatory variables, the Cox PH model reduces to the baseline hazard $h_0(t)$. The Cox PH assumption states that the hazard for a subject is proportional to the hazard for another subject in the same study where the proportionality constant, say θ, is independent of time (Kleinbaum and Klein, 2005), i.e.,

$$\theta = \exp\left[\sum_{i=1}^{p} \beta_i (\mathbf{X}_i^{'} - \mathbf{X}_i)\right].$$

The Cox PH model is appropriate for use when the PH assumption is met. When the hazard ratios vary with time, for example where hazards cross or when time varying confounding variables are present, the PH assumption may be violated, making it inappropriate to use the Cox PH model. If the Cox PH assumption is not met, variations of the Cox model can be used, for example the **extended Cox regression** or the **stratified Cox regression** depending on the context.

There are various approaches used to evaluate the reasonableness of Cox PH assumption. These include, inter alia, the graphical approach, goodness of fit tests as well as the time dependant variables assessment. The graphical approach is the most widely used technique to evaluate the Cox PH assumption. Given a set of categorised or coarse classified covariates as the predictors in a Cox PH model, the estimated $-\log_e(-\log_e(S(t)))$ survivor curves over different categories of covariates are compared over time t. The PH assumption is satisfied when *parallel curves* are obtained for $-\log_e(-\log_e(S(t)))$ survivor curves of different categories of the same covariate. The $-\log_e(-\log_e)$ survivor curves are popularly known as the *log-log* plots. A *log-log* survival curve is a transformation that results from taking the natural logarithm of an estimated probability curve twice. That is,

$$-\log_e(-\log_e(S)) = -\log_e\left(-\log_e\left(\exp\left[-\int_0^t h(u)du\right]\right)\right),$$

where $\int_0^t h(u)du$ is the cumulative hazard function resulting from the formula for the relationship between survival curves and hazard function, that is, $S(t) = \exp\left[-\int_0^t h(u)du\right]$.

2.2.2. Competing Risks Analysis

Analysis of more than one event in the same study is a variation of survival analysis known as ***competing risks*** analysis. A single customer can only experience one of the events and not both or gets censored. In this case, censorship occurs when a customer neither defaults nor pays off early such that the event of interest is never observed (Stepanova and Thomas, 2002). Censored subjects in this case are "good" customers. The KM approach may not be used in the presence of competing risks as it becomes very sensitive and may produce biased results. The modelling methodologies ideal for competing risks include the Cox PH model, parametric survival models and the Cumulative Incidence Curve (CIC). The Cox PH regression is widely used to model competing risks. Where each event type is modelled separately and other event types are treated as censored categories, the approach is called the "**cause specific**" method, this can be seen in Stepanova and Thomas (2002). Consider default as event type 1 and ES as event type 2. The cause specific approach fits 2 separate Cox regression models, one for each failure type. Time until default $\mathbf{T}_1$ is determined, and the rest of the observed lifetimes are assumed to be censored, including the subjects who entered into the ES group. An ES model, analogous to the default model, is based on the estimated time until ES, $\mathbf{T}_2$. A Cox PH model is fit in each case on $\mathbf{T}_1$ and $\mathbf{T}_2$. From literature, the predicted lifetime of a loan is thus $\mathbf{T} = \min(\mathbf{T}_1, \mathbf{T}_2$, term of the loan). In this analysis the cause specific hazard functions of the two events are

$$h_1(t) = \lim_{\Delta t \to 0} P(t \le T_1 < t + \Delta t | T_1 \ge t)/\Delta t,$$

and

$$h_2(t) = \lim_{\Delta t \to 0} P(t \le T_2 < t + \Delta t | T_2 \ge t)/\Delta t,$$

where the random variable T_1 denotes time to failure from default, T_2 denotes time to failure due to early repayment. $h_1(t)$ and $h_2(t)$ give the instantaneous failure rates at t for default and early repayment respectively. In general, given c events in an analysis, the Cox PH cause specific model is given by:

$$h_c(t, \mathbf{X}) = h_{0c}(t) \exp\left[\sum_{i=1}^{p} \beta_{ic} X_i\right].$$

In this analysis when $c = 1$ a default event is modelled, and $c = 2$, a model for early repayment is obtained. $\mathbf{X} = (X_1, X_2, ..., X_p)$ is a vector of explanatory variables included in the study. The β_{ic}'s are event specific regressions parameters. Furthermore, in cause specific models, the probabilities are calculated using the CIC. It estimates the "marginal probability" of each event where competing risks operate together in the same study. The marginal probability for each event type c at failure time t_i is computed as follows:

$$CIC_c(t_i) = \sum_{i=1}^{i} \widehat{S}(t_{i-1}) \widehat{h}_c(t_i),$$

where $\widehat{S}(t_{i-1})$ is the overall survival probability estimate of surviving previous time t_{i-1}. This computes subjects surviving all competing risks. The hazard estimate $\widehat{h}_c(t_i)$ for event type c is the proportion of subjects failing from event c at time t_i.

