Seasonal Warranty Prediction Based on Recurrent Event Data

Qianqian Shan*[1], Yili Hong†[2], and William Q. Meeker‡[1]

[1]Department of Statistics, Iowa State University, Ames, IA
[2]Department of Statistics, Virginia Tech, Blacksburg, VA

Abstract

Warranty return data from repairable systems, such as home appliances, lawn mowers, computers, and automobiles, result in recurrent event data. The non-homogeneous Poisson process (NHPP) model is used widely to describe such data. Seasonality in the repair frequencies and other variabilities, however, complicate the modeling of recurrent event data. Not much work has been done to address the seasonality, and this paper provides a general approach for the application of NHPP models with dynamic covariates to predict seasonal warranty returns. The methods presented here, however, can be applied to other applications that result in seasonal recurrent event data. A hierarchical clustering method is used to stratify the population into groups that are more homogeneous than the overall population. The stratification facilitates modeling the recurrent event data with both time-varying and time-constant covariates. We demonstrate and validate the models using warranty claims data for two different types of products. The results show that our approach provides important improvements in the predictive power of monthly events compared with models that do not take the seasonality and covariates into account.

Keywords: EM algorithm, Hierarchical clustering, Missing data, NHPP, Random effects, Seasonal dynamic covariates

*qshan@iastate.edu
†yilihong@vt.edu
‡wqmeeker@iastate.edu

1 Introduction

1.1 Background

Predictions of warranty returns, based on recurrent event data from repairable systems are often needed by manufacturing companies so they can help to make decisions on the supply of replacement parts, warranty reserves, pricing of the warranty plans and so on. Monthly predictions of warranty returns are particularly helpful when repairable systems have recurrence rates affected by the month of a year and geographical locations. For example, some products may have a higher recurrence rate in warmer months and in a warmer location due to higher average usage rate. The predictions could be more accurate and useful when the variabilities in seasonality and locations are taken into consideration. Meeker and Escobar (1998, Chapter 16), without giving details, describe how one might use a non-homogeneous Poisson process (NHPP) to make such predictions on the repairable systems, with the restrictive assumptions that all systems are independent and have the same recurrence rate function, $\nu(t)$. The assumptions of the simple NHPP model tend to be too strong for realistic complicated data structures when products have staggered entry, different failure patterns, and other system-to-system sources of variability. The purpose of this paper is to develop a general prediction methodology for applications with these complications. In addition to applications in warranty prediction, the methods are also applicable to many other applications such as the prediction of the number of recurrent visits to hospitals of patients in health care industry. We use hierarchical clustering to partition the available data into groups within which there are similar seasonal patterns, and then use the NHPP model with time-varying covariates and random effects to describe the recurrent event warranty data. We illustrate the methods with two different warranty prediction applications.

1.2 Related Literature and Our Work

The application of NHPP models to warranty prediction has been discussed extensively in many places in the literature. Rigdon and Basu (2000) present a general review of NHPP models and their applications including the power law process and kinds of tests for the validity of the models. Krivtsov (2007) gives NHPP models that provide alternatives to the commonly used power law and log-linear processes. Ross (2014) describes the NHPP model and shows how to simulate data from an NHPP model based on a homogeneous Poisson process (HPP). Fredette and Lawless (2007) describe mixed Poisson models for the prediction of the aggregated number of events at specified calendar times across a population of processes. Koutsellis et al. (2017) present a modified generalized renewal process model for warranty prediction of repairable systems with effects of production dates and replacement of defective components/subsystems.

In other related literature, Hamada et al. (2008, Section 6.4) and Ryan et al. (2011) apply NHPP models without covariates under a hierarchical Bayesian framework to describe the recurrent events on 48 shared-memory computer processors. Rai (2009) presents a warranty forecasting model with the monthly seasonality modeled by multiplicative seasonal indices based on data from a single representative production month. Wu (2012, Section 3.5) gives a

brief review of different types of coarse warranty data and methods. Xiao et al. (2015) develop nonparametric Bayesian methodology using a seasonal marked point process to predict hurricane occurrences. Cifuentes-Amado and Cepeda-Cuervo (2015) and Ngailo et al. (2016) use NHPP models with seasonality described by trigonometric functions of time in health diseases and seasonal rainfall events, respectively. Slimacek and Lindqvist (2016) use a piecewise constant rate of occurrence of failures (ROCOF) model with both observable and unobservable differences among repairable systems. Therneau et al. (2003) show that fitting survival models with random effects can be done efficiently via penalized likelihood estimation. Klein (1992) presents an expectation-maximization (EM) algorithm based on a profile likelihood for the semiparametric Cox model.

This paper focuses on developing a flexible warranty event prediction methodology by using the following:

- We develop a parametric recurrent event model to incorporate seasonal effects on the recurrence rates, which can improve the monthly warranty prediction significantly.
- We propose hierarchical clustering on the locations of the systems under warranty to differentiate among different seasonal patterns in the recurrent event processes.
- We take other available fixed covariates effects into consideration to further improve the predictive power of our model.
- We incorporate random effects into our model to describe heterogeneity not accounted for by the covariates.

1.3 Motivating Examples

All companies that offer a warranty for their products are required, by law, to put into reserves a sufficient amount of cash so that they will be able to pay their warranty claims. Warranty predictions are extremely important because there are penalties for not having enough cash in reserve and, of course, for holding too much in reserve. While it is common to use simple methods like the percentage of sales to predict warranty needs, it is now recognized that there is valuable information in warranty databases that can be used to predict warranty returns more accurately.

We apply our models to two product warranty applications. These data sets differ in terms of the type of products, number of systems, number of years of data, recurrence rates, and available covariates. For both applications, we hold out the last 12 months of data for model checking and use only the rest of the data to do exploratory analysis and model fitting.

For Product A, warranty/production information for 63,191 systems with 8,406 events from year 2011 to year 2016 is available for modeling. The Product A database contains variables such as in-service date of the systems, start and expiration date of the warranty contracts, country, model year, retail location, warranty price, model type, event date, and event cost. Approximately 10% of the systems have had at least one warranty-return event.

For Product B, warranty/production information for 33,645 systems with 18,972 events from year 2014 to year 2016 is available for modeling. Each record in the data set contains the start

and expiration date of the warranty contracts, product model year, retail location, warranty price, event date, and event cost. All of the Product B systems have a 24-month warranty term. Approximately 19% of the systems had at least one warranty-return event.

In both examples, the main objective is to generate point predictions and prediction intervals of future warranty returns.

1.4 Overview

The remainder of the paper is organized as follows. Section 2 provides general ways to do exploratory analysis of warranty data to help suggest the form of an appropriate model. A clustering methodology to identify different seasonal recurrence rate patterns is introduced. Section 3 describes the NHPP-based models to be used in this paper. Section 4 presents maximum likelihood estimation of the model parameters with and without random effects. Section 5 discusses point predictions for the number of future events and compares the point predictions results for different models of Product A. Prediction intervals of the point estimates are presented in Section 6. Section 7 describes the application of the same methodology to Product B. Section 8 studies the effects of missing data on clustering and explores two different ways to deal with the missing data. Section 9 discusses our conclusions and ideas for future work.

2 Exploratory Analysis

Exploratory analysis is often useful for providing insights into the structure of a data set and as an aid for model building. In this section, we explore the effects of the fixed covariates on the recurrence rate by examining the mean cumulative number of system recurrences for different levels of the covariates, and apply clustering analysis for identifying different seasonal recurrence rate patterns based on warranty locations.

2.1 The Mean Cumulative Function

Nonparametric methods provide useful tools to explore data without making strong assumptions. Kaplan and Meier (1958) introduced the nonparametric estimator of a survival function based on censored time-to-event data. Similar to the survival function for time-to-event data, the mean cumulative function (MCF), giving the mean number of events across a population of systems as a function of system age, provides a useful baseline model for recurrence data. Nelson (1988) describes how to compute a nonparametric estimate using recurrent event data. For more details see Lawless and Nadeau (1995), Meeker and Escobar (1998, Chapter 16) and Nelson (2003). The sample MCF can, for example, be used to compare the behavior of subpopulations defined by different levels of discrete covariates.

- If the different levels of a covariate have similar MCF curves, then the levels can be combined together for analysis.

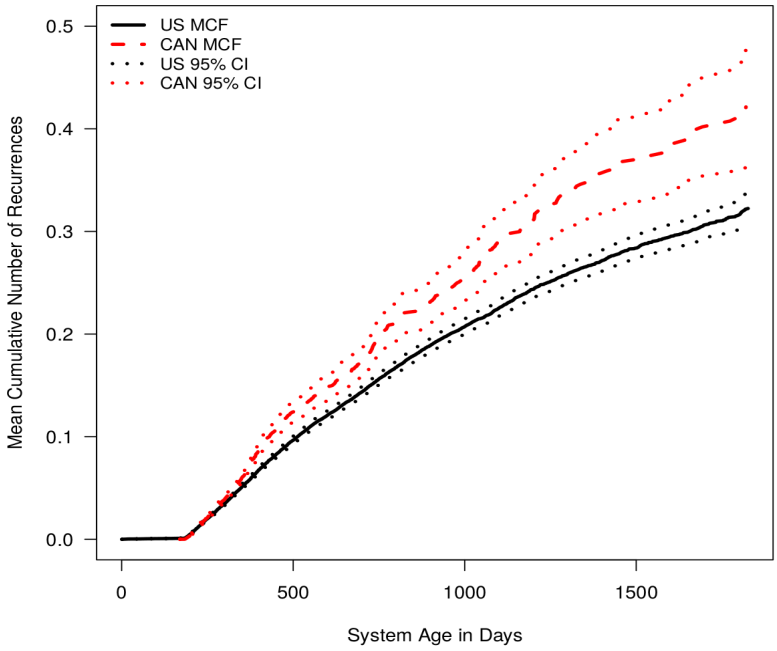

Figure 1: MCF versus age of systems in different countries of Product A with 95% pointwise CIs.

- If the different levels of a covariate have importantly different MCF curves, then terms could be added to the model to take account of the differences. This can be done by either adding the level information as a covariate of the model or by modeling the different levels separately.

Example 1. Exploratory Analysis for the Product A Data
Exploratory analysis based on MCF curves of different subpopulations defined by the covariate levels can help to check if the subpopulations have significantly different behavior. For example, by checking the MCF curves with confidence intervals (CIs) for different countries (Meeker and Escobar 1998, Chapter 16), we can gain insights about how the recurrent event process behaves, as illustrated in Figure 1. There are few reported events for the first 180 days in both countries because of the nature of the warranty contracts and the MCF curve for Canadian systems is higher than that for US systems. The MCF curves suggest that the warranty processes in the two countries are importantly different. The confidence intervals for the Canadian warranty process are wider because the size of the population and the number of recurrences are smaller.

□

The exploratory analysis using MCF curves provides information about the recurrence rate behaviors of the different covariate levels on the system age scale. Other tools are needed if the behavior of the recurrence rate is also affected by factors on the calendar time scale, for example, the clustering analysis tool as introduced in Section 2.2.

2.2 Data Clustering and Seasonality

2.2.1 Data for Clustering Analysis

The combination of climate differences and geographical locations can affect the usage of products and certain failure mechanisms, resulting in different seasonal patterns in different regions. For example, the usage rate of certain products could be higher during summer than in the winter in the northern US, while the seasonal pattern may be less pronounced in the southern US. Here we describe a data-based approach to group different locations across the US (or other geographical regions) into several clusters so locations within a cluster are, with respect to warranty report seasonality, more homogeneous. The variable to be clustered is the observed overall empirical monthly recurrence rate (the ratio of the number of claims in each month to the corresponding total number of repairable systems at risk) for each location, and we ignore the age effects on the number of events.

The following steps are used to construct the data to be used for clustering. For each location (for each state or province in our applications),

1. Use systems that have at least one event to compute the total number of events for each calendar month.
2. Compute the number of systems at risk for each calendar month.
3. Compute the empirical monthly recurrence rate as the ratio of the number of events to the number of systems at risk.
4. If the empirical monthly recurrence rates are computed for calendar months across multiple years, average the rates for each of the 12 months of the year.
5. For each location (e.g., US state and Canadian province), there will be 12 empirical recurrence rates for months from January to December. Use the rates as covariates (or features) to cluster the locations.

Note that if there are locations that have few events, we could observe their behavior and merge them to the locations with not dissimilar behavior until there is a substantial number of events in each location to do the clustering analysis.

2.2.2 Data with Missing Location Variables

It is common to have missing values in warranty (and other) databases. For our Product A data, 6.6% of the location variables are missing. This will affect our location-based clustering analysis. Every application is different, and there is no general approach for handling missing data that works for all. Three commonly used approaches are:

1. Group the missing values with a new category.
2. Impute the missing values based on information that is available.

3. Do a random assignment to replace missing values.

One can choose any one or a combination of the above methods depending on the specific available data, the missing mechanism, the missing percentage and so on. See Example 2 in Section 2.2.3 for an example of how we deal with the missing locations for Product A data.

