SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Nonparametric Modeling of Clustered Customer Survival Data

by

Iris Ivy M. Gauran
and
Erniel B. Barrios

UPSS Working Paper No. 2012-05
July 2012

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

Iris Ivy M. Gauran
University of the Philippines Diliman

Erniel B. Barrios
University of the Philippines Diliman

We postulate a nonparametric regression model to characterize clustered survival data, i.e., a random clustering effect is incorporated into the nonparametric version of Cox Proportional Hazards model.

There is evidence from the simulation study that clustered survival data can be better characterized though a nonparametric model. Predictive accuracy of the nonparametric model is affected by number of clusters and distribution of the random component accounting for clustering effect. As the functional form of the covariate departs from linearity, the nonparametric model is becoming more advantageous over the parametric method. Furthermore, the nonparametric model is better than the parametric model when the data is highly heterogeneous and/or there is misspecification error.

Keywords: *Survival Analysis; Clustered Data; Nonparametric Regression; Backfitting Algorithm; Random Effects; Generalized Additive Models*

1. Introduction

Interest towards Customer Relationship Management (CRM) began to grow in 1990s (Ling and Yen, 2001). CRM is a comprehensive strategy and process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer. It involves the integration of marketing, sales, customer service, and supply ---chain functions of the organization to achieve greater effectiveness in delivering customer value (Parvatiyar and Sheth, 2001). The critical role of CRM is to understand the dynamic

relationship between the client and the company since a developed relationship between the two can result in greater customer loyalty and profitability (Ngai, 2005).

On the other hand, customer churn is the focal concern of most companies because it figures directly on how long a customer stays with a company. Chandar, et. al. (2006) defined customer churn as the propensity of customers to cease doing business with a company in a given time period. In order to address this problem, companies must recognize the churners before they churn, thereby developing a model that predicts future churners is essential.

From the context of temporal dependencies in the dynamics of the phenomenon, the churn model is postulated from the theory of survival analysis where the time until the occurrence of a well-defined event is modelled. The event of interest is churn, and the model characterizing the time until churning is interpreted as the survival period. We formulate a nonparametric survival function embedded with random intercept term to account for clustering effect. This is carried out in the model building and estimation framework based on the Cox Proportional Hazards Model. A random term is added into the nonparametric survival function to account for clustering of data. The result is an additive model that is estimated using the backfitting algorithm,

Dubbed as a rapidly growing market in the early 2000s, the mobile telecommunications sector has already reached the state of saturation and fierce competition in many countries. Recently, the growth of mobile subscribers has slowed dramatically because the mobile penetration is approaching 100%. Thus, there is an observable shift in the focus of the telecommunications companies from customer acquisition to customer retention since the cost of acquiring new customers is higher than retaining existing customers. Upon

consideration of the churn rate of different industries, the telecommunications industry is one of the main targets of this hazard function, the churn rate in this industry ranges from 20 to 40 percent annually (Berson, et. al. 1999; Madden, et. al. 1999). Hence, it is important for companies to come up with models for customer churn prediction to be able to develop strategic promotional and other customer activation campaigns to mitigate the possibility of churning customers. Furthermore, the postulated model can be used as a component in the computation of the Customer Lifetime Value (defined as the total value gained by the company while the customer is still active). The method from this paper is useful for decision-makers of companies who offer non-contractual products and services.

2. Some Modeling Approaches

Survival analysis covers a collection of statistical techniques which model time-to-event data (Miller, 1981). It can be used for analyzing data representing lifetimes, waiting times or generally the occurrence of a well-defined 'event'. Kalbfleisch and Prentice (1980) notes that most methods for survival analysis are based on the following assumptions: (1) Given that all has happened up to time t, the failure mechanisms for different individuals act independently over the time interval $[t, t+dt)$ and; (2) For an individual alive and uncensored at t, the conditional probability of failing in $[t, t+dt)$ given all that has happened up to time t coincides with the conditional probability of failing in $[t, t+dt)$ given survival up to time t.

The objective of survival analysis is to draw inferences about the distribution of the survival times $T \geq 0$, the random variable that denotes the time at which the event occurs. Let $f(t)$ and $F(t)$ be the density and distribution function, respectively. The Survival Function is defined as

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

$S(t)$, a monotone decreasing function, is the probability that a subject will survive up to time t, The hazard rate function $\lambda(t)$ is defined as

$$\lambda(t) = \frac{\lim_{\Delta t \to 0} P(t < T < t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{1 - F(t)}.$$

The hazard function is also known as the instantaneous failure rate and age-specific failure rate. The component $\lambda(t)\Delta t$ can be interpreted as the instantaneous probability of having an event at time t given that one has survived (i.e. not had an event) up to time t. The functions $f(t), F(t), S(t)$ and $\lambda(t)$ give mathematically equivalent specifications of the distribution of T.

We consider the Cox Proportional Hazards Model model in this paper. It is the most popular survival regression model available since it does not makes any assumptions on the survival function as opposed to accelerated failure time models. The Cox Proportional Hazards Model specifies the hazard rate as $\lambda_{i,t} = \lambda_{0,t} e^{X_i \beta} = \lambda_{0,t} \times \exp(\beta_1 X_1 + \ldots + \beta_p X_p)$ and it assumes that the hazard for the ith individual at time t is the product of two factors: a baseline hazard function that is an unspecified nonnegative function over time and a linear combination of a set of p fixed covariates, which is exponentiated (Cox, 1972). The baseline function can be regarded as the hazard function for an individual whose covariates all have values 0. Also, the baseline hazard function can be interpreted as the average hazard at time t. In contrast to accelerated failure time model, the baseline hazard does not have to assume a particular distribution. This is a significant advantage over parametric survival models which are restricted by the shape of a particular distribution. The hazard rate at time t is thus the product of a scalar, $e^{X_i \beta}$ and the baseline hazard at time t. Equivalently, the covariates increase or decrease the hazard function by a constant relative to the baseline hazard function.