2.2.3. Mixture Models of Survival

Standard survival methods assume the empirical survival curve levels off at zero as time goes to $+\infty$. If the survival curve levels off to non-zero proportions, then the standard methodologies may be inappropriate. Empirical survival curves may level off to non-zero proportions in cases where some subjects in a study are not susceptible to the event(s) of interest. That is, given an extended observation period, the bulk of accounts may never default nor settle early. These are called long term survivors. Approaches to modelling lifetime data in the presence of long-term survivors are called **cure models** or **mixture models**.

Literature on mixture models is found in the works of Farewell (1982) as well as Sy and Taylor (2000). Mixture or cure models are designed to cater for a sizeable proportion of subjects who do not experience the event of interest at the end of the observation period. "*A KM survival curve that shows a long and stable plateau with heavy censoring at the tail maybe taken as empirical evidence of a cured fraction*" (Sy and Taylor, 2000, p. 228). A mixture of two populations is considered, the susceptible, denoted population **A** and the non-susceptible (long-term survivors) population **B**. A binary indicator is added to distinguish between subjects falling in the two populations. Let $Y = 1$ if the account defaults/pay-off early eventually and $Y = 0$ otherwise. Define $p = Pr(Y = 1)$, t = time to an event of subjects only in population **A**. The proportion of **B** = $1 - p$. The survivor function of the entire population (**A** + **B**) is given by:

$$S(t) = (1-p) + pS_A(t),$$

where $S_A(t)$ is the survivor function of population **A**. As evidenced by the empirical KM curve in Figure 1, the overall survival plot levels off at non-zero values. This indicates that the bulk of accounts are not susceptible to the events of interest. It makes business sense as most of the customers on the vehicle finance book are good customers and statistically, that prompts heavy censoring at the end of the study. A proportion (p) of good customers is chosen such that the overall survival curve levels of to p. In this case the minimum value of the survival curve was selected as $p = 0.49179045$. For each model, the hazard function was derived from the survivor function, adjusted for p and the corresponding CICs were calculated.

3. Data Structure

The aim of this paper is to analyse competing risks and long-term survivors in a consumer credit context. Loan data were obtained from a leading South African financial institution. All the information required was extracted for the period 01 April 2009 to 31 March 2014. The information includes details from the applicants and vehicle manufacturers. A total of 293 807 accounts were considered, with repayment terms ranging from 48 to 72 months. The available data set was randomly split using simple random sampling into a development and a validation data set in the ratio 80:20 respectively (Migut, Jakubowski and Stout, 2013).

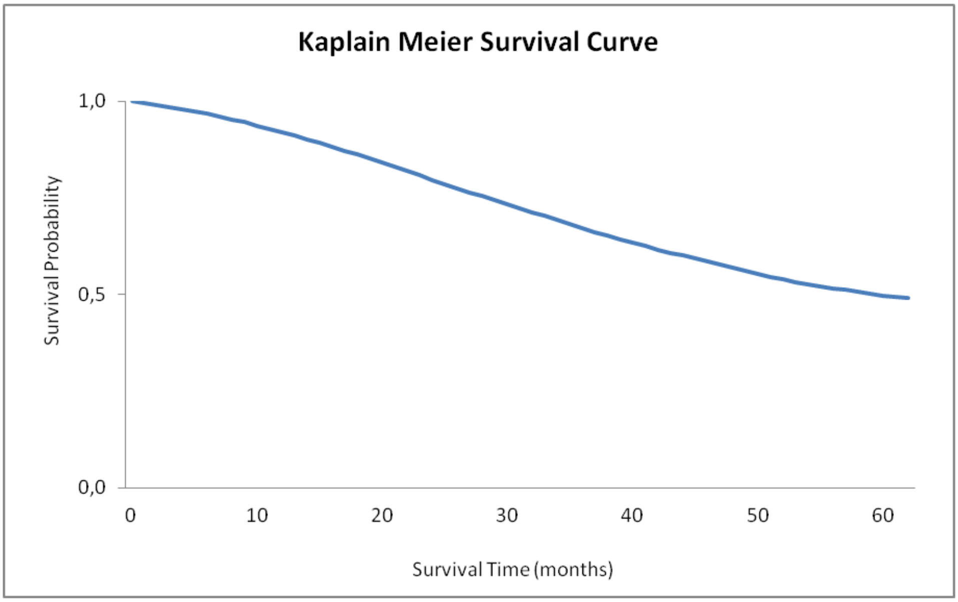


Figure 1: Kaplan Meier survival curve.

3.1. Univariate Analysis

Candidate covariates were selected and binned to ensure robustness of the models. The Weight of Evidence (WoE) transformation converts any variable into a numeric interval variable. Groups were assigned according to the risk of the group expressed by the logarithm of likelihood ratios, that is, a logarithm of a portion of, say, defaulted versus non-defaulted subjects (Jilek, 2008). WoE refers to the set of "goods" (customers who do not default), and the "bads" (customers who default or settle accounts early). The WoE (w_{ij}) is calculated as follows:

$$w_{ij} = \log_e \left(\frac{p_{ij}}{q_{ij}} \right),$$

where p_{ij} is the number of good risks in level/attribute j of variable/characteristic i divided by the total number of good risks who responded to i and q_{ij} is the number of bad risks in attribute j of characteristic i divided by the number of bad risks in attribute j who responded to characteristic i. The WoE curve should be a monotonic function across the categories of a covariate.