2.2.3 Hierarchical Clustering Analysis

In unsupervised clustering of different locations, the empirical recurrence rates for each month of a year per location are the observations. We need to specify a clustering method in order to identify clusters of similar location groups. Popular clustering methods include the K-means algorithm Lloyd (1982); Hartigan (1975); Hartigan and Wong (1979), K-medoid, hierarchical clustering Rousseeuw and Kaufman (1990) and so on. K-means and K-medoid methods require specifying the number of clusters and initial centers of each cluster. In contrast, hierarchical clustering only requires a measure of the similarity (or equivalently, dissimilarity) among observations and a definition of how the dissimilarity of clusters is measured (Hastie et al. 2009 and James et al. 2013). We adopt a hierarchical clustering analysis to take advantage of its convenience.

As there are no response variables to characterize each observation, a clear measure of the degree of similarity among the monthly recurrence rates in different locations also needs to be specified (e.g., see James et al. 2013). Possible choices of similarity measures include,

- Euclidean distance: compute the Euclidean distance for each pair of observations, and use it as the similarity measure for clustering.
- Correlation-based distance: compute the correlation between each pair of the observations and group together the locations with a large positive correlation.

The choice of a similarity measure depends on the data and the prior knowledge about the data generating process. In the clustering analysis for different seasonal patterns of recurrent event data, a measure based on correlation performs better when the empirical recurrence rates are low (generally less than 0.01) and the rates differences are not obvious. A measure based on Euclidean distance performs better when there exist obvious differences in recurrence rates across different locations. The largest dissimilarity of all the pairwise observations between two clusters is used to compare the dissimilarity of clusters.

The hierarchical clustering produces a dendrogram (or a tree-based diagram) in which each leaf represents one observation and the height (y-axis) is the specified distance metric. Cutting the dendrogram at different heights can split the observations into different clusters naturally. The selection of where to cut can be affected by factors such as the desired number of clusters, the minimum number of observations within each cluster, how many seasonal patterns exist in the data, and the expected degree of dissimilarity in the seasonal patterns after clustering analysis. The following example shows how the number of clusters can affect the prediction performance of models.

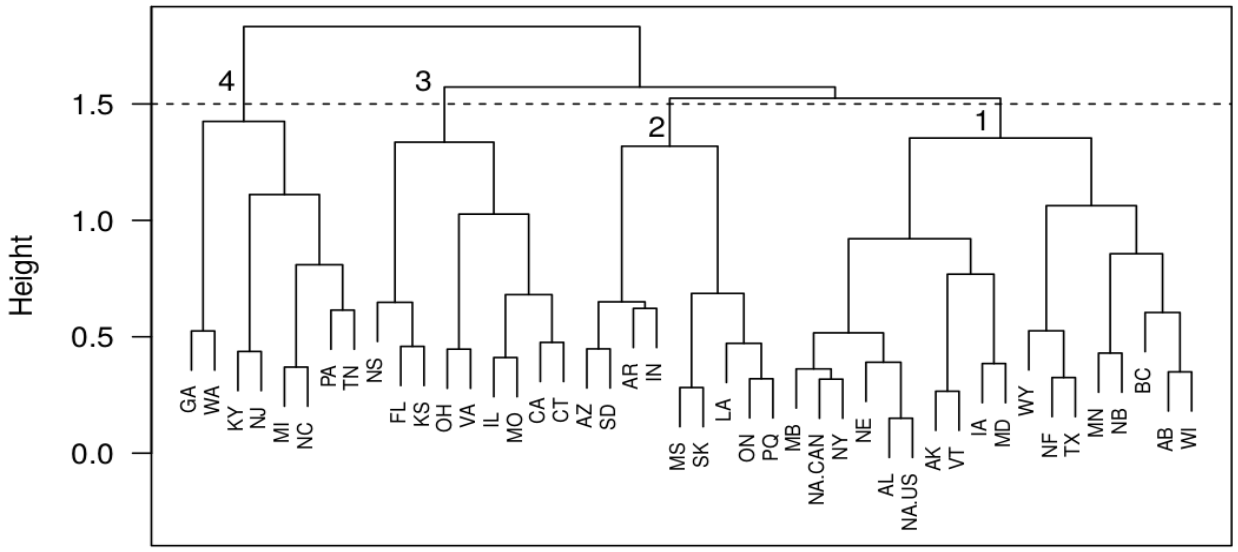

Figure 2: Dendrogram of correlation-based hierarchical clustering of Product A. The horizontal line indicates the cutoff location to divide the locations into different clusters. Denote the clusters from left to right as cluster 4, 3, 2 and 1, respectively.

Example 2. Clustering of Product A Seasonal Patterns
Because 4142 systems (approximately 6.6% of the Product A data) have missing location variables with no obvious missing patterns, we group these systems together by country and assign new location variables, NA.US and NA.CAN, for the US and Canada systems, respectively. For the purpose of comparison, we also use the random assignment method for the missing locations, and the detailed explanation and results are in A.3 of supplemental materials. We use event data after year 2014 to do clustering analysis as there are more events with more repairable systems at risk for most of the locations after 2014. Figure 2 shows the dendrogram of the hierarchical clustering results for Product A using correlation distance as similarity measure. Cutting the dendrogram horizontally at around 1.55 naturally separates the data into four clusters with near balanced number of locations and event counts in each cluster. Figures 3 and 4 show, respectively, the observed events and the number of systems at risk by clusters as a function of calendar date. Figure 5 shows the overall monthly empirical recurrence rates for the four clusters, and it indicates that the seasonal patterns vary considerably.

A sensitivity analysis on the choice of the number of clusters shows that cutting the dendrogram such that there are more than four clusters gives approximately the same results as using four clusters. But the amount of computing time required for model fitting can increase dramatically (e.g., there are 50 parameters to be estimated in a model with four clusters, and 122 parameters in the same model in the case of ten clusters). It will take even longer for the models with random effects involved, which requires a more complicated algorithm to find the parameter estimates. Given that more clusters do not improve the prediction performance substantially and that it will take a significantly longer time for model fitting, we choose the number of clusters that provides a good trade-off between model performance and computation costs.

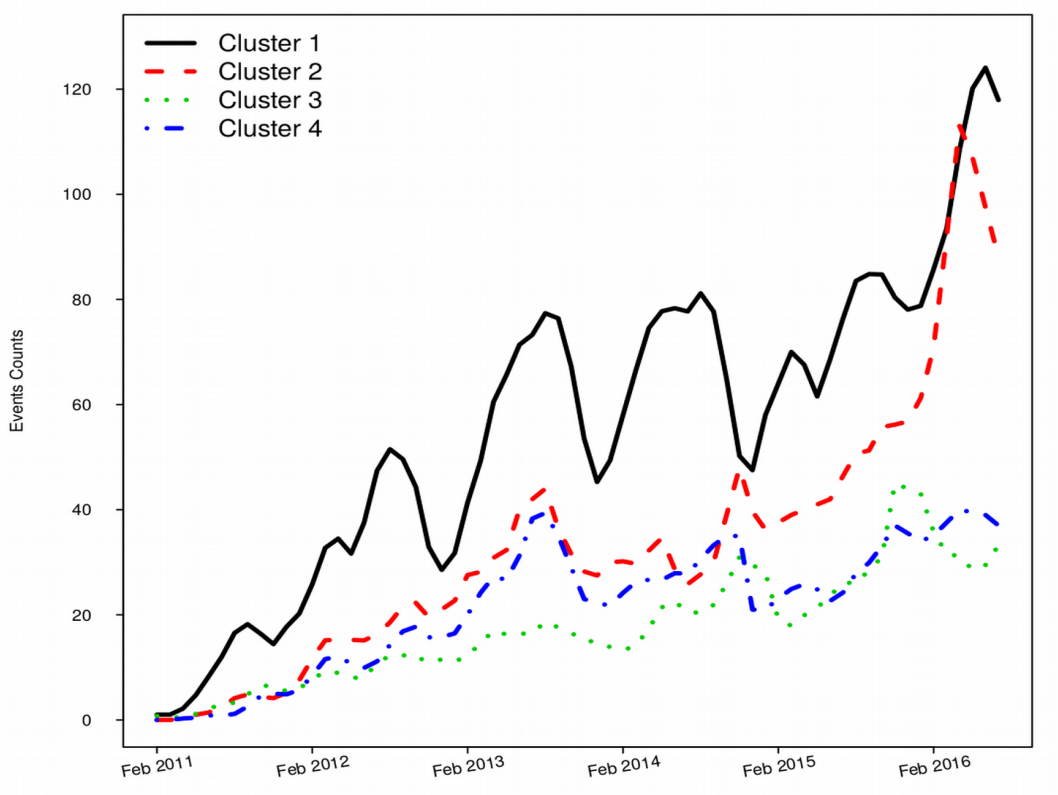

Figure 3: Observed event counts versus date for the different clusters of Product A.

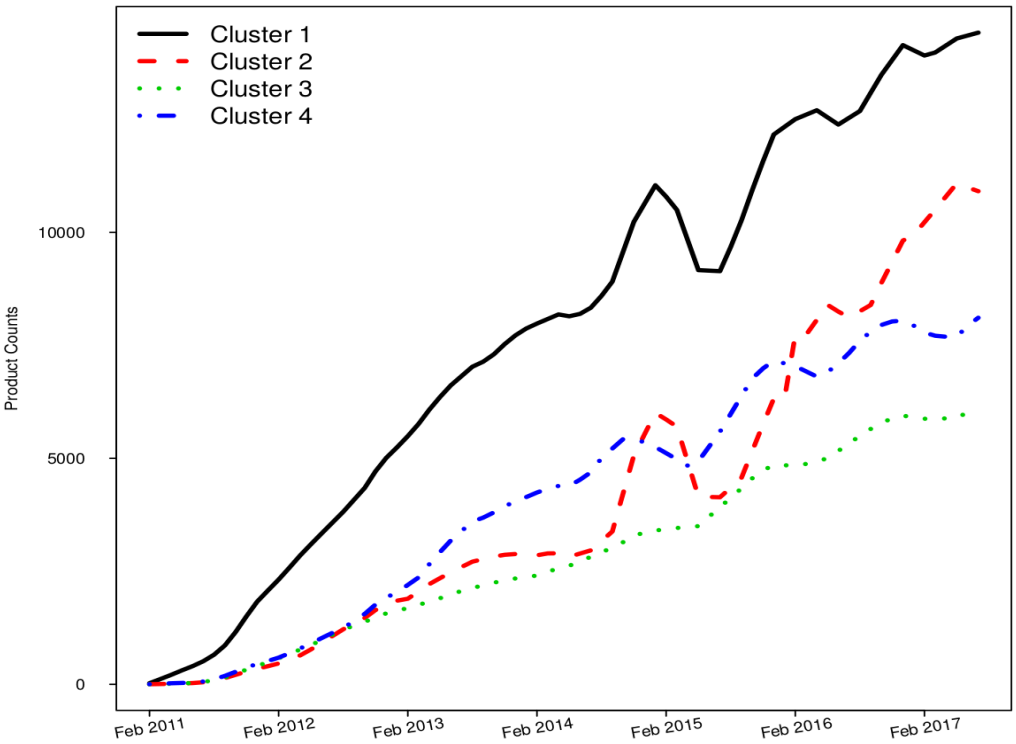

Figure 4: Number of systems at risk versus date for the different clusters of Product A.

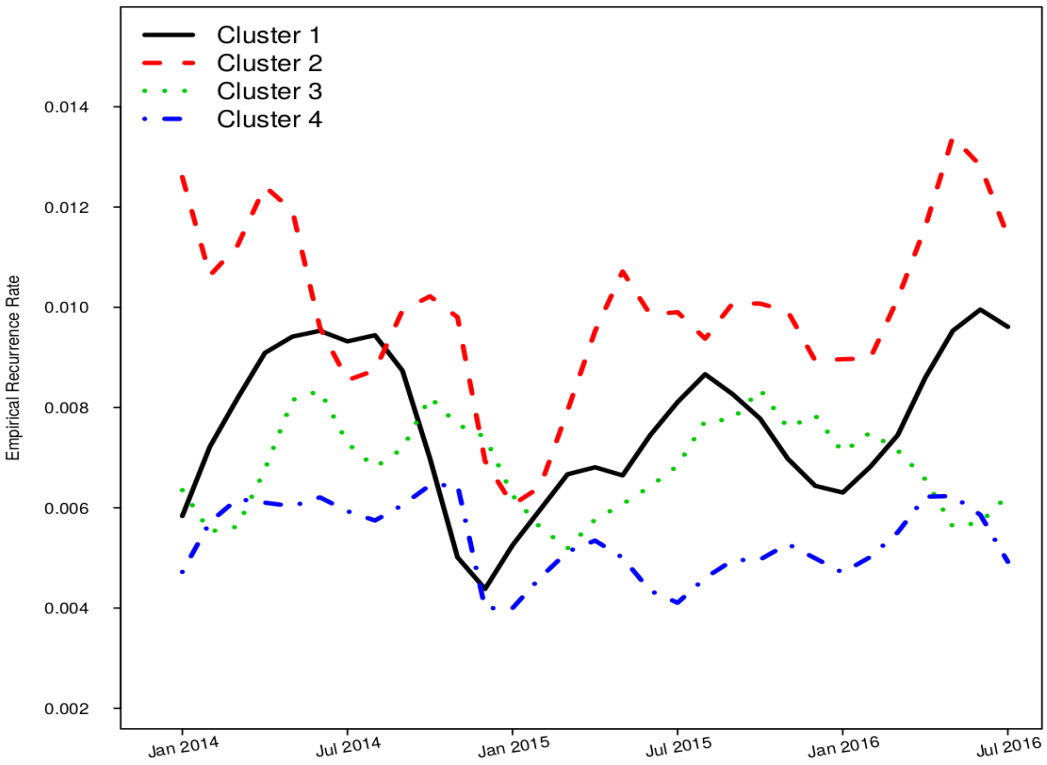

Figure 5: Empirical monthly recurrence rate by clusters of Product A systems since year 2014.