The Cox model is also referred to as the proportional hazard model since the hazard ratio for two subjects with covariate vector X_i and X_j, given by $\frac{\lambda_{i,t}}{\lambda_{j,t}} = \frac{\lambda_{0,t} e^{X_i \beta}}{\lambda_{0,t} e^{X_j \beta}} = \frac{e^{X_i \beta}}{e^{X_j \beta}}$, is constant over time. This implies that the covariates must have the same effect on the hazard at any point in time. It could also be noted that only time-independent covariates can be used in the model.

The nonparametric regression relaxes the usual assumption of linearity in the models above, instead, replacing it with the weaker assumption of a smooth regression function $f(X_1,\ldots X_p)$. The primary goal of nonparametric regression is to estimate the regression function between the response variable and the predictors directly instead of going through the estimation of the parameters of the model. This allows greater flexibility in uncovering the data structure that could be missed in a classical parametric regression. In building nonparametric regression model, smoothers summarizing the trend of a response measurement Y as a function of one or more predictor measurements $X_1, X_2, \ldots, X_p$, are used. Smoothing splines arise as the solution of the function $\hat{f}(X)$, which has continuous derivatives (up to at least second order) that minimizes the penalized sum of squares,

$$SS^*(h) = \sum_{i=1}^{n} [Y_i - f(X_i)]^2 + h \int_{x_{\min}}^{x_{\max}} [f''(x)]^2 dx$$

where h is a smoothing parameter. Smoothing splines enable a balance between fitting the data and avoiding fluctuations in the estimate by imposing a roughness penalty.

Nonparametric regression is not necessarily a universal solution to any modelling problem since it is difficult to implement when there many predictors to be included. Stone (1985) proposed a more restrictive method referred to as additive models. The additive

regression model given by $Y = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) + \varepsilon$ assumes that the functions $f(\cdot)$ are smooth and are not necessarily parametric in form. The effects of predictors can be separated into additive terms. This model is substantially more restrictive than the general nonparametric regression model, but less restrictive than the linear regression model which assumes that the functions $f(\cdot)$ are linear.

Hastie and Tibshirani (1986) discussed a generalization of the additive model, (Buja, et al, 1989) discussed further details along with some asymptotic results. Several approaches are proposed in fitting additive models as summarized by Mammen and Park (2006). These include classical backfitting, marginal integration method, smooth backfitting estimate and the local quasi-differencing approach. (Buja, et al, 1989) used the linear smoothers in estimating an additive model through the backfitting algorithm.

The classical backfitting by Hastie and Tibshirani (1990) used in this paper adopts the following algorithm:

(1) Initialize: $\alpha = E(Y), f_1^1 = f_2^1 = \ldots = f_p^1 = 0, m = 0$

(2) Iterate: $m = m + 1$

For $j = 1, \ldots, p,$ do $f_j = Y - \alpha - \sum_{k=1}^{j-1} f_k^m(X_k) - \sum_{k=j+1}^{p} f_k^{m-1}(X_k)$

$f_j^m = E(f_j \mid X_j)$

(3) Continue (2) until convergence.

(Buja, et al, 1989) showed that backfitting algorithm is the Gauss-Seidel iterative method of solving the normal equations of the additive model and that it is consistent and convergent for a class of smoothers including the cubic splines. Mammen, et al (1999) and (Opsomer, 2000) identified more general conditions under which the backfitting algorithm converges and produces consistent estimates.

3. Methodology

The churn prediction problem is formulated in the context of survival analysis. The origin of time corresponds to the commitment date of the customer. It is assumed that all of the observations chosen have the same commitment date for ease of administration of the follow-up period as well as comparability. The time scale is assumed to be discrete. Also, analysis of survival data assumes that the failure mechanisms for different individuals act independently over the time interval $[t-1, t)$, given survival up to time t - 1.

Survival/churn data are assumed to occur in clusters. Customers belonging to the same workplace, the same area, the same families, or even to the same circle of "friends", constitutes the clusters. Customers in the same clusters have the tendency to behave and act similarly relative to the product/service they subscribe to. A churn of an element within the cluster increases the likelihood of churn among other elements within the same cluster.

The interest is to explain the probability that the i^{th} observation in the k^{th} cluster will churn at time t in terms of $X_i^{(k)}$, a survival model that incorporate the influence of the covariates is employed. The data are represented as clustered survival information to take into account the homogeneity of the customers in the same cluster. However, in most of the survival data models described in the literature, heterogeneities between individuals have been taken into account only in the form of the observable covariates. Thus, to be able to model clustered survival data, the Cox Proportional Hazards Model was modified to include a cluster-specific random component.

The literature of survival analysis clearly indicates that parametric models will not be optimal when the data available is a composite of different clusters as in the case of data from telecommunication customers. Nonparametric methods are explored for better

characterization of the sector described above. If the functional form of the effect of $X_i^{(k)}$ on $h_i^{(k)}(t)$ is relaxed, it could possibly account for the varying effect of the covariates across the clusters.