Gini Statistic (GS), also known as the Somer's D, measures a variable's ability to differentiate risk when fitting a univariate (single input variable) logistic model. It measures uniformity of a distribution. The lower the GS, the more uniformly distributed the variable (Jilek, 2008). In the consumer credit context the GS statistic is used to measure how equal the event rates are across the attributes of a variable. The higher the GS, the higher the ability of the characteristic to differentiate risk. To be consistent with the bank model building standards, all variables with a GS of less than 4 percent were excluded in the model development. In the calculation of the GS, the attributes $i = 1, 2, \ldots, m$ are sorted in ascending order of their event rates. For each of the attributes the number of events is given by n_i^{event}, and the number of non-event by $n_i^{\text{non-event}}$. The total number of events is denoted N^{event} and the total number of non-events is given by $N^{\text{non-event}}$. Then the GS is calculated as follows:

$$GS = \left(1 - \frac{2 \times \sum_{i=2}^{m} \left(n_i^{\text{event}} \times \sum_{j=1}^{i-1} n_j^{\text{non-event}}\right) + \sum_{k=1}^{m} \left(n_k^{\text{event}} \times n_k^{\text{non-event}}\right)}{N^{\text{event}} \times N^{\text{non-event}}}\right) \times 100.$$

Individual covariates were then assessed for PH assumption for each event type. All covariates satisfying the univariate conditions outlined above were considered for further analysis. An example is given in Figure 2.

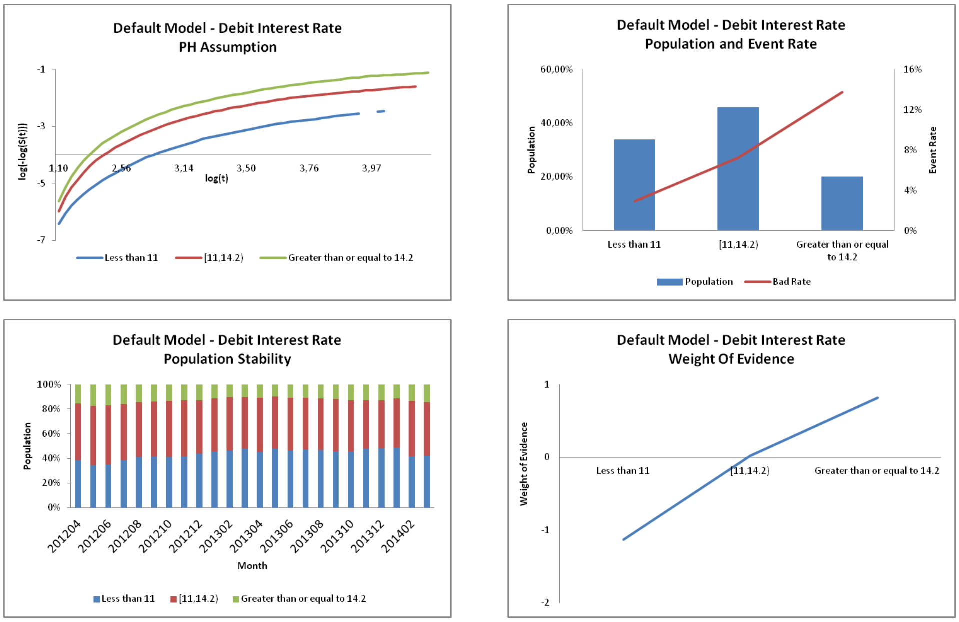


Figure 2: Univariate assessment plots – Default model.

The covariate Debit Interest Rate was assessed for all the univariate requirements discussed above. There is no evidence of crossing or overlapping hazards in the PH assumption plot. The lines are almost parallel, indicating that Debit Interest Rate satisfies the PH assumption. The population and event rate plot is satisfactory as each category has a population greater than 5 percent. The monotonic event rate and the WoE curves show that the variable has the ability to rank order. The population stability plot gives an intuitive assessment to show that there are no unreasonable trends in categories across the observation period. With all the conditions satisfied, Debit Interest Rate qualifies in the multivariate analysis stage for the default model.

3.2. Multivariate Data Analysis

This study uses stepwise regression to identify subsets of covariates befitting plausible models. Covariates are assessed in order to detect any multicollinearity that might be present using the Variance Inflation Factor (VIF) (Belsley, Kuh and Welsch, 2005), as well as correlation analysis to identify variables exhibiting pairwise correlation. It is imperative to assess multicollinearity among covariates before analysts conduct a multiple regression analysis (Mansfield and Helms, 1982). Highly correlated variables manifest in high VIF values. It is recommended that VIF values should lie below 3, if covariates are to be considered for model development. Any candidate covariates with a VIF greater than 3 were excluded from further analysis.

The final covariates selected for the default and ES models are detailed in Table 1 and Table 2 respectively.

Table 1: Final covariates – Default model.