□

3 General Models for Recurrence Rates

The Poisson process is commonly used for modeling of repairable systems, but it has the assumption that the numbers of events in non-overlapping time intervals are statistically independent. For such situations, it is natural to model event counts with an NHPP model with a non-constant recurrence rate as described by Meeker and Escobar (1998, Chapter 16). We employ and adapt the widely used NHPP model for our analysis of warranty recurrent event data.

3.1 Notation

Let $\mathrm{N}_i(t) = \mathrm{N}_i(0, t)$ denote the observed total number of events up to system age t for repairable system i, where t is the number of days since the system is put into service. Then the process *recurrence rate function* for system i is

$$\nu_i(t) = \lim_{\Delta t \to 0} \frac{\mathrm{E}[\Delta \mathrm{N}_i(t)]}{\Delta t}, \tag{1}$$

where $\Delta \mathrm{N}_i(t) = \mathrm{N}_i(t^- + \Delta t) - \mathrm{N}_i(t^-)$ is the number of events in $[t, t + \Delta t)$. We denote the parameter vector of a model as $\boldsymbol{\theta}$.

3.2 The Simple NHPP Model

The simple NHPP model assumes that all K systems have the same recurrence rate function and the function is defined as

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}), \quad i = 1, \cdots, K, \tag{2}$$

where $\nu_0(t;\boldsymbol{\theta})$ is a function depending only on system age t and parameters $\boldsymbol{\theta}$. Here we use the power law process

$$\nu_0(t;\boldsymbol{\theta}) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \tag{3}$$

with $\boldsymbol{\theta} = (\beta, \eta)^T$ and β without a subscript is the power law parameter. Subsequently, we will use β with a subscript to denote regression parameters. The simple NHPP model has strong assumptions that are rarely appropriate for modeling complicated recurrence data structures such as the recurrences in a warranty database.

3.3 NHPP Model with Common Seasonal Effects

The NHPP model with simple seasonality assumes that the rate function of each system has the same seasonal behavior over $M = 12$ months of each year,

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\beta,\eta)\exp\left(\sum_{m=1}^{M}\beta_m \mathrm{I}_{m,i}(t)\right), \tag{4}$$

where $\nu_0(t;\beta,\eta)$ is as defined in (3), $\boldsymbol{\theta} = (\beta, \eta, \beta_1, \cdots, \beta_M)^T$ and

$$\mathrm{I}_{m,i}(t) = \begin{cases} 1 & \text{if system } i \text{ is in calendar month } m \text{ at age } t \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We set one of the β_m values to be zero in order to have a full rank indicator matrix with its elements defined in (5). And the same rule applies to the rest of the models. Because the indicator $\mathrm{I}_{m,i}(\cdot)$ is obtained based on the number of days in service and the calendar date when the system is first put into service, it allows for systems to have staggered entry, as seen in typical warranty databases.

3.4 NHPP Model with Seasonal and Cluster Effects

As described in Section 2.2, the seasonal behavior will depend on the geographical location for some applications. We account for this by generalizing the seasonal time-varying covariates. In particular, by assuming that the seasonal recurrence rate patterns vary in both shapes and levels among the clusters, the recurrence rate function of system i is

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\beta,\eta)\exp\left(\sum_{m=1}^{M}\sum_{n=1}^{N}\beta_{m,n}\mathrm{I}_{m,n,i}(t)\right), \tag{6}$$

where N is the number of clusters, $\boldsymbol{\theta} = (\beta, \eta, \beta_{1,1}, \beta_{1,2}, \cdots, \beta_{M,N})^T$ and

$$\mathrm{I}_{m,n,i}(t) = \begin{cases} 1 & \text{if system } i \text{ is in calendar month } m \text{ at age } t \text{ and from cluster } n \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The model in (6) can be simplified if only the levels of the seasonal patterns change across different clusters. In this case,

$$\nu_i(t; \boldsymbol{\theta}) = \nu_0(t; \beta, \eta) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t) + \sum_{n=1}^{N} \beta'_n \mathrm{I}_{n,i} \right), \tag{8}$$

where $\mathrm{I}_{m,i}(\cdot)$ is as defined in (5), $\boldsymbol{\theta} = (\beta, \eta, \beta_1, \cdots, \beta_M, \beta'_1, \cdots, \beta'_N)^T$, and the time independent cluster indicator is

$$\mathrm{I}_{n,i} = \begin{cases} 1 & \text{if system } i \text{ is in cluster } n \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

3.5 NHPP Model with Seasonal, Cluster and Random Effects

If heterogeneity among systems cannot be completely explained by the Poisson process model with the adjustment of covariates, the incorporation of random effects of the repairable systems can be helpful to explain the system level variation. A model for the recurrence rate for system i conditional on the random effects is introduced in a manner that is similar to the use of frailty models in the survival analysis (Aalen 1988, Therneau et al. 2003, and Cook and Lawless 2007). Then the intensity is

$$\nu_i(t; \boldsymbol{\theta}, u_i) \;=\; u_i \, \nu_0(t; \beta, \eta) \exp\left(\sum_{m=1}^{M} \sum_{n=1}^{N} \beta_{m,n} \mathrm{I}_{m,n,i}(t) + \boldsymbol{x}_{i,fix}^T \boldsymbol{\beta}_{fix} \right). \tag{10}$$

Here,

- u_i denotes the i.i.d. random effects for system i. Because the u_i values are unknown and not observed, we assume that the random effects have an independent gamma distribution with mean 1 and variance ϕ, so the random effects are always positive and the distribution of the heterogeneity for individual systems can be reflected by the magnitude of the variance. The density function of u_i is

$$g(u_i; \phi) = \frac{u_i^{\phi^{-1}-1} \exp(-u_i/\phi)}{\phi^{\phi^{-1}} \Gamma(\phi^{-1})}. \tag{11}$$

- $\boldsymbol{x}_{i,fix}$ is a vector of fixed covariates that can help explain additional variability in the recurrence process, and $\boldsymbol{\beta}_{fix}$ is the corresponding column vector of regression coefficients. The fixed covariates can be identified in the exploratory analysis phase as described in Section 2.1 or by diagnostics based on model fitting and prediction performance.

In particular, $\boldsymbol{\theta} = \left(\beta, \eta, \beta_{1,1}, \cdots, \beta_{M,N}, \boldsymbol{\beta}_{fix}^T, \phi\right)^T$ is the parameter vector to be estimated.

3.6 Comparison of Different Models

The NHPP model in (10) can be treated as a general model from the perspective that all of the other models listed above can be viewed as a special case of it:

- Set $\phi = 0$, $\beta_{m,n} = 0$ for $m = 1, 2, \cdots, M$ and any $n = 1, 2, \cdots, N$, and $\boldsymbol{\beta}_{fix} = \mathbf{0}^T$, (10) reduces to the simple NHPP model in (2).
- Set $\phi = 0$, $\beta_{m,n_1} = \beta_{m,n_2}$ for $m = 1, \cdots, M$ and any $n_1, n_2 \in \{1, 2, \cdots, N\}$ and $\boldsymbol{\beta}_{fix} = \mathbf{0}^T$, (10) reduces to the NHPP model with simple seasonality in (4).
- Set $\phi = 0$ and $\boldsymbol{\beta}_{fix} = \mathbf{0}^T$, (10) reduces the NHPP model with seasonal and cluster covariates in (6). Further set $\beta_{m,n_1} = \mathrm{c}_{n_1-n_2} + \beta_{m,n_2}$ with $\mathrm{c}_{n_1-n_2}$ a constant related to $(n_1 - n_2)$ for $m = 1, 2, \cdots, M$ and $1 \leq n_2 < n_1 \leq N$, (10) reduces to (8).

4 Maximum Likelihood Estimation

4.1 Likelihood Function

By extending the approach in maximum likelihood estimation of the superimposed Poisson process likelihood which is used for the parameter estimation in Meeker and Escobar (1998), the total likelihood with the random effects u_i for system $i = 1, \cdots, K$ given the historical events data is

$$L\left(\boldsymbol{\theta}|\mathrm{DATA}\right) = \prod_{i=1}^{K} \int \left\{ \left[\prod_{j=1}^{r_i} \nu_i(t_{ij}; \boldsymbol{\theta}) \right] \exp\left[-\mu_i(0, t_{a_i}; \boldsymbol{\theta}) \right] \right\} g(u_i; \phi) \mathrm{d}u_i, \tag{12}$$

where $g(\cdot)$ is defined in (11), t_{ij} is the j^{th} observed event time for system i with $j = 1, 2, \cdots, r_i$, and t_{a_i} is the end-of-observation time or the end of warranty time for system i, whichever comes first.

We can re-write (10) as the multiplication of the random and non-random parts, $\nu_i(t_{ij}; \boldsymbol{\theta}) = u_i \nu_{b,i}(t_{ij}; \boldsymbol{\beta})$, where $\nu_{b,i}(t; \boldsymbol{\beta}) = \nu_0(t) \exp\left(\sum_{m=1}^{M} \sum_{n=1}^{N} \beta_{m,n} \mathrm{I}_{m,n,i}(t) + \boldsymbol{x}_{fix}^T \boldsymbol{\beta}_{fix} \right)$ and $\boldsymbol{\beta} = (\beta, \eta, \beta_{1,1}, \cdots, \beta_{M,N}, \boldsymbol{\beta}_{fix}^T)^T$. Similar to what is done in Lawless (1987), integrating over u_i for each system i in (12) gives the likelihood

$$L(\boldsymbol{\theta}|\mathrm{DATA}) = \prod_{i=1}^{K} \left[\prod_{j=1}^{r_i} \nu_{b,i}(t_{ij}; \boldsymbol{\beta}) \right]^{\delta_{ij}} \frac{\Gamma(\zeta_i)}{\phi^{1/\phi} \Gamma(1/\phi) \kappa_i^{\zeta_i}}, \tag{13}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$, $\zeta_i = r_i + 1/\phi$ and $\kappa_i = \mu_{b,i}(t_{a_i}; \boldsymbol{\beta}) + 1/\phi$ and $\mu_{b,i}(t_{a_i}; \boldsymbol{\beta})$ is short for $\mu_{b,i}(0, t_{a_i}; \boldsymbol{\beta}) = \int_0^{t_{a_i}} \nu_{b,i}(x; \boldsymbol{\beta}) \mathrm{d}x$. Details of the derivation are given in Section A.1 of the supplemental materials.

Note that when there are no random effects (i.e., $\phi = 0$), the likelihood function in (12) reduces to

$$L(\boldsymbol{\theta}|\text{DATA}) = \prod_{i=1}^{K} \left\{ \left[\prod_{j=1}^{r_i} \nu_i(t_{ij};\boldsymbol{\theta}) \right] \exp\left[-\mu_i(0,t_{a_i};\boldsymbol{\theta})\right] \right\}. \tag{14}$$

4.2 The EM Algorithm

For the model with random effects, we apply the EM algorithm based on the complete-data likelihood. The derivation of the formulas is based on the work of Klein (1992) for the semiparametric Cox model. If we could observe the random effects, $\boldsymbol{u} = (u_1, \ldots, u_i, \ldots, u_K)^T$, the complete-data log-likelihood of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$ up to a constant is,

$$\begin{aligned}
& \mathcal{L}(\phi, \boldsymbol{\beta}|\text{DATA}, \boldsymbol{u}) \\
= & \sum_{i=1}^{K}\sum_{j=1}^{r_i} \{\log(u_i) + \log\left[\nu_{b,i}(t_{ij};\boldsymbol{\beta})\right]\} - \sum_{i=1}^{K} u_i \mu_{b,i}(t_{a_i};\boldsymbol{\beta}) + \sum_{i=1}^{K} \log\left[g(u_i;\phi)\right] \\
= & \sum_{i=1}^{K} \{r_i \log(u_i) - \log\left[g(u_i;\phi)\right]\} + \sum_{i=1}^{K} \left\{ \sum_{j=1}^{r_i} \log\left[\nu_{b,i}(t_{ij};\boldsymbol{\beta})\right] - u_i \mu_{b,i}(t_{a_i};\boldsymbol{\beta}) \right\} \\
= & \mathcal{L}_1(\phi|\text{DATA}, \boldsymbol{u}) + \mathcal{L}_2(\boldsymbol{\beta}|\text{DATA}, \boldsymbol{u})
\end{aligned} \tag{15}$$

where

$$\mathcal{L}_1(\phi|\text{DATA}, \boldsymbol{u}) = -K\left\{ \frac{1}{\phi}\log(\phi) + \log\left[\Gamma\left(\frac{1}{\phi}\right)\right] \right\} + \sum_{i=1}^{K} \left[\left(r_i + \frac{1}{\phi} - 1 \right) \log(u_i) - \frac{u_i}{\phi} \right],$$

is the part of the likelihood that is related to the parameter ϕ and

$$\mathcal{L}_2(\boldsymbol{\beta}|\text{DATA}, \boldsymbol{u}) = \sum_{i=1}^{K} \left\{ \sum_{j=1}^{r_i} \log[\nu_{b,i}(t_{ij};\boldsymbol{\beta})] - u_i \mu_{b,i}(t_{a_i};\boldsymbol{\beta}) \right\},$$

is the part of the likelihood that is related to the parameters in $\boldsymbol{\beta}$.