3.1 Postulated Model

Given n clusters with m elements observed for T points, the postulated model is

$$\log h_i^{(k)}(t) = \lambda_k + \log \lambda_i^{(k)}(t) + f(X_i^{(k)}) + \varepsilon_i(t) \qquad (1)$$

where: i = 1, 2, …, m ; t = 1, 2, …, T

n is the total number of clusters

$h_i^{(k)}(t)$ is the i[th] value of the response variable within the kth cluster on time t

λ_k is the cluster-specific random intercept

$\lambda_i^{(k)}(t)$ is the baseline risk of the i[th] observation within the k[th] cluster at time t

$X_i^{(k)}$ is the value of the covariate of the i[th] observation within the k[th] cluster

$f(X_i^{(k)})$ is the smooth function of $X_i^{(k)}$ (nonparametric)

Assumptions:

1. $h_i^{(k)}(t)$ is a continuous response variable such that $h_i^{(k)}(t) = \lambda_i^{(k)}(t)\exp(f(X_i^{(k)}) + \lambda_k)$, $h_i^{(k)}(t) \in [0,1]$, the probability that the i[th] observation belonging in the k[th] cluster will churn on time t given that the customer survived until time $t-1$.
2. $X_i^{(k)}$ is a qualitative variable (represented by dummy variables) or quantitative variable (continuous or discrete). The postulated model can be extended to more than one covariate so that the model becomes

$$\log h_i^{(k)}(t) = \lambda_k + \log \lambda_i^{(k)}(t) + f(X_{i1}^{(k)}) + f(X_{i2}^{(k)}) + \ldots + f(X_{ip}^{(k)} + \varepsilon_i(t)$$

or similarly,

$$y_i^{(k)}(t) = \lambda_k + \log \lambda_i^{(k)}(t) + f(X_{i1}^{(k)}) + f(X_{i2}^{(k)}) + \ldots + f(X_{ip}^{(k)} + \varepsilon_i(t) \qquad (2)$$

where $y_i^{(k)}(t) = \log h_i^{(k)}(t)$ and $f(X_{i1}^{(k)})$, $f(X_{i2}^{(k)})$,..., $f(X_{ip}^{(k)})$ are smooth functions of the explanatory variables. Since Equation (1) is an additive model, additional predictors can be interpreted as adding more terms. Thus, without loss of generality, we consider only one covariate into the model.

3. λ_k is the cluster-specific random variable
4. $\lambda_i^{(k)}(t)$ is the baseline risk of the ith individual within the kth cluster at time t. The expression $\log \lambda_i^{(k)}(t)$ is equal to the sum of the function $\beta e^{-\beta t}$ and $\delta_i(t)$, the error term.
5. Clusters are independent. The clusters are assumed to be independent, mutually exclusive and mutually exhaustive. This means that each observation must belong to exactly one cluster. Cluster independence also implies that only the elements within the cluster can possess dependencies but elements between clusters are independent.
6. Cluster sizes are equal.
7. The functional form of the covariate $f(X_i^{(k)})$ is allowed to vary across clusters.
8. The information about the contributions of the term (i.e. existence of a dominating term or not) in the model are available at the onset of the study.

3.2 Estimation Procedure

Since the formulated model has additive components, the backfitting algorithm is used to sequentially estimate the compoents of Model (3).

$$\log E[h_i^{(k)}(t)] = \lambda_k + \log \lambda_i^{(k)}(t) + f(X_i^{(k)}) \qquad (3)$$

The modified Cox Proportional Hazard Model was transformed into an additive model so that the components can be estimated separately with the appropriate estimation method. The three components of the model are estimated one at a time, i.e., the most important term comes first. Backfitting algorithm provides good estimates among the model terms estimated early on (Santos and Barrios, 2012). The estimation procedure is summarized as follows:

1. In the case wherein the covariate dominates the postulated model, λ_k and $\log\lambda_i^{(k)}(t)$ are ignored first and $f(X_i^{(k)})$ is estimated nonparametrically using smoothing splines.
2. The partial residual $\hat{e}_i$ is computed as $(y_i^{(k)}(t) - \hat{f}(X_i^{(k)}))$. At this point, the partial residual contains information on λ_k and $\log\lambda_i^{(k)}(t)$.
3. Ignore $\log\lambda_i^{(k)}(t)$ further, the partial residual $\hat{e}_i$ it will be used to estimate λ_k. Since λ_k is a random effect, a mixed effects model is used to compute the estimate $\hat{\lambda}_k$. The residual is computed again as $(y_i^{(k)}(t) - \hat{f}(X_i^{(k)}) - \hat{\lambda}_k)$. This residual is a function of $\log\lambda_i^{(k)}(t)$ alone.
4. Again, the residual will be used to estimate $\log\lambda_i^{(k)}(t)$. Finally, the residual is computed as follows: $[y_i^{(k)}(t) - \hat{f}(X_i^{(k)}) - \hat{\lambda}_k - \hat{\log\lambda_i^{(k)}}(t)]$. The estimate for the component to be estimated in the next step is left out for the computation of the residual to ensure that the residual will contain information on that term.
5. Iterate the procedure above until convergence.

3.3 Simulation Studies

We designed a simulation study to evaluate the predictive ability of the survival model for clustered data. Each data set was generated on n clusters wherein each cluster contains 20 observations. The response variable $h_i^{(k)}(t)$ was computed following Equation (4):

$$\log[h_i^{(k)}(t)] = a * \lambda_k + \log \lambda_i^{(k)}(t) + \beta * f(X_i^{(k)}) + m * \varepsilon_i(t) \quad (4)$$

or,

$$y_i^{(k)}(t) = a * \lambda_k + 0.5e^{-0.5t} + b * \delta_i(t) + \beta * f(X_i^{(k)}) + m * \varepsilon_i(t) \quad (5)$$

where:

λ_k	is the cluster-specific random intercept
$\log \lambda_i^{(k)}(t)$	is the natural logarithm of the baseline risk of the ith observation at time t
$X_i^{(k)}$	is the value of the covariate of the ith observation within the kth cluster
$f(X_i^{(k)})$	is any functional form of $X_i^{(k)}$
m	is used to introduce misspecification error, known to induce bias in parametric modeling
a	is used to incorporate the minimal or dominating effect of the cluster-specific component
b	is used to incorporate the minimal or dominating effect of the error term in the baseline risk
β	is used to incorporate the minimal or dominating effect of the covariate
$\delta_i(t)$	is used to include error term component of the baseline risk

The response variable $y_i^{(k)}(t)$ was computed by adding the values of the different components generated using the simulation boundaries. These components were simulated

such that $y_i^{(k)}(t)$ possess a common characteristic innate within the cluster. Table 1 shows the simulation boundaries employed in the study.