Variable	Category	Description
Debit Interest Rate	1	Less than 11
	2	[11,14.2)
	3	≥ 14.2
Deposit to Loan	1	No deposit paid
	2	Deposit paid
Dwelling Type Code	1	Customer is a tenant
	2	Customer living with parents
	3	Customer owns a residential property
Equipment Category Code	1	Demo and Pre-owned vehicles
	2	Brand new Vehicles
Marital Status Code	1	The customer is married
	2	Single, Divorced, Widowed, Other

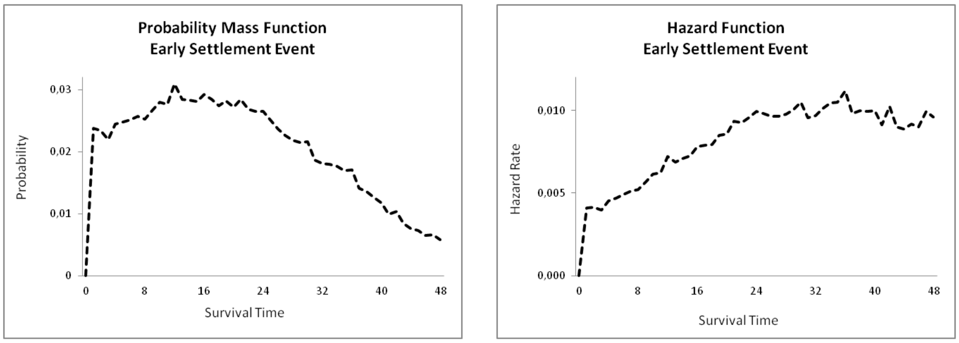
3.3. Empirical Hazard Functions

Survival analysis assumes that neither of the events can happen at the point of entry. Thus, the survival probability at time 0 is equal to 1 and conversely, the hazard function at time 0 is equal to 0. As the survival time in this study is discrete, we expect the events to start occurring at month 1 onwards. Figures 3 and 4 show the empirical probability mass (a) and hazard (b) functions for ES and default events respectively. For both events, the functions start at zero as there are no events recorded at the entry points. The probability mass function of ES is the number of accounts settling early at any t, from 0 to 48, relative to the total number of early settlements in the book. It is a function of the total number of ES events which shows the probability distribution of the early settlement event over time. The probability distribution functions were determined for both events. The hazard function is the instantaneous rate of occurrence of an event. This is a function of the accounts at risk at any t. The hazard functions were calculated and plotted for each event as well.

Table 2: Final covariates – Early settlement model.

Variable	Category	Description
Debit Interest Rate	1	Less than 12.55
	2	$\geq$ 12.55
Deposit to Loan Ratio	1	Less than 0.45
	2	$\geq$ 45
Dwelling Type Code	1	Customer is a tenant
	2	Owner or lives with parents
Equipment Category Code	1	Vehicle older than 5 years
	2	Demo vehicles or less than 5 years
	3	New and Light Utility Vehicles
Original Term	1	Tenure less than 60 months
	2	Tenure $\geq$ 60 months

The probability mass function in Figure 3 (a) increases steadily in the first 18 months, decreases at a steady rate beyond 18 months and diminishes towards 48 months.



(a)

(b)

Figure 3: Early settlement event – Hazard function.

This is also reflected in the hazard function in Figure 3 (b) that is on an increasing trend for the first 24 months and stabilizes thereafter, at around 0.07. For the first year from the entry point, the risk of early settlement increases with increasing time then more or less stabilises for the next year, then starts decreasing thereafter. As the vehicle ages, the chances of early settlement increases as customers upgrade to new vehicle models. However, as the accounts approach maturity, the risk of early settlement lessens as it becomes easier to complete the originally agreed term of repayment and customers opt to complete the repayment normally instead of settling early, to avoid penalty charges associated with the event.

With reference to Figure 4 (a), the probability mass function of the default event increases sharply in the first 18 months of the loans.

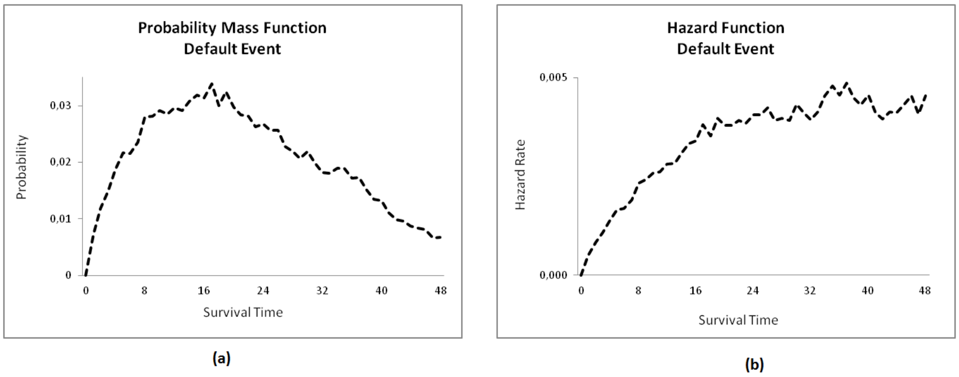


Figure 4: Default event – Hazard function.