Simple calculations show that the distribution of u_i, conditional on the observed event process data, has a Gamma(ζ_i, κ_i) distribution, where ζ_i and κ_i are the same as in (13) and they are shape and rate parameters, respectively. The expected complete-data log-likelihood in (15) is obtained by replacing the u_i values in $\mathcal{L}_1(\cdot)$ and $\mathcal{L}_2(\cdot)$ with their expected values,

$$\begin{aligned}
\widehat{\mathcal{L}}_1(\phi|\text{DATA}, \widehat{\boldsymbol{u}}) &= -K\left\{ \frac{1}{\phi}\log(\phi) + \log\left[\Gamma\left(\frac{1}{\phi}\right)\right] \right\} \\
&+ \sum_{i=1}^{K} \left\{ \left(r_i + \frac{1}{\phi} - 1 \right) \left[\psi(\zeta_i) - \log(\kappa_i)\right] - \frac{\zeta_i/\kappa_i}{\phi} \right\},
\end{aligned} \tag{16}$$

where $\psi(\cdot)$ is the digamma function derived from $\text{E}[\log(u_i)|\text{H}(t_{a_i})] = \psi(\zeta_i) - \log(\kappa_i)$.

$$\widehat{\mathcal{L}}_2(\boldsymbol{\beta}|\text{DATA}, \widehat{\boldsymbol{u}}) = \sum_{i=1}^{K} \left\{ \sum_{j=1}^{r_i} \log\left[\nu_{b,i}(t_{ij};\boldsymbol{\beta})\right] - \left(\frac{\zeta_i}{\kappa_i}\right) \mu_{b,i}(t_{a_i};\boldsymbol{\beta}) \right\}. \tag{17}$$

In the maximization step, we maximize (16) and (17) with respect to the parameters ϕ and $\boldsymbol{\beta}$, and in the expectation step, we update the expected values of u_i. The EM algorithm proceeds as follows,

1. Obtain initial estimates of $\boldsymbol{\beta}$ by setting $u_i = 1$ for all systems (or equivalently, $\phi = 0$) and pick a nonzero initial value of ϕ to avoid infinite values of ζ_i and κ_i.
2. Update ζ_i and κ_i using the current values of $\boldsymbol{\beta}$, ϕ and $u_i = (\zeta_i/\kappa_i)$.
3. Update the estimates of ϕ and $\boldsymbol{\beta}$ by maximizing (16) and (17), respectively.
4. Repeat Step 2 and 3 until convergence.

Example 3. Model Fitting for Product A We fit the following thirteen models labeled from (18.1) to (18.13) to the Product A data and check the model prediction performance of different combinations of the seasonal effects, cluster effects, fixed covariates and random effects:

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \tag{18.1}$$

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t)\right) \tag{18.2}$$

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t) + \boldsymbol{x}_{i,fix}^T \boldsymbol{\beta}_{fix}\right) \tag{18.3}$$

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t) + \sum_{n=1}^{N} \mathrm{I}_{n,i}\beta_n\right) \tag{18.4}$$

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t) + \sum_{n=1}^{N} \mathrm{I}_{n,i}\beta_n + \boldsymbol{x}_{i,fix}^T \boldsymbol{\beta}_{fix}\right) \tag{18.5}$$

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M}\sum_{n=1}^{N} \beta_{m,n} \mathrm{I}_{m,n,i}(t)\right) \tag{18.6}$$

$$\nu_i(t;\boldsymbol{\theta}) = \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M}\sum_{n=1}^{N} \beta_{m,n} \mathrm{I}_{m,n,i}(t) + \boldsymbol{x}_{i,fix}^T \boldsymbol{\beta}_{fix}\right) \tag{18.7}$$

$$\nu_i(t;\boldsymbol{\theta}) = u_i\, \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t)\right) \tag{18.8}$$

$$\nu_i(t;\boldsymbol{\theta}) = u_i\, \nu_0(t;\boldsymbol{\theta}) \exp\left(\sum_{m=1}^{M} \beta_m \mathrm{I}_{m,i}(t) + \boldsymbol{x}_{i,fix}^T \boldsymbol{\beta}_{fix}\right) \tag{18.9}$$

Table 1: Summary computation time of different models for the Product A data.

No.	Model Components	Time (hours)
1	Simple NHPP	0.18
2	NHPP with Common Seasonal Effects	0.28
3	NHPP with Country and Common Seasonal Effects	0.37
4	NHPP with Cluster and Common Seasonal Effects	0.44
5	NHPP with Cluster, Common Season and Country Effects	0.53
6	NHPP with Cluster and Seasonal Interactions	3.23
7	NHPP with Cluster and Seasonal Interactions and Country Effects	3.29
8	NHPP with Common Season and Random Effects	2.26
9	NHPP with Common Season, Country and Random Effects	2.56
10	NHPP with Cluster, Common Season and Random Effects	3.90
11	NHPP with Cluster, Common Season, Country and Random Effects	4.51
12	NHPP with Cluster and Seasonal Interactions and Random Effects	36.5
13	NHPP with Cluster and Seasonal Interactions, Country and Random Effects	37.5

$$\nu_i(t;\boldsymbol{\theta}) = u_i\,\nu_0(t;\boldsymbol{\theta})\exp\left(\sum_{m=1}^{M}\beta_m \mathrm{I}_{m,i}(t) + \sum_{n=1}^{N}\mathrm{I}_{n,i}\beta_{n,i}\right) \tag{18.10}$$

$$\nu_i(t;\boldsymbol{\theta}) = u_i\,\nu_0(t;\boldsymbol{\theta})\exp\left(\sum_{m=1}^{M}\beta_m \mathrm{I}_{m,i}(t) + \sum_{n=1}^{N}\mathrm{I}_{n,i}\beta_{n,i} + \boldsymbol{x}_{i,fix}^T\boldsymbol{\beta}_{fix}\right) \tag{18.11}$$

$$\nu_i(t;\boldsymbol{\theta}) = u_i\,\nu_0(t;\boldsymbol{\theta})\exp\left(\sum_{m=1}^{M}\sum_{n=1}^{N}\beta_{m,n}\mathrm{I}_{m,n,i}(t)\right) \tag{18.12}$$

$$\nu_i(t;\boldsymbol{\theta}) = u_i\,\nu_0(t;\boldsymbol{\theta})\exp\left(\sum_{m=1}^{M}\sum_{n=1}^{N}\beta_{m,n}\mathrm{I}_{m,n,i}(t) + \boldsymbol{x}_{i,fix}^T\boldsymbol{\beta}_{fix}\right), \tag{18.13}$$

where $\boldsymbol{x}_{i,fix}$, a vector of length 1, denotes the fixed country effect for Product A (i.e., $\boldsymbol{x}_{i,fix}$ equals 1 if the product is from Canada, or 0 otherwise) and $\boldsymbol{\beta}_{fix}$ is the corresponding regression coefficient. Table 1 gives the computing time needed to fit each model. The fitting was done in R with a computer having a single 2.6GHz core and 128GB of main memory. The computing time is generally longer when the EM algorithm is involved, and it will also depend on the choice of the convergence criterion of the algorithm.

□

5 Point Predictions for the Number of Future Events

Prediction of the number of recurrences for system i in a future time-in-service interval $[t_1, t_2)$ is based on the estimated expected value of the random variable $N_i(t_1, t_2)$.

- When there are no random effects, $N_i(t_1, t_2)$ has a Poisson distribution with mean $\mu_{b,i}(t_1, t_2; \boldsymbol{\beta}) = \int_{t_1}^{t_2} \nu_i(x; \boldsymbol{\beta})\mathrm{d}x$.

- When there are random effects in the model, $N_i(t_1, t_2)$ has a negative binomial distribution with mean $[\zeta_i/\kappa_i]\,\mu_{b,i}(t_1, t_2; \boldsymbol{\beta})$ and probability function

$$\begin{aligned} & \Pr[N_i(t_1, t_2) = n | \text{DATA}, \boldsymbol{\theta}] \\ = \quad & \frac{\Gamma(n + \zeta_i)}{\Gamma(\zeta_i) n!} \left[\frac{\mu_{b,i}(t_1, t_2; \boldsymbol{\beta})}{\mu_{b,i}(t_1, t_2; \boldsymbol{\beta}) + \kappa_i}\right]^n \left[\frac{\kappa_i}{\mu_{b,i}(t_1, t_2; \boldsymbol{\beta}) + \kappa_i}\right]^{\zeta_i}. \end{aligned} \tag{19}$$

  That is, $N_i(t_1, t_2)$ has a $NB(\zeta_i, \kappa_i/\left[\mu_{b,i}(t_1, t_2; \boldsymbol{\beta}) + \kappa_i\right])$ distribution (see Section A.2 of the supplemental materials for more details).

Although the recurrence rate function contains time-dependent covariates, month-by-month integration is possible because the time dependent covariates remain unchanged within each calendar month. The total expected number of events in a future month is the sum of the expected number of events for each system at risk. Similarly, the total expected cumulative number of events for all systems up to a specified future month is the sum of the cumulative number of events for each system at risk. A point prediction for these quantities can be obtained by replacing $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$ with the maximum likelihood estimates.

In order to compare the prediction accuracy of different models, we compute the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percent error (MAPE) of prediction errors for the hold-out data. Denote the non-negative observed and predicted monthly number of events as Y_h and $\widehat{Y}_h$, $h = 1, 2, \cdots, H$, respectively, then

$$\begin{aligned} \text{RMSE}^H &= \sqrt{\frac{\sum_{h=1}^H (Y_h - \widehat{Y}_h)^2}{H}} \\ \text{MAE}^H &= \frac{1}{H}\sum_{h=1}^H |Y_h - \widehat{Y}_h| \\ \text{MAPE}^H &= \frac{1}{H}\sum_{h=1}^H \frac{|Y_h - \widehat{Y}_h|}{Y_h} \times 100. \end{aligned} \tag{20}$$

Example 4. Comparisons of point predictions on Product A
Table 2 gives a comparison of the prediction performances of different models on the hold-out data. By incorporating the cluster information and assuming different seasonal effects for different clusters, the predictions on the hold-out data can be improved with smaller prediction errors. Model 7 has the best prediction performance, especially for the first 6 months of the hold-out data. This model assumes both the shapes and levels of the seasonal patterns in the recurrence rates are different across clusters and that the recurrence rates vary in different countries. The incorporation of the random effects does not improve the prediction performance for this specific example. Figures 6 and 7 show the fitted Model 7 of the monthly and cumulative event counts, respectively. The fitted model deviates from the observed counts between January 2013 and January 2015. The agreement is better after January 2015.

In order to select an appropriate model, one should take the business goal, data size, model fitting time, model prediction performance, computation costs, and other factors into consideration. For example, if our goal is to find a model with reasonably good prediction performance with as little computation costs as possible, Model 7 (sometimes even Model 3) could be a

Table 2: Summary results of point predictions on different models for the Product A hold-out data. The superscript 6 and 12 indicate the RMSE/MAE/MAPE of the first 6 and 12 months, respectively. The model with the smallest RMSE/MAE/MAPE values is marked in **bold**.