Table 1. Boundaries of Simulation Study

1.	Distribution of λ_k	a. $\lambda_k \sim \text{Poisson}(\mu_k)$, mean increases by 5 for the succeeding clusters b. $\lambda_k \sim \text{N}(0, \mu_k)$, variance increases by 5 for the succeeding clusters
2.	Number of Clusters	a. Small –3 clusters b. Large – 10 clusters
3.	Functional Form of $f(X_i^{(k)})$	a. $f(X_i^{(k)}) = \beta X_i^{(k)}$ b. $f(X_i^{(k)}) = \exp(\beta X_i^{(k)})$ c. $f(X_i^{(k)}) = \beta (X_i^{(k)})^2$ d. $f(X_i^{(k)})$ is linear in some clusters, exponential in some clusters
4.	Distribution of ε	a. $\varepsilon \sim \text{N}(0,1)$ b. $\varepsilon \sim \text{N}(0,10)$
5.	Misspecification m in the model	a. $m = 1$ b. $m = 5$
6.	Contribution of the terms in the model	a. Equal Contribution b. Dominating Cluster Component c. Dominating Baseline Risk d. Dominating Covariate
7.	Distribution of $\delta_i(t)$	$\delta_i(t) \sim$Exponential(1)
8.	Distribution of $X_i^{(k)}$	$X_i^{(k)} \sim$Weibull(1,2)

The overall behavior of the model cannot be understood without knowing the specification of the baseline hazard function. Thus, $\log \lambda_i^{(k)}(t)$ is assumed to be equal to $\beta e^{-\beta t} + \delta_i(t)$ where $\beta = 0.5$. The scale parameter of the exponential distribution, which equal to the reciprocal of the parameter, is the constant churn rate. Since the response variable is

computed for each consumer, the component $\delta_i(t)$ is included in the computation of $\log \lambda_i^{(k)}(t)$ to emphasize the individual differences in the baseline risk of churning.

The cluster-specific component λ_k was generated to incorporate the clustering concept, i.e., observations belonging in the same cluster are homogeneous. This implies that the propensities to churn of the observations belonging in the same cluster are similar. Two distributions of λ_k were investigated, Poisson and Normal. The parameter of Poisson is allowed to vary across clusters while the variance of Normal increases for the succeeding clusters. Poisson distribution is used to represent cluster endowments that are skewed (e.g., the presence of extreme behavior) while the normal distribution ought to represent cluster characteristic that is symmetric within an "average" behavior (e.g., average number of SMS sent in a given month). Even though the number of observations in each cluster is constant, the number of clusters is allowed to vary. This simulation boundary addresses the fact that adding clusters not individual observations leads to an increase in efficiency (Arceneaux and Nickerson, 2009).

The functional form (f) of the covariate could be Case (1) linear in all clusters; Case (2) exponential in all clusters; Case (3) quadratic in all clusters or Case (4) linear in some clusters and exponential in the remaining clusters. These were done to assess the implications of deviations from linearity on the predictive ability of the nonparametric model vis-à-vis the parametric model.

For the settings wherein $f(X_i^{(k)})$ is linear in some clusters and exponential in the remaining clusters, four mixed cases were considered for the scenario with small number of clusters as well as when the number of clusters is large. These combinations were chosen among all possible combinations to aid interpretation and to accommodate the fact that

neighboring clusters tend to behave similarly than clusters which are far apart. Table 2 presents the mixed cases for Case (4).

Table 2. Specifications of the Mixed Linear and Exponential Function among the clusters

	Small Number of Clusters (k=3)	Large Number of Clusters (k=10)
Mixed 1	$\exp(\beta_1 X_i^{(1)}) + \exp(\beta_2 X_i^{(2)}) + \beta_3 X_i^{(3)}$	$\exp(\beta_1 X_i^{(1)}) + \ldots + \exp(\beta_5 X_i^{(5)}) + \beta_6 X_i^{(6)} + \ldots + \beta_{10} X_i^{(10)}$
Mixed 2	$\beta_1 X_i^{(1)} + \exp(\beta_2 X_i^{(2)}) + \exp(\beta_3 X_i^{(3)})$	$\beta_1 X_i^{(1)} + \ldots + \beta_5 X_i^{(5)} + \exp(\beta_6 X_i^{(6)}) + \ldots + \exp(\beta_{10} X_i^{(10)})$
Mixed 3	$\exp(\beta_1 X_i^{(1)}) + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)}$	$\beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \exp(\beta_4 X_i^{(4)}) + \ldots + \exp(\beta_7 X_i^{(7)} + \beta_8 X_i^{(8)} + \beta_9 X_i^{(9)} + \beta_{10} X_i^{(10)}$
Mixed 4	$\beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \exp(\beta_3 X_i^{(3)})$	$\exp(\beta_1 X_i^{(1)}) + \exp(\beta_2 X_i^{(2)}) + \exp(\beta_3 X_i^{(3)}) + \beta_4 X_i^{(4)} + \ldots + \beta_7 X_i^{(7)} + \exp(\beta_8 X_i^{(8)}) + \exp(\beta_9 X_i^{(9)}) + \exp(\beta_{10} X_i^{(10)})$

Two distributions of $\varepsilon_i(t)$ were also examined. The first involves minimal variation while the other takes into account the possibility of high variability or heterogeneity among the observations. The recognition of clustering of data into the model helps address the problem of heterogeneity, see for example Ruru and Barrios (2003).