The same trend is reflected in the corresponding hazard function in Figure 4 (b) which increases sharply up to 18 months and generally stabilises beyond the 18-month point of 0.004. This is due to the fact that at the point of application, the selected customers have low risk of default but, as time goes on, customers experience various social and economic events leading to default and the risk of default increases. This trend is experienced in the first one and half years of loans for the vehicle finance product. As the accounts grow older than 18 months, the rate of default decreases, accounts passing this point have a lower chance of default. This is attributable to the fact that most customers improve their financial status with time and the original repayment amount becomes insignificant with time and hence the chance of default diminishes as time approaches 48 months. Also note that the trajectory is different between ES and default. For ES, the probability of early settlement starts high at 0.023 in month 2 while for default it gradually increasing starting from around 0.003 in month 2.

4. Model Fitting

The baseline categories were manually selected as opposed to automatic selection. This was done to optimise the volumes in the baseline to ensure maximum statistical significance for as many variables as possible. For LR, in each case, accounts used for model development were allowed at least 48 months to perform. This is the "waiting" performance period to maturity. The account-level probabilities were calculated based on the event observed at month 48. Both LR models were fitted satisfactorily.

Unlike the LR development data selection process, the Cox regression model does not consider a "waiting" period. All accounts are eligible for inclusion in the development. Accounts recently entered into the study also play a very important role. If not absorbed into any of the events then they can be classified as censored and form part of the risk set according to time spent in study. Two cause specific PH models were fitted separately for each event (ES and default). A CIC was calculated in each case to determine the marginal probabilities of an event at a fixed workout period of 48 months in the presence of competing risks and long term survivors.

4.1. Goodness of Fit Measures

Following the model fitting process, the LR models were tested for goodness of fit using the Hosmer and Lemeshow test. It measures how well predicted events align with the observed events. The population is subdivided into decile groups according to the probabilities. Each decile group is compared on the expected versus observed events. Low values of the Hosmer and Lemeshow statistic and high p-values (greater than 0.05) indicate a good fit of the observed versus predicted event rates. For the default model, the Chi-square statistic of 15.2447 was obtained, with a corresponding p-value of 0.0546 see Table 3. This indicates that this model does fit the data.

Table 3: Default model: Hosmer and Lemeshow Goodness-of-Fit test.

Chi-Square	DF	Pr > ChiSq
15.2447	8	0.0546

For the ES Logistic regression model, a Chi-square statistic of 3.0120 was obtained, associated with a high p-value of 0.9336, see Table 4.

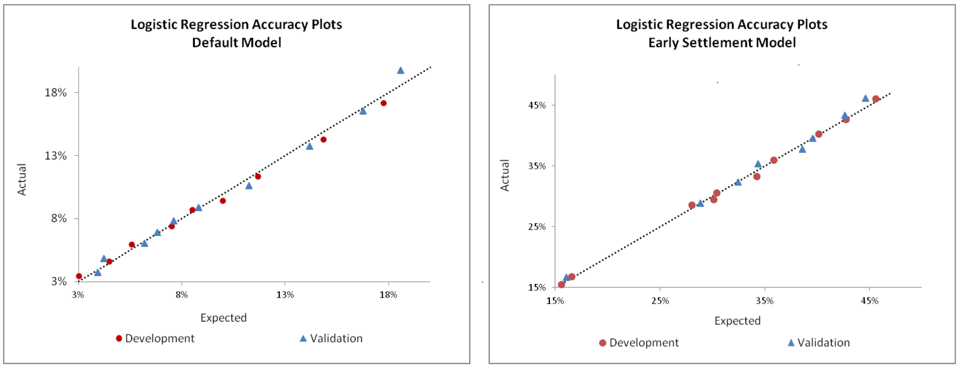
Table 4: ES model: Hosmer and Lemeshow Goodness-of-Fit test.

Chi-Square	DF	Pr > ChiSq
3.0120	7	0.9336

This indicates that the observed and expected ES rates are similar by population deciles and the model fits very well. The Hosmer and Lemeshow test partitions data into decile groups according to risk levels. For each risk group, the actual versus observed event rates were calculated based on the total population. This was done to determine the ability of the model to rank order risk and to establish how accurate the model is in predicting risk.

Thus, the values of actual and expected event rates were plotted across the range of risk. For the models to be accurate, the actual versus expected plots should not deviate significantly from the 45 degree diagonal. To check the ability of the model to rank order risk, the points should lie in increasing order of the risk group. Accuracy and rank ordering metrics were determined for both development and validation data sets. The accuracy plots are provided in Figure 5 (a) and (b). All the points in both data sets are satisfactorily rank ordered and they all lie close to the 45 degree diagonal.