No.	$RMSE^6$	$RMSE^{12}$	MAE^6	MAE^{12}	$MAPE^6$	$MAPE^{12}$
1	58.2	46.2	46.5	38.7	16.4	15.3
2	45.7	38.8	39.9	33.6	15.0	14.1
3	39.4	31.8	33.1	28.0	12.6	11.8
4	41.2	34.9	34.5	29.5	13.1	12.5
5	38.9	32.8	32.9	27.8	12.5	11.8
6	39.7	34.1	34.0	29.6	12.8	12.4
7	**37.0**	**31.9**	**30.9**	**27.2**	**11.6**	**11.5**
8	50.0	39.9	41.3	34.7	15.6	14.6
9	40.8	34.3	34.2	29.0	13.0	12.3
10	42.7	36.1	36.2	30.8	13.7	13.0
11	40.6	34.2	34.1	29.0	13.0	12.3
12	41.4	35.7	35.9	31.1	13.6	13.1
13	38.9	33.4	32.9	28.8	12.5	12.2

good trade-off between a complicated model and computation time. As we will see in the Product B example in Section 7, more complicated models sometimes lead to better performance, however, it will also take a longer time for model fitting due to the need to estimate the random effects. As is usually the case, the modeling process requires judgment combined with experimentation and sensitivity analysis.

□

6 Prediction Intervals

6.1 Prediction Interval Basics

Prediction intervals (PIs) for random variables and the calibration of PIs are introduced in literature such as Beran (1990), Meeker and Escobar (1998), Lawless and Fredette (2005), Fredette and Lawless (2007), and Fonseca et al. (2014). In our applications, the random variable of interest, Y, is the total number of monthly events or the cumulative number of events up to a specified month across all systems at risk.

Producing prediction intervals requires the distribution of the sum of the number of events across systems within specified intervals. Under the NHPP model without random effects, the number of events in non-overlapping intervals has a Poisson distribution and the sum of independent Poisson random variables also has a Poisson distribution. In contrast, when the model includes random effects $\boldsymbol{u}$, the sum is a convolution of K negative binomial distributions, which does not have a closed form.

Teerapabolarn (2014) shows that, when $\{\zeta_i \cdot \mu_{b,i}(\cdot\,;\boldsymbol{\beta})/\left[\mu_{b,i}(\cdot\,;\boldsymbol{\beta}) + \kappa_i\right]\}$ is small for each i, the distribution of the sum of independent negative binomial random variables can be approximated by a Poisson distribution with mean, $\sum_i \left[(\zeta_i/\kappa_i)\mu_{b,i}(\cdot;\boldsymbol{\beta})\right]$, where the sum is across all

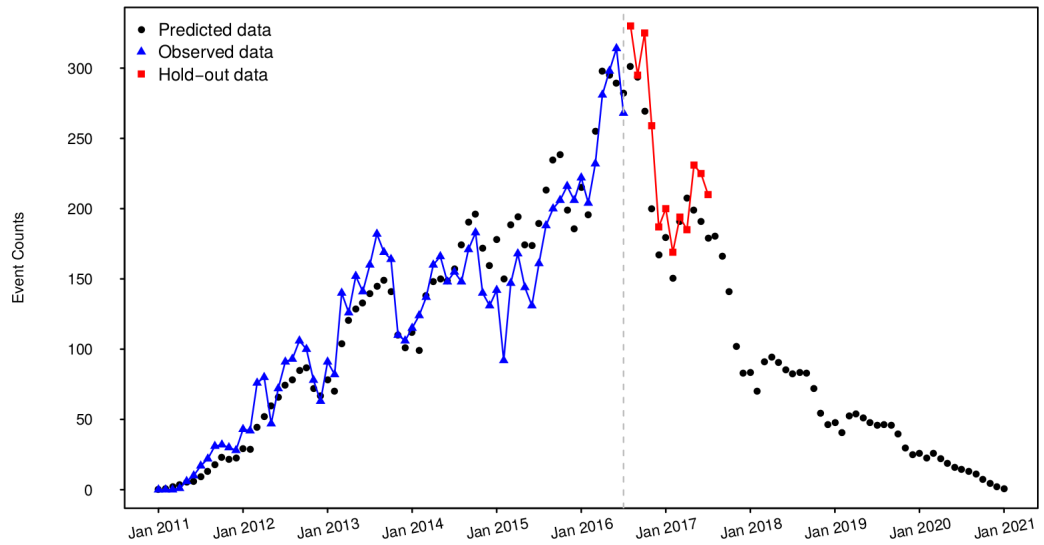

Figure 6: Monthly prediction of the event counts for Product A based on Model 7.

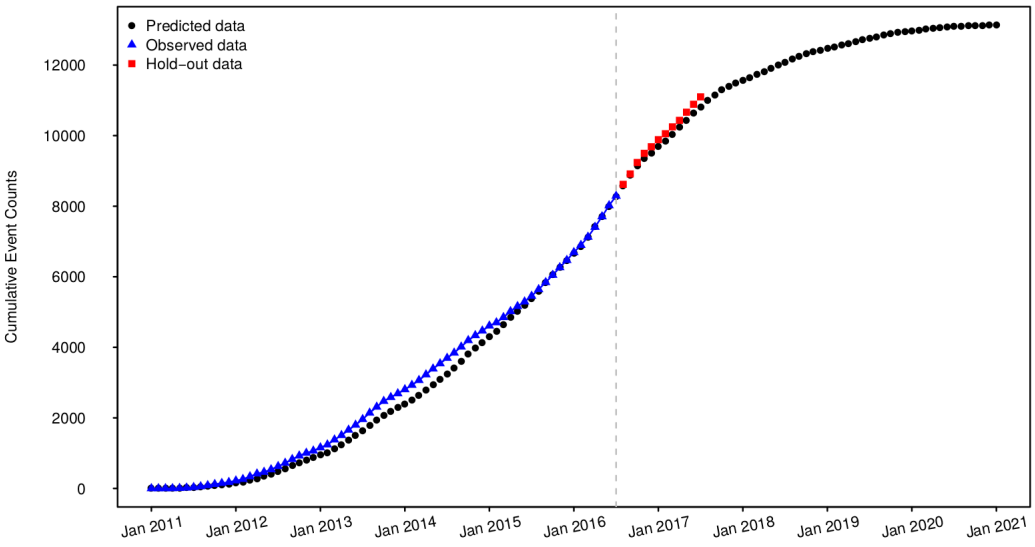

Figure 7: Cumulative prediction of the event counts for Product A based on Model 7.

systems at risk. We use this approximation for the distribution function of random variables when there are random effects in the model.

Here we compare plug-in prediction intervals, simple normal-approximation prediction intervals, and calibrated prediction intervals procedures for a random variable Y with the distribution function $G(Y;\boldsymbol{\theta})$.

6.2 Plug-in Prediction Intervals

The simple plug-in prediction interval is obtained by simply using the quantiles of $G(Y;\boldsymbol{\theta})$. Specifically, a two-sided $100\,(1-\alpha)\,\%$ plug-in prediction interval of a random variable Y is $[L,\, U]$ so that,

$$G\left[L < Y \leq U; \widehat{\boldsymbol{\theta}}\right] = 1 - \alpha. \tag{21}$$

The actual coverage probability of this procedure will generally be less than $(1-\alpha)$ because plug-in method ignores the uncertainty in $\boldsymbol{\theta}$. It is generally good practice to choose L and U such that $L = Y_{\alpha/2}$ and $U = Y_{1-\alpha/2}$.

6.3 Normal Approximate Prediction Intervals

A normal approximate $100(1-\alpha)\%$ prediction interval for Y assumes that $Y \dot{\sim} Normal(\widehat{Y}, se^2_{\widehat{Y}})$, so the prediction interval is,

$$[L,\, U] = \widehat{Y} \pm z_{1-\alpha/2}\widehat{se}_{\widehat{Y}}, \tag{22}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, $\widehat{Y}$ is the point prediction, and $\widehat{se}_{\widehat{Y}} = \sqrt{\widehat{\mathrm{Var}}(\widehat{Y})}$, where $\widehat{\mathrm{Var}}(\widehat{Y})$ is estimated variance of Y.

6.4 Calibrated Prediction Intervals

Bootstrap procedures to calibrate prediction intervals have been described by literature such as Beran (1990), Meeker and Escobar (1998), Lawless and Fredette (2005), and Fredette and Lawless (2007). Xu et al. (2015) give the following algorithm which is a simulation implementation of the general prediction calibration method described in Section 3 of Lawless and Fredette (2005).

1. Simulate the model estimates $\boldsymbol{\theta}_i^*$ with $i = 1, \cdots, B$ using a parametric bootstrap method.
2. Sample Y_i^* from the distribution function of the random variable, $G(Y;\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is the ML estimate of the parameters from the original data.
3. Compute $w_i = G(Y_i^*;\boldsymbol{\theta}_i^*)$ for $i = 1, \cdots, B$.
4. Let w_L and w_U be the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the empirical distribution of $(w_1, \cdots, w_B)$.
5. Solve L and U from $w_L = G(L;\widehat{\boldsymbol{\theta}})$ and $w_U = G(U;\widehat{\boldsymbol{\theta}})$.

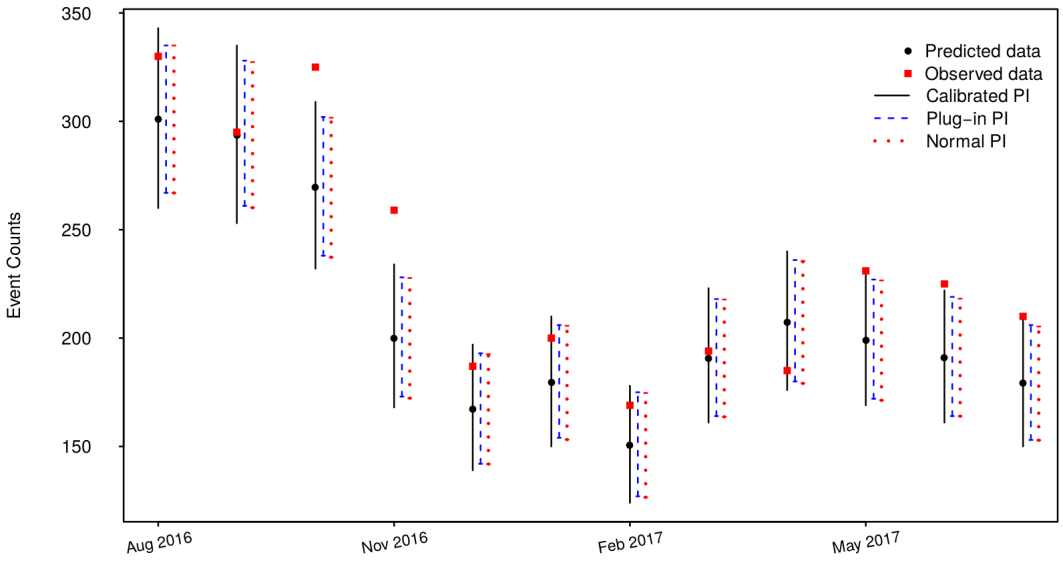

Figure 8: Monthly predictions of the event counts for Product A on the hold-out data based on Model 7.

In Step 1, the parameters $\boldsymbol{\theta}_i^*$ are estimated from the simulated data sets based on $\widehat{\boldsymbol{\theta}}$, which would require a huge amount of computation time in our application. We approximate this procedure by simulating the model estimates $\boldsymbol{\theta}^*$ from the asymptotic multivariate normal distribution of the ML estimates. That is, let $\mathcal{L}(\boldsymbol{\theta})$ denote the total log likelihood of a specified model from K independent systems. Then,

$$\widehat{\boldsymbol{I}}_{\widehat{\boldsymbol{\theta}}} = -\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\widehat{\boldsymbol{\theta}}} \tag{23}$$

is the observed Fisher information matrix for $\boldsymbol{\theta}$ evaluated at the ML estimate $\hat{\boldsymbol{\theta}}$. Then draws from the multivariate normal distribution MVN $\left(\hat{\boldsymbol{\theta}}, \widehat{\boldsymbol{I}}_{\hat{\boldsymbol{\theta}}}^{-1}\right)$ can be used in the calibration process. Because of the large amount of data, the approximation will be good.

Example 5. Prediction Intervals for Product A
The monthly and cumulative event predictions from Model 7 for Product A are shown in Figures 8 and 9, respectively. The calibrated prediction intervals are based on $B = 5{,}000$ simulations. Asymptotic theory (e.g., Beran 1990) suggests that the calibrated prediction interval procedure has coverage probabilities that will be close to the nominal $(1 - \alpha)$ confidence interval. For this example, 9 out of 12 observed event counts are within the calibrated prediction intervals and all observed cumulative event counts are within the calibrated prediction intervals. The plug-in and normal approximate predictions intervals are narrower when compared with the calibrated ones.