The constant m were included to explore the effect on the predictive ability of the model had there been misspecification error. The value of m equal to 1 represents proper specification of the error term. Conversely, the value of 5 for m represents a case of model misspecification.

Four cases were identified wherein the contribution of the three components λ_k, $X_i^{(k)}$ and $\log \lambda_i^{(k)}(t)$ are varied. As shown by Santos and Barrios (2012), the backfitting algorithm works best if the dominating term in the model is incorporated first in the model. To be able to determine if the three terms are of equal importance or if one term is

dominating, the proportions contributed by λ_k, $\log\lambda_i^{(k)}(t)$ and $f(X_i^{(k)})$ are computed based on the value of the response variable in Equation (5). The constants a, b and β were chosen to be able to vary the contribution of the terms in the model. For instance, in the case wherein each of the three terms have equal importance, the value of a, b and β would yield proportions that are more or less equal. This means that for different simulated data sets, the proportions contributed by the three terms are, on the average, almost equal. Note that the chosen values of β for the mixed case is dependent on whether the functional form in a given cluster is linear or exponential. Table 3 shows the chosen values of a, b and β.

Table 3. Values of a, b and β depending on the functional form of the covariate and the contribution of the terms in the model

Contribution		Linear	Quadratic	Exponential	Mixed
Equal Importance	a	0.1	0.1	0.1	0.1
	b	5	5	10	10
	β	1	0.5	0.01	0.01/1
Dominating Cluster Component	a	0.5	1	1	1
	b	5	5	5	5
	β	0.5	0.5	0.5	0.5
Dominating Baseline Risk	a	0.05	0.05	0.05	0.05
	b	5	10	10	10
	β	0.5	0.1	0.005	0.005/0.05
Dominating Covariate	a	0.5	0.05	0.1	0.05
	b	5	5	5	5
	β	5	1	0.5	0.5/5

Lastly, the values of the covariate $X_i^{(k)}$ were generated from the Weibull distribution with shape and scale parameter equal to 1 and 2, respectively. The Weibull distribution was

chosen because it is used to model a variety of life behaviors, specifically the characteristics of the covariates.

For each of the case wherein the contribution of the terms are varied (i.e. equal contribution of terms, dominating cluster-specific component, dominating baseline risk and dominating covariate), a total of 112 settings are divided into 14 scenarios with 8 settings each. Thus, there were a total of 448 settings.

For each setting, a total of 100 replicates were generated. These simulations are conducted to compare the existing parametric method and the nonparametric method through their median absolute percentage error and root mean square error (RMSE). The formulas are given by:

$$MAPE = \underset{i,k,t}{median} \left| \frac{y_i^{(k)}(t) - \hat{y}_i^{(k)}(t)}{y_i^{(k)}(t)} \right| * 100 \text{ and } RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(y_i^{(k)}(t) - \hat{y}_i^{(k)}(t)\right)^2}{n}}$$

The median absolute percentage error was computed instead of the mean percentage error (MAPE) because the mean is strongly affected by outliers (i.e. very small values of $y_i^{(k)}(t)$). The median absolute percentage error computed for the 100 replicates where averaged and the values presented here are the mean of those median absolute percentage errors. This will be denoted by MAPE for brevity.

4. Results and Discussion

This section presents the results of the simulation study to gain insights about the predictive ability of the proposed nonparametric model compared to the parametric method. The comparison is based on three simulation boundaries namely the number of clusters, the distribution of the cluster component and the functional form of the covariate. The

assessment of the predictive ability is also categorized on the contribution of the components of the model because this affects the sequence of the backfitting algorithm.

For each of the abovementioned simulation boundaries, the results were presented by the value of m , an index of the extent of misspecification error and the variance of the error term. This is so because the predictive abilities of the parametric and nonparametric methods being compared are drastically altered once the variance of the error term is high and if bias is induced because of misspecification.

4.1 Number of Clusters

The number of clusters is classified into two groups: small (3 clusters) and large (10 clusters). The data sets were simulated such that the cluster sizes are equal.

4.1.1 Equal Importance of the terms in the Model

Table 4 shows that MAPE for the nonparametric method is lower than the parametric method especially when there is no misspecification error. If there is misspecification, the parametric method is slightly better when the error term has more homogeneous distribution.

Table 4. MAPE by Number of Clusters when the terms are of equal importance

Number of Clusters	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Small	N(0,1)	m=1	11.73	16.20
		m=5	65.75	63.60
	N(0,10)	m=1	78.03	103.52
		m=5	96.04	145.91
Large	N(0,1)	m=1	7.07	19.28
		m=5	44.85	58.47
	N(0,10)	m=1	73.88	94.52
		m=5	88.65	148.44

4.1.2 Dominating Cluster-Specific Component in the Model

In Table 5, the MAPE for the nonparametric method is lower than the parametric method when the number of clusters is large. If the number of clusters is small, it is only better than the parametric model for the case where the variance of the error term is large and there is misspecification error.

Table 5. MAPE by Number of Clusters when the cluster component is the dominating term in the model

Number of Clusters	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Small	N(0,1)	m=1	57.59	33.54
		m=5	169.68	70.86
	N(0,10)	m=1	135.09	75.63
		m=5	109.97	140.32
Large	N(0,1)	m=1	15.83	41.50
		m=5	31.24	82.58
	N(0,10)	m=1	57.13	84.44
		m=5	78.46	109.78

In a heterogeneous data, models that account for more clusters have the tendency to adjust the model within the cluster, thereby enhancing the predictive ability relative to an overall model (i.e., without clustering effect).

The parametric model performs well in small clusters regardless of the error distribution and the extent of misspecification. The nonparametric model that accounts for the cluster-specific effect will not make a significant difference from the parametric model that ignores clustering if there are only few clusters to deal with.