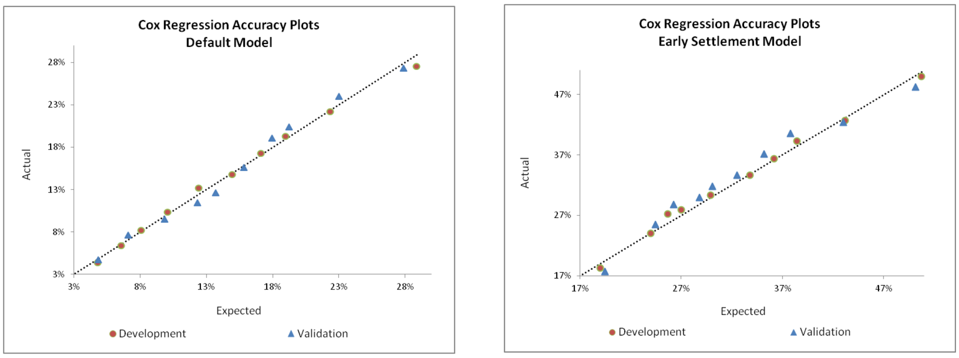
The accuracy measures were developed for the Cox PH models per event type. This was performed on both the development and validation data sets. Accounts were ranked separately for each data set into deciles based on their default cumulative probabilities at month 48. For each decile group the actual and expected observations were determined. The default rate in each group was determined based on the total volumes. For the models to be accurate, the actual versus expected



(a) (b)

Figure 5: Accuracy plots – Logistic models.

plots should not deviate significantly from the 45 degree diagonal. The accuracy plots for the default and ES models are provided in Figure 6 (a) and (b) respectively. Both of the models predicted the events well over the range.



(a) (b)

Figure 6: Accuracy plots – Cox PH models.

4.2. The Competing Risk Approach

The default model was constructed independently with the ES event treated as a censored event. Similarly the ES model was constructed with the default treated as a censored event. In the consumer credit risk context, these events co-exist based on which event occurs first. At the portfolio level these two risk events "compete" to absorb accounts.

With the logistic regression methodology fitted above, the (ES/default) models remain independent as they directly produce the probability of an event. The proportional hazards model on the other hand produces the hazard functions first, then makes use of the CIC to accommodate or to allow for the presence of the other event in a competing risks scenario.

5. Model Comparison and Validation Metrics

Survival analysis methods highlight the evolution of the target variable (default, early settlement) over time. This is reflected in Figures 3 and 4 wherein the hazard curve is plotted against each survival time point. This is not obtainable from a logistic approach. Logistic regression, as used in this study, enforces subjects to be observed in a fixed horizon before inclusion in the analysis and therefore the evolution of events is not clearly defined over time.

For any subjects whose maximum observation period falls short of the workout period, the subject is disqualified and discarded in logistic regression but included in survival methods as a censored observation. This implies that proportional hazards method uses more information and thus more stable estimates are obtained.

In the South African consumer credit context survival methods have only been recently introduced. The conventional modelling methodologies which include empirical roll rates and logistic regression fell short by failing to detect early warning signs and symptoms prior to recession and under-performance in the global banking system in 2008/2009 leading to national distress and disorientation. Babajide, Olokoyo and Adegboye (2015) successfully used a proportional hazards model on Nigerian bank data to determine how bank failure can be predicted far ahead of its occurrence. This helps financial institutions with more vigilant lending strategies and meet financial obligations accordingly, as they fall due.

Logistic regression with a logit link function has been the cornerstone in financial models. However, the symmetrical sigmoidal logistic function may not be achieved with heavy censoring due to the presence of long term survivors. Because of the asymmetrical shape of the logistic function, the use of the logit function is potentially compromised. The introduction of survival models with cure/mixture models corrects for the presence of long term survivors.

Due to its ability to incorporate censored observations, survival methods make use of the most recent information, there is no need to enforce subjects to a specific horizon period to perform before inclusion in the study. To this effect, survival methods are reactive to changes in the book construct as the models capture real time changes as they occur.

The Receiver Operating Characteristic (ROC) curves and area under the ROC curve (AUC), metrics were used to compare performance of Logistic versus Cox PH regression models in a consumer credit setting. A discussion of each model variant under Section 4.1. With particular reference to Figures 5 and 6, reflecting accuracy of the default Logistic, ES Logistic, default Cox and ES Cox models respectively, it is evident that all models were satisfactorily accurate with each having the ability to rank order risk. Table 5 supports the notion. This is detailed in Section 5.1 and Section 5.2.

5.1. The ROC Curves

The ROC curve plots the sensitivity (true positive rate) against 1-specificity (true negative rates) of the models at various cut-off values of risk. For the default event, sensitivity refers to a fraction of accounts in default that the model correctly identifies as defaulted. The same goes for the ES event.

Specificity refers to a fraction of accounts not in default that the model correctly identifies as not in default. The ROC curves for the LR and Cox PH regression models are provided in Figure 7.