□

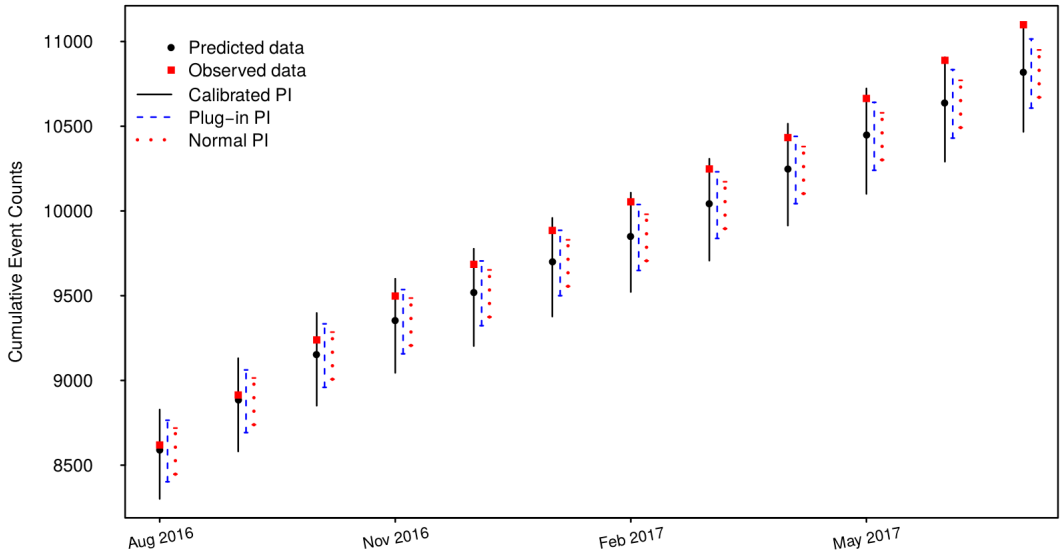

Figure 9: Cumulative predictions of the event counts for Product A on the hold-out data based on Model 7.

7 Models and Predictions for Product B

In this section, we present a second example based on warranty data from Product B, initially described in Section 1.3. The models used in this section are similar to those that were applied to Product A, while the seasonal patterns and fixed covariates differ.

7.1 Exploratory Analysis

Similar to what we have done for Product A, Figure 10 gives the MCF for different model years with 95% pointwise confidence intervals. Although we observed only the mean cumulative number of recurrences per system up to the first year for data of model year 2016, we could tell that the curves of the two different model years behave differently, which indicates that the model year information should be taken into account for modeling.

7.2 Clustering for the Seasonal Models

Figure 11 shows the dendrogram of the hierarchical clustering results for Product B based on warranty information of the most recent year. The observed events and the number of systems at risk by clusters as a function of calendar time are shown in Figures 12 and 13, respectively. The empirical monthly recurrence rates of the two clusters have similar shapes but different levels as shown in Figure 14.

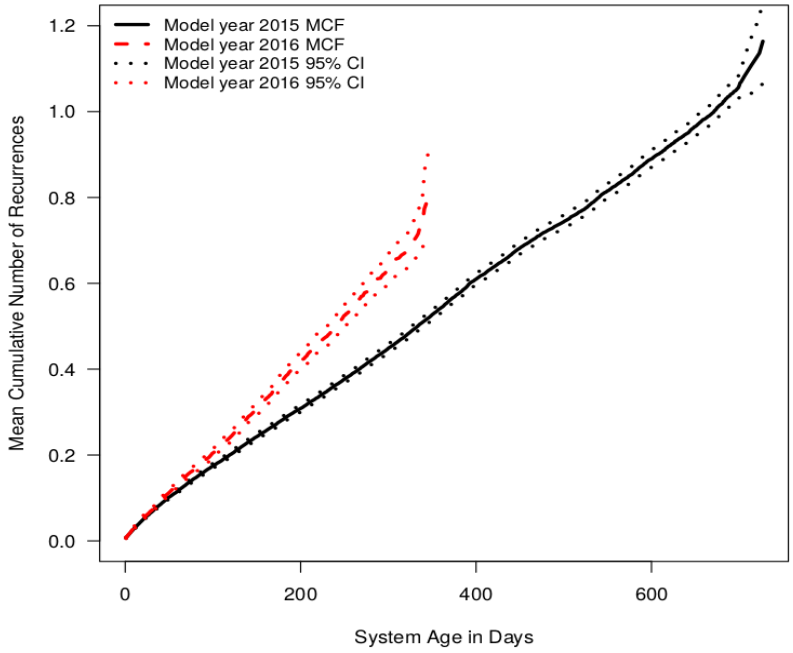

Figure 10: MCF versus system age for systems in model years 2015 and 2016 of Product B with 95% pointwise CIs.

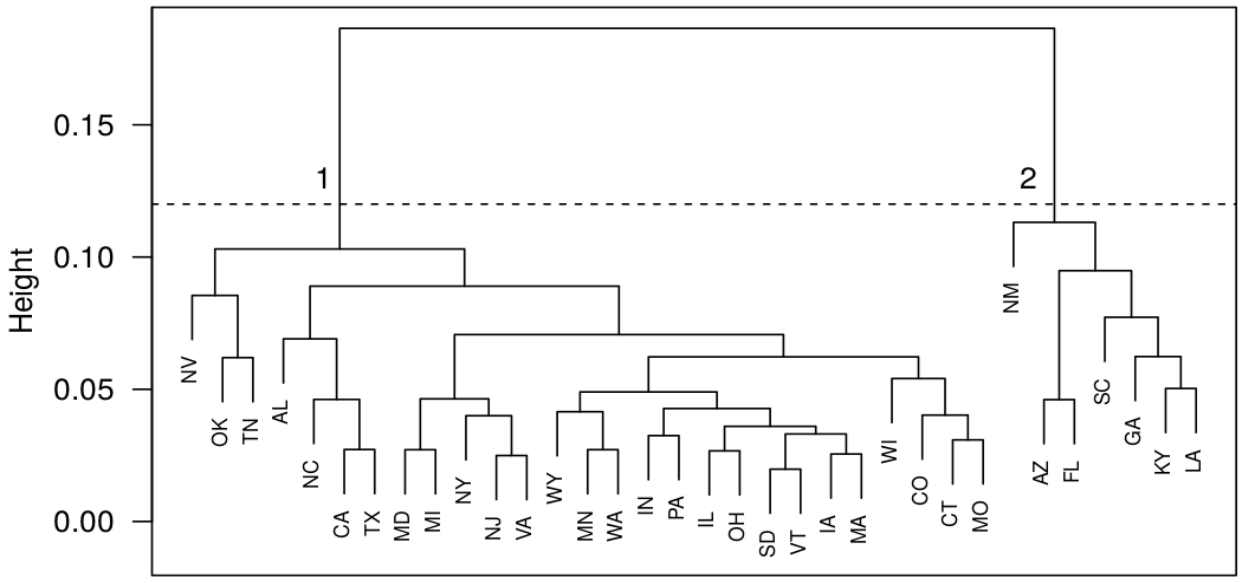

Figure 11: Dendrogram of hierarchical clustering for Product B. The horizontal line indicates the cutoff location to split the observations into different clusters. From left to right, the cluster numbers are 1 and 2, respectively.

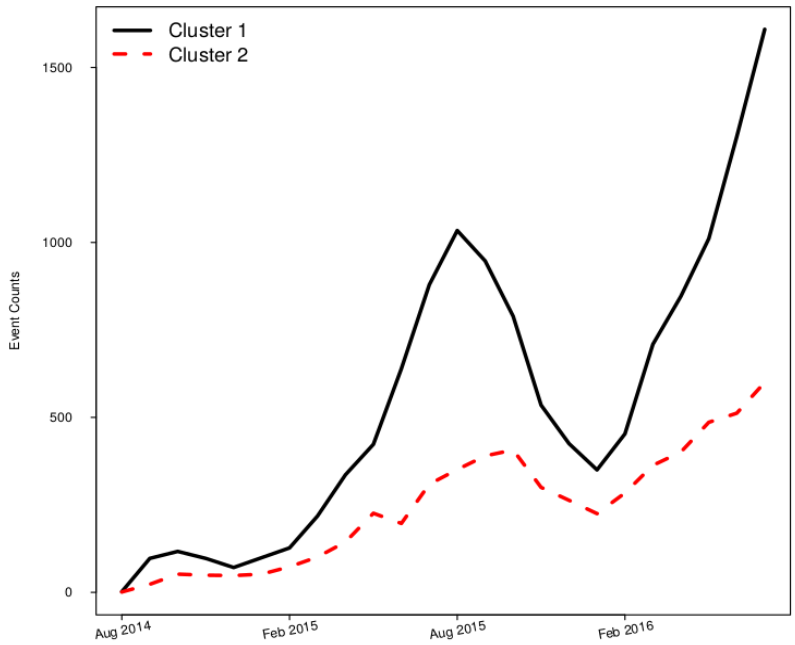

Figure 12: Observed event counts as a function of date for the two different clusters of Product B systems.

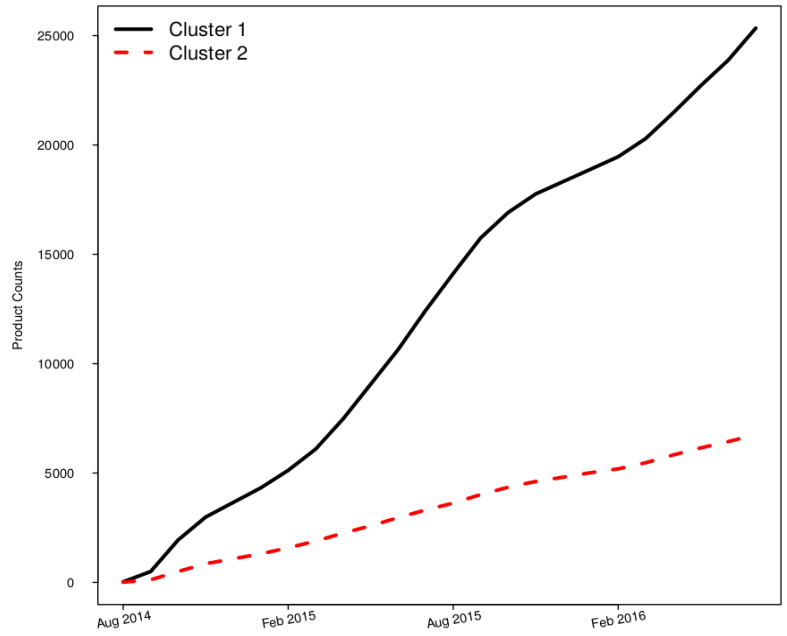

Figure 13: Number of systems at risk as a function of date for the two different Product B clusters.

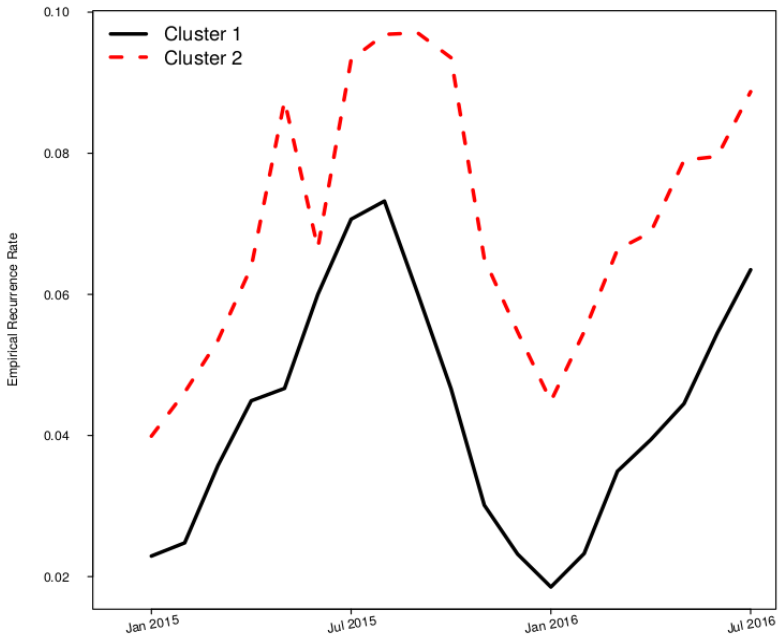

Figure 14: Empirical monthly recurrence rate for Product B since year 2015.

7.3 NHPP Model Fitting

In this subsection, we fit the thirteen models in (18) with $\boldsymbol{x}_{i,fix}$, a vector of length 1, being the product model year indicator (i.e., $\boldsymbol{x}_{i,fix}$ equals 1 if system i has model year 2015 or 0 otherwise). Table 3 shows the model prediction performance for the hold-out data. Model 9 with the common seasonal covariates, model year effects, and random effects provides the best predictions among all of the models. The model with cluster factors does not improve the predictions. An explanation is that we only have two years of data for clustering and model fitting. When doing prediction, we implicitly assume that the seasonal patterns in the future behave like the past. Prediction accuracy may suffer if the future seasonal patterns behave in a different manner. The model fitting of monthly and cumulative event counts based on Model 9 are shown in Figures 15 and 16, respectively.

7.4 Prediction Intervals

Figures 17 and 18 give the prediction intervals of the monthly and cumulative events, respectively, while the calibrated prediction intervals are based on $B = 5{,}000$ simulations. 6 out of 8 observed counts fall within the calibrated prediction intervals for the first eight months of hold-out data. The numbers of events for three out of the last four months, however, fall outside of the prediction intervals. Delayed event reports could be a reason why the monthly predictions are much higher than the observed events for June and July of 2017. The higher number of observed events for April and May could be because more of the product warranties expire in these two months compared with previous years, which encourages people to use the warranty shortly before the expiration dates.