4.1.3 Dominating Baseline Risk Component in the Model

Table 6 shows that the nonparametric method is better than the parametric method for all the settings. For both methods, it can also be observed that lower MAPE is generally achieved when the cluster size is large, the variance of the error term is small and there is no misspecification present. Both methods generally have lower MAPE values if there is no misspecification error and the variance of the error term is small.

Table 6. MAPE by Number of Clusters when the baseline risk is the dominating term in the model

Number of Clusters	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Small	N(0,1)	m=1	12.13	17.06
		m=5	72.12	72.33
	N(0,10)	m=1	81.23	127.66
		m=5	96.55	149.45
Large	N(0,1)	m=1	9.64	17.91
		m=5	64.62	71.38
	N(0,10)	m=1	80.15	121.16
		m=5	90.86	149.28

4.1.4 Dominating Covariate in the Model

In Table 7, the MAPE for the nonparametric method is lower than the parametric method only if the number of clusters is large. The parametric method yields lower MAPE values if there are a small number of clusters. In particular, the MAPE values for the nonparametric method are considerably high when the error term is large or if there is misspecification error.

Table 7. MAPE by Number of Clusters when the covariate is the dominating term in the model

Number of Clusters	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Small	N(0,1)	m=1	58.88	16.62
		m=5	155.47	47.73
	N(0,10)	m=1	138.04	58.36
		m=5	110.96	141.26
Large	N(0,1)	m=1	7.10	15.51
		m=5	43.33	47.45
	N(0,10)	m=1	63.47	64.69
		m=5	88.07	137.42

For small number of clusters, the advantage of the nonparametric model vanishes as the covariate becomes dominant. This is so because the dominating component in the model is primarily parametric in nature.

A general behavior which is observable among Tables 4 to 7 is that as the number of cluster increases, the MAPE of both methods generally decreases. This is consistent with the observation of Arceneaux and Nickerson (2009) that adding clusters and not observations within a cluster, results to an increase in efficiency in modeling clustered data.

The four scenarios of the contributions of the terms in the model pointed out that dominating baseline risk results to the important evidence on the advantage of the nonparametric method in terms of predictive ability. This is explained by the sequence in the backfitting algorithm. When the terms are contributing equally to the model, the first component entered and estimated is the baseline risk. Thus, as noted by Santos and Barrios (2012), we can expect good estimates among the model terms estimated early on.

4.2 Distribution of λ_k

Two distributions of cluster-specific component λ_k are employed in the study namely: Poisson distribution and Normal distribution. This study deals with clustered data thereby implying that the data sets were simulated such that there is an apparent homogeneity of the cluster-component in each group.

4.2.1 Equal Importance of the terms in the Model

Table 8 shows that the nonparametric method yields better predictive ability than the parametric method, for all the settings. For both methods, it is also evident that whenever there is no misspecification error, the MAPE values are lower. This means that regardless of the variance of the error term, so long as there is no misspecification present, the MAPE values are lower than that which has misspecification.

For the nonparametric and parametric methods, a Poisson distributed λ_k consistently produces lower MAPE than a normally distributed λ_k. This can be explained by the fact that there are only 20 observations within each cluster. Thus, the distribution of the random variation of cluster effects will be more likely skewed with fairly small cluster size.

Table 8. MAPE based on the Distribution of the Cluster Component when the terms in the model are of equal importance

Dist'n. of λ_k	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Poisson	N(0,1)	m=1	14.86	26.41
		m=5	43.08	45.84
	N(0,10)	m=1	22.74	44.54
		m=5	67.51	76.24

Normal	**N(0,1)**	**m=1**	73.53	80.82
		m=5	92.21	148.24
	N(0,10)	**m=1**	83.47	123.62
		m=5	92.48	146.10

4.2.2 Dominating Cluster Component in the Model

As shown in Table 9, the parametric method generally provides better values of MAPE than the nonparametric method. If λ_k follows a normal distribution and there is a misspecification error, the nonparametric method provides lower MAPE values.

Both the Poisson distributed and normally distributed cluster component would yield lower values when the variance of the error term is small. There is a considerable increase in the values of the MAPE once the variance is high regardless of the presence of misspecification error. In general, a Poisson distributed λ_k yields lower MAPE than its normal counterpart.

Table 9. MAPE based on the Distribution of the Cluster Component when the cluster component is the dominating term in the model

Dist'n. of λ_k	**Error Term Dist'n.**	**Misspec. Error**	**Method**	
			Nonparametric	**Parametric**
Poisson	**N(0,1)**	**m=1**	39.70	36.24
		m=5	61.09	39.44
	N(0,10)	**m=1**	107.14	113.85
		m=5	139.84	114.00
Normal	**N(0,1)**	**m=1**	69.53	45.97
		m=5	88.36	106.04
	N(0,10)	**m=1**	126.07	120.22
		m=5	100.08	144.06

4.2.3 Dominating Baseline Risk in the Model

Table 10 reveals that the nonparametric method generally provides better MAPE values than the parametric method. For both methods, it can be noted that the values of the MAPE are lower when there is no misspecification present and if the error term has a small variance. Also, it is notable that there are lower MAPE values if the distribution of the cluster-specific component is Poisson than if it is normal.

Table 10. MAPE based on the Distribution of the Cluster Component when the baseline risk is the dominating term in the model

Dist'n. of λ_k	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Poisson	N(0,1)	m=1	18.87	29.50
		m=5	63.66	62.14
	N(0,10)	m=1	24.66	40.43
		m=5	73.09	81.57
Normal	N(0,1)	m=1	80.44	118.18
		m=5	93.68	149.19
	N(0,10)	m=1	86.20	139.52
		m=5	93.73	149.54

4.2.4 Dominating Covariate in the Model

In the scenario wherein the covariate dominates the model, it could be seen in Table 10 that the parametric method generally provides lower MAPE values than the nonparametric method. The nonparametric method yields better values when there the variance of the error term is large and if there is misspecification error.