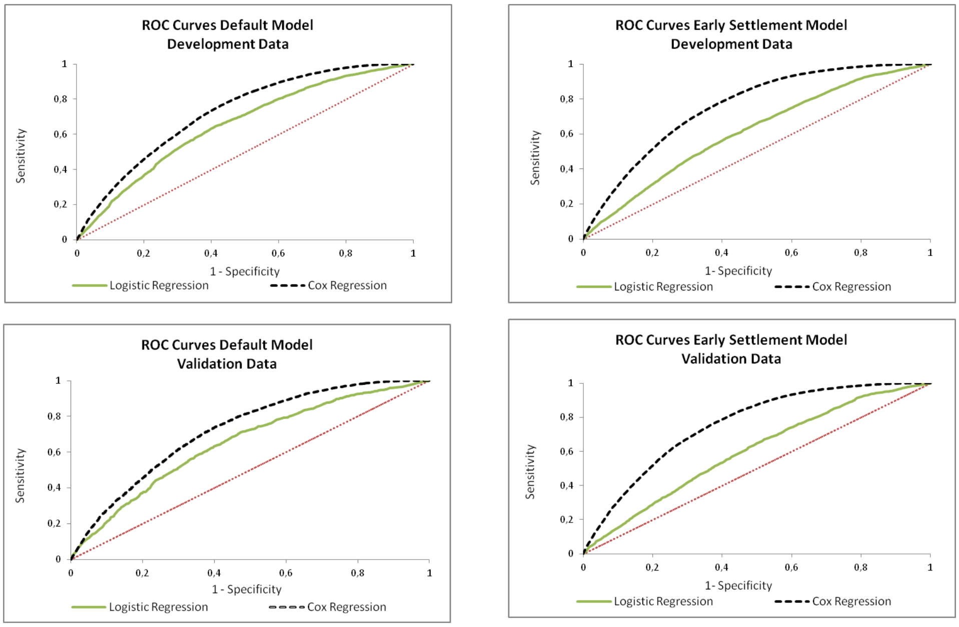


Figure 7: ROC curves.

The vertical axis represents sensitivity and the horizontal axis represents 1- specificity values at each cut-off point. Both axes range from 0 to 1. The diagonal divides the ROC Cartesian plane. Curves above the diagonal line represent good classification model whereas points below the line represent poor results. Points along the diagonal represent a random model. In this case all the curves lie above the diagonal indicating that good classification models were developed in this study. When comparing models, a better model is the one whose ROC curve lie closer to the upper end of the ROC space. In both models, it is clearly seen that Cox PH models perform better than the LR models in the development and validation data. For this particular sample and bank, the data suggest that it is better to use Cox regression than LR in a lifetime data analysis. The Cox PH model also has the advantage of describing the evolution of the hazards over time.

5.2. The Gini Statistic and the Area Under the ROC Curve

The overall model Gini Statistics (GS) as well as the corresponding overall Area Under the Curves (AUC) were calculated for each model as a generalised measure to quantify the ability of the models to differentiate risk. The results are given in Table 5 for each model.

The lower GSs for LR models are attributable to the use of older vintages for development as the accounts should be allowed a sufficient performance period (48 months in this case) before they can be considered for modelling. The fact that the overall GS for the Cox PH models are higher, suggest

Table 5: Area under ROC curves and Gini statistic.

Model	Default		Early Settlement	
Statistic	AUC	GS	AUC	GS
Cox Development	72.40	44.78	75.60	51.27
Cox Validation	72.30	44.50	75.59	51.17
Logistic Development	65.60	31.10	61.20	22.40
Logistic Validation	65.50	30.90	60.20	20.40

that Cox PH performs better than LR. The Cox PH model strength is enhanced by the inclusion of censored observations and the use of the most recent data in Cox PH regression.

The lower AUCs for LR models are also attributable to the use of older vintages and are in agreement with the Gini statistic. The closer the area is to the perfect model of area = 1, the better the model. The AUC can be represented by the overall model GS values and it has been stated that the Cox models have higher GS and subsequently higher AUC compared to LR models. Comparing the logistic model AUCs in Table 5, the default model is estimated better using Logistic regression than the ES model. This is also reflected in the ROC plots in Figure 7. The Early Settlement ROC plots lie closer to the diagonal line as compared to the default model curves.

For completeness the standard errors of GSs were calculated for each model, as it may be useful for assessing the reliability of models (Greene and Milne, 2010). The standard deviations were calculated by re-sampling 100 samples and calculating the GS standard errors. The means and standard deviations as well as the 95% confidence intervals of the GSs are shown in Table 6. Greene and Milne (2010) showed that a 30 sample resampling approach produced estimates of the Gini statistic and standard deviation which were similar to the more complicated ordinary Least squares based estimates, which means the 100 samples should give reasonable results.

Table 6: Re-sampled Gini statistic 95 % confidence intervals.

Model	Default			Early Settlement		
Statistic	GS	std dev	95% CI	GS	std dev	95% CI
Cox Development	44.76	2.12	(44.34, 45.18)	51.26	1.65	(50.94, 51.58)
Cox Validation	44.43	2.74	(43.89, 44.97)	51.15	1.81	(50.80, 51.50)
Logistic Development	31.11	5.26	(30.08, 32.14)	22.41	6.13	(21.21, 23.61)
Logistic Validation	30.88	5.78	(29.75, 32.01)	20.39	7.21	(18.98, 21.80)

The 100 sample estimates of the GS 95% confidence intervals for the Cox models do not overlap with the logistic models suggesting the two models differ significantly. The fact that the overall GS for the Cox PH models are higher with lower standard errors, suggest that Cox PH performs better than LR. The Cox PH model strength is enhanced by the inclusion of censored observations and the use of the most recent data in Cox PH regression.