Table 3: Summary results of point predictions on different models for the Product B hold-out data. The model with the smallest RMSE/MAE/MAPE values is marked in **bold**.

No.	$RMSE^6$	$RMSE^{12}$	MAE^6	MAE^{12}	$MAPE^6$	$MAPE^{12}$
1	606.5	454.5	547.1	369.7	43.0	31.2
2	226.2	235.1	194.4	197.7	12.1	17.2
3	101.9	194.7	94.0	146.4	6.9	15.3
4	225.1	236.9	193.9	200.2	12.1	17.5
5	121.8	199.0	112.0	155.2	7.9	15.6
6	224.2	236.6	193.1	200.2	12.1	17.6
7	120.4	198.9	110.9	155.2	7.9	15.6
8	210.2	229.1	182.0	192.8	11.5	17.1
9	**84.2**	**192.4**	**76.1**	**137.5**	**6.4**	**15.1**
10	206.5	228.9	179.3	193.1	11.4	17.2
11	120.6	198.1	110.9	154.3	7.9	15.5
12	204.5	227.7	177.4	192.0	11.3	17.1
13	115.3	196.4	106.0	151.9	7.5	15.3

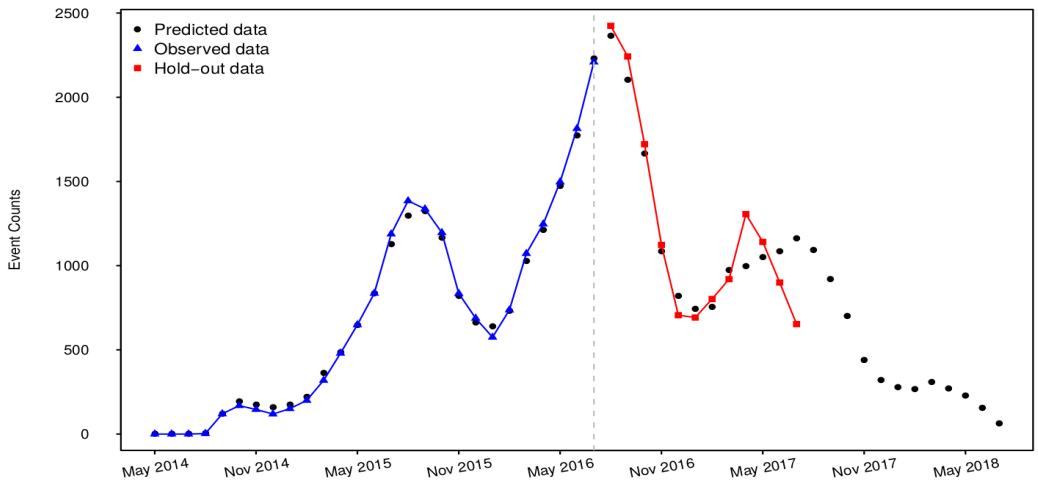

Figure 15: Monthly prediction of the event counts for Product B based on Model 9.

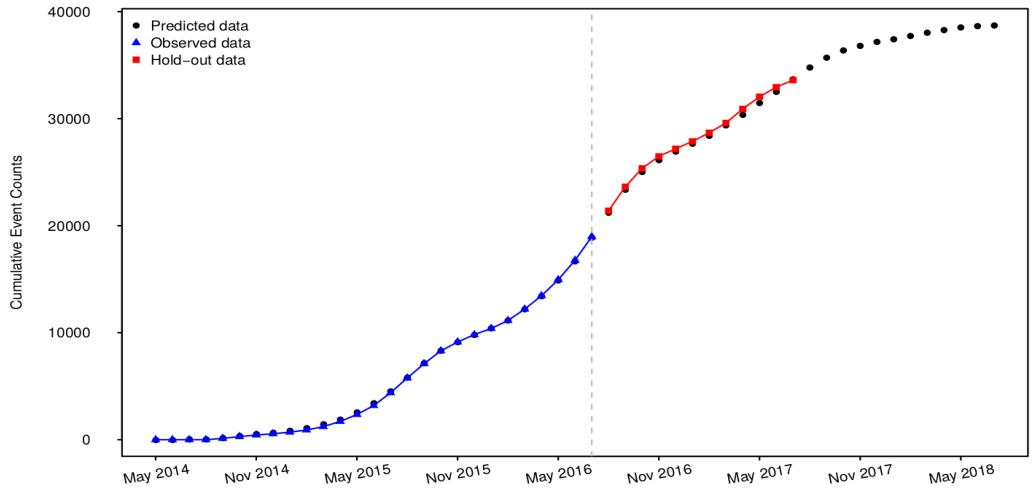

Figure 16: Cumulative prediction of the event counts for Product B based on Model 9.

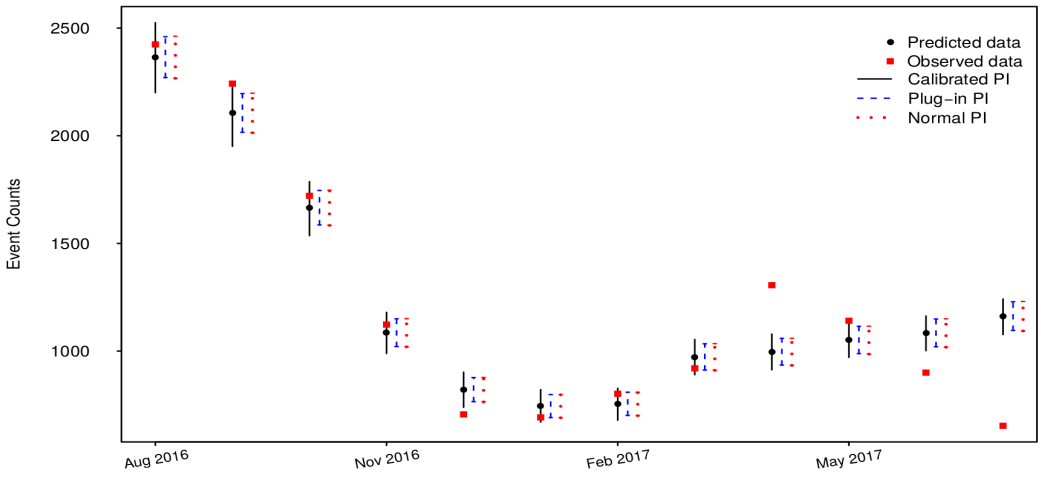

Figure 17: Monthly prediction of the event counts on the hold-out data based on Model 9 for Product B.

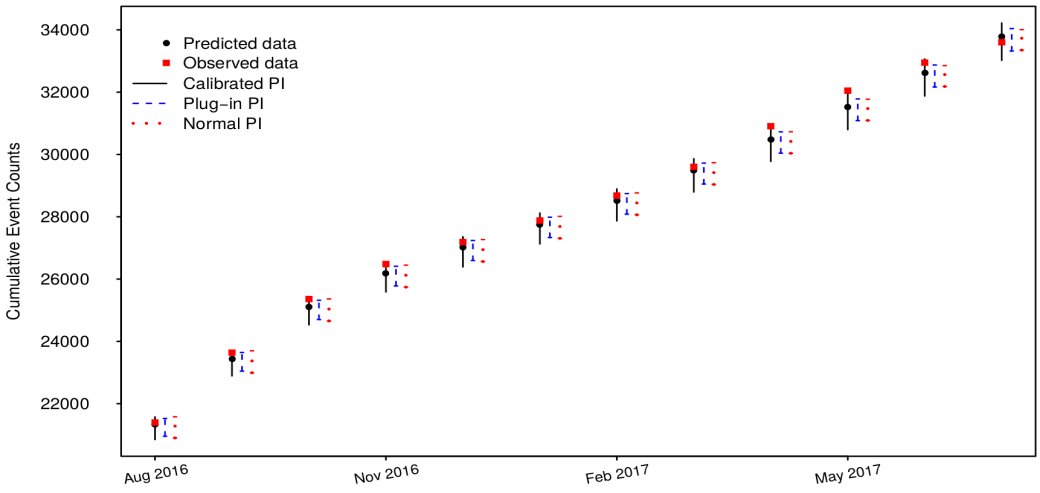

Figure 18: Cumulative prediction of the event counts on the hold-out data based on Model 9 for Product B.

8 Simulation to Study Larger Amounts of Missing Data

In this section, we study the effects of missing data on the prediction performance of new data: we randomly erase between 10% and 30% of location variables from the original data of Product A, and use these data to fit the models based on two different ways to handle missing locations data when clustering:

1. Assign all the missing locations to a new location category.
2. Make a random weighted assignment for the missing locations.

In order to compare the prediction metrics such as the RMSE on the hold-out data, we keep using four clusters and compare the prediction performance of Model 7 with different missing percentages. Figures 19 and 20 show the scatter plots of the 6-month and 12-month RMSE values based on the above two ways of handling the missing locations. We experiment with 10%, 15%, 20%, 25% and 30% missing locations. The random assignment methods shown in A.3 of the supplemental materials are used to fill the missing locations. We experiment by repeatedly doing both the random erasing and assignment of the location data 20 times. Scatter plots of other prediction metrics such as MAE and MAPE are presented in the supplemental materials in Figures 22, 23, 24, and 25. The results of this simulation study can be summarized as:

1. The model using the original data with the 6.6% missing locations assigned to separate categories by their country information has better prediction performance than the models with higher missing percentages.
2. When the missing percentage is relatively small (e.g., less than 15%), the average prediction metrics from two different ways to handle the missing data are close to each other. The random assignment of missing data gives prediction metrics with less variation over the 20 experiments. When the missing percentage is larger, there is no best way to handle the missing data.
3. When the missing percentage increases to around 30%, all of the performance metrics have similar prediction performance with Model 3 in Section 5. An explanation for this is that when we randomly erase 30% location variables and randomly assign a location to each of the missing values, we introduce noise for the seasonality clustering as the actual seasonality pattern in the data may be corrupted by the random assignment or the separate categories of missing locations. The clustering may not help any more, and the model will behave more like Model 3, where all repairable systems have the same seasonality pattern.
4. In general, the model has better prediction performance in terms of the 6-month prediction metrics when the percentage of missing locations is small(e.g., 10%). The differences in the 12-month prediction metrics among different missing percentages are relatively small.

5. When the missing percentage increases from 20% to 30%, the prediction performance degrades. For example, the MAPE[12] values for the 25% missing data are slightly smaller than that of 20%, while the RMSE[6] values for 25% are larger than that of 20%.

9 Concluding Remarks

In this paper, we introduce a general model for the recurrence rate of repairable systems that can be used to predict the number of future events. The model can be applied to various applications depending on the characteristics of the recurrent event processes. Our approach allows the use of covariates that may affect the recurrence rate (for example, the different seasonal trends for different locations), and provides better prediction results than the simple NHPP models. In particular, the use of the cluster and seasonality information improves the predictions of future monthly events for more useful decisions in industry.

Possible extensions of our current work include:

1. In this paper, we model the seasonal trends in the recurrence rate based on the calendar month and we assume implicitly that each month has the same number of business days. The number of business days varies from month to month because of holidays and the number of weekends in a month. Taking the number of business days of each month and the exact calendar entry dates of the product warranty into the model could lead to more accurate modeling and prediction of events.

2. In some applications, claims are not reported immediately after the system failure. This introduces extra variability in the time-dependent seasonal patterns in the model and can lead to inaccuracies for data near the data-freeze date. Also, there can be spikes of warranty claims near to the end-of-warranty date. Such factors might be included in the prediction model.

3. Our paper focused on the prediction of future events. Sometimes it is important to predict future costs as well. Our model can be extended to a compound mixed Poisson process like that described in Grandell (1997). Marked point processes can also be used for claim cost prediction, as described in Brémaud (1981) and Karyagina et al. (1998). Information about failure modes could be helpful for claim cost prediction.