The advantage of parametric model over nonparametric model is explained by the fact that the dominating component in the model is primarily parametric in nature. Thus, the enhancements introduced in the nonparametric model will no longer matter.

Table 11. MAPE based on the Distribution of the Cluster Component when the covariate is the dominating term

Dist'n. of λ_k	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Poisson	N(0,1)	m=1	49.33	26.84
		m=5	80.07	42.97
	N(0,10)	m=1	82.63	37.42
		m=5	118.73	52.21
Normal	N(0,1)	m=1	79.77	58.20
		m=5	104.39	136.31
	N(0,10)	m=1	126.32	68.53
		m=5	94.64	142.37

4.3 Functional Form of the Covariate Effect

Four functional forms of f are considered in this study: a linear function $\beta X_i^{(k)}$, an exponential function $\exp(\beta X_i^{(k)})$, a quadratic function $\beta(X_i^{(k)})^2$ and the mixed linear and exponential case.

4.3.1 Equal Importance of the terms in the Model

Table 12 shows that for the linear, quadratic and mixed case, the nonparametric procedure is superior across all settings of the error term variance and the misspecification error.

Table 12. MAPE for the varying functional form of the covariate when the terms are of equal importance

Functional Form	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Linear	N(0,1)	m=1	17.13	36.46
		m=5	62.23	66.91
	N(0,10)	m=1	80.01	106.74
		m=5	90.39	148.93
Quadratic	N(0,1)	m=1	17.98	36.24
		m=5	62.38	70.56
	N(0,10)	m=1	79.68	112.67
		m=5	92.94	149.11

Exponential	**N(0,1)**	**m=1**	22.01	34.67
		m=5	59.12	58.69
	N(0,10)	**m=1**	78.36	86.91
		m=5	92.96	136.67
Mixed	**N(0,1)**	**m=1**	18.62	35.24
		m=5	50.84	57.77
	N(0,10)	**m=1**	77.87	102.31
		m=5	92.53	148.87

There is an observable increase in the MAPE value as the variance of the error term is increased and the misspecification error introduced. Among the functional forms, the lowest MAPE values are usually observed in the scenario wherein the covariate is linear in all the clusters. The values of the linear case are comparable to the mixed case in terms of predictive ability. The advantage of nonparametric method is observed even in cases where the functional form is linear and becomes more apparent as the functional form deviates from linearity.

4.3.2 Dominating Cluster Component in the Model

In the scenario wherein the λ_k is the dominating term in the model, the nonparametric method produces better MAPE values only for the linear and exponential case. The results are presented in Table 13.

Table 13. MAPE for the varying functional form of the covariate when the cluster component is the dominating term in the model

Functional Form	**Error Term Dist'n.**	**Misspec. Error**	**Method**	
			Nonparametric	**Parametric**
Linear	**N(0,1)**	**m=1**	10.20	72.11
		m=5	39.79	78.54
	N(0,10)	**m=1**	60.50	90.80
		m=5	89.93	144.39
Quadratic	**N(0,1)**	**m=1**	144.01	75.07
		m=5	136.55	72.14

	N(0,10)	**m=1**	145.26	78.30
		m=5	119.36	119.15
Exponential	**N(0,1)**	**m=1**	6.51	66.87
		m=5	24.01	72.62
	N(0,10)	**m=1**	39.69	78.70
		m=5	81.34	121.75
Mixed	**N(0,1)**	**m=1**	88.31	77.82
		m=5	125.72	78.43
	N(0,10)	**m=1**	109.78	83.46
		m=5	92.22	122.52

The parametric method is better for all settings if the functional form considered is quadratic. This is also the case for the mixed case except for the setting wherein the misspecification error is induced and the variance of the error term is large.

In the case of quadratic or mixed functional form, more curvature are contained in the model thereby posing a constraint in the smoothing procedure in nonparametric model estimation.

4.3.3 Dominating Baseline Risk in the Model

The nonparametric procedure is superior for functional forms of the covariate, regardless of the variance of the error term and the misspecification error. The corresponding MAPE values are presented in Table 14.

Table 14. MAPE for the varying functional form of the covariate when the baseline risk is the dominating term in the model

Functional Form	**Error Term Dist'n.**	**Misspec. Error**	**Method**	
			Nonparametric	**Parametric**
Linear	**N(0,1)**	**m=1**	25.44	40.40
		m=5	73.82	87.18
	N(0,10)	**m=1**	86.73	138.81
		m=5	94.86	149.54
Quadratic	**N(0,1)**	**m=1**	19.87	33.01
		m=5	65.89	68.25

	N(0,10)	**m=1**	81.57	126.01
		m=5	92.90	149.32
Exponential	**N(0,1)**	**m=1**	22.78	36.51
		m=5	71.29	75.82
	N(0,10)	**m=1**	84.98	131.14
		m=5	93.72	149.35
Mixed	**N(0,1)**	**m=1**	21.07	33.71
		m=5	66.90	67.94
	N(0,10)	**m=1**	82.48	126.50
		m=5	93.61	149.34

Moreover, for all the settings, the quadratic form consistently produced better MAPE values compared to the linear, exponential and mixed scenario. This goes to show that the nonparametric procedure can address the problem of parametric modeling as it departs from the assumption of linearity.