6. Summary and Conclusion

In this paper we analysed competing risks in a consumer credit context with two events of interest, default and early settlement. These events were modelled using statistically sound techniques. The bulk of accounts under investigation were not susceptible to the events of interest. The data typically had long term survivors with heavy censoring at the end of the observation period. The data structure is analogous to lifetime data in other domains of study such as engineering and biomedical research, making the statistical methodologies versatile. Models were adjusted to accommodate long term survivors. The performance of the models was compared using overall model receiver operating characteristic curves. Clearly Cox regression outperforms Logistic regression as evidenced by higher Gini statistics and better receiver operating characteristic curves in both default and early settlement models for this dataset. This analysis was conducted in SAS®.

In all models, there is strong empirical support for the results as evidenced by the actual versus predicted analyses. The models predicted and correctly classified events in the validation set. The models can be used to determine and compare survival prognosis of different risk groups in a consumer credit context. However, LR uses older vintages in model building, therefore it becomes more difficult to capture the most recent activities as the dependant variable in LR is binary and does not consider time. The use of survival methods to model credit risk data is motivated by the existence of lifetime loans which can be observed from the point of origin to an event of interest. Survival methods thus, estimate not only if, as in Logistic regression, but also when borrowers will default. This enhances flexibility as the model generates probabilities of each event happening at various points in time. For any given observation period, some customers default and some pay-off earlier than the originally agreed term. Where the event occurs before the end of the observation period, the lifetime of such credits are observable. For customers who do not default or pay-off early, before the end of the observation period, it is not possible to observe the time instant when the event occurs.

Censoring allows the response variable to be incompletely determined for some accounts. Unlike in conventional statistical methodologies, censored accounts are not discarded in survival analysis, but contribute information to the study. Censoring is the defining feature of survival analysis, making it distinct from other kinds of analysis. Logistic regression in particular tends to ignore censoring information. The response variable is binary and it should be fully observable. Although in terms of predictive performances the models are substantially similar, survival analysis gives more valuable information such as a whole predicted survival function rather than a single predicted survival probability. Survival analysis is superior to Logistic regression in that, a better credit granting decision is made if supported by the estimated survival times.

References

BABAJIDE, A. A., OLOKOYO, F. O., AND ADEGBOYE, F. B. (2015). Predicting bank failure in Nigeria using survival analysis approach. *Journal of South African Business Research*, **2015**.

BELLOTTI, T. AND CROOK, J. (2007). Credit scoring with macroeconomic variables using survival analysis. *International Journal of Bank Marketing*, **54**, 276–278.

BELSLEY, D. A., KUH, E., AND WELSCH, R. E. (2005). *Regression Diagnostics: Identifying*

Influential Data and Sources of Collinearity, volume 571. John Wiley & Sons: Hoboken, New Jersey.

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34** (2), 187–220.

DEVARAJAN, K. AND EBRAHIMI, N. (2011). A semi-parametric generalization of the Cox proportional hazards regression model: Inference and applications. *Computational Statistics & Data Analysis*, **55** (1), 667–676.

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

GREENE, H. J. AND MILNE, G. R. (2010). Assessing model performance: The Gini statistic and its standard error. *Journal of Database Marketing & Customer Strategy Management*, **17** (1), 36–48.

JENKINS, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK.*

JILEK, O. (2008). Mathematical applications in credit risk modelling. *Journal of Applied Mathematics*, **1** (1), 432–438.

KLEINBAUM, D. G. AND KLEIN, M. (2005). *Survival Analysis. A Self-learning Approach.* Springer: New York, USA.

LOTTES, I. L., DEMARIS, A., AND ADLER, M. A. (1996). Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology*, **24** (3), 284–298.

MANSFIELD, E. R. AND HELMS, B. P. (1982). Detecting multicollinearity. *The American Statistician*, **36** (3a), 158–160.

MIGUT, G., JAKUBOWSKI, J., AND STOUT, D. (2013). Developing scorecards using STATISTICA Scorecard. Technical report, Statsoft Polska/Statsoft Inc. http://documentation.statsoft.com/portals/0/formula%20guide/STATISTICA%20Scorecard%20Formula%20Guide.pdf.

ODD, O. A., ANDERSEN, P. K., BORGAN, Ø., GILL, R. D., AND KEIDING, N. (2009). History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, **5** (1), 1–28.

STEPANOVA, M. AND THOMAS, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, **50** (2), 277–289.

SY, J. P. AND TAYLOR, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56** (1), 227–236.

XUE, X., KIM, M. Y., GAUDET, M. M., PARK, Y., HEO, M., HOLLENBECK, A. R., STRICKLER, H. D., AND GUNTER, M. J. (2013). A comparison of the polytomous logistic regression and joint Cox proportional hazards models for evaluating multiple disease subtypes in prospective cohort studies. *Cancer Epidemiology Biomarkers & Prevention*, **22** (2), 275–285.

Manuscript received, 2015-10-23, revised, 2016-07-06, accepted, 2016-07-16.