4. Two-dimensional warranty policies are widely used (e.g., in the North American automobile market). For example, the observation of a warranty contract will end when the mileage of a product reaches 36 thousand miles or three years after the purchase date, whichever comes first. A model based on both system age and usage would be more appropriate. However, usage data are often not complete as the usage information may not be available until there is a claim and some of the systems may not have any claims before the end of the warranty period. Also, automobiles with claims may not be a representative sample of the entire population. Lawless and Crowder (2010) proposed joint models on the age and usage dimensions for the warranty data for dependence assessment and model parameter estimation. Some work has also been done on using

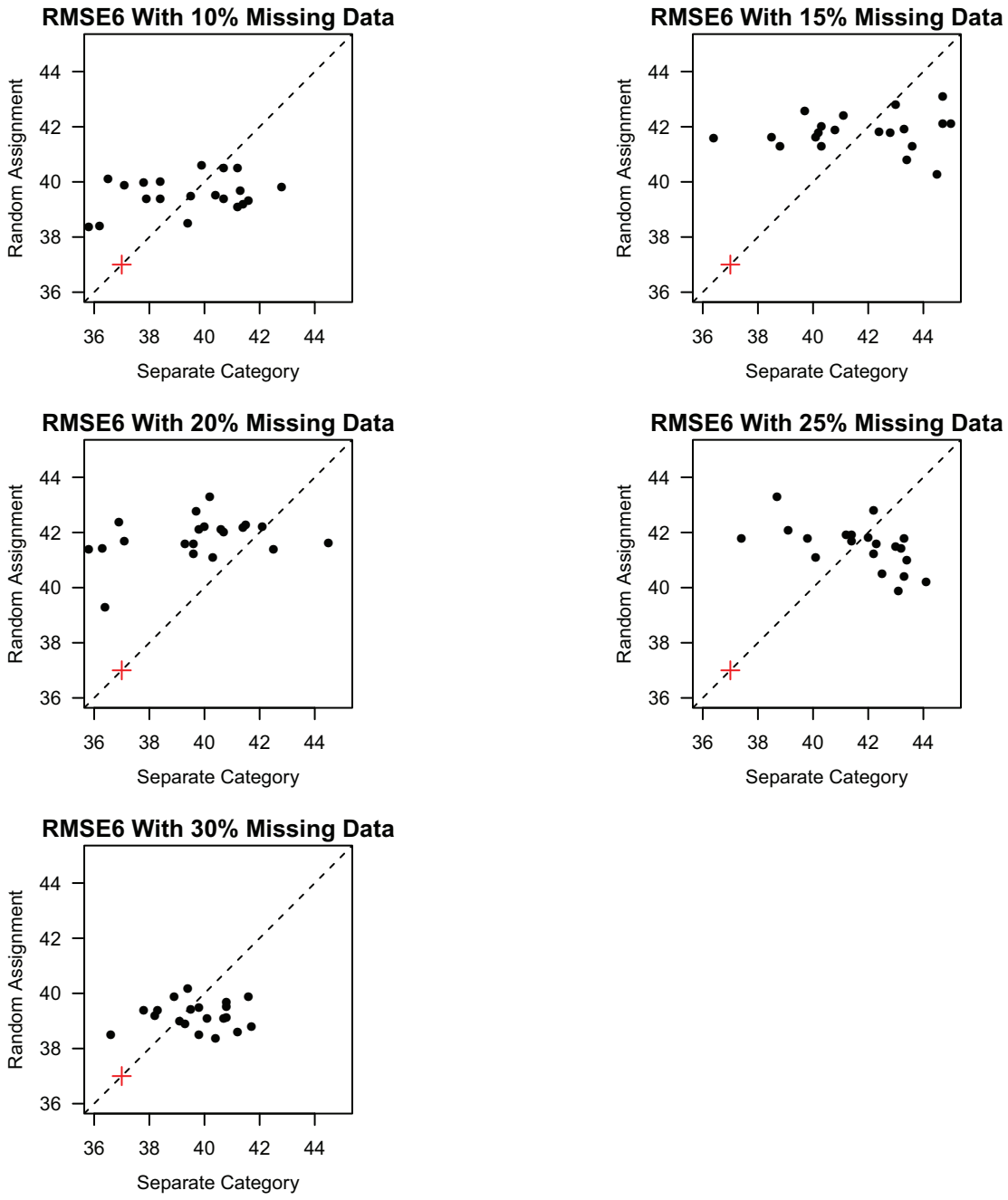

Figure 19: Scatter plots of 20 experimental prediction results with weighted random assignment and separate categories for RMSE^6. The cross marks indicate the prediction metrics in Model 7 of Section 5.

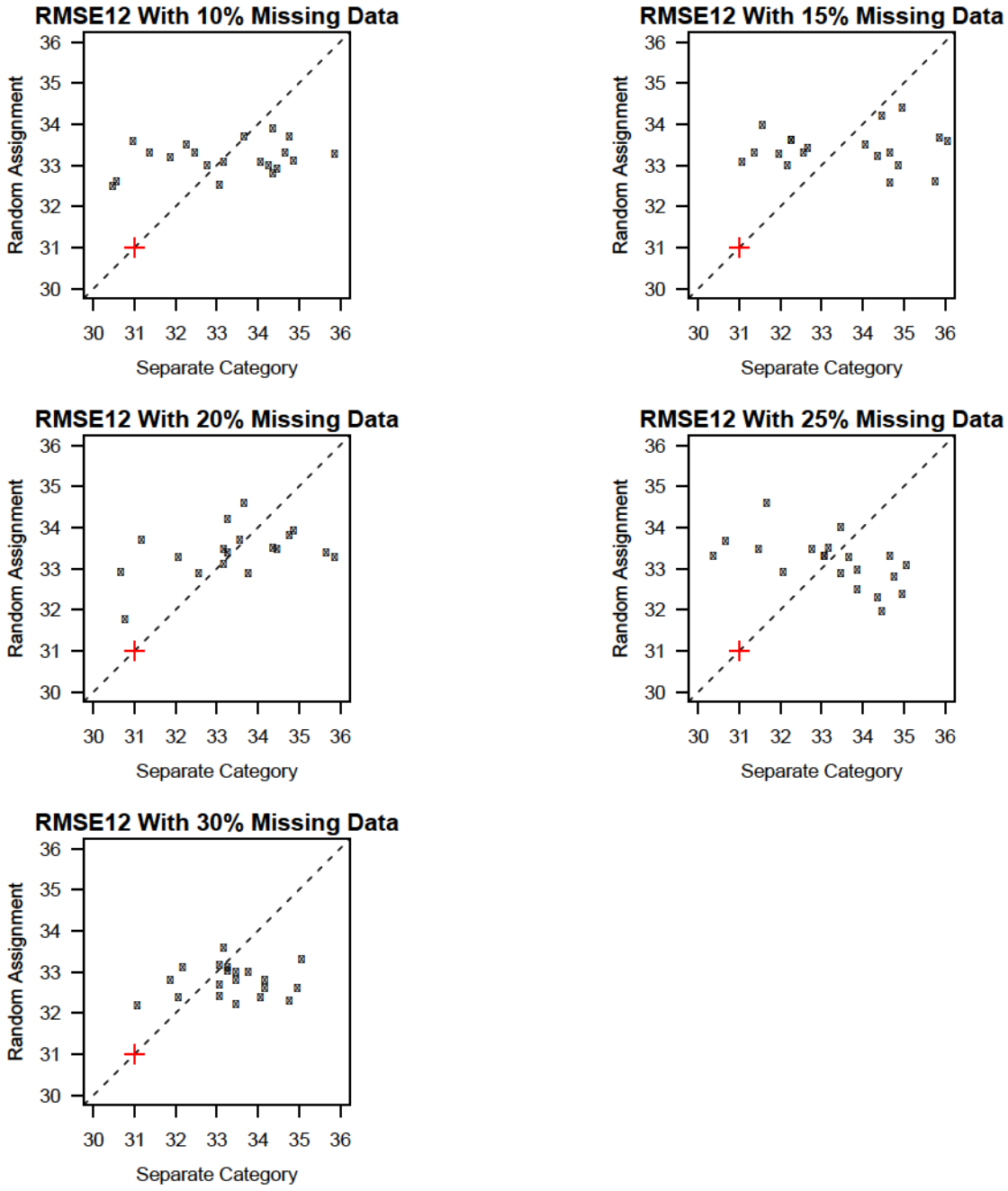

Figure 20: Scatter plots of 20 experimental prediction results with weighted random assignment and separate categories for RMSE^{12}. The cross marks indicate the prediction metrics in Model 7 of Section 5.

a synthesized scale based on both age and usage as described in Ahn et al. (1998) and Duchesne and Lawless (2000).

5. The investigation of NHPP models with time-varying covariates and random effects under the Bayesian framework would be useful. The hierarchical modeling together with tools such as Markov Chain Monte Carlo (MCMC) can then be used conveniently for estimating model parameters and producing prediction intervals.

Acknowledgments

The authors gratefully acknowledge the help of After, Inc. for supporting and assisting in the work presented in this paper. We appreciate the insightful comments from the editor, associate editor and reviewers very much. The work by Hong was partially supported by National Science Foundation Grant CMMI-1904165 to Virginia Tech.

References

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine 7*, 1121–1137.

Ahn, C.-W., K.-C. Chae, and G. M. Clark (1998). Estimating parameters of the power law process with two measures of failure time. *Journal of Quality Technology 30*, 127–132.

Beran, R. (1990). Calibrating prediction regions. *Journal of the American Statistical Association 85*, 715–723.

Brémaud, P. (1981). *Point Processes and Queues: Martingale Dynamics*, Volume 50. Springer.

Cifuentes-Amado, M. V. and E. Cepeda-Cuervo (2015). Non-homogeneous Poisson process to model seasonal events: Application to the health diseases. *International Journal of Statistics in Medical Research 4*, 337–346.

Cook, R. J. and J. Lawless (2007). *The Statistical Analysis of Recurrent Events*. Springer.

Duchesne, T. and J. Lawless (2000). Alternative time scales and failure time models. *Lifetime Data Analysis 6*, 157–179.

Fonseca, G., F. Giummole, and P. Vidoni (2014). Calibrating predictive distributions. *Journal of Statistical Computation and Simulation 84*, 373–383.

Fredette, M. and J. F. Lawless (2007). Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. *Technometrics 49*, 66–80.

Grandell, J. (1997). *Mixed Poisson Processes*, Volume 77. CRC Press.

Hamada, M. S., A. Wilson, C. S. Reese, and H. Martz (2008). *Bayesian Reliability*. Springer Science & Business Media.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.

Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*, 100–108.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*, Volume 112. Springer.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 457–481.

Karyagina, M., W. Wong, and L. Vlacic (1998). Life cycle cost modeling using marked point processes. *Reliability Engineering & System Safety 59*, 291–298.

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics 48*, 795–806.

Koutsellis, T., Z. Mourelatos, M. Hijawi, H. Guo, and M. Castanier (2017). Warranty forecasting of repairable systems for different production patterns. *SAE International Journal of Materials and Manufacturing 10*(2017-01-0209), 264–273.

Krivtsov, V. V. (2007). Practical extensions to NHPP application in repairable system reliability analysis. *Reliability Engineering & System Safety 92*, 560–562.

Lawless, J. and M. Fredette (2005). Frequentist prediction intervals and predictive distributions. *Biometrika 92*, 529–542.

Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association 82*, 808–815.

Lawless, J. F. and M. J. Crowder (2010). Models and estimation for systems with recurrent events and usage processes. *Lifetime Data Analysis 16*, 547–570.

Lawless, J. F. and C. Nadeau (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics 37*, 158–168.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory 28*, 129–137.

Meeker, W. Q. and L. A. Escobar (1998). *Statistical Methods for Reliability Data*. John Wiley & Sons.

Nelson, W. (1988). Analysis of repair data. In *1988. Proceedings., Annual Reliability and Maintainability Symposium, IEEE*.

Nelson, W. B. (2003). *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, Volume 10. SIAM.

Ngailo, T., N. Shaban, J. Reuder, E. Rutalebwa, and I. Mugume (2016). Non homogeneous Poisson process modelling of seasonal extreme rainfall events in Tanzania. *International Journal of Science and Research 5*, 1858–1868.

Rai, B. K. (2009). Warranty spend forecasting for subsystem failures influenced by calendar month seasonality. *IEEE Transactions on Reliability 58*, 649–657.

Rigdon, S. E. and A. P. Basu (2000). *Statistical Methods for the Reliability of Repairable Systems*. Wiley.

Ross, S. M. (2014). *Introduction to Probability Models*. Academic Press.

Rousseeuw, P. J. and L. Kaufman (1990). Finding groups in data. *Series in Probability & Mathematical Statistics 34*, 111–112.

Ryan, K. J., M. S. Hamada, and C. S. Reese (2011). A Bayesian hierarchical power law process model for multiple repairable systems with an application to supercomputer reliability. *Journal of Quality Technology 43*, 209–223.

Slimacek, V. and B. Lindqvist (2016). Reliability of wind turbines modeled by a Poisson process with covariates, unobserved heterogeneity and seasonality. *Wind Energy 19*, 1991–2002.

Teerapabolarn, K. (2014). Poisson approximation for independent negative binomial random variables. *International Journal of Pure and Applied Mathematics 93*, 779–781.

Therneau, T. M., P. M. Grambsch, and V. S. Pankratz (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics 12*, 156–175.

Wu, S. (2012). Warranty data analysis: A review. *Quality and Reliability Engineering International 28*, 795–805.

Xiao, S., A. Kottas, B. Sansó, et al. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *The Annals of Applied Statistics 9*, 353–382.

Xu, Z., Y. Hong, and W. Q. Meeker (2015). Assessing risk of a serious failure mode based on limited field data. *IEEE Transactions on Reliability 64*, 51–62.