4.3.4 Dominating Covariate in the Model

If the covariate is the dominating term in the model, Table 15 presents varying patterns among the MAPE values of the different functional forms of $X_i^{(k)}$. For the linear scenario, the nonparametric method yields better MAPE values for all the settings. In the case wherein the functional form of $X_i^{(k)}$is quadratic, mixed and exponential, the nonparametric method yields superior results only when there is misspecification of $\varepsilon_i(t)$ and if the variance is high. The nonparametric method is also better for the exponential scenario wherein there is no misspecification and the variance of the error term is small. Thus, the nonparametric method provides better MAPE values for the extremely good and extremely bad exponential settings.

Lastly, the linear scenario provides better MAPE values compared to the other specified functional forms. This is consistent with the result on the case where the cluster effect dominates the model.

Table 15. MAPE for the varying functional form of the covariate when the covariate is the dominating term in the model

Functional Form	Error Term Dist'n.	Misspec. Error	Method	
			Nonparametric	Parametric
Linear	**N(0,1)**	**m=1**	8.17	38.04
		m=5	24.61	41.47
	N(0,10)	**m=1**	42.13	51.25
		m=5	84.98	122.00
Quadratic	**N(0,1)**	**m=1**	129.32	30.82
		m=5	136.47	56.06
	N(0,10)	**m=1**	106.63	78.43
		m=5	124.09	142.76
Exponential	**N(0,1)**	**m=1**	17.91	28.43
		m=5	55.58	47.46
	N(0,10)	**m=1**	74.51	65.05
		m=5	93.34	143.77
Mixed	**N(0,1)**	**m=1**	76.61	31.91
		m=5	119.79	47.04
	N(0,10)	**m=1**	124.51	62.21
		m=5	98.55	141.71

5. Conclusions

The postulated nonparametric model for clustered survival data given n clusters with m elements observed for T points, is formulated as follows

$$\log h_i^{(k)}(t) = \lambda_k + \log \lambda_i^{(k)}(t) + f(X_i^{(k)}) + \varepsilon_i(t)$$

where: $h_i^{(k)}(t)$ is the ith value of the response variable within the kth cluster on time t

i = 1, 2, …, m ; t = 1, 2, …, T

n is the total number of clusters

λ_k is the cluster-specific random intercept

$\lambda_i^{(k)}(t)$ is the baseline risk of the ith observation at time t

$X_i^{(k)}$ is the value of the covariate of the ith observation within the kth cluster

$f(X_i^{(k)})$ is the smooth function of $X_i^{(k)}$

The cluster effect is accounted for by the parameter λ_k whereas the varying coefficient of the covariates is characterized by $f(X_i^{(k)})$. In addition, $\lambda_i^{(k)}(t)$ is incorporated in the model to address the possibly distinct baseline risk of each customer at time t. The postulated method was transformed into an additive model and the backfitting algorithm was used to estimate the parameters sequentially.

The simulation study confirmed that the clustered survival data can be better characterized by a nonparametric procedure. The nonparametric model is advantageous over the parametric method in cases where the model deviates from the assumptions of linearity. Regardless of the number of clusters, the nonparametric method generally provides better estimates than the parametric method due to the incorporation of cluster endowments and between-cluster heterogeneity into the model. Furthermore, the nonparametric model is superior in cases of large number of clusters, where the cluster heterogeneity is aptly accounted by the model. Finally, if the variance of the error term is high and there is misspecification error, the nonparametric method yields better results than the parametric method.

References

Arceneaux, K. and Nickerson, D. (2009). Modeling Certainty with Clustered Data: A comparison of Methods. *Political Analysis*, **17**, 177-190.

Berson, A., Smith, S. and Therling, K. (1999). *Building Data Mining Applications for CRM*. New York: Mcgraw-Hill.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear Smoothers and Additive Models, *Annals of Statistics*, 17(2):453-510.

Chandar, M., Laha, A. and Krishna, P. (2006). Modeling Churn Behavior of bank customers using predictive data mining techniques. *National Conference on Soft Computing Techniques for Engineering Applications* (SCT – 2006).

Cox, D. R. (1972). Regression models and life–tables (with discussion). *Journal of the Royal Statistical Society: Series B 34*, 187–220.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models (with discussion). *Statistical Science*, **1**, 297-318.

Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. London: Chapman and Hall.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Ling, R. and Yen, D. (2001). Customer Relationship Management: An analysis framework and implementation strategies. *Journal of Computer Information Systems, 41 (3)*, 82-97.

Madden, G., Savage, S. and Coble-Neal, G. (1999). *Subscriber Churn in the Australian ISP Market. Information Economics and Policy, 11,* 195-207.

Mammen, E., Linton, O., and Nielsen, J. (1999). The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions, *Annals of Statistics*, 27: 1003-1490.

Mammen, E. and Park, B. (2006). A Simple Backfitting Method for Additive Models. *Annals of Statistics*. **34**, 2252-2271.

Miller, R. (1981). *Survival Analysis*. John Wiley and Sons, New York, USA.

Ngai, E. (2005). Customer Relationship Management Research (1992-2002): An Academic Literature Review and Classification. *Marketing Intelligence and Planning, 23 (6)*, 582-605.

Opsomer, J. (2000). Asymptotic Properties of the Backfitting Estimators, *Journal of Multivariate Analysis*, 73: 166-179.

Parvatiyar, A. and Sheth, J. (2001). Customer Relationship Management: Emerging practice, process and discipline. *Journal of Economics and Social Research, 3* (2), 1-34.

Ruru, Y. and Barrios, E. (2003). Poisson Regression Models of Malaria Incidence in Jayapura, Indonesia. *The Philippine Statistician,* **52**, 27-38.

Santos, E. and Barrios, E. (2012) Decomposition of Multicollinear Data and Time Series using Backfitting and Additive Models. *Communications in Statistics – Simulation and Computation*, 41(9):1693-1710.

Stone, C. (1985). Additive Regression and other Nonparametric Models. *Annals of Statistics*, **13**, 689-705.